Lakefront property prices and water quality: Influence of quality evidenced from sales in the Twin Cities region

BY

David Sedgwick

Thesis Project
Submitted in partial fulfillment of the
Requirements for the degree of
Master of Science in Predictive Analytics
May 2019

Joe Wilck, First Reader
Sophia Vaughan, Second Reader

Abstract

Minnesota is the land of 10,000 lakes (technically 11,842 > 10 acres)! This bold proclamation can be seen everywhere from license plates to tourism pamphlets; it reveals how much water matters to the state on not only a commercial level, but as the very identity of the community. At the heart of Minnesota is the Twin Cities region. This study seeks to identify if a relationship exists between water quality and prices of property with proximity to lakefront. A hedonic pricing method (HPM) was implemented across 5,584 properties located on 75 lakes within the Twin Cities. HPM allowed for controlling of housing factors outside of water quality, while also providing a mechanism for comparing varying degrees of lake water quality against a housing baseline that omitted water features. The generalized least square estimator that was selected based on performance was trained against an extensively cleansed dataset that had both temporal dimensions flattened using aggregate with averages, and spatial dimensions reduced based on distance to water filters. Overall, the results depict a clear picture that there exists a statistically significant relationship between water quality and lakefront property prices in the Twin Cities region.

Table of Contents

Introduction

House prices have entered a range where property attributes no longer support market valuations, resulting in the potential to pay significantly for perceived premiums. One such perceived premium is being located on a recreational body of water, when in fact that water may be impaired. While the water potentially has value, it does not support a significantly increased property value from a quantitative perspective if it proves to be of low quality. A buyer may be expecting bountiful fishing, swimming, recreation, and an overall healthy wildlife population supported by the lake. Instead they may experience sparse fish, swimmers itch, algae blooms, and a lack of wildlife. This lack of awareness could potentially occur due to an absence of investigation by buyers, who rely instead on aesthetics, reputation, and the threat of competitive bids to make emotional purchase decisions.

The objective of this research is to test the hypothesis that water quality affects purchase price for residential lakefront properties within the seven county Twin Cities region. This is especially important during a time when market prices have increased at a rapid clip, contributing to illogical purchase behavior.

The questions that will be answered are:

1. How does water quality influence lakefront property valuation?
2. What are the most influential water quality factors on valuation?
3. Is there a method available for improving any of these factors, and by extension valuation?

Finding an answer to these questions with a high degree of confidence requires the implementation of predictive analytics. The reason this is the case is that the number of variables required to perform a deterministic evaluation of whether a relationship exists between water quality and property price is considerable. This opens up the potential for influencing factors that have not been accounted for, which can compromise the association between dependent and independent variables. Predictive analytic methodologies offer mechanisms to control these and other factors, given the availability of suitable datasets for training and evaluation.

Statement of the Problem

The topic being explored is the effect of water quality on lakefront property prices. While on the surface this may seem to be a topic that can be assessed using intuition, the problem lies in controlling for confounding and lurking variables, avoiding under/overfitting, dealing with autocorrelation, multicollinearity, and heteroscedasticity. Intuition alone has the potential to result in misleading cost-benefit conclusions.
There are scenarios where it is difficult to decipher if a lake's higher property value is attributed to larger, more expensive houses built along its shores or if the higher property values are a result of the water quality itself. This is where intuition and simplistic comparison breakdown. The evaluation must be conducted with predictive algorithms that control for the multitude of factors influencing property price beyond actual water quality. This will provide an assessment of price fluctuation in a way that results in confidence that the causal factor is indeed water quality. The expression "correlation does not imply causation" rings ever true!

Justification

The genesis for the idea behind this thesis resulted from my real-world experience of shopping for a lakefront property. This is an endeavor that I commenced in 2014, at first on a casual basis, then becoming more serious as time progressed. The observation that I made was that there is an apparent disassociation between the water quality of the lakes and the monetary value assigned to the houses built within close proximity of their shores.

This climaxed in 2016 as I prepared to make an offer on a lakefront property in Crystal, Minnesota. During tours of the house I marveled at the beauty of the construction, the layout of rooms, the large yard with a boathouse and firepit. But at the edge of the yard something less pleasant struck me. The lake itself seemed to have a greenish hue to it. While performing additional due diligence I discovered data from the MPCA (Minnesota Pollution Control Agency) that revealed that this lake had been classified as an impaired body of water to such a degree that swimmers should expect to develop an itch from the severe algae blooms that develop during the summer. At this point it was spring, which masked the future green soup this water would turn into during the warmer, sunnier months. This shocked me as the house, as well as the surrounding houses, was large, beautiful, and expensive! Revealing another correlation in that areas with a higher socioeconomic status tend to have larger lawns with greater maintenance that produces increased levels of lawn fertilizer runoff leading to algae blooms that are fed by nutrient loading.

It was at this point that I contemplated using the power of predictive analytics to munge through the hundreds of lakefront properties around the Minneapolis - St. Paul metropolitan area in hopes of preventing the ineffective use of time and energy that I had spent on this house. Justification for this thesis lies in its potential to produce findings that identify not only over-priced properties located on impaired water, but also under-valued properties located on high quality water. In addition, based on the type of water quality issue, it may be possible to identify if these bodies of water have hope for remediation with implementation of best management practices, or if they are flawed in such a way that there is no foreseeable improvement.

Moving beyond personal experience, there is a societal fiscal impact regarding the value of lakes in Minnesota. Local, county, and state governments are responsible for writing policies regulating land development and conversely water quality and quantity. If in fact lakes do not have a monetary benefit, the economic and political reality is that they will not be protected at the opportunity cost of developing land for commercial and agricultural use. If lakes do not have value, budgets will not be allocated for the costly projects required to sustain healthy, and rehabilitate impaired, water. Furthermore, the value of property prices as an extension of water quality has an impact on state and local taxes. If it is shown that improved water quality results in revenue from tourism, as well as increased tax revenue generated from higher property values, there will be a return on investment to be made for legislatures.

Moving even further beyond the fiscal impact, there is a spiritual element, a sense of identity that is drawn from Minnesota lakes. (Nichols, 2014) said it best

"We are inspired by water—hearing it, smelling it in the air, playing in it, walking next to it, painting it, surfing, swimming, or fishing in it, writing about it, photographing it, and creating lasting memories along its edge . . . We know instinctively that being by water makes us healthier, happier, reduces stress and brings us peace"

Being able to measure and report upon the monetization of these lakes is significant in ensuring that stewardship of such a resource remains ever diligent.

Review of the Literature

In order to perform a thorough, accurate, and effective analysis in preparation of answering the questions posed by this paper, a review of existing research will be conducted. This review will select studies from credible sources across universities and government institutions. The predominant approach for performing such analysis is the use of a hedonic pricing method. Hedonic pricing models are a form of multiple regression analysis that have become popular for estimating whether non-market amenities affect the price paid for market goods, and by extension the implicit value of that amenity's properties. This lends itself well to gauging the value of environmental factors such as water quality, which are purchased as part of a property rather than as a standalone product.

Often the most effective ways to design a functional study is by walking backwards in time to see where prior studies experienced shortcomings, why the shortcoming was an issue that required correction, and how it was corrected. In the spirit of this exercise it is appropriate to review the first well known study performed on the association of water quality and property price. This study was performed by (David, 1968) on artificial lakes in Wisconsin in which water quality was represented using dummy variables ["poor", "moderate", "good"]. These categories were based on expert opinion of lake pollution derived from inspection of the bodies of water. Although estimates using regression analysis showed statistically significant findings supporting the fact that lakefront property values increased as the water quality category improved, it is difficult to differentiate between the categories. This difficulty arises from the fact that the water quality measures used for categorization were based on government ratings determined by the subjective individual opinion of those from the Department of Conservation. The preferred method is one that implements measures of water quality based on quantified metrics (e.g. total phosphorus and Secchi depth). Another concern with this study is its disregard for omitted variables. The author does mention pulp and paper production as contributing to poor water quality, but David does not test whether her water quality parameter was capturing undesirable indirect factors caused by the factory. Factors such as odor and noise emitted from the factory.

Given the shortcomings inherent in subjective evaluation, additional review will be conducted utilizing empirical studies that are based on objective measures produced by hedonic regression modeling. These studies should also discuss techniques for accounting for lurking variables.

The pioneering study that utilized hedonic methods to demonstrate the effect of water quality on lakefront property prices was by (Rosen, 1974). That study showed that the unit price of a good, that is comprised of characteristics with varying degrees of quality, is a function of the levels of quality. As a result, characteristics that are desired by consumers increase the function due to buyers bidding up unit prices. It is because of this that the slope of the function with respect to the characteristic, such as water quality, illustrates a consumer’s willingness to pay for the characteristic.

An example of a study utilizing hedonic methods to measure the impact of environmental attributes on property values is (Leggett & Bockstael, 2000). This study tracked 1,183 waterfront property transactions along the Anne Arundel coastline in Chesapeake Bay over a four-year period (1993-1997). Water quality was measured as inversely related to the level of fecal coliform bacteria in the water. The result was that water quality had a statistically significant effect on waterfront prices. Additionally, the study reveals that homeowners along the Chesapeake Bay exhibit an inclination towards funding projects that will reduce levels of fecal coliform bacteria in the bay, in turn improving water quality and increasing property prices. This represents a cognitive connection between water quality, value, and a resulting willingness to invest.

In addition to discussing the merits of using hedonic methods to empirically measure the value of water quality, the study discusses the fault inherent in such an approach, pointing out the ambiguous nature of hedonic applications. Given the importance of identifying not only a method's strength, but its weaknesses, this study proves to be vital. The authors even go so far as to quote another study by (Small, 1974) referencing hedonic methods:

> "I have entirely avoided in this comment the important question of whether the empirical difficulties, especially correlation between pollution and unmeasured neighborhood characteristics, are so overwhelming as to render the entire method useless."

They however stop short of such an opinionated view, taking the stance that by being aware of these concerns they can account for them. Below is a summary of the four factors that were identified in this study as common challenges found in hedonic based methods. By documenting them it will be possible to ensure they are taken into consideration for the purposes of model design:

1) As part of model specification functional form is arbitrary. This can be addressed by implementing a flexible functional form, an approach that has its own risks,

namely the possibility of specification error. Misspecification occurs when the algebraic form or the choice of predictor variables does not accurately represent the real-world process being modeled. In this case water qualities influence on lakefront property prices with the possibility of a nonlinear form.

2) Establishing the boundaries of the housing market that is being evaluated as a response to water quality is difficult. When too small a market is chosen there is the potential for a loss of efficiency, however selecting only a subset of the housing market may actually yield better results. There are also scenarios where a body of water proves to be ideal, such a scenario is described by (Leggett & Bockstael, 2000)

   "Maryland's Anne Arundel County, located on the western shore of the Chesapeake Bay, is especially well suited for a hedonic analysis of water quality. Within 40 miles of both Baltimore and Washington, DC, the number of waterfront properties in the county is substantial. These waterfront locations are valued for their boat access to the Chesapeake Bay, for *in situ* recreational (swimming, wildlife viewing, fishing, and boating) experiences, and for aesthetic reasons. The irregularity of the Anne Arundel coastline (which inhibits mixing), together with the multiplicity and geographic dispersion of sources of water pollution, produces considerable variation in water quality."

3) Multicollinearity poses problems for selection of predictor variables. Often house structural variables are correlated with one another, as are neighborhood variables. The tendency is to then select a subset of variables in hopes of eliminating collinear predictor variables. When the predictor variables are not themselves the object of interest, or where they are all proxies for the same exogenous effect, this is not detrimental. When these conditions are not true

however, it is possible the result will be a biased coefficient estimate in the form of a lurking variable.

4) An example of lurking variable bias comes in the form of close proximity point and nonpoint source pollution that adversely affect property prices. An example of a point source is a factory, an example of a nonpoint source is a feedlot or stormwater runoff. These sources have influences on a multitude of environmental factors beyond water quality. An illustration of this concept is a factory that has an unsightly presence in the neighborhood. This factory may not only affect water quality, but introduce congestion, visual degradation, light, noise, smell, and air pollution. It is possible that even if this factory did not emit substances that negatively impact water quality, property prices would still suffer in response to these other pollutants. This compromises the predictive power of a model that omits these variables since it is not possible to identify which pollutant(s) has a causal relationship with lakefront property price. As such it is important to identify if such predictor variables exist, and if so, control them. Otherwise the coefficients on the water quality variables may be negatively biased to such a degree that the null hypothesis of water quality not affecting the purchase price for Twin Cities region lakefront properties is incorrectly rejected.

Of the reviewed studies one was built using non-technical methods, while two others were built on technical methods. What has not been reviewed is a hybrid study that utilizes both technical numeric measures, along with more easily interpreted non-technical categorical measures of water quality. One such study that compares technical and non-technical measures was performed by (Bin & Czajkowski, 2013).

The technical and non-technical variables used for hedonic analysis from that study are shown in Table 1.

Table 1. Technical and non-technical variables (Bin & Czajkowski, 2013)

48 *Bin and Czajkowski*

Table 1
Summary Statistics for the Hedonic Data

Variable	Mean	Std. Dev.	Minimum	Maximum
2004 House sales price Q2 values	937,294.72	851,731.89	89,907.39	7,224,467.27
District 1 (=1)	0.35	0.48	0.00	1.00
District 2 (=1)	0.19	0.39	0.00	1.00
District 3 (=1)	0.03	0.16	0.00	1.00
District 4 (=1)	0.43	0.49	0.00	1.00
Sold in 2000 (=1)	0.23	0.43	0.00	1.00
Sold in 2001 (=1)	0.21	0.40	0.00	1.00
Sold in 2002 (=1)	0.22	0.41	0.00	1.00
Sold in 2003 (=1)	0.21	0.41	0.00	1.00
Sold in 2004 (=1)	0.13	0.33	0.00	1.00
Lot square footage (in thousands)	26.22	45.22	2.60	883.69
Total housing square footage (in thousands)	2.69	1.50	0.56	13.39
Number of bathrooms	2.75	1.18	1.00	10.00
Concrete block exterior walls (=1)	0.53	0.50	0.00	1.00
Fireplace (=1)	0.37	0.48	0.00	1.00
Pool/patio enclosure (=1)	0.24	0.42	0.00	1.00
In-ground pool (=1)	0.54	0.50	0.00	1.00
Boat lift (=1)	0.33	0.47	0.00	1.00
Waterfront dock (=1)	0.77	0.42	0.00	1.00
Home special feature (=1)	0.29	0.45	0.00	1.00
Percent of population that is white*	94.98	7.29	41.72	99.08
Percent of population that is age 65 or over*	32.76	7.99	14.81	51.30
Percent of households that are owner occupied*	80.67	8.91	36.59	92.57
National 30-year fixed interest rate	6.93	0.88	5.43	8.71
Location grade (%)	80.82	7.81	63.00	88.00
Location grade B (=1)	0.48	0.50	0.00	1.00
Location grade C (=1)	0.09	0.29	0.00	1.00
Location grade D (=1)	0.34	0.47	0.00	1.00
Water visibility (%)	49.10	13.84	31.20	77.80
Water visibility fair (=1)	0.70	0.46	0.00	1.00
Water visibility good (=1)	0.20	0.40	0.00	1.00
Salinity (ppt)	15.75	7.41	1.00	30.40
Salinity fair (=1)	0.96	0.19	0.00	1.00
pH	8.01	0.12	7.80	8.20
Dissolved oxygen (mg/L)	6.42	0.49	5.70	7.70

Note: The number of observations is 510. * denotes the 2000 census track level data. The national 30-year fixed interest rate is measured at the month and year of sale.

The non-technical categorical measure comes in the form of 'location grades' that are made available to homebuyers in urban coastal housing markets of South Florida. Of note is that a grade of A was not used since no median annual value achieved that value. Whereas F was used as the base category, resulting in the inclusion of three water quality dummy variables corresponding to the letter grades of B, C, and D.

The results show that water quality has an effect on waterfront property prices. In the comparison between technical and non-technical measures of water quality, the authors found that technical measures provide a better prediction of property prices than the non-technical ‘location grade’. The belief is that these results are useful for policymakers as they assess their level of investment in protecting coastal waterways.

Table 2 and Table 3 show the models implemented for non-technical and technical measures from the study.

