**The exploration of influential factors on residential property price using data mining algorithms**

-- The study of real estate market of Piercy County, Washington

Seattle University

BUAN 5510 Capstone Project

Group: Walt Ames, Qianhui Guo, Shuai Ma, Ankita Pathak

2019-08-12

Executive Summary

Residential property prices are a key indicator of economic well-being of a community, factors which have a positive or negative impact have been studied in depth using a wide variety of methods. Insight into these factors can be actionable information for governments, developers, and community members alike. For this research we are using the standard data used by the Pierce County Assessor-Treasurer’s Office as well as additional features extracted from Geographical Information Systems (GIS) data. The methodologies we utilize include; K-Means Clustering, Decision Trees, Random Forests, and Neural Networks in order to identify the most important determinants for residential property price. Of the aforementioned models the Random Forest proved to be the most accurate based on Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). The most crucial factors for home price as identified by the Random Forest include; square footage, quality, bathrooms, year built, land size, amount of crime, and bedrooms. The importance of crime indicates the feature we created using GIS data improved the performance of the model compared to the standard dataset. While the features we created improved the models we are still limited by the amount of missing or erroneously recorded data, additionally many of the features in the model are correlated. Moving forward we would like to add more features to the model including more details regarding education, crime, income, and neighborhood demographics. For example, the types of crime may affect housing prices differently as could school rankings, and median household income. While limited, our research provides some insight into the drivers of residential property prices in Pierce County Washington using machine learning algorithms and provides a benchmark for further study concerning this increasingly important measure of economic well-being.

Table of Contents

[1. Problem Statement 5](#_Toc16605707)

[a. Main study question 5](#_Toc16605708)

[b. Supporting study questions 5](#_Toc16605709)

[c. The importance of this research 5](#_Toc16605710)

[2. Literature Reviews 6](#_Toc16605711)

[a. The hedonic pricing models 6](#_Toc16605712)

[b. The Impact of Crime on property value 7](#_Toc16605713)

[c. The Impact of Environmental attributes on property value 8](#_Toc16605714)

[3. Description of Datasets 9](#_Toc16605715)

[a. Geographical Boundary 9](#_Toc16605716)

[b. Time Scope 10](#_Toc16605717)

[c. Data Size 11](#_Toc16605718)

[d. Entity Relationship Diagram (ERD) 12](#_Toc16605719)

[e. Data Resource 12](#_Toc16605720)

[f. Data Dictionary 13](#_Toc16605721)

[4. Data Pre-processing 13](#_Toc16605722)

[a. Missing Values 13](#_Toc16605723)

[b. Invalid Values 13](#_Toc16605724)

[c. Duplicate Values 14](#_Toc16605725)

[d. Data from Geographical Information System (GIS) Data Base 14](#_Toc16605726)

[e. Outliers 16](#_Toc16605727)

[f. Correlation Analysis 16](#_Toc16605728)

[g. Data Integration 20](#_Toc16605729)

[h. Feature Selection 21](#_Toc16605730)

[i. Model Construction 21](#_Toc16605731)

[j. Clustering 22](#_Toc16605732)

[5. Data Mining Models 26](#_Toc16605733)

[a. Data Mining Workflow 26](#_Toc16605734)

[b. Clustering 26](#_Toc16605735)

[c. Decision Tree 27](#_Toc16605736)

[d. Random Forest 28](#_Toc16605737)

[e. Artificial Neural Network 29](#_Toc16605738)

[6. Discussion and Evaluations 30](#_Toc16605739)

[f. Measures 30](#_Toc16605740)

[g. Decision Tree 31](#_Toc16605741)

[h. Random Forest 34](#_Toc16605742)

[i. Artificial Neural Network 37](#_Toc16605743)

[j. Model Comparison 39](#_Toc16605744)

[7. Conclusion 41](#_Toc16605745)

[a. Influential Features 41](#_Toc16605746)

[b. Business Implication 48](#_Toc16605747)

[c. Limitation 50](#_Toc16605748)

[d. Further Direction 50](#_Toc16605749)

[References 52](#_Toc16605750)

[Appendices 54](#_Toc16605751)

[a. Statistical summary – Before the outlier treatment 54](#_Toc16605752)

[b. Statistical summary – After the outlier treatment 54](#_Toc16605753)

[c. Sample Codes 55](#_Toc16605754)

[d. Data Dictionary Details 68](#_Toc16605755)

# Problem Statement

## Main study question

What are the influential factors on residential property prices in Pierce County, Washington?

## Supporting study questions

1. What are the residential building’s sales prices in 2018 Pierce County?
2. What are the possible influential physical factors to residential properties?
3. What are the possible influential environmental factors to residential properties?
4. Which variables and/or records should be normalized, grouped, or removed?
5. Which data mining models should be applied?

## The importance of this research

Increasing real-estate prices lead to more spending from consumers and higher economic growth. However, when property prices drop sharply, it hinders consumer confidence, construction, and lower economic growth. At the same time, property prices directly affect the living quality of residents, especially for those who do not own their residence. Property prices are an important metric for governments, developers, as well as residents, we are going to explore what factors influence the property value.

Allocating budget and investment for an optimized outcome is always a difficult task. This research aims to allow the stakeholders to understand what resources they should utilize towards real estate. Should the developer who intends to build a new apartment exhaust all budget for a parcel in downtown for a regular building, or spend less money for a parcel in the suburbs for a higher quality building? Should the government who would like to improve the local housing market spend budget to build more police stations and hire more policemen for a lower crime rate, or initiate a project to reallocate the wastewater treatment plant to a remote region? Should an apartment buyer spend $10,000 more on an apartment annually, only because the apartment is close to a top tier hospital? The result from this study helps answer these questions by analyzing the interaction between the property's features and its market price.

First, we will review some related literature to understand previous research on real-estate price and make an evidence-based suggestion concerning which factors might influence the residential property price in Pierce County. Then, we will perform data wrangling and cleansing to combine all the features we need in one table for analysis by machine learning and statistical models. We will then select the optimal data mining model by comparing different data mining techniques and models, in order to describe what property features would affect its market price.

# Literature Reviews

## The hedonic pricing models

In the paper of *The Composition of Hedonic Pricing Models* the authors introduce the hedonic pricing models and review the hedonic pricing model studies in real estate market. The paper examines approximately 125 real estate market studies from 1995 to 2005 and summarizes the top selected features in hedonic price model. The study reveals that the top twenty selected features in hedonic pricing model are: Lot Size, log(Lot Size), House Square Feet, log(House Square Feet), Brick, Age, Number of Stories, Number of Bathrooms, Number of Bedrooms, Full Baths, Fire Place, Air-Conditioning, Basement, Garage Space, Deck, Pool, Distance, Time on the Market, Time Trend. Many studies include the marketing and economical features, such as the period of a property listed on the market. For this feature, most of the study results show the listing period is not statistically significant and some results show the negative impact. It is reasonable that the longer a property listed, it is possible for the price to decrease since few people express interest to close a deal. Besides the single features, this paper also summarizes the eight categories of selected features, which are Construction & Structure, House Internal Features, House External Amenities, Environmental – Nature, Environmental – Neighborhood & Location, Environmental – Public Service, Marketing, Occupancy, & Selling, and Financial. This paper provides a frame structure to inform futures studies in this field.

This study reveals that the characters provided by the natural environment, such as good lake view, always post a positive impact on properties’ price. The location generally plays the most important role in determining the price. The location can be measured as zip code, distance to CBD or town center, and the sign of the location feature can be both positive and negative, influenced by the character of the district. The number of trees is also included in many studies. Trees usually mean the wooded lot rather than the vacant lot. The trees consistently post a positive impact on properties price, indicating the good natural environment can help increase the real estate’s market price.

## The Impact of Crime on property value

Numerous previous researches in the property literature shows that crime has a negative impact on property value. Thaler (1978) found that house value decreased by 3 percent with a one-standard-deviation increase in property crime. This effect is much worse in a recent finding, where Steve (2004) found that property value reduced by 10 percent with the same change in property crime. Moreover, by adopting the hedonic regression model, total, property, and violent crime all have a different negative impact on housing values (Tita et al. 2006). Their estimates also show that an increase in level and numbers of crime in a neighborhood is more likely to lower housing values, controlling for structural housing and neighborhood characteristics. Such an effect is much greater on a poor neighborhood, relative to middle-class and wealthy neighborhood.

However, the negative impact of crime on property value is varied when mixing different external factors, such as neighborhood income level (Tita et al. 2006), and job accessibility level (Ceccato and Wilhelmsson 2018). Ceccato and Wilhelmsson (2018) found that only when a place is close to the central business district (CBD) with a high level of accessibility, there is no impact of burglary on housing values in Stockholm. Otherwise, burglary discounted housing values. Such an effect is even greater in the locations which are far from CBD and with poor accessibility.

## The Impact of Environmental attributes on property value

Numerous studies confirm that environmental attribute is a significant determinant of property values. Weicher and Hartzell (1982) uncovered that those environmental amenities including transportation, proximity to school, shops, police station and fire station significantly affect home prices in 59 Metropolitan areas of the United States. GIS techniques along with hedonic model have been applied to examine the relationship between location externalities and house prices in the past decade (Oxford 2002, C. M. Hui et al. 2007), showing the importance of environmental characteristics. The accessibility to central locations is negatively correlated with housing prices (C. M. Hui et al. 2007) but the size of the shopping center has a positive effect instead (Des Rosiers et al 1996). In terms of school, both proximity to school and school quality plays an important factor in determining housing values (Weicher and Hartzell 1982, Black 1999). By adopting hedonic housing price regression, Black (1999) found a one-point increase in elementary school test scores could increase up to $70 million in residential property prices in Massachusetts. However, the proximity to a hospital was not preferable by citizens, as Tan (2011) found that house prices decrease by 15.1 percent if the house is located less than 20 minutes away from the hospital through semi-logarithmic regression model in Klang Valley, Malaysia.

Besides, the sea view and better air have the opposite effect in Hong Kong(C. M. Hui et al. 2007). Similar findings were gained by many researchers. Cassal and Mendelsohn (1985) revealed that the view of mountain, lake, and Puget Sound is associated with the increase of residential housing values in Seattle. Kendree and Rauch(1990) also regarded the view of a golf course as important. Moreover, the proximity to a lake could lead to an increase in housing value (McMillan 1980). But recently Boyle et al (2014) reported an inconsistency with previous research and concluded that walkability has no significant effect on house prices using fixed effects regression model.

# Description of Datasets

## Geographical Boundary

The municipal area of Pierce County sets the boundary of this data set. All data are collected within the Pierce County region.