Table 2. Non-technical hedonic model (Bin & Czajkowski, 2013)

Table 2

ML Estimation Results for the Spatial Hedonic Model—Nontechnical Water Quality Measure

Variable	Model I Coeff.	Model I Std. Error	Model II Coeff.	Model II Std. Error	Model III Coeff.	Model III Std. Error	Model IV Coeff.	Model IV Std. Error
Constant	10.108[a]	2.442	12.551[a]	2.437	13.800[a]	0.635	12.966[a]	0.660
Sold in 2001 (=1)	–0.168[c]	0.098	–0.178[c]	0.093	–0.196c	0.109	–0.209[b]	0.104
Sold in 2002 (=1)	0.011	0.121	–0.143	0.118	–0.042	0.124	–0.187	0.120
Sold in 2003 (=1)	–0.207	0.166	–0.265[c]	0.159	–0.242	0.175	–0.295[c]	0.167
Sold in 2004 (=1)	–0.117	0.160	–0.219	0.154	–0.170	0.163	–0.264[c]	0.157
Lot square footage	0.005[a]	0.001	0.004a	0.001	0.005a	0.001	0.004[a]	0.001
Lot square footage squared	–5.6e–06[a]	1.4e–06	–4.1e–06[a]	1.4e–06	–5.5e–06[a]	1.4e–06	–4.1e–06[a]	1.4e–06
Total housing square footage	0.486[a]	0.048	0.411a	0.047	0.486a	0.048	0.412a	0.047
Total housing square footage squared	–0.027[a]	0.005	–0.022[a]	0.005	–0.027[a]	0.005	–0.022[a]	0.005
Number of bathrooms	0.099	0.071	0.124[c]	0.068	0.104	0.071	0.128[c]	0.068
Number of bathrooms squared	–0.008	0.009	–0.011	0.008	–0.009	0.009	–0.011	0.009
Concrete block exterior walls (=1)	–0.001	0.040	–0.008	0.038	0.005	0.040	–0.003	0.038
Fireplace (=1)	–0.005	0.045	0.013	0.043	–0.010	0.045	0.008	0.043
Pool/patio enclosure (=1)	–0.176[a]	0.053	–0.145[a]	0.051	–0.174[a]	0.053	–0.143[a]	0.050
In-ground pool (=1)	0.122[a]	0.046	0.097[b]	0.044	0.123[a]	0.046	0.097[b]	0.044
Boat lift (=1)	0.039	0.042	0.027	0.040	0.041	0.042	0.030	0.040
Waterfront dock (=1)	0.055	0.048	0.083c	0.046	0.053	0.048	0.081[c]	0.046
Home special feature (=1)	0.024	0.044	0.016	0.042	0.020	0.044	0.015	0.042

Note: Models II and IV show the results of the district-level fixed effects models. Dependent variable is natural log of sales price. [a, b,] and [c] denote significance at the 0.01, 0.05, and 0.10 levels, respectively.

The Impact of Water Quality on Property Values

Table 2 (continued)
ML Estimation Results for the Spatial Hedonic Model—Nontechnical Water Quality Measure

Variable	Model I Coeff.	Model I Std. Error	Model II Coeff.	Model II Std. Error	Model III Coeff.	Model III Std. Error	Model IV Coeff.	Model IV Std. Error
Percent of population that is white	–0.001	0.004	–0.011[a]	0.004	–0.001	0.004	–0.011[a]	0.004
Percent of population that is age 65 or over	–0.003	0.003	–0.011[a]	0.004	–0.003	0.003	–0.011[a]	0.004
Percent of households that are owner occupied	0.002	0.003	0.029[a]	0.006	0.002	0.003	0.029[a]	0.006
National 30-year fixed interest rate	–0.231[a]	0.067	–0.255[a]	0.065	–0.234[a]	0.067	–0.257[a]	0.065
Location grade	0.103[c]	0.062	0.015	0.061				
Location grade squared	–0.001[c]	4.0e–04	–1.3e–04	4.0e–04				
Location grade B (=1)					–0.050	0.100	–0.068	0.097
Location grade C (=1)					–0.058	0.128	–0.122	0.123
Location grade D (=1)					0.093	0.076	0.003	0.074
LAMBDA	0.139[b]	0.054	0.151[a]	0.054	0.133[b]	0.054	0.147[a]	0.054
Log likelihood	–288.792		–264.441		–288.828		–264.410	
Akaike info. criterion	625.584		582.881		627.655		584.820	
Schwarz criterion	727.209		697.213		733.515		703.383	

Note: Models II and IV show the results of the district-level fixed effects models. Dependent variable is natural log of sales price. [a, b,] and [c] denote significance at the 0.01, 0.05, and 0.10 levels, respectively.

Table 3. Technical hedonic model (Bin & Czajkowski, 2013)

Table 3
ML Estimation Results for the Spatial Hedonic Model—Technical Water Quality Measures

Variable	Model I Coeff.	Model I Std. Error	Model II Coeff.	Model II Std. Error	Model III Coeff.	Model III Std. Error	Model IV Coeff.	Model IV Std. Error
Constant	–627.130[a]	206.457	–379.431[a]	204.832	–249.591[c]	129.456	27.476	140.267
Sold in 2001 (=1)	0.029	0.112	–0.032	0.109	0.118	0.097	–0.020	0.096
Sold in 2002 (=1)	–0.038	0.126	–0.170	0.123	–0.286b	0.118	–0.381[a]	0.115
Sold in 2003 (=1)	0.364[c]	0.205	0.151	0.200	0.034	0.179	–0.188	0.177
Sold in 2004 (=1)	0.082	0.165	–0.098	0.161	–0.093	0.158	–0.261[c]	0.155
Lot square footage	0.003[b]	0.001	0.002[b]	0.001	0.003[a]	0.001	0.003[a]	0.001
Lot square footage squared	–3.2e–06[b]	1.4e–06	–2.8e–06[b]	1.3e–06	–3.8e–06[a]	1.4e–06	–3.4e–06[b]	1.3e–06
Total housing square footage	0.404[a]	0.046	0.349[a]	0.045	0.407[a]	0.046	0.357[a]	0.045
Total housing square footage squared	–0.021[a]	0.004	–0.017[a]	0.004	–0.021[a]	0.004	–0.018[a]	0.004
Number of bathrooms	0.114[c]	0.067	0.121[c]	0.065	0.102	0.067	0.108[c]	0.065
Number of bathrooms squared	–0.010	0.008	–0.010	0.008	–0.009	0.008	–0.010	0.008
Concrete block exterior walls (=1)	0.016	0.038	–0.001	0.037	0.018	0.038	–0.004	0.037
Fireplace (=1)	–0.003	0.042	0.018	0.041	–0.004	0.042	0.018	0.041
Pool/patio enclosure (=1)	–0.135[a]	0.050	–0.103[b]	0.049	–0.124b	0.051	–0.108[b]	0.049
In-ground pool (=1)	0.139[a]	0.043	0.104[b]	0.042	0.131[a]	0.044	0.098[b]	0.042
Boat lift (=1)	0.046	0.040	0.042	0.038	0.045	0.040	0.038	0.038
Waterfront dock (=1)	0.050	0.045	0.073[c]	0.044	0.074[c]	0.045	0.086[b]	0.044
Home special feature (=1)	0.003	0.042	1.3e–04	0.040	–0.003	0.042	0.001	0.040

Note: Models II and IV show the results of the district-level fixed effects models. Dependent variable is natural log of sales price. [a, b] and [c] denote significance at the 0.01, 0.05, and 0.10 levels, respectively.

Table 3 (continued)
ML Estimation Results for the Spatial Hedonic Model—Technical Water Quality Measures

Variable	Model I Coeff.	Model I Std. Error	Model II Coeff.	Model II Std. Error	Model III Coeff.	Model III Std. Error	Model IV Coeff.	Model IV Std. Error
Percent of population that is white	0.004	0.004	–0.002	0.004	0.002	0.004	–0.004	0.004
Percent of population that is age 65 or over	–0.008[a]	0.003	–0.010b	0.004	–0.008[a]	0.003	–0.009[b]	0.004
Percent of households that are owner occupied	0.005	0.003	0.017[a]	0.006	0.003	0.003	0.016[b]	0.006
National 30-year fixed interest rate	–0.239[a]	0.064	–0.275[a]	0.061	–0.233[a]	0.064	–0.268[a]	0.062
Water visibility	0.067[a]	0.025	0.062[b]	0.024				
Water visibility squared	–0.001b	2.3e–04	–0.001b	2.2e–04				
Salinity	0.029[c]	0.016	0.017	0.016				
Salinity squared	3.4e–04	4.2e–04	4.3e–04	4.1e–04				
pH	1.614[a]	0.518	0.991[c]	0.514	0.675[b]	0.326	–0.023	0.353
pH squared	–0.001[a]	3.3e–04	–0.001[c]	3.2e–04	–4.1e–04[b]	2.0e–04	2.5e–05	2.2e–04
Dissolved oxygen	–1.506	1.053	–1.438	1.010	–4.250[a]	1.171	–3.447[a]	1.137
Dissolved oxygen squared	0.109	0.080	0.107	0.077	0.299[a]	0.088	0.247[a]	0.086
Water visibility fair (=1)					0.433[a]	0.140	0.281[b]	0.139
Water visibility good (=1)					0.411[a]	0.150	0.197	0.152
Salinity good (=1)					–0.245	0.157	–0.185	0.152
LAMBDA	0.173[a]	0.054	0.182[a]	0.053	0.151[a]	0.054	0.158[a]	0.054
Log likelihood	–260.573		–239.446		–261.255		–241.133	
Akaike info. criterion	581.146		544.892		580.509		546.266	
Schwarz criterion	708.179		684.628		703.307		681.767	

Note: Models II and IV show the results of the district-level fixed effects models. Dependent variable is natural log of sales price. [a], [b], and [c] denote significance at the 0.01, 0.05, and

Although intuition says that the non-technical measures would be easier for homebuyers to understand and therefore be used more effectively, actual results indicate that waterfront consumers were savvy and effective in their interpretation of technical measures of water quality. Higher values of all technical measures of water quality, excluding DO (dissolved oxygen), increase property values significantly. This may be explained by lower DO levels not always being associated with water pollution. Lower DO levels may indicate groundwater influence OR the presence of excess nutrients, while higher DO indicates surface water and adequate oxygen concentration available in the water column. Higher DO is going to be better for aquatic life, but low DO is not necessarily from poor water quality.

In a study conducted by (Krysel, Boyer, Parson, & Welle, 2003) it was shown that water clarity has a statistically significant positive relationship with lakefront properties located in the Mississippi Headwaters Region of northern Minnesota. The recommendation made is that changes in lake water clarity will result in millions of dollars in property values---

lost or gained---in this lake region of Minnesota. Clearly, for economic reasons alone---not to mention the ecological, health, and social benefits at stake---it is important to protect the water quality of all Minnesota's lakes. In fact, current the Minnesota Pollution Control Agency is in the process of assessing when it is best to invest in protection versus restoration of certain water bodies throughout the state.

Another way of tracking lakefront property value as an effect of water quality is to assess the impact that an invasive species such as milfoil has on property prices. (Horsch & Lewis, 2009) use hedonic analysis to estimate the effects of Eurasian watermilfoil (myrophyllum spicatum) across 170 lakes in northern Wisconsin in terms of property values. The finding is that lakes invaded by milfoil experience on average a thirteen percent decrease in land values post invasion.

Using data from 3,186 real estate transactions collected between 1999 and 2010 from the Wisconsin counties of Dunn and Barron, (Kashian & Kasper, 2010) were able to show how property prices on impaired lakes have not kept pace with non-impaired lakes in the same market. They implemented hedonic analysis to obtain implicit prices of lakeshore while controlling for housing and real estate characteristics (i.e. bedrooms, square footage, bathrooms, etc.).

Both Tainter Lake and Menomin Lake suffer from severe blue green algae (cyanobacteria) blooms that not only greatly reduce water clarity but are thick enough to make fishing and recreational activities extremely difficult during the summer. By contrast Red Cedar Lake, Beaver Dam Lake, Chetek Lake, and Prairie Lake all provide healthy ecosystems for recreation and fishing. The findings show a staggering difference in lakefront price per foot, as shown in Figure 1. The authors make the argument that not only does this adversely affect valuations for the homeowners, but that it impacts the community's ability to generate tax revenue as well as support increased economic activity that may result in additional jobs. As such it is the authors' belief that investing in protecting the lakes from future damage, in conjunction with repairing existing damage, is an economically sound policy. Figure 1 identifies the incremental property value per foot of shoreline for each of the lakes included in the study, as determined by the hedonic model results.

Figure 1. Lakes within the 7 county region (Kashian & Kasper, 2010)

The studies above all point to the positive relationship between water quality and property prices. They do so by highlighting varying techniques that can be implemented to boost statistical significance of the findings, while also exposing important shortcomings that must be accounted for in hedonic methods. The lessons learned will be applied to an analysis of a region of lakefront properties that have not been investigated: the seven county Twin Cities region. Specifics on how this will be accomplished are provided in the following section.

Methods

Statistical Theory

As a result of extensive research, it has been determined that the authoritative econometric approach for identifying the valuation of individual environmental amenities that compose market products is the hedonic pricing method. The hedonic pricing method (HPM going forward) is a valuation technique that utilizes linear regression to remove 'hedonic' amenities from market products such as lakefront properties (i.e. features that consumers derive pleasure from). The remaining price change from period to period is attributed to inflation. Meaning that the implicit value of the hedonic amenities, in this case water quality, is represented simplistically with the equation:

(original lakefront value – inflation – baseline property value (i.e. without water features) = water quality value)

The remainder is the indirect value placed by consumers on water quality as a portion of what they are willing to pay for lakefront properties. In summary, HPM relates the product price to the characteristics that it is comprised of, resulting in the ability to estimate the influence these characteristics have on the product price that is supported by the market (Freeman, 1993).

Given that HPM is a technique rather than a specific form of regression analysis, it does not have unique libraries or packages built for it. Rather it highlights the method of price estimation and interpretation that is implemented as part of regression analysis. This technique identifies implicit value by determining price differentials between properties on lakes with varying levels of water quality. The valuations are made meaningful when controls are implemented for other property characteristics, as will be done as part of this study.

Procedure

The origination of the datasets for this study was a MinneMUDAC competition held in November 2016. Since that time the site has been retired, along with the datasets that were posted as part of the competition. Alternative access for those datasets is provided in the following sections 1a, 1b, and 1c. (MinneMUDAC: Dive into Water (Data), 2016).

1) How the data will be collected

a. MetroGIS Regional Tax Parcel Dataset (MetroGIS, n.d.)

   i. The MinneMUDAC datasets originally provided that cover tax parcel data from 2002-2014 are no longer available on the competition site but are now hosted on the Amazon Web Services (AWS) Simple Storage Services (S3) buckets below.

      1. 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014

   ii. The raw sources of those datasets can be downloaded from the MetroGIS site. The steps undertaken on the raw datasets by the competition committee are as follows:

      1. Converted shapefile file attributes to tabular data, no need to process geographic information system (GIS) data
      2. Converted Parcel Polygons and Points to Latitude/Longitude points
      3. Appended Year field

4. Excluded all parcels that used MultiPolygon shapes due to the difficulty of accurately and logically creating a centroid for them. This was a small fraction of the total data set (less than 0.1%)

b. Metropolitan Council Environmental Services – Environmental Information Management Systems – Lake Monitoring Data (EIMS, n.d.)
   i. The MinneMUDAC dataset originally provided that covers lake monitoring data from 1999-2014 is no longer available on the competition site but is now hosted on S3.
   ii. As an academic exercise in curiosity, the original dataset was approximately reproduced using EIMS AdvancedSearch. The query search criteria used to reproduce the dataset interactively from EIMS is outlined below, as well as the URL for downloading that dataset, Appendix A (note: fields have changed since the original dataset was created in 2016).
      1. Advanced Search
         a. By Location
            i. County
               1. Anoka
               2. Carver
               3. Dakota
               4. Hennepin
               5. Ramsey

6. Scott
7. Washington

b. By Date Range

i. 1/1/1999 – 12/31/2014

c. Nutrient

i. Phosphorus

1. Total Phosphorus, Filtered (mg/L)

d. Physical

i. Light/Transparency

1. Secchi Depth (m)

ii. Observation

1. Physical Condition
2. Recreational Suitability

e. Summary

i. Rating

1. Lake Grade, Seasonal

2. The steps undertaken by the competition committee on the raw dataset are as follows:

a. Converted Monitor Station Points to Latitude/Longitude points.

c. Water proximity reference table

i. The original dataset is no longer available on the MinneMUDAC site but is now hosted on [S3.]

ii. The competition committee created the xref table by calculating the distance from a tax parcel centroid to the nearest Lake Monitoring Station provided by the Metropolitan Council Environmental Services. Once the nearest station was identified the distance calculation from the tax parcel centroid to the edge of the lake containing Lake Monitoring Station was performed.

iii. Lake shapefiles and metadata downloads were used as data sources for creating the xref table.

1. MCES Lake Monitoring Sites (MCES Lake Monitoring Sites, n.d.)
2. Census 2010 Geography (Census 2010 Geography - Blocks, Block Groups, Tracts, TAZs, Counties, County Subdivisions and Water, n.d.)

2) How the data will be prepared

a. Data will be interrogated and prepared using a variety of tools as described in the subsequent tools section. This insight will be used for engineering a final merged cohesive dataset that is optimal for model training and testing.

b. Data inspection will occur by first loading data into dataframes using the Pandas library in Python. The objective will be to conduct sampling, feature subsetting, and cleansing.

c. Output from each of the three datasets will be saved to flat text files that will then be merged together using Pandas.

d. It is the resulting merged tabular text-based dataset that will be loaded for modeling. This simple approach has been selected given the manageable size and structure of the raw datasets, which in aggregate are < 1GB with a readily manipulated relational form. That is to say a more complex data processing solution will not be required (e.g. Hadoop, Cassandra, Oracle, MySQL, AWS Redshift).

3) How the data will be modeled

a. By using HPM with regression analysis in Python scikit-learn and statsmodels, the aforementioned data will provide evidence to either reject or fail to reject the null hypothesis that water quality does not have an effect on lakefront property prices in the Twin Cities region.

b. Code will be hosted on Github with notebook editing and computing occurring in a JupyterLab environment.

c. Datasets will be stored in AWS S3.

4) How the analysis relates to the research questions

a. How does water quality influence lakefront property valuation?

   i. The implicit value of water quality will be identified indirectly using hedonic regression analysis with controls to account for other influential property attributes.

   ii. This will ultimately reveal an indirect association (or lack thereof) between water quality and lakefront property price.

b. What are the most influential water quality factors on valuation?

        i. HPM relates the product price to the characteristics that it is comprised of, resulting in the ability to estimate the influence said characteristics have on the product price that is supported by the market.
        ii. This will provide a mechanism for weighing the water quality factors in terms of influence.
    c. Is there a method available for improving any of these factors, and by extension valuation?
        i. The resulting influential factors will provide direction for researching treatment options.

Measurement Techniques

The subsequent sections are meant to describe how the methods are used for executing the project analysis and findings. A thorough description of what the values are and what their corresponding interpretations are, is provided as part of the Results section along with supporting definitions provided in the appendices.

Problem statement attributes

- Dependent (response) variable
    - Single (=1)
    - Continuous
- Independent (predictor) variables
    - Multiple (>1)
    - Continuous, discrete, and categorical

- The objective is to establish a linear relationship between response and multiple predictor variables

Selection of model type

Given the structure of the problem statement to be solved, the appropriate form of model is multiple linear regression. Furthermore, the hedonic pricing method will be utilized as a technique of interpreting regression analysis results.

Validation / Evaluation of model

Now that a model type has been selected it is necessary to define the potential errors inherent in that type of model, catch them if they exist, and handle if applicable. This will ensure our model findings are trustworthy. Furthermore, statistical measures will be reviewed in order to ensure that the findings are statistically significant. The result will be conclusions that are both trustworthy and significant.

- Validation of the 5 core linear regression assumptions (see Appendix B for working definitions)
    1. Linear relationship exists
    2. Multivariate normality exists
    3. Multicollinearity does not exist
    4. Autocorrelation does not exist
    5. Heteroscedasticity does not exist
- Evaluation of model performance (see Appendix C for working definitions)
    - Measures of predictive power
        - Hypothesis
            - H_0 – null hypothesis

        - Water quality does not have an effect on lakefront property prices in the Twin Cities region.
    - H_A – alternate hypothesis
        - Water quality has an effect on lakefront property prices in the Twin Cities region.
- Significance test
    - Significance level
        - p-value
        - 95% confidence interval does not include zero
            - $\alpha = 0.05$
            - If p-value < α
            - Reject the null
            - There is a relationship between water quality and lakefront property
        - 95% confidence interval does include zero
            - $\alpha = 0.05$
            - If p-value > α
            - Fail to reject the null
            - There is no relationship between water quality and lakefront property
- R-squared
- F-statistic

- Model Selection
    - Cross-validation to calculate root mean squared error
    - RMSE will be used as the measure for comparison of models
        - R-squared will be used as a tie breaker

Tools

The software systems and functionality that they provide are as follows:

- Language
    - Python – 3.6.5
- Coding notebook and computing environment
    - JupyterLab – 0.32.1
- Code version control
    - Git hosted GitHub
- Datasets
    - Hosted on S3
    - Flat text files manipulated directly
    - Flat text files loaded into Pandas DataFrames
- Libraries
    - Pandas
    - Numpy
    - Statsmodels
    - Scikit-learn
    - Matplotlib
    - Yellowbrick

Results

Overview

Decisions on functional form and temporal duration will be guided by literature research, which shows that performing analysis on a superfluous feature space across numerous points in time not only adds complexity to the model but can weaken results. This aligns with the law of parsimony, a popular principle in the field of data analytics stating that simpler solutions are more likely to be correct than complex ones.

When it comes to selecting features that best represent water quality in a hedonic equation, there is no universal consensus on a list of standard features given the variance of existing datasets. There is however a pattern showing the indicator that influences consumers valuations of lakefront property above all others is water clarity, represented by the Secchi disk measurement. Similarly, there are no accepted best practices when it comes to temporal duration of water quality measurements. In recent studies it has become common to use water quality values from a single year, e.g. (Netusil, Kincaid, & Chang, 2014) (Walsh, 2009). The reasoning is that findings may be of reduced significance in longer studies due to the increased likelihood of unobserved influences having an impact on property valuations (Michael, Boyle, & Bouchard, 2000).

Figure 2 and Figure 3 provide a visual scope of the study area.

Figure 2. Twin Cities 7 county region

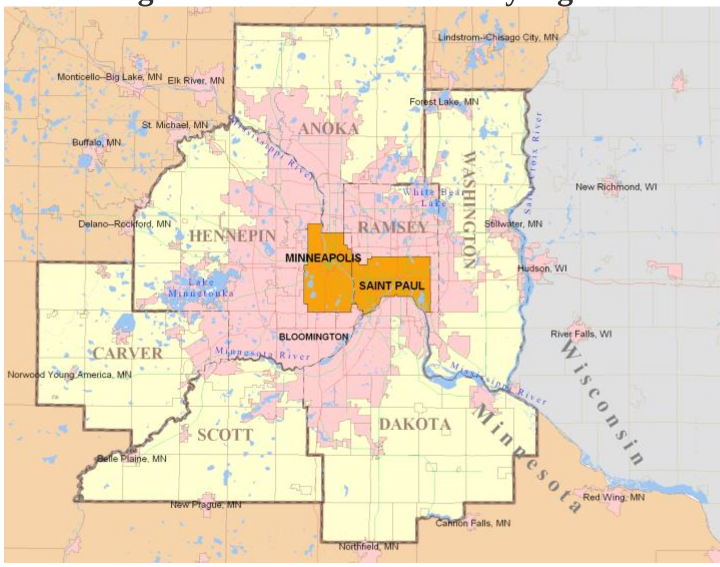
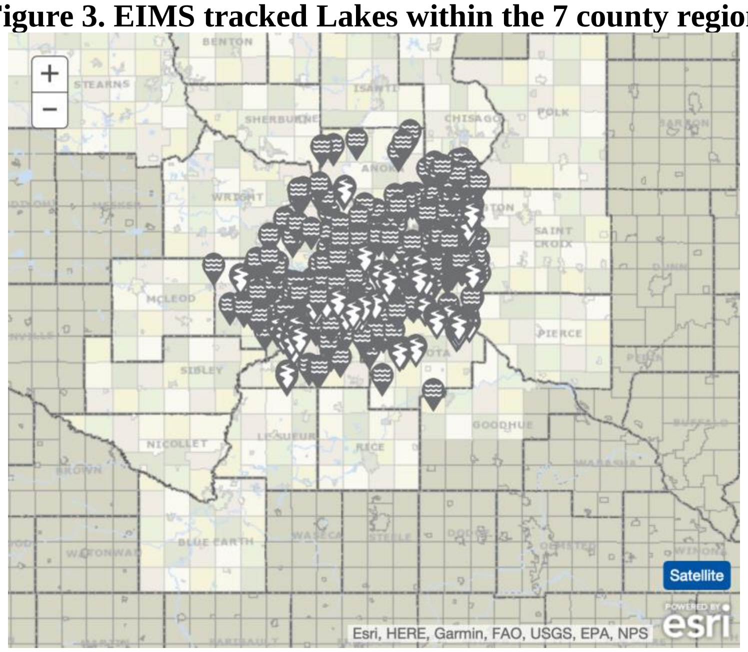


Figure 3. EIMS tracked Lakes within the 7 county region



Datasets

The data sources used in the analysis are referenced in the prior Procedure section. They are Tax Parcel data, Lake Monitoring data, and Water Proximity reference table. The Water Proximity reference table identifies the distance from lakefront property parcels to the nearest lake and monitoring site. This reference table will be used to connect the Tax Parcel data with the Lake Monitoring data.

The raw files will undergo two preparation steps; feature selection, followed by data cleansing. The final data preparation task will be to merge the resulting individual datasets into a single cohesive master dataset ideal for model training and testing as illustrated in Figure 4.

Figure 4. Dataset join variables

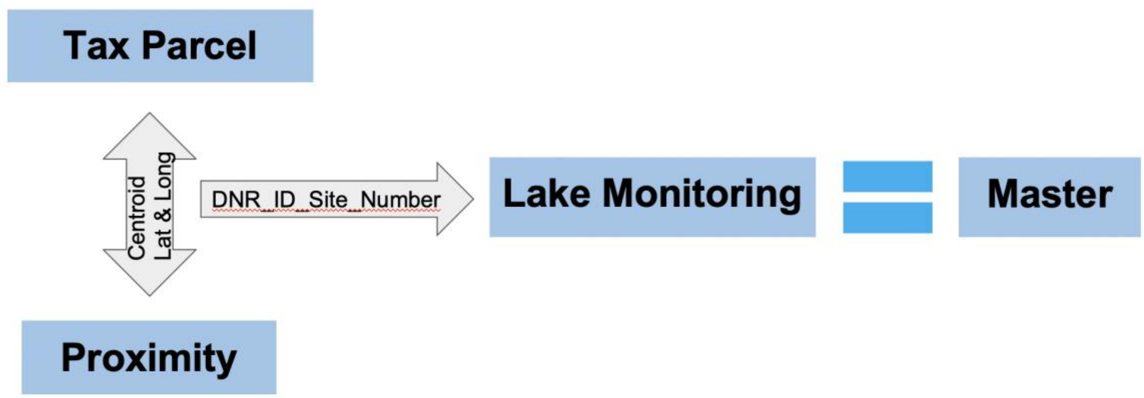


Table 4. JupyterLab Notebooks: Supporting code and visuals
https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis

Phase	Notebook
Data Preparation	data_1_2_3_subset.ipynb
Data Preparation	taxparcel_1_clean.ipynb
Data Preparation	lakemonitoring_2_clean.ipynb
Data Preparation	master_4_merge.ipynb
Validation	model_validation.ipynb
Training	model_train_eval.ipynb
Evaluation	model_train_eval.ipynb
Interpretation	model_final_interpret.ipynb
Conclusion	model_final_conclusion.ipynb

Data Preparation

1) Tax Parcel data (Appendix D) – aggregation of the original datasets (1 year per file for years 2002-2014, excluding 2003 due to incomplete fields) has 23,942,414 parcels with an average of 71.46 features (excluding 2003 due to incomplete fields). Following guidance from the reviewed literature it was decided to perform dimension reduction to keep only the features that will control for lakefront property price independent of water quality. In addition, temporal duration will be reduced. The most recent year (2014) which has 2,116,399 parcels with 74 features will undergo subsetting, with the resulting subset cleansed and used for model training.

    a. Feature subsetting (Appendix E)

        i. Based on the study by (Boyle, 1998) a subset of features that have the highest likelihood of providing predictive power, along with features that serve as join variables, have been selected.

            1. The result is a reduction of 63 features, with 11 of the original 74 remaining. Reference Data Dictionary 1.

            2. Observations with null values on critical features were deleted at this time, namely centroid_long. Since it is used for joins having null values is not possible. Only a small number of observations needed to be deleted, six in total.

    b. Data cleansing (Appendix E)

        i. Given that the focus of this study is valuation of residential lakefront property, it is necessary to remove property types that will skew valuations. As such, the feature DWELL_TYPE (dwelling type) is used as a filter to exclude records that are not of a residential variety, preventing commercial properties with differing profiles from influencing the model.

1. With this filter applied the number of parcel observations was reduced from 2,116,402 to 771,149.

ii. The next data cleansing step was removing sites that had an EMV_TOTAL (estimated market value of land + building) value of less than $25,000. This was done to control for anomalies such as foreclosure, condemned, and abandoned properties. In order to use this column as a greater than or equal to filter predicate, the data type had to be converted to integer.

1. With this filter applied the number of parcel observations was reduced from 771,149 to 762,188.

iii. In order to ensure a significant model, features with excessive null values are to be identified in Figure 5. By first filtering observations that are not applicable to analysis in prior steps there was a byproduct benefit of reducing the number of null observations. This means the subsequent deletion of rows with null values will have less of an impact on integrity of the dataset since fewer pertinent observations will be lost.

Figure 5. Null values by tax parcel variables

```
      (762188, 11)
ACRES_DEED            0
BASEMENT         192074
COUNTY_ID             0
```

```
DWELL_TYPE            0
EMV_TOTAL             0
FIN_SQ_FT             0
GARAGE           194576
GARAGESQFT       216308
YEAR_BUILT            0
centroid_lat        0
centroid_long       1
```

iv. A commonly followed guideline is that if a column has a greater frequency than 70% null values the column will be dropped. Otherwise, assuming that there is a sizeable sample and that the frequency is less than 30%, the rows with null values will be deleted. Given that none of the columns have greater than 70% null value (greatest is GARAGESQFT at 28.3%), combined with a large sample size (762,188) and less than 30% frequency, the approach to be taken is deleting records rather than dropping columns. The additional consideration is that the dimension space is not large (11), meaning additional reduction is likely to compromise the model's goodness-of-fit. The resulting deletion of rows with null values produced a dataset with 358,683 observations across 11 features. Large enough to move forward with model training without a significant concern regarding sample size.

v. Following deletion of records with null column values, it was discovered that GARAGESQFT had values of 'None' in addition to the expected numeric entries. These records were also deleted to ensure integrity of the feature.

1. The resulting dataset now has 331,191 observations across 11 features.

vi. The final cleansing operation was converting GARAGESQFT data type from object to float64. This is possible now that the data has been cleansed by removing non-numeric NaN and 'None' values.

vii. Given that the 2014 sample size has shrunk from 2,116,402 to 331,191 properties, and those have yet to be narrowed down to lakefront specific properties, it was decided to append tax parcel data from 2012 and 2013 in order to create a larger sample size.

1. The resulting raw dataset now has 6,324,826 observations across 11 features.

viii. Preparing the combined 2012-2014 tax parcel years entails running through the cleansing steps above that were originally run against the single 2014 year, with two additional preparation steps. Previously properties on the low value side of the curve were deleted. Conversely, this step eliminates properties on the high side of the curve that would also adversely affect the outlier sensitive OLS regression analysis. In total 1,547 properties having EMV_TOTAL values greater than $1,500,000 were removed. The second additional preparation step is dropping GARAGE and BASEMENT due to a combination of a high number of nulls along with low predictive potential given that GARAGESQFT will be retained.

1. The resulting cleansed dataset now has 1,548,188 observations across 9 features covering years 2012-2014.

c. Transformed file stored in S3

Table 5. Data Dictionary 1: Tax parcel

Variable	Data Type	Description
ACRES_DEED	float64	The deeded acreage of the parcel. (numeric field with two decimal places
COUNTY_ID	int64	Three digit FIPS and State standard county code
DWELL_TYPE	object	Type of dwelling (e.g. single family, duplex, etc.)
EMV_TOTAL	int64	Total estimated market value (land + building)
FIN_SQ_FT	int64	Finished square footage
GARAGESQFT	float64	Square footage of garage
YEAR_BUILT	int64	Year the building was built
centroid_lat	float64	Latitude of the Parcel centroid
centroid_long	float64	Longitude of the Parcel centroid

2) Lake Monitoring data (Appendix F) – original dataset (1999-2014) has 48,257 site observations across 33 features. The featureset dimensions have been reduced to those with predictive power in terms of lakefront property price. In addition, temporal duration has been shortened to align with Tax Parcel years. The resulting three most recent years (2012-2014) with a subset of features will be used for modeling.

a. Feature subsetting (Appendix G)

   i. Based on the study by (Boyle, 1998) a subset of features that have the highest likelihood of providing predictive power, along with features that serve as join variables, have been selected.

      1. The result is a reduction of 22 features, with 11 of the original 33 remaining. Reference Data Dictionary 2.

b. Data cleansing (Appendix G)

   i. Based on visual inspection of the Figure 3 map it appears that sites outside of the 7 county region were included in the lake monitoring data. Upon performing a list of values for the COUNTY feature, this suspicion was confirmed in that sites across 11 counties are included in the data. As such, the feature COUNTY is used as a filter to exclude records that are not in the 7 county region.

      1. With this filter applied the number of site observations was reduced from 48,258 to 47,511.

   ii. During inspection of the data it became apparent that the DNR_ID_Site_Number column has a trailing [-01, -02, -03] on all eight-digit DNR Site IDs. This will prevent the lake monitoring dataset from joining with the water proximity reference table since the xref table makes use of the standard eight-digit convention. In order to bring lake monitoring data into compliance with that convention, the trailing [-01, -02, -03] were stripped from all DNR_ID_Site_Number fields.

iii. In order to ensure a significant model, features with excessive null values are to be identified in Figure 6.