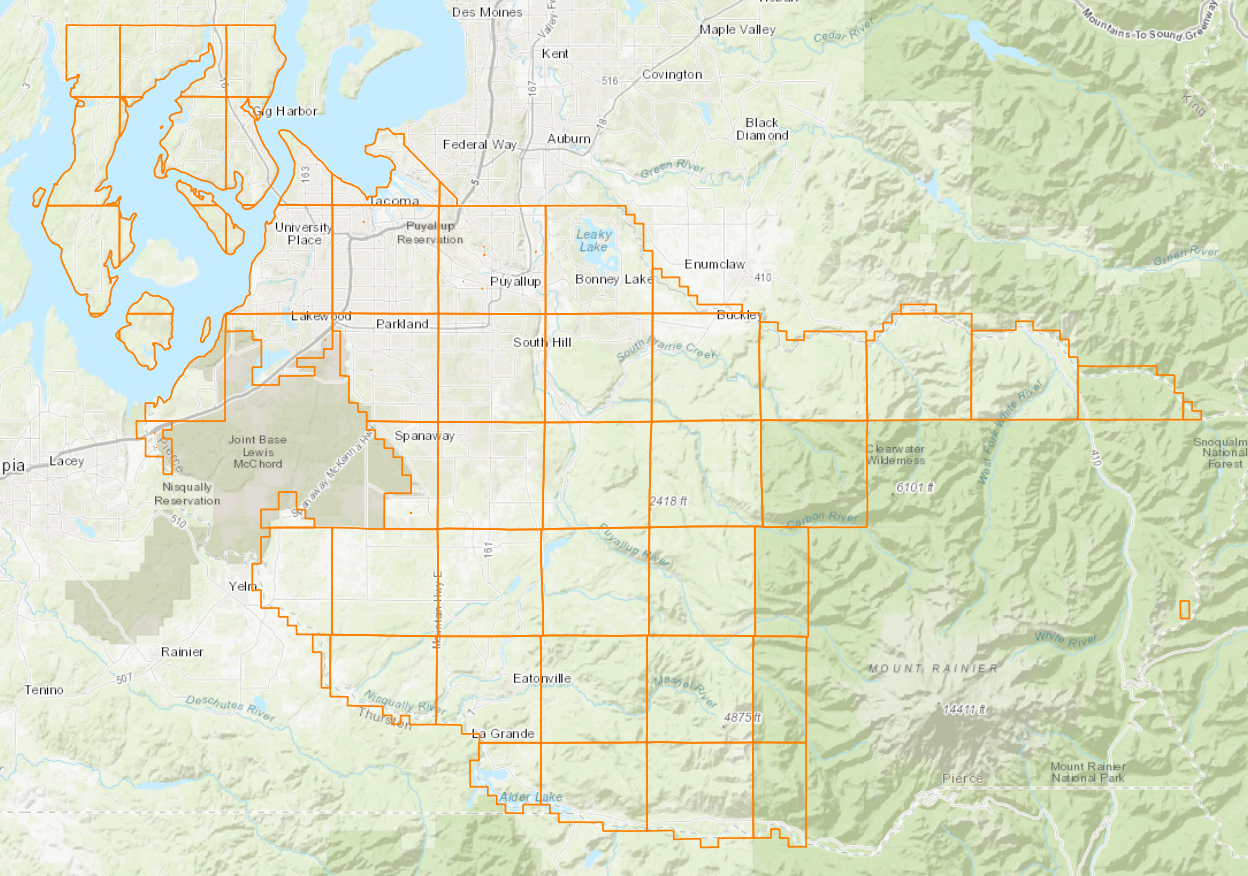


Figure 1: The Geographical Boundary

## Time Scope

The data listed on the pierce county government’s website are updated weekly. The data included in this study was retrieved from the website on 2019-07-04. The date range of this data sets are:

Table 1: The date range of data set

|  |  |
| --- | --- |
| Data set | Date Range |
| Appraisal Account (based on “appraisal date”) | 2011-09-12 ~ 2019-06-25 |
| Seg merge (based on the “completed date”) | 1991-07-11 ~ 2019-06-12 |
| Tax Account | Up to 2019 |
| Tax Description | Up to 2019 |
| Improvement | Up to 2019 |
| Improvement Builtas (based on “year built”) | 1861 ~ 2019 |
| Improvement Details | Up to 2019 |
| Land Attribute | Up to 2019 |
| Sale (based on “sale date”) | 1997-01-01 ~ 2019-06-04 |
| Crime | 2018-06-01 ~ 2019-05-31 |
| GIS Data | Up to 2019 |

## 

## Data Size

Table 2: The size of data set

|  |  |  |  |
| --- | --- | --- | --- |
| Table | Columns/Fields | Rows | Flat File Size |
| Appraisal Account | 24 | 331342 | 58.5 MB |
| Seg merge | 6 | 170439 | 6.9 MB |
| Tax Account | 28 | 328297 | 55.8 MB |
| Tax Description | 3 | 5199 | 78.0 MB |
| Improvement | 25 | 346979 | 38.1 MB |
| Improvement Builtas | 26 | 344252 | 42.9 MB |
| Improvement Details | 5 | 228644 | 93.8 MB |
| Land Attribute | 3 | 552542 | 21.7 MB |
| Sale | 12 | 524625 | 65.5 MB |
| Crime | 3 | 39 | 7.92 KB |
| GIS Data | 12 | 13587 | 483 KB |