Figure 6. Null values by lake monitoring variables

```
                (47511, 11)
LAKE_NAME                              0
COUNTY                                 0
DNR_ID_Site_Number                     0
START_DATE                             0
Seasonal_Lake_Grade_RESULT         44470
Physical_Condition_RESULT          17875
Recreational_Suitability_RESULT    18635
Secchi_Depth_RESULT                13051
Total_Phosphorus_RESULT             4514
longitude                              0
latitude                               0
```

iv. The guideline for dropping/deleting columns/rows does not apply in the case of this dataset. The reason this is the case is that it is a timeseries, with site observations scattered across a timeline. As a result, it is not the individual row for a given point in time that matters as much as the aggregate of points describing the lake over a range of time.

v. Exploring the data revealed that Seasonal_Lake_Grade_Result has null values for 93.6% of records. Upon first review this appears to be a column that should be dropped. It is not until further inspection that it becomes clear that this feature is sparse by design since a value is only assigned annually apart from the other water quality attribute tracking. That is to say, all other water quality attributes are NaN when a seasonal lake grade is recorded, and inversely seasonal lake grade is NaN when other water quality attributes are recorded. This is by design, as shown in Figure 7.

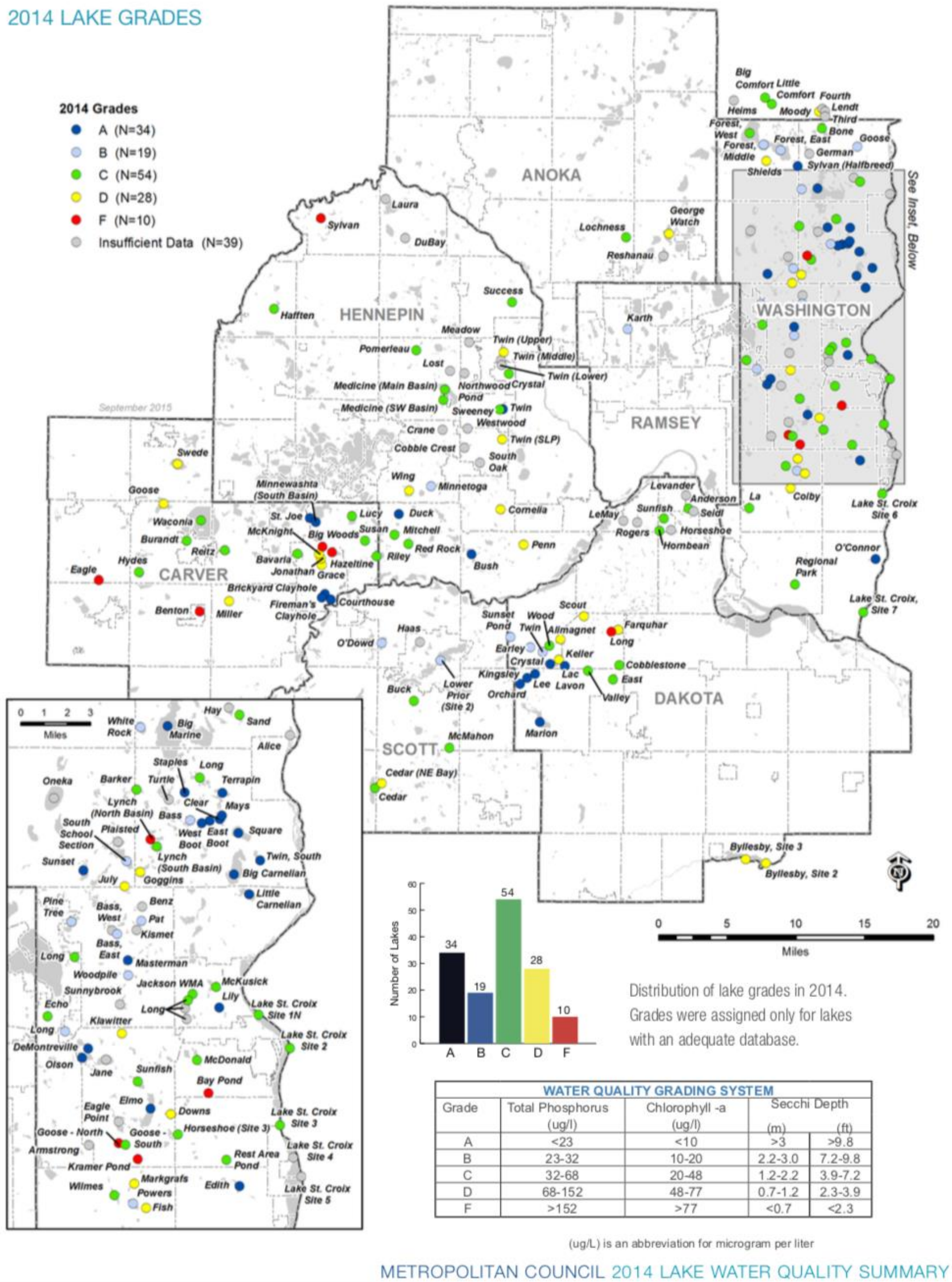
Figure 7. Showing inverse relationship

Out[10]:

LAKE_NAME	COUNTY	DNR_ID_Site_Number	START_DATE	Seasonal_Lake_Grade_RESULT	Physical_Condition_RESULT	Recreational_Suitability_RESULT	Secchi_Depth_RESULT	Total_Phosphorus_RESULT
Acorn Lake	Washington	82010200	2006-04-16	NaN	1.0	5.0	1.00	0.156
Acorn Lake	Washington	82010200	2006-05-01	2.0	NaN	NaN	NaN	NaN
Acorn Lake	Washington	82010200	2006-05-02	NaN	1.0	5.0	0.66	0.107
Acorn Lake	Washington	82010200	2006-05-16	NaN	2.0	5.0	0.66	0.141
Acorn Lake	Washington	82010200	2006-05-30	NaN	2.0	5.0	0.50	0.029

vi. Each lake is assigned a lake grade using an A through F grading system (coded 4-0 respectively) as originally developed by Council staff in 1989 (Metropolitan Council, 2014). The objective of the lake grade system is to provide a tool for assessing lakes on a regional basis. The grading system allows comparisons of lake water quality across the metro area yet is understandable to the public and non-technical audiences. The grading system uses percentile ranges of the summer-time (May-September) average values for three water quality indicators: total phosphorus, chlorophyll-a, and Secchi depth. Total phosphorus is a key nutrient measure; chlorophyll-a is a measure of algal abundance; and Secchi depth is a measure of water clarity. The lake's water quality grade is calculated as the average grade for the three individual parameter grades. Only lakes with a sufficient quantity of data are assigned a lake grade, as shown in Figure 8, along with the criteria used for the grading system.

Figure 8. Lakes assigned grades within the 7 county region (Metropolitan Council, 2014)



WATER QUALITY GRADING SYSTEM

Grade	Total Phosphorus (ug/l)	Chlorophyll -a (ug/l)	Secchi Depth (m)	(ft)
A	<23	<10	>3	>9.8
B	23-32	10-20	2.2-3.0	7.2-9.8
C	32-68	20-48	1.2-2.2	3.9-7.2
D	68-152	48-77	0.7-1.2	2.3-3.9
F	>152	>77	<0.7	<2.3

c. Timeseries data indexing and aggregation (Appendix G)

    i. In the case of timeseries, measurements recorded of the features composing lake water quality have a diminishing value as they move backwards in time. The reason is that numerous lurking variables outside of the recorded dataset change in a way that influences property prices. A prime example is that prior to the selected 2012-2014 year range there was a massive housing crisis. If data had been used from this period it would give the impression that home prices were plummeting, regardless of the state of water quality. The best way to account for this is to index, aggregate, and average lake monitoring observations on a recent range of dates that aligns with tax parcel date ranges. This also mitigates inflation as a material factor given the small range of time.

    ii. The first step in this process is to convert the START_DATE data type from string to datetime, including a datetime index that replaces the standard dataframe index.

    iii. The following step is to eliminate observations prior to 2012, creating a range from 2012-2014. This date range aligns with the tax parcel dataset.

        1. With this filter applied the number of site observations was reduced from 47,511 to 6,420.

i. The next step is to index, aggregate, and average the features. This was accomplished using the GroupBy function within Pandas on the DNR_ID_Site_Number attribute, followed by a mean operation. By collapsing the timeseries observations in this way the sparse data has been significantly reduced while maintaining the features deemed to have predictive potential. Thus, avoiding the need to compromise the dataset by either dropping columns or deleting a large number of rows. It will also allow for a cleaner join operation with the Water Proximity reference table.

    1. Following the GroupBy and mean operations the resulting dataset now has 174 observations across 8 features.

iv. As suspected, the frequency of rows with Seasonal_Lake_Grade_Result null values dropped significantly once the features were aggregated based on the DNR_ID_Site_Number attribute. Null frequency was reduced from 93.6% to 18.4%, meaning it now falls well within the guideline criteria stating not to drop a column that has less than 70% null values. Additionally, with a frequency of less than 30% it is generally deemed safe to delete the rows that have null values without a significant risk of introducing sampling bias. The resulting dataset has 141 averaged observations across 8 features.

d. Transformed file stored in S3

Table 6. Data Dictionary 2: Lake monitoring

Variable	Data Type	Description
DNR_ID_Site_Number	float64	The eight-digit Hydrological Unit Code (HUC) used by Department of Natural Resources (DNR) lake basin identification subdivision number
Seasonal_Lake_Grade_RESULT	float64	4 =A 3 =B 2 =C 1 =D 0 =F
Physical_Condition_RESULT	float64	1 = Crystal Clear 2 = Some Algae Present 3 = Definite Algal Presence 4 = High Algal Color 5 = Severe Algal Bloom
Recreational_Suitability_RESULT	float64	1 = Beautiful 2 = Minor Aesthetic Problem 3 = Swimming Impaired 4 = No Swimming, Boating OK 5 = No Aesthetics Possible
Secchi_Depth_RESULT	float64	The Secchi disk is a measure of water clarity
Total_Phosphorus_RESULT	float64	Under natural conditions phosphorus (P) is typically scarce in water. Human activities, however, have resulted in excessive loading of phosphorus into many freshwater systems. This can cause water pollution by promoting excessive algae growth, particularly in lakes.
longitude	float64	Coordinates of the Site
latitude	float64	Coordinates of the Site

3) Water proximity reference table (Appendix H) – the original dataset has 2,688,766 observations across 10 features. The featureset dimensions have been reduced to those required for merging Tax Parcel data with the Lake Monitoring data.

    a. Feature subsetting (Appendix I)

i. A subset of features required for merging Tax Parcel data with Lake Monitoring data were selected.

1. The result is elimination of 6 features, with 4 of the original 10 remaining. Reference Data Dictionary 3.

b. Data cleansing

i. None required, all 2,688,766 observations retained

c. Transformed file stored in S3

Table 7. Data Dictionary 3: Water proximity

Variable	Data Type	Description
Monit_SITE_CODE	float64	The eight-digit Hydrological Unit Code (HUC) used by Department of Natural Resources (DNR) lake basin identification subdivision number (if one exists) or whole lake number.
centroid_long	float64	Longitude of the Parcel centroid truncated to 5 digits. If using as a Key keep it as a String/Text type since different systems handling floating points differently.
centroid_lat	float64	Latitude of the Parcel centroid truncated to 5 digits. If using as a Key keep it as a String/Text type since different systems handling floating points differently.
Distance_Parcel_Lake_meters	float64	Distance of the parcel centroid in meters to the nearest lake containing a monitoring site. To keep compute time low assumption was that most tax parcels are comparatively small to the size of a monitored lake and so a simple point to nearest lake edge was calculated rather than edge to edge.

4) Master data (Appendix J) – The merged dataset is the culmination of extensive data interrogation and preparation steps undertaken on the Tax

Parcel, Lake Monitoring, and Water Proximity datasets. It represents 27,446 properties comprised of 18 attributes on 59 distinct lakes observed from 2012 until 2014. Reference Data Dictionary 4 for the final attributes.

a. Feature engineering

    i. Tax Parcel data was merged with the Water Proximity xreference table on ['centroid_long', 'centroid_lat'] columns.

    ii. The resulting dataset was then merged with Lake Monitoring water quality data on Monit_SITE_CODE against DNR_ID_Site_Number. Resulting in the Master dataset.

        1. The master dataset has 310,389 observations across 19 features.

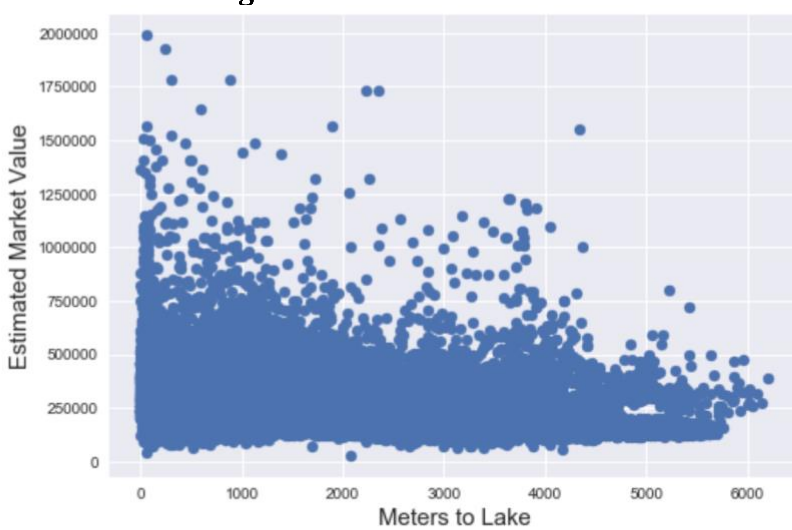
    iii. After inspecting the data, it became clear that there were both redundant columns and duplicate observations with each property having a record for each of the three years. First the redundant columns were dropped, then the annual observations were combined into a single record with an average of EMV_TOTAL computed as an integer (EMV_TOTAL being the only column with differing values between years).

        1. The resulting dataset has 84,200 observations across 16 features.

    iv. Now that Tax Parcel is merged with Water Proximity it is possible to perform the final preparation task. That task is filtering parcel observations to those with a proximity of less than 200 meters to the lake. This will allow us to model the influence water quality has on property valuations with proximity to lakefront rather than on general property valuations.

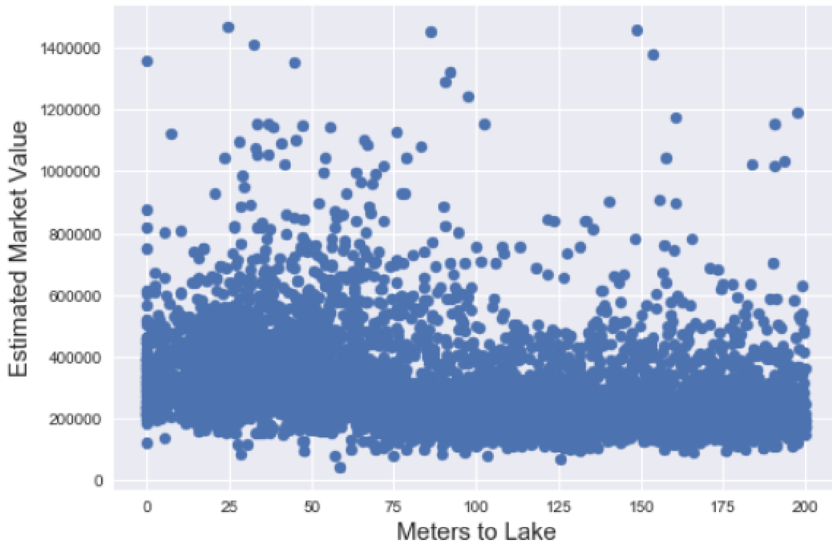
1. The filter distance value was selected based on the visualization showing how estimated market value of properties drops off at an increased pace as the distance moves beyond ~200 meters. This represents the disassociation of water and property value, making the observations beyond ~200 meters likely to skew model fitting. Shown in Figure 9 and 10.

Figure 9. EMV vs Meters to Lake



2. With this filter applied the number of property observations was reduced from 84,200 to 5,777.

Figure 10. EMV vs Meters to Lake < 200 meters



b. Transformed file stored in [S3]()

Table 8. Data Dictionary 4: Master

Variable	Data Type	Description
ACRES_DEED	float64	The deeded acreage of the parcel. (numeric field with two decimal places
COUNTY_ID	int64	Three digit FIPS and State standard county code
DWELL_TYPE	object	Type of dwelling (e.g. single family, duplex, etc.)
EMV_TOTAL	int64	Total estimated market value (land + building)
FIN_SQ_FT	int64	Finished square footage
GARAGESQFT	int64	Square footage of garage
YEAR_BUILT	int64	Year the building was built
centroid_lat	float64	Latitude of the Parcel centroid
centroid_long	float64	Longitude of the Parcel centroid
Distance_Parcel_Lake_meters	float64	Distance of the parcel centroid in meters to the nearest lake containing a monitoring site.
DNR_ID_Site_Number	int64	The eight-digit Hydrological Unit Code (HUC) used by Department of Natural Resources (DNR) lake basin identification subdivision number
Seasonal_Lake_Grade_RESULT	float64	4 =A 3 =B 2 =C 1 =D 0 =F
Physical_Condition_RESULT	float64	1 = Crystal Clear 2 = Some Algae Present 3 = Definite Algal Presence 4 = High Algal Color 5 = Severe Algal Bloom
Recreational_Suitability_RESULT	float64	1 = Beautiful 2 = Minor Aesthetic Problem 3 = Swimming Impaired 4 = No Swimming, Boating OK 5 = No Aesthetics Possible
Secchi_Depth_RESULT	float64	The Secchi disk is a measure of water clarity
Total_Phosphorus_RESULT	float64	Excessive loading of phosphorus can cause water pollution by promoting excessive algae growth, particularly in lakes

Modeling

Validation

Appendix K. – JupyterLab notebook

The first step in validation is to ensure that the features and observations are optimal. As part of creating the master dataset, certain features were required for filtering and joins that no longer serve a purpose. Subsetting occurred on 3 such features ['centroid_lat', 'centroid_long', 'DWELL_TYPE'], resulting in 13 features remaining.