## Entity Relationship Diagram (ERD)

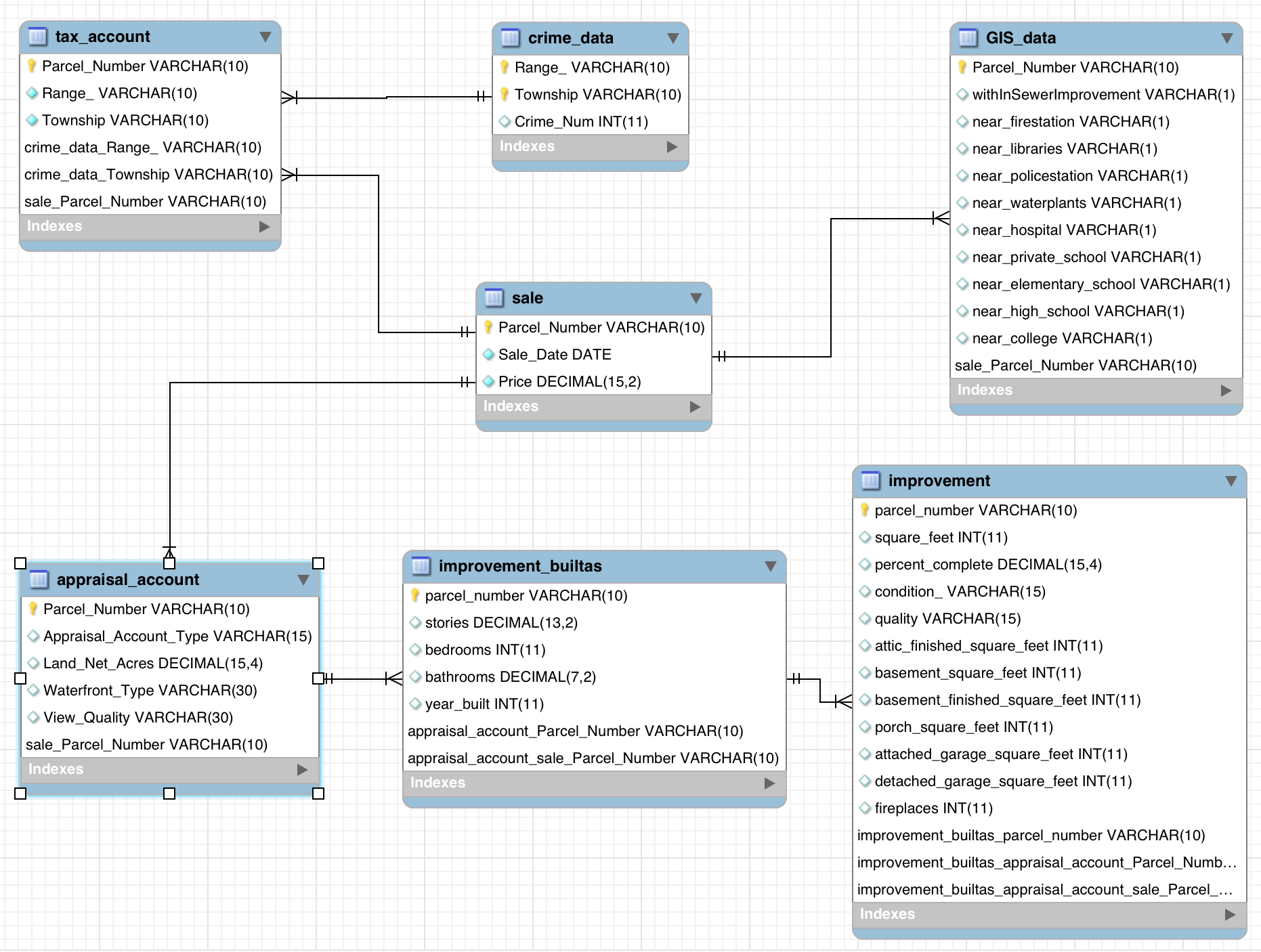


Figure 2: ERD Model

## Data Resource

Data contained in this dataset are from two sources: *Pierce County Assessor ‐ Treasurer[[1]](#footnote-1)* and *Crime Data[[2]](#footnote-2)* . Both are public datasets listed on Pierce County government website.

The source of *Pierce County Assessor ‐ Treasurer* has the data of “appraisal account”, “seg merge”, “tax account”, “tax description”, “improvement”, “improvement builtas”, “improvement details”, “land attribute” and “sale”.

The source of *Crime Data* has the data of “crime”.

## Data Dictionary

The data dictionary summarizes the variables’ “name”, “description”, “valid values”, “data type”, “length”, “example values”, “measure”, and “null ratio”. The full data dictionary is included in the appendix of this research paper.

# Data Pre-processing

## Missing Values

The missing values are filled or removed based on the following principles:

1. The variables/features with null ratio greater than 50% are dropped.
2. Missing values for “Number of Fireplace”, “Attic Square”, “Basement Size”, “Basement Net Size”, “Porch Size”, “Attached Garage Size”, “Detached Garage Size” are filled with “0”.

## Invalid Values

The missing values are filled or removed based on the following principles:

1. The records with invalid value, such as crime\_number < 0 and percent\_complete > 1.0, are dropped.
2. The records of outliers are dropped. The identification of outliers follows the statistical principles that outliers are records which go beyond the 1.5 times interquartile boundary.
3. The records of adjusted\_year\_built are change to the same value to the year\_built when adjusted\_year\_built is smaller than the year\_built

## Duplicate Values

All duplicate values are removed from the data set.

## Data from Geographical Information System (GIS) Data Base

This study includes the data from GIS database. The GIS data from Piercy County is to identify the external factors which have the potential to influence the property price. The following features are identified in the GIS database and added into the final master table as the extra columns.

(1) Hospital: parcels which are within 10 miles distance to a hospital are labeled as “1” in the feature of “near\_hospital”

(2) Private School: parcels which are within 3 miles distance to a private school are labeled as “1” in the feature of “near\_private\_school”

(3) Elementary School: parcels which are within 3 miles distance to an elementary school are labeled as “1” in the feature of “near\_elementary\_school”

(4) High School: parcels which are within 3 miles distance to a high school are labeled as “1” in the feature of “near\_high\_school”

(5) College: parcels which are within 3 miles distance to a college are labeled as “1” in the feature of “near\_college\_school”

(6) Fire Station: parcels which are within 1 miles distance to a fire station are labeled as “1” in the feature of “near\_fire\_station”

(7) Library: parcels which are within 1 miles distance to a library are labeled as “1” in the feature of “near\_library”

(8) Police Station: parcels which are within 1 miles distance to a police station are labeled as “1” in the feature of “near\_police\_station”

(9) Water Plants: parcels which are within 1 miles distance to a water plants are labeled as “1” in the feature of “near\_water\_plants”

(10) Sewer Improvement: parcels which are within the sewer improvement area are labeled as “1” in the feature of “withinSewerImprovement”

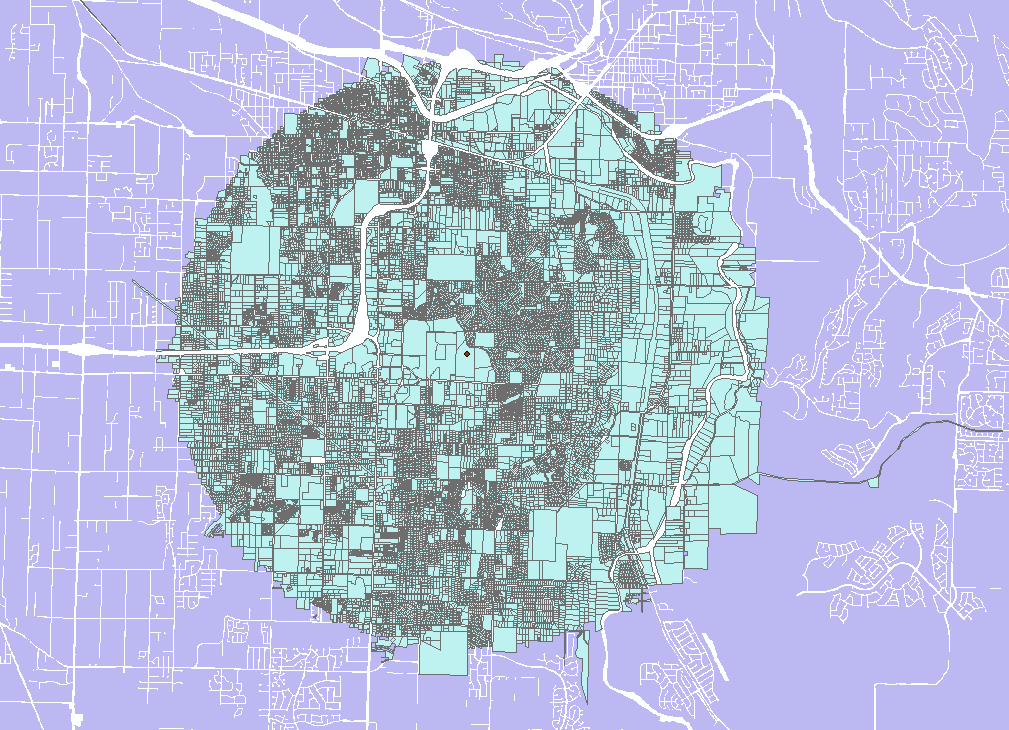


Figure 3: Parcels which are within three-mile distance to the school

## Outliers

We treated for outliers using the interquartile range (IQR) rule for outliers which means we multiplied the IQR by 1.5 and added this figure to the third quartile and subtracted it from the first quartile. In cases when the lower threshold went below legal limits or zoning regulations, we used the Pierce County minimums. For example, the minimum legal square footage for a residence is 120 square feet so, we used this square footage as the lower threshold. We applied this treatment only to numeric features with values well beyond the upper threshold, the list includes; Sale Price between $20,000 and $800,000, Land Net Acres below .7 acres, Square Feet between 120 and 4500, Bedrooms less than or equal to 6, Bathrooms between 1 and 4, Stories between 0 and 4, and Fireplaces under 4. By eliminating outliers, we ensure there is no erroneous data or extreme values which would skew our models.

## Correlation Analysis

(1) Categorical Variables

Table 3:Chi-Square Test for Categorical Variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable 1** | **Variable 2** | **Test Result** | **Conclusion** | **Implication** |
| near\_firestation | near\_healthcare | X-squared = 743.67, df = 1, p-value < 2.2e-1 | Reject the null Hypothesis at 1% significance level. | The properties which are near Fire stations is associated with that near the healthcare center |
| near\_policestation | near\_waterplants | X-squared = 84.281, df = 1, p-value < 2.2e-16 | Reject the null Hypothesis at 1% significance level | The properties which are near police station is associated with that near the water plants. |
| near\_libraries | near\_firestation | X-squared = 1092.5, df = 1, p-value < 2.2e-16 | Reject the null Hypothesis at 1% significance level | The properties which are near a library is associated with that near the fire stations. |
| near\_libraries | near\_policestation | X-squared = 3472, df = 1, p-value < 2.2e-16 | Reject the null Hypothesis at 1% significance level. | The properties which are near a library is associated with that near the police stations. |
| view\_quality | waterfront\_type | X-squared = 3.7973, df = 1, p-value = 0.05133 | Cannot reject the null Hypothesis at 5% significance level. | The View Quality of a property doesn’t have any relationship with the waterfront type of a property |

This study also applies the “pairs.panels( )” function in r to figure out how correlated each pair of categorical variables is. As the below figure reveals, most of categorical variables are correlated.

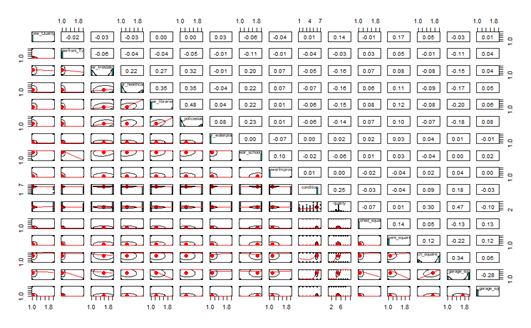


Figure 4: Paired Chi-Square Analysis in R

Based on the assumptions we made, we also use pairs.panels( ) in r to figure out how correlated each pair of categorical variables is. So, from the graph, we found out that there aren’t highly correlated pairs of variables among categorical variables.

(2) Numerical Variables

There are some interior characteristics showing some moderate correlations.

The square feet of a property are moderately correlated with the stories (r = 0.56). The square feet also have a slight strong positive relationship with the number of bathrooms in a property (r = 0.66). Besides, a property with larger square feet is probably with more bedrooms showing a correlation coefficient of 0.56. This makes sense as the larger house is always associated with more bathrooms and bedrooms.