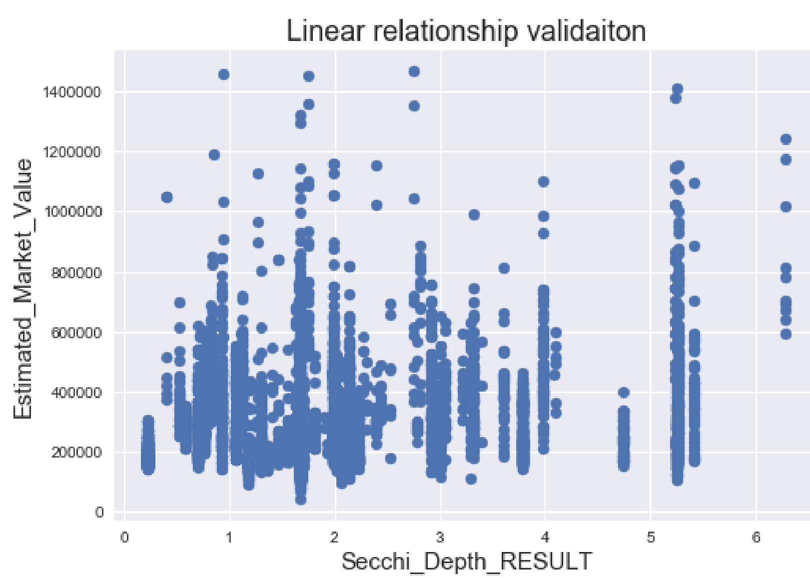
Validation of the 5 core linear regression assumptions (see Appendix B for working definitions)

1) Linear relationship exists
    a. During visual inspection additional items requiring cleaning were identified
        i. 193 observations existed that had a 0 value for GARAGESQFT. These were deleted, resulting in 5,584 observations.
        ii. ACRES_DEED was highly sparse and therefore dropped. At first glance this may seem like a significant loss of model predictive performance, however given that the acres measurement was not specific to shoreline, rather overall plot acreage, it had greatly reduced value as a predictor. Resulting in 12 features.

iii. Physical_Condition_RESULT and Recreational_Suitability_RESULT exhibited poor linear characteristics and were dropped. Resulting in 10 features.

b. The conclusion is that visual inspection shows clear linear relationships between predictor variables and the EMV_TOTAL (estimated market value) response variable. Below is one such example with a positive relationship, as Secchi_Depth_RESULT increases, clusters of EMV_TOTAL follow suit. This is as expected given that the Secchi depth represents clarity of water, a greater Secchi value indicates clearer water. The full set of plots are recorded in the supporting JupyterLab notebook – model_validation.ipynb with an example shown in Figure 11.

Figure 11. Linear relationship EMV vs. Secchi Depth

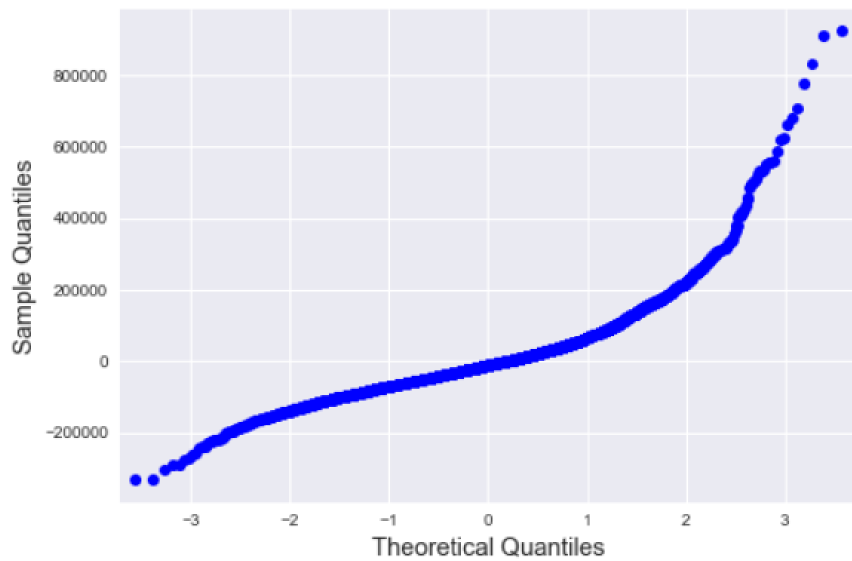


2) Multivariate normality exists

a. An OLS model was fitted using statsmodels in Python for assessing additional assumptions, such as multivariate normality.

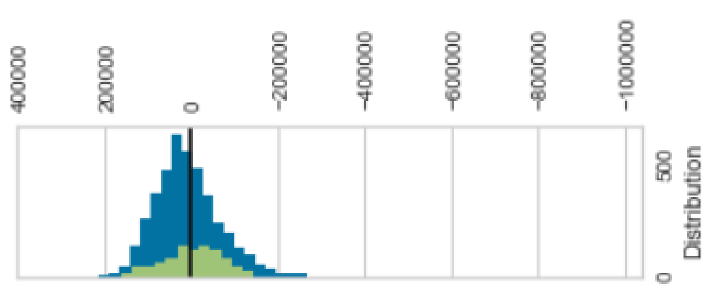
   i. The Q-Q plot shows a violation of multivariate normality.

Figure 12. Q-Q plot of residuals



   ii. The Histogram from the OLS sklearn fitted model shows residual distribution that is close to being normally distributed. Both for the train (blue) and test (green) datasets

Figure 13. Histogram residual distribution



b. The conclusion is that while this assumption is not satisfied, the tests give varying subjective impressions of how far off it actually is. Given that multivariate normality is important for making predictions on future observations, while HPM is concerned with extracting coefficients from current observations, the violation is not a threat to compromise findings of this thesis.

3) Multicollinearity does not exist

    a. Variance inflation factor (VIF) was used to check for multicollinearity. A value of 1 indicates a lack of correlation between features.

        i. The first test showed high correlation on Seasonal_Lake_Grade_RESULT and Secchi_Depth_RESULT, as well as COUNTY_ID and DNR_ID_Site_Number.

        Figure 14. VIF values for variables pre-fix

```
const                           10227.930990
COUNTY_ID                        2551.802210
FIN_SQ_FT                           2.652140
GARAGESQFT                          1.282843
YEAR_BUILT                          1.234743
DNR_ID_Site_Number               2550.591319
Seasonal_Lake_Grade_RESULT          8.704986
Secchi_Depth_RESULT                 5.490671
Total_Phosphorus_RESULT             2.788617
EMV_TOTAL                           2.845488
Distance_Parcel_Lake_meters         1.227400
dtype: float64
```

            1. By dropping the non-technical Seasonal_Lake_Grade_RESULT, the VIF scores fall into a range indicating an acceptable level of correlation. This is explained by the fact that the parameters used to calculate lake grade are a composite of Secchi depth, total phosphorus, and chlorophyll-a.

2. COUNTY_ID and DNR_ID_Site_Number are correlated since sites are numbered in ranges based on county. Even though there are 75 distinct sites (lakes), they cluster into 3 groups (by county). Neither of these variables will be included in model formulas, alleviating any multicollinearity concerns. There are now 9 features.

Figure 15. VIF values for variables post-fix

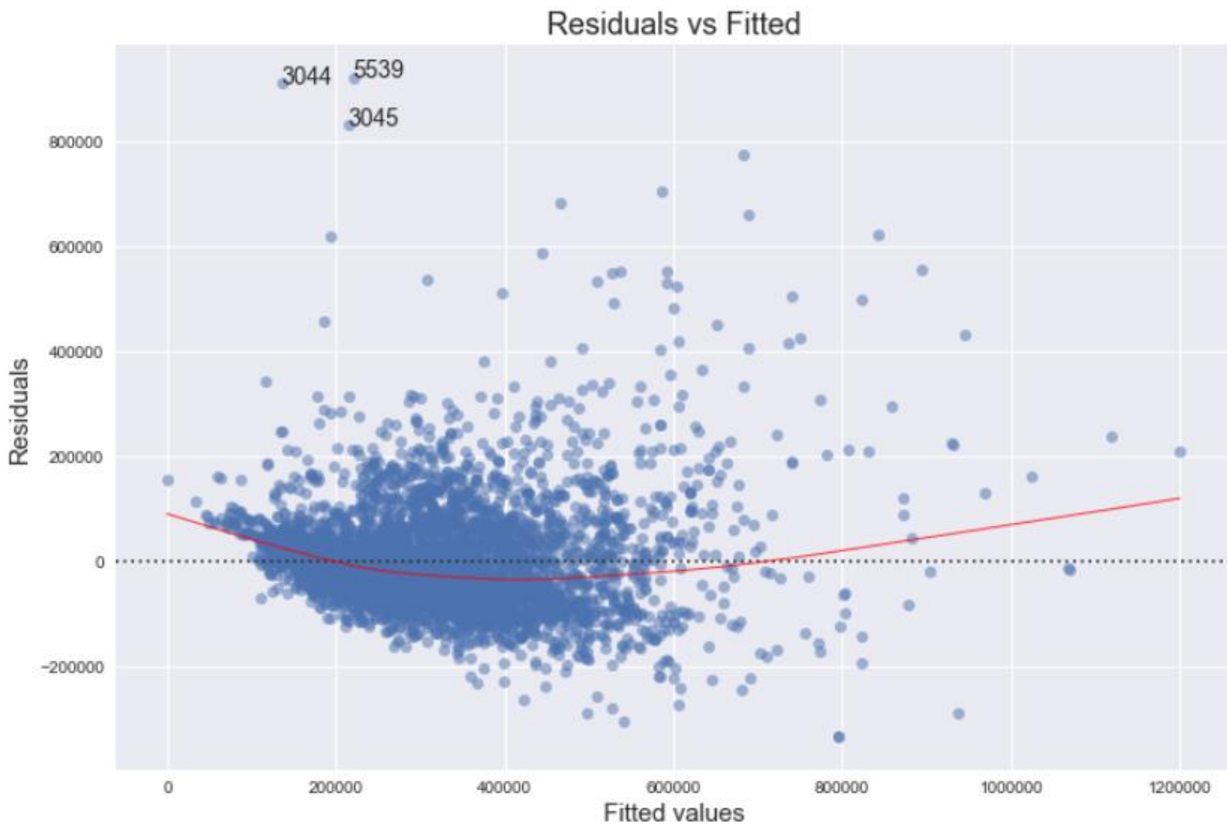
```
const                          10227.849860
COUNTY_ID                       2551.655363
FIN_SQ_FT                          2.651073
GARAGESQFT                         1.280128
YEAR_BUILT                         1.232673
DNR_ID_Site_Number              2550.416834
Secchi_Depth_RESULT                1.487639
Total_Phosphorus_RESULT            1.443684
EMV_TOTAL                          2.840603
Distance_Parcel_Lake_meters        1.224536
dtype: float64
```

b. The conclusion is that multicollinearity does not exist when lake grade is separated from Secchi depth and total phosphorus. This aligns with the approach this thesis will take for model interpretation, the approach being that the technical and non-technical features that are used to measure water quality will be fitted as separate models.

4) Heteroscedasticity does not exist

a. Assessing the Residuals versus Fitted scatterplot in Figure 16 for the OLS model it is clear that heteroscedasticity does exist.

Figure 16. Scatterplot OLS residuals vs. Fitted



i. Breusch-Pagan test supports the scatterplot with a p-value less than 0.05, meaning that we reject the null hypothesis of no evidence of heteroscedasticity.

ii. Suggested courses of action to convert a heteroscedastic to homoscedastic model are using Generalized Least Squares, Weighted Least Squares, log transformation of response and predictor variables to compress the scales of measurement, or polynomial transformations.

b. The conclusion is that since heteroscedasticity exists each of these methods for remedying it will be attempted during the training phase. It is worth noting that in the use case of pricing homes it is not unusual to see higher variance amongst more expensive homes, resulting in the familiar left-to-right cone shape on the scatterplot. The reason is that customizations reign supreme in high-end homes, meaning quantified features such as finished square feet contribute less proportionately to the overall value. This results in a greater likelihood of misspecification.

5) Autocorrelation does not exist
    a. The Durbin-Watson test indicates that autocorrelation exists for all models with scores in the sub 1.2 range, 2.0 being an ideal score. The lone exception is the GLSAR estimator which had a score of 2.298
    b. Autocorrelation is especially important in analysis of timeseries datasets. Given that the element of time has been flattened by aggregation and averaging of observations, autocorrelation is not a primary concern.
6) Transformed file shown in Data Dictionary 5 is stored in S3

Table 9. Data Dictionary 5: Master post Validation

Variable	Data Type	Description
COUNTY_ID	int64	Three digit FIPS and State standard county code
EMV_TOTAL	int64	Total estimated market value (land + building)
FIN_SQ_FT	int64	Finished square footage
GARAGESQFT	int64	Square footage of garage
YEAR_BUILT	int64	Year the building was built
DNR_ID_Site_Number	int64	The eight-digit Hydrological Unit Code (HUC) used by Department of Natural Resources (DNR) lake basin identification subdivision number
Distance_Parcel_Lake_meters	float64	Distance of the parcel centroid in meters to the nearest lake containing a monitoring site.
Secchi_Depth_RESULT	float64	The Secchi disk is a measure of water clarity
Total_Phosphorus_RESULT	float64	Under natural conditions phosphorus (P) is typically scarce in water. Human activities, however, have resulted in excessive loading of phosphorus into many freshwater systems. This can cause water pollution by promoting excessive algae growth, particularly in lakes.

Training

Appendix L. – JupyterLab notebook

A vigorous regimen for training was followed using both of the top regression libraries available in Python, statsmodels and scikit-learn. The methodology for using each was mirrored in that the dataset was split into a response and a predictor dataset. Additionally, in the first scenario for models using logarithmic transformation, the log operation was performed against the response variable. Then each was split again between train and test, 80% / 20% respectively.

In the second scenario for models using logarithmic transformation, the log operation was performed against the entire dataset. This resulted in 100 observations having the predictor variable Distance_Parcel_Lake_meters approach infinity. These were converted to null values and then the observations were deleted, which is why the count of Table 11 and Table 12 differ. The resulting dataset was then split into a response and predictor, each with log values. Finally, each was split again between train and test, 80% / 20% respectively.

Shuffle was turned off when performing the splits to ensure all models were fitting on the same property observations. Although this results in less accurate fits, it provides a consistent split of observations for performance comparisons, and ultimately selecting a model. For all models, COUNTY_ID and DNR_ID_Site_Number were omitted.

Table 10. Training Libraries

library	estimator
statsmodels	OLS
statsmodels	WLS
statsmodels	GLSAR
statsmodels	OLS with log transformation of response
statsmodels	OLS with log transformation of response & predictors
sklearn	OLS
sklearn	OLS with log transformation of response
sklearn	LASSOCV
sklearn	KNN
sklearn	OLS with nonlinear polynomial terms

Table 11. Train / Test datasets – original and with log transformed Response variable

dataset	count	features
df_master_y_train	4467	EMV_TOTAL
df_master_y_test	1117	EMV_TOTAL
df_master_y_train_log	4467	ln(EMV_TOTAL)
df_master_y_test_log	1117	ln(EMV_TOTAL)
df_master_X_train	4467	constant, FIN_SQ_FT, GARAGESQFT, YEAR_BUILT, Secchi_Depth_RESULT, Total_Phosphorus_RESULT, Distance_Parcel_Lake_meters
df_master_X_test	1117	constant, FIN_SQ_FT, GARAGESQFT, YEAR_BUILT, Secchi_Depth_RESULT, Total_Phosphorus_RESULT, Distance_Parcel_Lake_meters

Table 12. Train / Test datasets – log transformed Response and Predictor variables

dataset	count	features
df_master_X_train_log	4386	ln(EMV_TOTAL)
df_master_X_test_log	1097	ln(EMV_TOTAL)
df_master_y_train_log	4386	constant, ln(FIN_SQ_FT, GARAGESQFT, YEAR_BUILT, Secchi_Depth_RESULT, Total_Phosphorus_RESULT, Distance_Parcel_Lake_meters)
df_master_y_test_log	1097	constant, ln(FIN_SQ_FT, GARAGESQFT, YEAR_BUILT, Secchi_Depth_RESULT, Total_Phosphorus_RESULT, Distance_Parcel_Lake_meters)

Evaluation

Appendix L. – JupyterLab notebook

Now that the models have been fit against identical train/test datasets it is possible to evaluate their performance in Table 13. This will allow for the selection of the highest performing model, which will then be used for additional in-depth analysis interpreting the effects of water quality predictors on lakefront property prices. (see Appendix C for working definitions)

Table 13. Model Performance tests

library	estimator	RMSE	R-squared	F-statistic	Prob (F-statistic)
statsmodels	OLS	$96,734.03	0.577	1063	0.00
statsmodels	WLS	$97,549.01	0.570	806.2	0.00
statsmodels	**GLSAR**	**$89,761.59**	**0.636**	**1007**	**0.00**
statsmodels	ln_y(OLS)	0.30	0.595	1309	0.00
statsmodels	ln_X_y(OLS)	0.28	0.652	988.1	0.00
sklearn	OLS	$96,734.03	0.577	1063	0.00
sklearn	ln_y(OLS)	0.30	0.595		
sklearn	LASSOCV	$96,629.37	0.578		
sklearn	KNN(w=distance)	$92,354.07	0.614		
sklearn	quadratic(OLS)	$105,641.37	0.495		

A combination of R-squared and RMSE will be used for selecting the model with the best fit. R-squared will act as the first filter for ranking, followed by RMSE as the deciding factor. This will allow for models with log transformation to be compared to non-transformed models. In the case of statsmodels, the summary output R-squared value will not be used since it is based on fit of the training data, not test. Instead, a manually calculated R-squared value based on test data fit is used across all models to create an equivalent point of comparison.

Out of the 10 models only 3 have R-squared values above 0.600, statsmodels.GLSAR, statsmodels.ln_X_y(OLS), and KNN. KNN is disqualified immediately based on not having interpretable coefficients. This is a requirement for the hedonic pricing method given that its purpose is the determination of not just the overall model’s significance, but individual water quality feature influence. KNN was included out of curiosity rather than as a true candidate for the HPM exercise. Given that the 2 remaining models have R-squared values within a few points of one another, RMSE is the next determining factor.

This however is not a viable measurement since it is a comparison between log transformed vs. non-transformed models. Which leaves the tie-breaker to be Occam's razor, i.e. whichever is simpler. This makes the GLSAR estimator the winner given its readily interpretable coefficients, it having the lowest RMSE score, and it being significant at the 0.01 level. Indicating rejection of the null hypothesis stating that there is no relationship between water quality and lakefront property valuation.

It is not surprising that this model performed above the others given they are all standalone models whereas GLSAR is an ensemble technique that utilizes boosting. Meaning it is actually a sequential series of two models that produce the end estimates; first OLS is fitted, then as an output the rho score representing lag in the residuals is input into GLSAR, improving its ability to fit the data. The benefit does not stop at goodness-of-fit, GLSAR uses a generalized least square algorithm that improves handling of autocorrelation and heteroscedasticity.

The intent of the thesis is to not only select the highest performing model with significance, but also one where the significance findings are trustworthy. When validating the five assumptions of linear regression there were two assumptions that raised red flags which were not able to be resolved through pre-model techniques, namely autocorrelation and heteroscedasticity. In the case of autocorrelation, the lone exception to this was the GLSAR model which had a Durbin-Watson score of 2.298. Indicating that the generalized least squared estimator handled autocorrelation in an acceptable manner given proximity to the ideal score of 2. As a point of comparison, the OLS model had a Durbin-Watson score of 1.165, indicating the presence of autocorrelation.

In the case of heteroscedasticity, the generalized least squared estimator utilizes heteroscedasticity-robust standard errors. This is possible due to the identification of the feasible GLS estimator taken as an input from the OLS regression residuals that were used to model the relation of errors with independent variables. Ultimately, this allowed for the correct specification of the form of variance to be used in GLSAR modeling, mitigating concerns over heteroscedasticity.

In closing, for the next modeling step, statsmodels.GLSAR will be used, providing the best performance combined with the highest level of statistical trustworthiness.

Interpretation of Findings

Appendix M. – JupyterLab notebook

With a thorough analysis completed against a well-prepared set of data it is now possible to sift through the modeling output to gain an understanding of what is occurring in regard to water quality influence on lakefront property prices. The first step is establishing a data profile using descriptive statistics.

Table 14. Descriptive statistics of Master

Feature	Count	Median	Mean	Std. Dev.	Min.	Max.
COUNTY_ID	3	NA	NA	NA	37	163
FIN_SQ_FT	2084	2,093	2,221	942	0	9,040
GARAGESQFT	749	576	661	298	12	3,903
YEAR_BUILT	126	1978	1976	21.4	1870	2013
DNR_ID_Site_Number	75	NA	NA	NA	19000600	82051400
Secchi_Depth_RESULT	75	1.669	2.001	1.221	.223	6.276
Total_Phosphorus_RESULT	75	.034	.072	.085	.010	.697
EMV_TOTAL	3877	269,583	309,768	153,735	43,166	1,466,200
Distance_Parcel_Lake_meters	5358	96.3	97.3	58.7	0	199.9
Seasonal_Lake_Grade_RESULT	21	2.5	2.4	1.1	0	4

The key takeaways from the descriptive statistics is that the master dataset covers 5,584 properties residing on 75 lakes with highly dispersed feature values. The statistics cover both the technical and non-technical water quality features, which will be assessed as two separate models. Both FIN_SQ_FT and Distance_Parcel_Lake_meters have minimum values of 0, this prompted further exploration to ensure there isn't a sizeable set of properties with these values. The investigation showed only one property had FIN_SQ_FT of 0, and that its other features contained normal looking values. Therefore, it was kept. Distance_Parcel_Lake_meters has 100 observations with values of 0, however this may very well be a valid value for properties that are directly on the shore. Once again, no observations were deleted.

As part of interpreting the GLSAR model an additional dataset was constructed, shown in Table 15 (hosted on S3). The intent of this dataset is to test non-technical predictor variables, namely seasonal lake grade. Comparing models trained on technical vs. non-technical predictor variables is a concept championed by the (Bin & Czajkowski, 2013) study. This study also pointed out controlling for inflation and boundaries of the body of water included in the study. For this thesis inflation was accounted for by using a small range of time with aggregation. Boundaries are addressed given the bodies of water are small metro lakes. The challenge with boundaries in the authors studies came into play with large lakes such as Lake Michigan, or open bays such as Chesapeake Bay.

Table 15. Train / Test datasets – Non-technical

dataset	count	features
df_master_grade_y_train	4467	EMV_TOTAL
df_master_grade_y_test	1117	EMV_TOTAL
df_master_grade_X_train	4467	constant, FIN_SQ_FT, GARAGESQFT, YEAR_BUILT, Seasonal_Lake_Grade_RESULT, Distance_Parcel_Lake_meters
df_master_grade_X_test	1117	constant, FIN_SQ_FT, GARAGESQFT, YEAR_BUILT, Seasonal_Lake_Grade_RESULT, Distance_Parcel_Lake_meters

Table 16 and 17 contain the trained coefficients for both the technical and non-technical GLSAR models, respectively.

Table 16. GLSAR model with Technical water quality features

variable	coef	std err
constant	-1,024,000	158,000
FIN_SQ_FT	92.7107	1.726
GARAGESQFT	94.8084	4.882
YEAR_BUILT	539.0569	80.706
Secchi_Depth_RESULT	18,560	1674.264
Total_Phosphorus_RESULT	184,600	25,400
Distance_Parcel_Lake_meters	-559.8175	23.9

Table 17. GLSAR model with Non-Technical water quality features

variable	coef	std err
constant	-949,500	159,000
FIN_SQ_FT	91.7202	1.726
GARAGESQFT	90.0347	4.888
YEAR_BUILT	526.0648	81.165
Seasonal_Lake_Grade_RESULT	4,186.8806	1,568.689
Distance_Parcel_Lake_meters	-572.4190	23.916

In the next step a comparison between the model estimates is provided in Table 18. This will help in establishing an understanding of which model, technical or non-technical, is more effective in explaining water quality influence on property valuations.

Table 18. GLSAR technical vs non-technical model performance

library	estimator	RMSE	R-squared	F-statistic	Prob (F-statistic)
statsmodels	GLSAR – tech	$89,761.59	0.636	1007	0.00
statsmodels	GLSAR – nontech	$85,128.02	0.672	1168	0.00

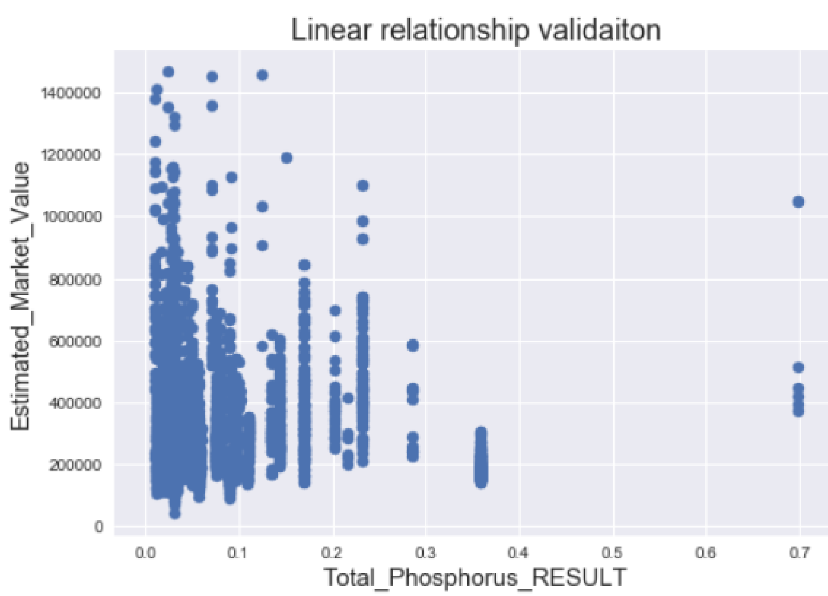
There is a slight edge with the non-technical seasonal lake grade model providing superior estimates, evident from a lower RMSE with a higher R-squared score. Both models are statistically significant at the 0.01 level. Furthermore, individually all predictor variables are statistically significant at the 0.01 level.

Now that coefficients have been captured and the models are confirmed as being statistically significant, we are able to estimate the dollar impact that unit changes in the water quality predictor variables have on lakefront property prices. The following interpretation covers both technical and non-technical models.

- Seasonal lake grade: for every 1-point grade increase (on a scale of 0-4), property value within 200 meters of a lake increases by $4,186.88.
- Distance to the lake: given property within 200 meters of a lake, every 1 meter further from the lake a property is located results in that property value decreasing by
    - $539.06 in the technical model.
    - $572.42 in the non-technical model.

- Secchi depth: for every 1-foot improvement in clarity of water, property value within 200 meters of a lake increases by $18,560.
- Total phosphorus: for every increase of 1 milligram per liter of water, property value within 200 meters of a lake increases by $184,600. Given that this would be an unrealistically large increase in phosphorus it is more plausible to interpret it as for every increase of 0.1 milligram per liter of water, property value within 200 meters of a lake increases by $18,460.
    - This coefficient is both counterintuitive and contradicts visual inspection of the negative relationship that exists between phosphorus levels and property prices. As phosphorus levels increase lake water becomes impaired with algae blooms. Common sense tells us this will result in a decrease in market value of properties located within 200 meters of that lake, as depicted in Figure 17.

Figure 17. Negative relationship of EMV vs. TP



- Diving into this contradiction further requires creation of a dataset that isolates Total_Phosphorus_RESULT as a water quality predictor of property values. It will be paired with FIN_SQ_FT which will act as a control for the influence house characteristics have on property prices. This dataset is described in Table 19.

Table 19. Train / Test datasets – Total Phosphorus water quality predictor

dataset	count	features
df_master_tp_y_train	4467	EMV_TOTAL
df_master_tp_y_test	1117	EMV_TOTAL
df_master_tp_X_train	4467	constant, FIN_SQ_FT, Total_Phosphorus_RESULT
df_master_tp_X_test	1117	constant, FIN_SQ_FT, Total_Phosphorus_RESULT

A GLSAR model was fit against this dataset, however that model also failed to capture the negative relationship that exists between total phosphorus and property values within 200 meters of a lake. As mentioned previously, GLSAR takes input from a fitted OLS model, it was within this OLS model that a negative relationship was identified. In that model the individual Total_Phosphorus_RESULT variable is not statistically significant at the 0.05 level, see Table 20. The model as a whole however is statistically significant at the 0.01 level, see Table 21. Setting statistical insignificance aside for a theoretical interpretation, the coefficient indicates that for every total phosphorus increase of 0.1 milligram per liter of water, property value within 200 meters of a lake decreases by $1,749.

Table 20. OLS model with only Total Phosphorus as a water quality predictor

variable	coef	std err	P>|t|
constant	73,730	5576.152	0.00
FIN_SQ_FT	103.6839	1.995	0.00
Total_Phosphorus_RESULT	-17,490	20,100	0.383

Table 21. OLS model performance with only TP as a water quality predictor

library	estimator	RMSE	R-squared	F-statistic	Prob (F-statistic)
statsmodels	OLS	$104,985.94	0.501	1423	0.00

Interpretation of the findings for seasonal lake grade, distance to the lake, and Secchi depth all pass statistical tests as well as common sense review. When it comes to total phosphorus the picture is not so clear, given the contradicting indicators any interpretations made based on this feature are inconclusive. It is appropriate to reassess technical water quality modeling with the omission of total phosphorus given its ambiguity, shown in Table 22. This analysis was completed in model ‘housing+water_secchi’ as part of the JupyterLab notebook identified in Appendix N.

Table 22. GLSAR model with only Secchi as a Technical water quality feature

variable	coef	std err
constant	-989,200	159,000
FIN_SQ_FT	91.8700	1.725
GARAGESQFT	92.6389	4.890
YEAR_BUILT	537.6672	80.999
Secchi_Depth_RESULT	12,160	1478.615
Distance_Parcel_Lake_meters	-564.8928	23.933

Table 23. GLSAR technical vs Secchi only technical model performance

library	estimator	RMSE	R-squared	F-statistic	Prob (F-statistic)
statsmodels	GLSAR – tech(tp+secchi)	$89,761.59	0.636	1007	0.00
statsmodels	GLSAR – tech(secchi)	$86,850.83	0.659	1187	0.00

Table 23 shows that the technical model omitting Total_Phosphorus_RESULT has superior performance. This is not surprising given the mixed signals that it produced. Under the new model shown in Table 22, interpretation for Secchi depth is that for every 1-foot improvement in clarity of water, property value within 200 meters of a lake increases by $12,160. Interpretation for distance to the lake is that given property within 200 meters of a lake, for every one meter further away from the lake a property is located its value decreases by $564.89.

While coefficients provide valuable specific insights regarding interpretation of individual property attributes, hedonic pricing methods gives the bigger picture. In order to produce the HPM findings a series of additional steps are undertaken, resulting in Table 24.

- housing_baseline model median predicted EMV_TOTAL
    - GLSAR estimator is used to fit a model against a training dataset that has a subset of features representing properties without water attributes.
        - FIN_SQ_FT, GARAGESQFT, YEAR_BUILT
    - That model is then used to make predictions against the entire set of observations. The output is a median property price that will be used as a baseline for comparing the effect of water quality on property estimates. Median versus mean is considered a best practice in real estate estimates given the presence of outliers.
- housing+water_secchi model median predicted EMV_TOTAL

- GLSAR estimator is used to fit a model against a training dataset that has a subset of features representing lakefront properties with technical water quality attributes.
    - FIN_SQ_FT, GARAGESQFT, YEAR_BUILT, Secchi_Depth_RESULT, Distance_Parcel_Lake_meters
- That model is then used to make predictions against the entire set of observations. The output is a median property price and a median Secchi score that will be used as a delta for comparing the effect of water quality on property estimates against the baseline.
- (housing+water_secchi) – housing_baseline = HPM value derived from water quality using technical measurements
    - Inflation is mitigated given aggregation and averaging of EMV_TOTAL across a small window of time.
- These results show the lift of water quality based on a median Secchi score from the entire population of observations. In order to determine if lift changes given higher or lower quality of water, filters were applied against the Secchi_Depth_RESULT variable to produce a top 21% and bottom 21%. The model was then used to make EMV_Total predictions against the subsets of data representing high and low water quality.
    - It is important to note that while Secchi disk readings can be a predictor of water quality, they are not a conclusive indicator of poor versus good water quality. For example, there may the

presence of cyanobacteria aphanizomenon which is almost translucent, this would allow greater water clarity (higher Secchi score), but lower water quality.

- The top and bottom output are a median property price and a median Secchi score for comparison to the median property price and median Secchi score for the entire population of observations. Additionally, a sum of property prices is used to determine the impact on potential property tax revenue.

- housing+water_grade model median predicted EMV_TOTAL
    - GLSAR estimator is used to fit a model against a training dataset that has a subset of features representing lakefront properties with non-technical water quality attributes.
        - FIN_SQ_FT, GARAGESQFT, YEAR_BUILT, Seasonal_Lake_Grade_RESULT, Distance_Parcel_Lake_meters
    - That model is then used to make predictions against the entire set of observations. The output is a median property price and a median Secchi score that will be used as a delta for comparing the effect of water quality on property estimates against the baseline.
    - (housing+water_grade) – housing_baseline = HPM value derived from water quality using non-technical measurements
        - Inflation is mitigated given aggregation and averaging of EMV_TOTAL across a small window of time.