Moreover, there also shows that stories and bedrooms, bedrooms and bathrooms, as well as the pair of bathrooms and the year a property built, are all moderately associated with each other.

When we look into the correlation coefficient between the sales price and other variables, we found there is almost no relationship between crime number and sales price. But, the square feet of a property are strongly correlated with sales price (r = 0.65), showing that bigger house is associated with higher prices. There’s also a moderate correlation between the sales price and bathrooms where the correlation coefficient is 0.53. Lastly, number of bathrooms is highly correlated with stories (r = 0.66).

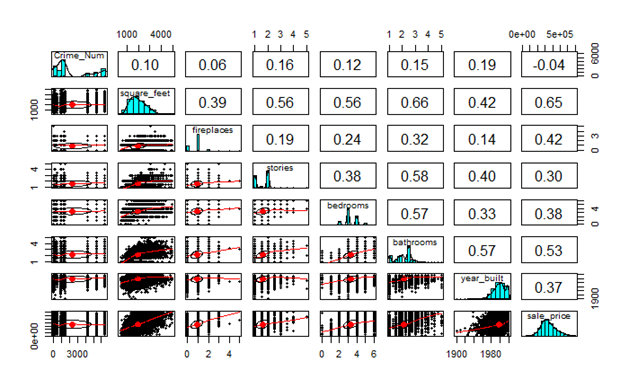


Figure 5: Correlation Analysis in R

## Data Integration

The data integration follows the following steps:

(1) Import the raw data, the csv file from Pierce County data warehouse, into SQL server.

(2) Import the geographical data, the shape file from Pierce County GIS website, into ArcGIS.

(3) Select the features from data in SQL server by creating and joining “views”.

(4) Use GIS spatial join and selection to identify parcels which are close to the educational facilities, public facilities, or within the sewer improvement area.

(5) Use GIS spatial join and selection to summarize the crime data in each district.

(6) Export the GIS data to SQL server

(7) Join the data from Pierce County treasury database and GIS website in SQL server

(8) Export the master dataset into R remove the outliers and invalid data.

The following figures shows the workflow of data integration process:

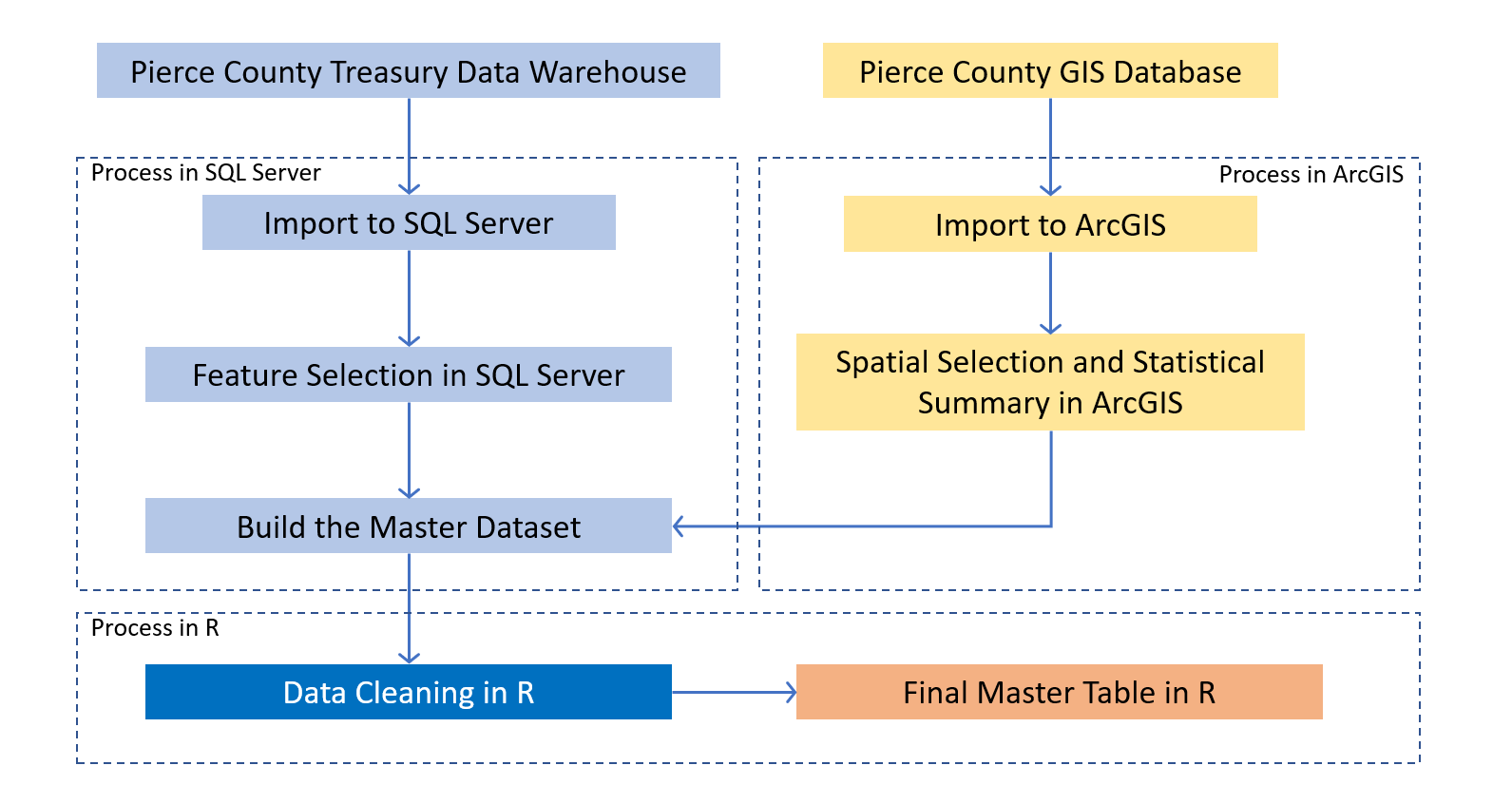


Figure 6: Workflow of data integration

## Feature Selection

This study only selects the data of residential property sales record in 2018. The selected features are attributed to two categories: (1) the property’s intrinsic features, including Land Area, View Quality, Waterfront, House Size, Percent Complete, Condition, Quality, Attic Square, Basement Size, Basement Net Size, Porch Size, Attached Garage Size, Detached Garage Size, Number of Fireplace, Stories, Number of Bedrooms, Number of bathroom, and Year Built, and (2) the property’s external features, including Near Hospital, Near Private School, Near Elementary School, Near High School, Near College, In Sewer Improvement Area, Near Fire Station, Near Library, Near Police Station, Near Water Plants.