- These results show the lift of water quality based on a median lake grade from the entire population of observations. In order to determine if lift changes given higher or lower quality of water, filters were applied against the Seasonal_Lake_Grade_RESULT variable to produce a top 21% and bottom 22%. The model was then used to make EMV_Total predictions against the subsets of data representing high and low water quality.
- The top and bottom output are a median property price and a median lake grade for comparison to the median property price and median lake grade for the entire population of observations. Additionally, a sum of property prices is used to determine the impact on potential property tax revenue.

Interpretation of the results in Table 24 shows that by both technical and non-technical measures simply having proximity to water provides a negligible level of lift, 1% and 1.1% respectively. However, when isolated to the top 21% bodies of water based on technical and non-technical measures, lift jumps 15.3% and 9.2%. Conversely, having proximity to the bottom 21% and 22% (technical and non-technical) causes property value to decrease by 1.8% and 2.1%, respectively.

When filtering property observations to represent either the top or bottom sections by water quality, it is possible that house specific predictor variables could be an uncontrolled influencer on the increase in property price. A plausible scenario that may exist is one where larger houses are built on more desirable lakes, bringing into question how much the increase in value should be attributed to water versus the house itself. In order to account for the housing factor, median finished square feet of houses were tracked in the model subsets. This reveals that the difference in size of house between the baseline and filtered subset was small; i.e. a max of a 2.9% increase from baseline to top for Secchi, and a max of a 3% increase from baseline to top for lake grade. In the case of properties with the bottom Secchi depths, median finished square feet was actually greater than the baseline and the top subsets. Based on this it is reasonable to infer that house size was an immaterial factor in explaining property value lift, leaving water quality as the likely influencing factor.

The final calculation was that of a sum of property prices for each model subset. These numbers are used to represent the gain or loss of tax revenue potentially realized from higher or lower property values that result from changing levels of water quality. Specifically, this serves as a comparison between the best-case total taxable property value and the worst-case total taxable property value, based on the difference between the top and bottom model subsets.

Table 24. HPM water quality effect on property value (distance < 200m)

model	filter	rank% of total	median $ predicted property price	median water quality	median square feet	lift $ above baseline	lift % above baseline	sum $ predicted property prices	water quality impact on property value for tax revenue
housing_baseline	**none**	**100%**	**$290,113.66**	**NA**	**2093’**	**$0**	**0%**	**$1,695,935,170**	
housing+water_secchi	none	100%	$293,004.28	1.67’	2093’	$2,890.62	1%	$1,719,111,386	
top_housing+water_secchi	>=2.5’	top 21%	$334,410.64	3.79’	2153’	$44,296.98	15.3%	$407,744,026	$58,460,213
bot_housing+water_secchi	<=1.1’	bot 21%	$284,914.54	0.70’	2189’	-$5,199.12	-1.8%	$349,283,813	$58,460,213
housing+water_grade	none	100%	$293,247.45	2.56	2093’	$3,133.79	1.1%	$1,717,921,701	
top_housing+water_grade	>=3.5	top 21%	$316,967.89	4.0	2156’	$26,854.23	9.2%	$385,500,816	$13,188,738
bot_housing+water_grade	<=1.5	bot 22%	$284,143.44	1.0	2101’	-$5,970.22	-2.1%	$372,312,078	$13,188,738

Appendix N. – JupyterLab notebook

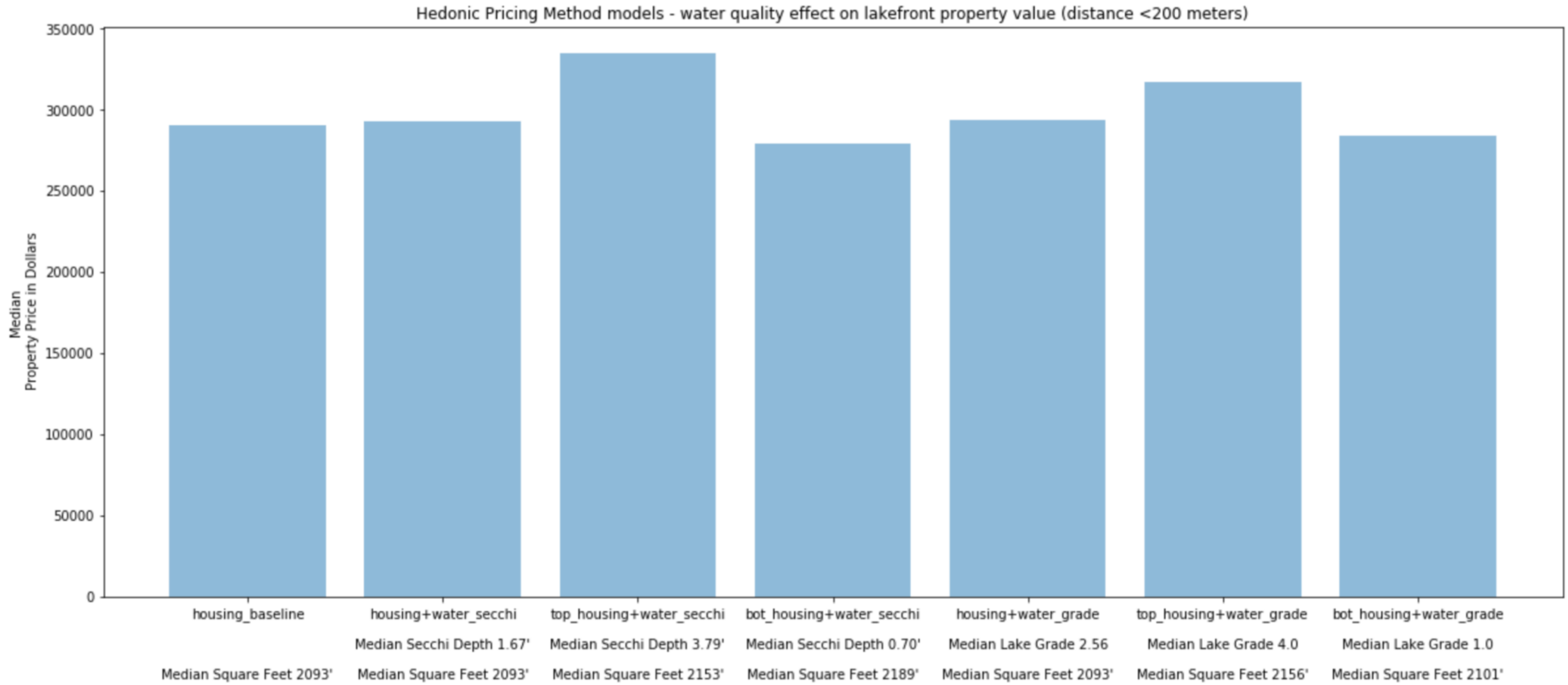
Conclusions

The major finding of this research is that water quality, measured either by technical or non-technical features, significantly affects residential lakefront property prices within three counties of the Twin Cities region. Namely, Dakota County, Ramsey County, and Washington County. The original thesis set out to identify the relationship between water quality and lakefront property prices in all seven of the Twin Cities counties. However, due to sparse data the list of counties was paired down to the aforementioned three. The relationship uncovered was positive, with an inflection point occurring as water quality reaches an impaired state. Meaning that being located within 200 meters of an impaired lake actually decreased property value below what a property would be valued at with no water whatsoever. Conversely, proximity to a lake with a high quality of water provides meaningful lift to the property value.

Technical (Secchi - water clarity) had a slight edge over non-technical (lake grade) measurements in terms of lift on property values, as shown previously in Table 24. This reveals that consumers make a determination on whether to pay a premium or not to pay a premium for living near water based on what meets the eye. That is to say, superior Secchi depth indicate a lake has higher water clarity but is not a true measure of the health of the water. Whereas lake grade is a composite view of Secchi depth, total phosphorus, and chlorophyll-a. If consumers were evaluating health of the lake versus simply making a determination based on how clear the water is, lake grade would exhibit a higher degree of lift than Secchi depth alone. The findings of this study indicate that is not the case.

Ultimately this reveals that the clarity of water, or lack thereof, is a significant factor in consumers determination of what premium a body of water demands in the housing market. A point succinctly illustrated by Figure 18.

Figure 18. HPM technical vs. non-technical models for three county Twin Cities region



Recommendations

Based on the findings of this research it is evident that clarity of water has a significant influence on estimated market value, which in turn is used for calculating property taxes. Outside of the soft benefits clean lakes provide their communities, there is a case to be made for the hard benefit of increased or decreased tax revenue resulting from improvement or degradation of water quality. The property value spread between the top 21% and bottom 21% properties based on Secchi depth is $58,460,213, as shown in Table 24. That is a significant difference in potential tax revenue that could be realized as a result of restoring and maintaining water clarity. One such improvement project was undertaken on the Minneapolis Chain of Lakes in the heart of Minnesota (Huser, Brezonik, & Newman, 2011). The project entailed using aluminum sulfate as a chemical treatment to reduce total phosphorus levels, which in turn significantly improved water clarity as represented by increased Secchi depth. Additionally, more cost-effective best management practices exist, such as establishing wetland barriers that act as filters for runoff entering waterways.

It is the recommendation of this thesis that such projects be explored for additional high value lakes within the Twin Cities, pending cost/benefit assessments. In conjunction, improving the consistency of data collection mechanisms by the MCES, MDNR, and MPCA will allow for more accurate modeling. This is easier said than done, when it comes to the accuracy of property price modeling it is known to be a challenging undertaking, exhibited by Zillow's willingness to host a competition with a $1 million grand-prize (Zillow Prize, 2019).

References

Bin, O., & Czajkowski, J. (2013). The impact of technical and non-technical measures of water quality on coastal waterfront property values in South Florida. *Marine Resource Economics, 28*, 43-63.

Boyle, K. J. (1998). *Lakefront Property Owners' Economic Demand for Water Clarity in Maine Lakes.* University of Maine. Orono: Maine Agriculture and Forest Experiment Station.

Census 2010 Geography - Blocks, Block Groups, Tracts, TAZs, Counties, County Subdivisions and Water. (n.d.). Retrieved from Minnesota Geospatial Commons: https://gisdata.mn.gov/dataset/us-mn-state-metc-society-census2010tiger

David, E. L. (1968). Lakeshore Property Values: A Guide to Public Investment in Recreation. *Water Resources Research, 4*, 697-707.

EIMS. (n.d.). Retrieved from Metropolitan Council: https://eims.metc.state.mn.us/AdvancedSearch

Freeman, A. M. (1993). *The measurement of environmental and resource values: theory and practice.* Washington, DC: Resources For the Future Press.

Horsch, E. J., & Lewis, D. J. (2009). The effects of aquatic invasive species on property values: evidence from a quasi-experiment. *Land Economics, 85*, 391-409.

Huser, B., Brezonik, P., & Newman, R. (2011). Effects of alum treatment on water quality andsediment in the Minneapolis Chain of Lakes,Minnesota, USA. *Lake and Reservoir Management*, 220-228. Retrieved from https://doi.org/10.1080/07438141.2011.601400

Kashian, R., & Kasper, J. (2010). *Tainter Lake & Lake Menomin: The Impact of Diminishing Water Quality on Value.* University of Wisconsin - Whitewater. Whitewater: UWW - Fiscal and Economic Research Center.

Krysel, C., Boyer, E. M., Parson, C., & Welle, P. (2003). *Lakeshore Property Values and Water Quality: Evidence from Property Sales in the Mississippi Headwaters Region.* Mississippi Headwaters Board and Bemidji State University. Bemidji: Legislative Commission on Minnesota Resources.

Leggett, C. G., & Bockstael, N. E. (2000). Evidence of effects of water quality on residential land prices. *Journal of Environmental Economics and Management, 39*, 121-144.

MCES Lake Monitoring Sites. (n.d.). Retrieved from Minnesota Geospatial Commons: https://gisdata.mn.gov/dataset/us-mn-state-metc-env-mces-lake-monitoring-sites

MetroGIS. (n.d.). Retrieved from Minnesota Geospatial Commons: https://gisdata.mn.gov/organization/us-mn-state-metrogis?q=MetroGIS+Regional+Parcel+Data&sort=title_string+asc

Metropolitan Council. (2014). *Water Quality Management - Lake Monitoring & Assessment.* Retrieved from Metropolitan Council - Wastewater & Water: https://eims.metc.state.mn.us/Documents/GetDocument/918

Michael, H. J., Boyle, K. J., & Bouchard, R. (2000). Does the Measurement of Environmental Quality Affect Implicit Prices Estimated from Hedonic Models? *Land Economics, 76*(2), 283-298.

MinneMUDAC: Dive into Water (Data). (2016, November 5). Retrieved from MinneAnalytics: http://minneanalytics.org/event/analytics-student-event-dive-into-water-data/

Netusil, N. R., Kincaid, M., & Chang, H. (2014). Valuing water quality in urban watersheds: A comparative analysis of Johnson Creek, Oregon, and Burnt Bridge Creek, Washington. *Water Resources Research, 50*(5), 4254-4268.

Nichols, W. J. (2014). *Blue Mind: The Surprising Science That Shows How Being Near, in, on, or under Water Can Make You Happier, Healthier, More Connected, and Better at What You Do.* New York, NY, USA: Little, Brown and Company.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy, 82*, 34-55.

Small, K. A. (1974). Air pollution and property values: further comment. *Rec. Econom. Statist, 57*, 105-107.

Walsh, P. (2009). *Hedonic Property Value Modeling of Water Quality, Lake Proximity, and Spatial Dependence in Central Florida.* Department of Economics. Orlando: University of Central Florida.

Zillow Prize. (2019). Retrieved from Zillow: https://www.zillow.com/promo/zillow-prize/

Appendices

Appendix A. Metropolitan Council EIMS query result dataset

Provided in the appendices due to the URL being too long to be embedded as a hyperlink

http://eims.metc.state.mn.us/Download?startDate=01-01-1999&endDate=12-31-
2014&parameterIds=48;50;358;361;362;1019&counties=Anoka;Carver;Dakota;Hennepi
n;Ramsey;Scott;Washington&siteIds=82004900-01;82005204-01;02013300-
01;02009100-01;02004200-01;10000500-01;70012001-01;82016300-01;02000600-
01;27018402-
02;MI0394;MI0251;MI0143;UM8156;UM8477;VR0156;SC0003;UM8267;UM8391;U
M8310;NM0018;MI0035;BE0020;SC0233;MI0085;SA0082;UM8716;VR0206;RUM00
06;CR0009;BL0035;CA0017;SA0001;BE0050;VR0219;SC0234;UM8128;UM8218;UM
8368;UM8178;EA0008;SA0016;RI0013;MH0017;UMSP8208;SA0051;SD0003;BELT0
005;CR0006;WR0047;BA0022;BR0003;BS0019;VA0010;VR0020;FC0002;RUM0007;
WI0010;SW0015;SI0001;CWS0203;CM0030;SI0007;PU0039;02065400-01;82011602-
01;27009800-01;19002400-01;10004800-01;10008500-01;70002600-01;70002600-
02;70007200-01;19007500-01;62006900-01;70007600-01;82012200-01;27010000-
01;19007600-01;10006300-01;10002800-01;10005900-01;10010900-01;10010500-
01;27005700-01;19003100-01;82008900-01;10005400-01;02000700-01;10007000-
01;82003000-01;82033400-01;82009700-01;19044600-01;10008800-01;27007800-
01;19002500-01;19006500-01;27063400-01;02008000-01;82015900-01;82008000-
01;82002000-01;10002900-01;70002100-01;19003200-01;10008900-01;82013300-
01;19003700-01;82008200-01;10003100-01;19002300-01;82011000-01;82009400-
01;27013700-01;02008400-01;82005400-01;19002100-01;82009002-01;10006600-
01;10005200-01;02000900-01;82016200-01;10012100-01;19009500-01;27007600-
01;82008700-01;27010700-01;10008000-01;82009200-01;10000900-01;10004200-
01;02000400-01;27007000-01;27013400-01;10000600-01;10007800-01;10006800-
01;27006500-01;82014000-01;62005400-01;10011000-01;10010700-01;82001000-
01;82002300-01;19004100-01;10005800-01;19002601-01;82007700-01;19034800-
01;82006500-01;02007900-01;02004500-01;82015300-01;10009300-01;10001600-
01;82002100-01;19002800-01;02000500-01;27004202-01;10008600-01;82004600-
01;70006900-01;19003300-01;19002700-01;70006100-01;27004700-01;10001900-
01;10006900-01;10010300-01;82012000-01;82012500-01;82012600-01;82031800-
01;82015100-01;02013000-01;10001300-01;82036800-01;27009100-01;10009500-
01;10001400-01;10010800-01;10008400-01;82012400-01;82012300-01;27003501-
02;27003502-01;27104501-01;82010100-01;82030500-01;82013200-01;27003501-
01;27069300-01;70005000-01;82015900-03;27010400-02;82015900-02;82000400-
01;70009100-01;10000200-01;82051400-01;19002000-01;82007400-03;19005000-
01;82010900-01;62005800-01;82011800-01;82013700-01;82000200-01;82004200-
02;82004200-01;82014800-01;70005400-01;27017500-01;10012700-01;19008000-
01;19019800-01;27011601-01;82010300-01;62007200-01;27008800-01;10001100-
01;27010400-01;19045600-01;19045100-01;62004901-01;10000700-01;19002200-

01;27011102-01;82005204-02;27005300-01;27010200-01;10010400-01;27065600-01;82013000-01;82001900-01;27066100-01;82002602-01;82002500-01;82010600-01;19008200-01;82007200-01;27062700-01;10000900-02;10022600-01;10022500-01;10021800-01;10021700-01;10021600-01;82004800-01;82003400-01;82006400-01;82006800-01;82005202-01;82000900-01;02002200-01;70009500-01;62003600-01;82005900-01;82006700-01;19034900-01;19002900-01;70001800-01;27071100-01;82010400-01;27009400-01;82009300-01;82010700-01;19008800-01;19009400-01;82004500-01;82003300-01;82002100-03;82002100-02;82000200-02;27005800-01;27002802-01;27018401-01;82030800-01;82004400-01;82001100-01;82010800-01;82001502-01;27000400-01;10001000-01;19005100-01;62004902-02;19003000-01;19004700-01;82001400-01;70007400-01;82001700-01;27015300-01;27009200-01;27003400-01;27008600-01;82046200-01;82001800-01;02058500-01;62018700-01;62007100-01;27004201-01;82005300-01;82002800-01;82005600-01;82003500-01;27008902-01;82001600-01;27004203-01;82036500-02;82036500-01;82011900-01;82013401-01;82003600-01;82039900-01;82010200-01;82031300-01;82011700-01;27064500-01;82003100-01;82007600-01;02002600-01;02003400-01;27017100-01;82013500-01;82011200-02;82011200-01;82048800-01;70009800-01;27019900-01;82048200-01;82011302-02;82000500-02;27014702-01;27016900-01;27011900-01;82011301-01;99001184-02;82009900-02;82007400-02;82009900-01;82000500-01;27017901-01;27011101-01;27009300-01;62000100-01;82003800-01;02007200-01;10001500-01;82030100-01;27016500-01;27012700-01;10005300-01;10009300-03;10009300-02;62000200-02;62000200-01;10001800-01;02008100-01;27012500-01;82006200-01;27008200-01;82001501-01;27002900-01;27067000-01;27067500-01;02007100-01;27001500-01;27008600-02;19011700-01;27008500-01;82014700-01;62020400-01;82005202-02;70007200-02;10001200-01;10004100-01;27018700-01;70007600-02;82000100-08;82000100-03;82000100-02;82000100-07;82000100-06;82000100-05;82000100-04;82000100-01;27012900-01;70009100-02;27014600-01;19004200-02;19004200-01;19000600-03;27007300-01;27012300-01;27006700-01;27010300-01;10024900-01;82028700-01;27073400-01;70006500-01;27006900-01;70007800-01;UM8139;UM8455;UM8344;02001300-01;99002487-01;10000200-04&format=tsv

Appendix B. Detailed information regarding validation of linear regression assumptions

- Validation of linear regression assumptions
  - Linear relationship
    - Description
      - Linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects
    - Why it’s important
      - If there is no linear relationship it will not be possible to calculate coefficients that fit the relationship using a linear regression model
    - Catch

        - Scatterplot of dependent variable versus independent variable
    - Handle
        - Either change the scope of observations, transform the variables, or use a non-linear model (e.g. nonlinear least squares, generalized additive model, generalized linear models)
- Multivariate normality
    - Description
        - Multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed
    - Why it’s important
        - When the residuals are not normally distributed, then the hypothesis that they are a random dataset, takes the value NO
        - This means that the regression model does not explain all trends in the dataset
    - Catch
        - Checked with a histogram or a Q-Q plot
    - Handle
        - If the residuals are not normally distributed a non-linear transformation of the response variable can be attempted (e.g. log transformation)
        - It is possible the data does not have a linear relationship and that a non-linear model is required to capture the trend
- Multicollinearity
    - Description
        - A phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy
    - Why it’s a problem
        - It violates the assumption that independent variables are not too highly correlated with each other
        - It can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
    - Catch
        - The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model
            - A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables.
    - Handle

        - Remove predictor variables with high VIF values
        - Centering the data (that is deducting the mean of the variable from each score)
        - Use Partial Least Squares Regression (PLS) or Principal Components Analysis
  - Autocorrelation
    - Description
      - It is a characteristic of data in which the correlation between the values of the same variables is based on related objects. It occurs when the residuals are not independent from each other, and is most commonly found in timeseries regressions
    - Why it's a problem
      - It violates the assumption of error term independence, which underlies linear regression models
      - It is a problem because its presence means that useful information is missing from the model. Such information might explain the movement in the dependent variable more accurately
    - Catch
      - Scatterplot of the residuals versus the time measurement for that observation
      - Durbin-Watson test
        - Since d is approximately equal to 2(1 − r), where r is the sample autocorrelation of the residuals, d = 2 indicates no autocorrelation. The value of d always lies between 0 and 4. If the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation.
    - Handle
      - Investigate the omission of a key predictor variable
      - If this does not aid in reducing AR, a more involved variable transformation is required. Three such methods are:
        - Cochrane-Orcutt Procedure
        - Hildreth-Lu Procedure
        - First Difference Procedure
  - Heteroscedasticity
    - Description
      - It is present when the size of the error term differs across values of an independent variable
    - Why it's a problem
      - It violates the assumption of homoscedasticity (meaning "same variance") of residuals that is central to linear regression models

    - The ordinary least squares estimators are still linear and unbiased, but are no longer best; there is another form that produces smaller variances
    - The standard errors are biased. Because the standard error is central to conducting significance tests and calculating confidence intervals, biased standard errors lead to incorrect conclusions about the significance of the regression coefficients
      - Meaning tests tends to produce p-values that are smaller than they should be. This effect occurs because heteroscedasticity increases the variance of the coefficient estimates but the OLS procedure does not detect this increase. Consequently, OLS calculates the t-values and F-values using an underestimated amount of variance. This problem can lead you to conclude that a model term is statistically significant when it is actually not significant.
  - Catch
    - Scatterplot of the least squares residuals versus fitted values
    - Scale-Location plot
    - Durbin-Watson test
      - Since d is approximately equal to $2(1 - r)$, where r is the sample autocorrelation of the residuals, $d = 2$ indicates no autocorrelation. The value of d always lies between 0 and 4. If the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation.
  - Handle
    - Use generalized least squares to obtain our parameter estimates. This involves keeping the functional form intact, but transforming the model in such a way that it becomes a heteroscedastic model to a homoscedastic one
    - Introduce weighted least squares to the regression. WLS assigns each data point a weight based on the variance of its fitted value. The idea is to give small weights to observations associated with higher variances to shrink their squared residuals. Weighted regression minimizes the sum of the weighted squared residuals.

Appendix C. Detailed information regarding evaluation of model performance

- Evaluation of model performance
  - Measures of predictive power
    - Hypothesis
      - H_0 – null hypothesis

    - Water quality does not affect purchase price for residential lakefront properties within the seven county Twin Cities region
  - H_A – alternate hypothesis
    - Water quality affects purchase price for residential lakefront properties within the seven county Twin Cities region
- Significance test
  - In order to make the determination to either reject or fail to reject the null hypothesis a level of significance is established as a statistical threshold
  - Significance level
    - p-value
      - Represents the probability that the coefficient is actually zero
    - 95% confidence interval does not include zero
      - $\alpha = 0.05$
      - If p-value < α
      - Reject the null
      - There is a relationship between water quality and lakefront property valuation
    - 95% confidence interval does include zero
      - $\alpha = 0.05$
      - If p-value > α
      - Fail to reject the null
      - There is no relationship between water quality and lakefront property valuation
- R-squared
  - R-squared is the proportion of variance explained
  - It is the proportion of variance in the observed data that is explained by the model, or the reduction in error over the null model
    - The null model just predicts the mean of the observed response, and thus it has an intercept and no slope
  - R-squared is between 0 and 1
    - Higher values are better because it means that more variance is explained by the model
- Root mean squared error
  - RMSE is the square root of the mean square error. It is the most easily interpreted statistic since it has the same units as the quantity plotted on the vertical axis.
  - Key point: The RMSE is thus the distance, on average, of a data point from the fitted line, measured along a vertical line.

        - The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient. One can compare the RMSE to observed variation in measurements of a typical point
        -
    - F-statistic
        - The F-Statistic: Variation Between Sample Means / Variation Within the Samples. The F-statistic is the test statistic for F-tests. In general, an F-statistic is a ratio of two quantities that are expected to be roughly equal under the null hypothesis, which produces an F-statistic of approximately 1.

Appendix D. Tax Parcel data – Counts

1) Tax Parcel data row count (2002-2014, excluding 2003 due to incomplete fields)
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis

python taxparcel_count_rows.py
{'2009_metro_tax_parcels.txt': 2088219, '2007_metro_tax_parcels.txt': 2025484, '2011_metro_tax_parcels.txt': 2100698, '2005_metro_tax_parcels.txt': 1968481, '2013_metro_tax_parcels.txt': 2106917, '2014_metro_tax_parcels.txt': 2116402, '2002_metro_tax_parcels.txt': 1236819, '2008_metro_tax_parcels.txt': 2109722, '2010_metro_tax_parcels.txt': 2097801, '2006_metro_tax_parcels.txt': 2007924, '2012_metro_tax_parcels.txt': 2101529, '2004_metro_tax_parcels.txt': 1982418}
23942414

2) Tax Parcel data field count (2002-2014, excluding 2003 due to incomplete fields)
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis

python taxparcel_count_avg_fields.py
{'2009_metro_tax_parcels.txt': 72, '2007_metro_tax_parcels.txt': 72, '2011_metro_tax_parcels.txt': 70, '2005_metro_tax_parcels.txt': 70, '2013_metro_tax_parcels.txt': 70, '2014_metro_tax_parcels.txt': 74, '2002_metro_tax_parcels.txt': 75, '2008_metro_tax_parcels.txt': 72, '2010_metro_tax_parcels.txt': 71, '2006_metro_tax_parcels.txt': 70, '2012_metro_tax_parcels.txt': 70, '2004_metro_tax_parcels.txt': 71}
71.41666666666667

Appendix E. Tax Parcel data – Data Preparation

1) Sample of fields and a parcel observation from the original dataset. Also used to identify the delimiter.

head -n 2 2014_metro_tax_parcels.txt
ACRES_DEED|ACRES_POLY|AGPRE_ENRD|AGPRE_EXPD|AG_PRESERV|BASEMENT|BLDG_NUM|BLOCK|CITY|CITY_USPS|COOLING|COUNTY_ID|DWELL_TYPE|EMV_BLDG|EMV_LAND|EMV_TOTAL|FIN_SQ_FT|GARAGE|GARAGESQFT|GREEN_ACRE|HEATING|HOMESTEAD|HOME_STYLE|LANDMARK|LOT|MULTI

_USES|NUM_UNITS|OPEN_SPACE|OWNER_MORE|OWNER_NAME|OWN_ADD_L1|OWN_ADD_L2|OWN_ADD_L3|PARC_CODE|PIN|PLAT_NAME|PREFIXTYPE|PREFIX_DIR|SALE_DATE|SALE_VALUE|SCHOOL_DST|SPEC_ASSES|STREETNAME|STREETTYPE|SUFFIX_DIR|Shape_Area|Shape_Le_1|Shape_Leng|Shape_STAr|Shape_STLe|TAX_ADD_L1|TAX_ADD_L2|TAX_ADD_L3|TAX_CAPAC|TAX_EXEMPT|TAX_NAME|TORRENS|TOTAL_TAX|UNIT_INFO|USE1_DESC|USE2_DESC|USE3_DESC|USE4_DESC|WSHD_DIST|XUSE1_DESC|XUSE2_DESC|XUSE3_DESC|XUSE4_DESC|YEAR_BUILT|Year|ZIP|ZIP4|centroid_lat|centroid_long
30.0|26.71|||N||||SAINT FRANCIS|ELK RIVER|N|003|AGRICULTURAL|0.0|132100.0|132100.0|0.0|N|||N|N|||||0|||JONES TRUSTEE RAYMOND|23725 NACRE ST NW|ELK RIVER|MN, 55330|0.0|003-333425210001|||||0.0|15|0.0||||||||23725 NACRE ST NW|ELK RIVER|MN, 55330|1080.0|N|JONES TRUSTEE RAYMOND||1671.0||AGRICULTURAL||||UPPER RUM RIVER WMO|||||0.0|2014|55330||45.39768|-93.46219

2) Feature subsetting steps
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
        i. data_1_2_3_subset.ipynb
3) Data cleansing step
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
        i. taxparcel_1_clean.ipynb

Appendix F. Lake Monitoring data – Counts

1) Lake Monitoring data row count (1999-2014)
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis

python lakemonitoring_count_rows.py
{'mces_lakes_1999_2014.txt': 48258}

2) Lake Monitoring data field count (1999-2014)
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis

python lakemonitoring_count_fields.py
33

Appendix G. Lake Monitoring data – Data Preparation

1) Sample of fields and a site observation from the original dataset. Also used to identify the delimiter.
    a. First cleanup the deprecated MAC OS 9 line endings of a carriage return that cause rows to appear as one long line to MAC GNU tools (e.g. head, tail, vim, wc -l, less, etc). Note Python does not have issues interpreting the carriage return as a line ending.

tr '\r' '\n' < mces_lakes_1999_2014.txt > mces_lakes_1999_2014_ret.txt

head -n 2 mces_lakes_1999_2014_ret.txt
PROJECT_ID DATA_SET_TITLE LAKE_NAME CITY COUNTY DNR_ID_Site_Number MAJOR_WATERSHED WATER_PLANNING_AUTHORITY LAKE_SITE_NUMBER

START_DATE START_HOURMIN24 END_DATE END_HOURMIN24 SAMPLE_DEPTH_IN_METERS Seasonal_Lake_Grade_RESULT Seasonal_Lake_Grade_QUALIFIER Seasonal_Lake_Grade_Units Physical_Condition_RESULT Physical_Condition_QUALIFIER Physical_Condition_Units Recreational_Suitability_RESULT Recreational_Suitability_QUALIFIER Recreational_Suitability_Units Secchi_Depth_RESULT_SIGN Secchi_Depth_RESULT Secchi_Depth_QUALIFIER Secchi_Depth_Units Total_Phosphorus_RESULT_SIGN Total_Phosphorus_RESULT Total_Phosphorus_QUALIFIER Total_Phosphorus_Units longitude latitude
7108 Citizen Assisted Monitoring Program (CAMP) for Lakes Acorn Lake Oakdale Washington 82010200-01 Lower St. Croix River Valley Branch WD 1 2006-04-16 0:00 2006-04-16 0:00 0 0-4 Categorical Calculated Seasonally: 4 good & 0 bad 1 Approved 1-5 Categorical: 1 good & 5 bad 5 Approved 1-5 Categorical: 1 good & 5 bad 1 Approved0.156 Approved mg/L -92.97171054 45.01655642

2) Feature subsetting step
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
        i. data_1_2_3_subset.ipynb
3) Data cleansing and timeseries data merging steps
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
        i. lakemonitoring_2_clean.ipynb

Appendix H. Water Proximity reference table – Counts

1) Water proximity reference table row count
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis

python xreftable_count_rows.py
{'Parcel_Lake_Monitoring_Site_Xref.txt': 2688767}

2) Water proximity reference table field count
    a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis

python xreftable_count_fields.py
10

Appendix I. Water Proximity reference table – Data Preparation

1) Sample of fields and a parcel observation from the original dataset. Also used to identify the delimiter.

head -n 2 Parcel_Lake_Monitoring_Site_Xref.txt
Parcel_PIN Monit_MAP_CODE1 Monit_SITE_CODE Monit_LAKE_SITE Distance_Parcel_Monitoring_Site_meters Lake_Hydroid Distance_Parcel_Lake_meters centroid_long centroid_lat Parcel_pkey
 19007900-01 19007900 1 2815.4927104148851 110517277058 2571.5267922258381 -93.11451 44.94

283 2163034

2) Feature subsetting steps
   a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
      i. data_1_2_3_subset.ipynb
3) Data cleansing step
   a. NA

Appendix J. Master data – Data Preparation

1) Tax Parcel, Lake Monitoring, and Water Proximity merge steps
   a. https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
      i. master_4_merge.ipynb

Appendix K. Model Validation – Modeling

1) https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
   a. model_validation.ipynb

Appendix L. Model Training & Evaluation – Modeling

1) https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
   a. model_train_eval.ipynb

Appendix M. Model Final Interpretation – Modeling

1) https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
   a. model_final_interpret.ipynb

Appendix N. Model Final Conclusion – Conclusions

1) https://github.com/wickedsedg/PropertyPrices_WaterQuality_thesis
   a. model_final_conclusion.ipynb