## Model Construction

This study builds two models with selected features which are the “Full Model Predictor” and the “Reduced Model Predictor”.

(1) The Full Model Predictor (Model 1) is constructed as:

Sale Price ~ Land Area + View Quality + Waterfront + House Size + Percent Complete + Condition + Quality + Attic Square + Basement Size + Basement Net Size + Porch Size + Attached Garage Size + Detached Garage Size + Number of Fire Place + Stories + Number of Bedrooms + Number of bathroom + Year Built + Hospital + Private School + Elementary School + High School + College + In Sewer Improvement + Fire Station + Library + Police Station + Water Plants.

The features of Hospital, Private School, Elementary School, High School, College, In Sewer Improvement, Fire Station, Library, Police Station, Water Plants are from the GIS database.

(2) The Reduced Model Predictor (Model 2) is constructed as:

Sale Price ~ Land Area + View Quality + Waterfront + House Size + Percent Complete + Condition + Quality + Attic Square + Basement Size + Basement Net Size + Porch Size + Attached Garage Size + Detached Garage Size + Number of Fire Place + Stories + Number of Bedrooms + Number of bathroom + Year Built.

All features included in the reduced model are from the treasury data warehouse.

## Clustering

Two types of clustering are practiced in this study: Hierarchy cluster (with complete linkage and average linkage) and K-Means cluster. This study divides the dataset into three clusters with Hierarchy cluster method and divides the dataset into three and five clusters with K-Means clustering method.

As shown below, the clusters from Hierarchy clustering method are very biased. For both the Full Model Predictor and the Reduced Model Predictor, two clusters contain too small amount of records to proceed further machine learning algorithm. The clustering results from K-Means methods is more even in terms of records number in each cluster.

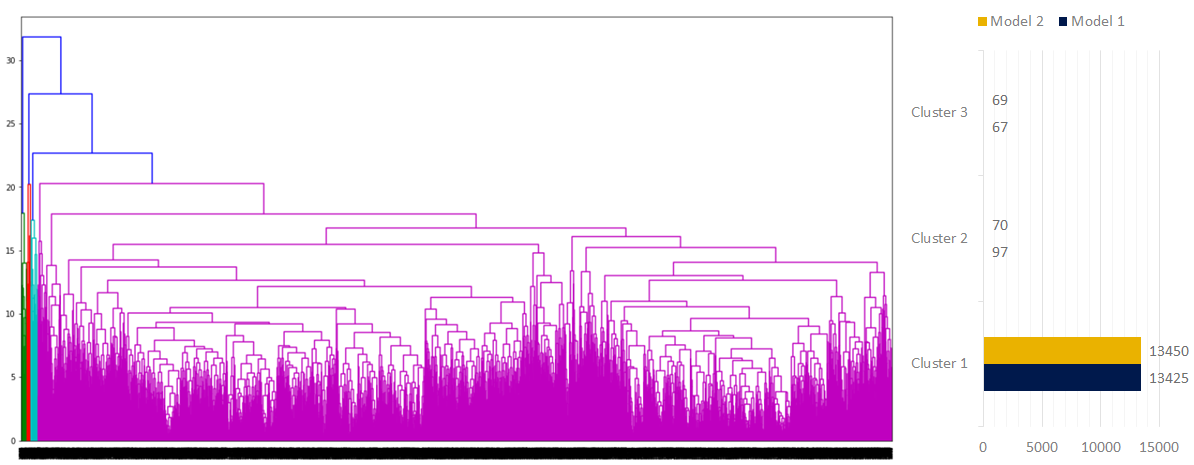


Figure 7: Hierarchy cluster with complete linkage - Full Model Predictor

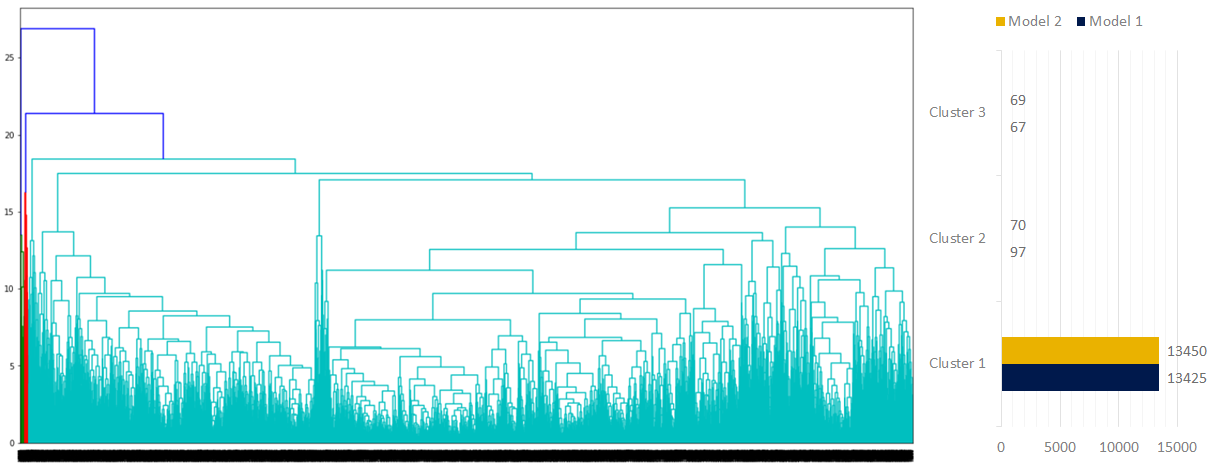


Figure 8: Hierarchy cluster with complete linkage - Reduced Model Predictor

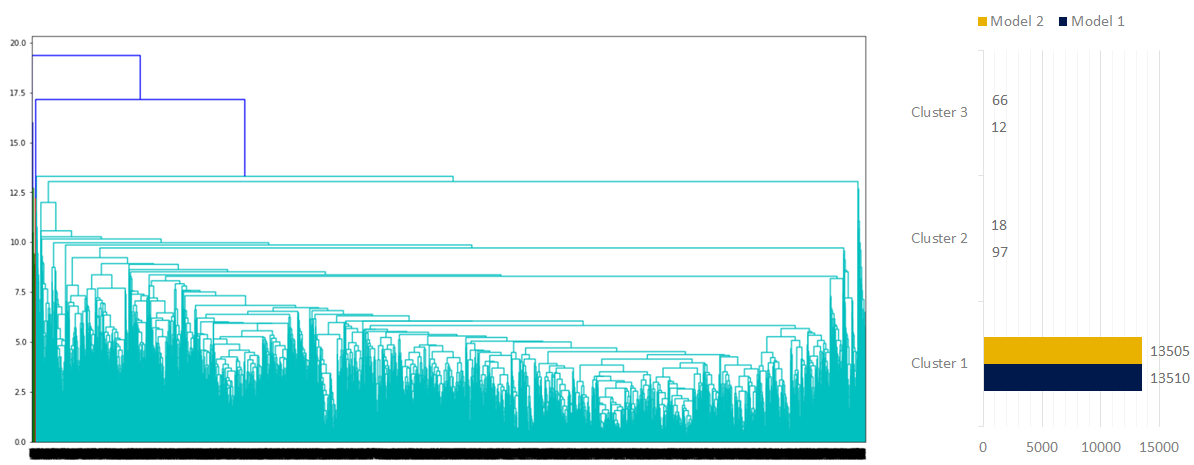


Figure 9: Hierarchy cluster with average linkage - Full Model Predictor



Figure 10: Hierarchy cluster with complete linkage - Reduced Model Predictor

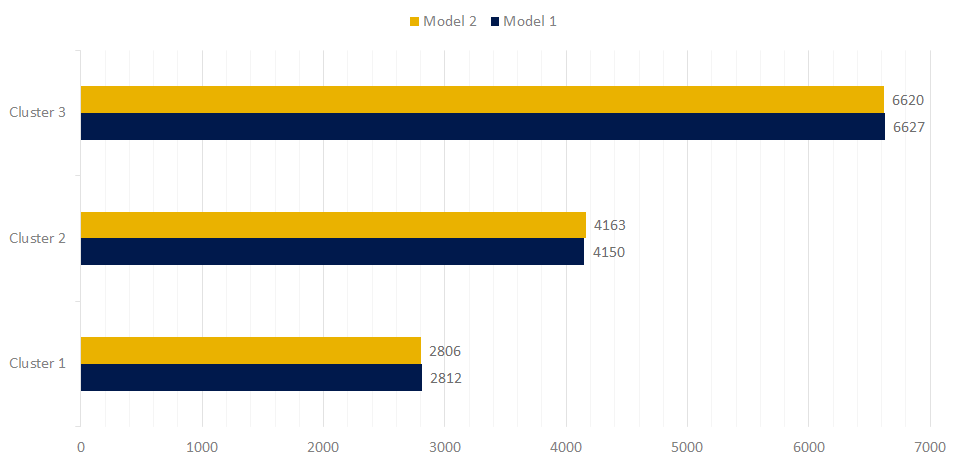


Figure 11: K-Means clustering with K = 3

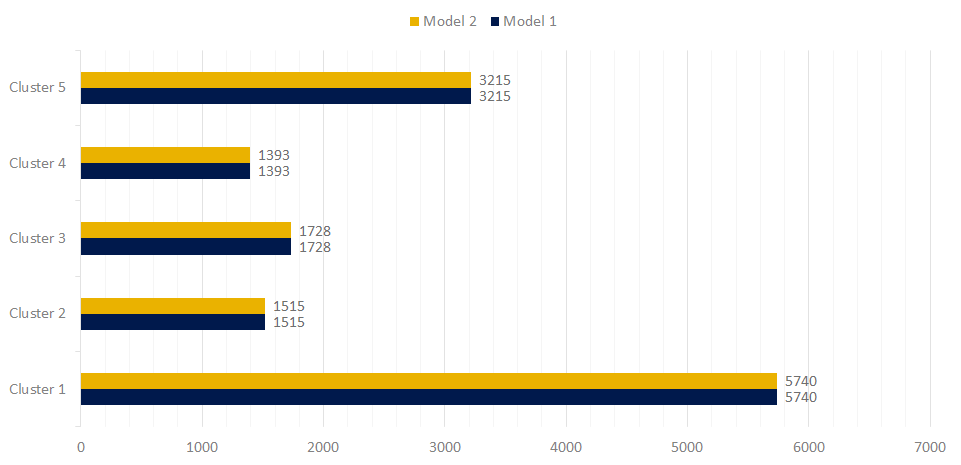


Figure 12: K-Means clustering with K = 5

# Data Mining Models

## Data Mining Workflow

This study uses the clustering results from K-Means method, and applies the models of Decision Tree, Random Forest, and Neural Networks into the full data set and clustered data set with K=3 and K=5. The workflow of this study is shown as below.

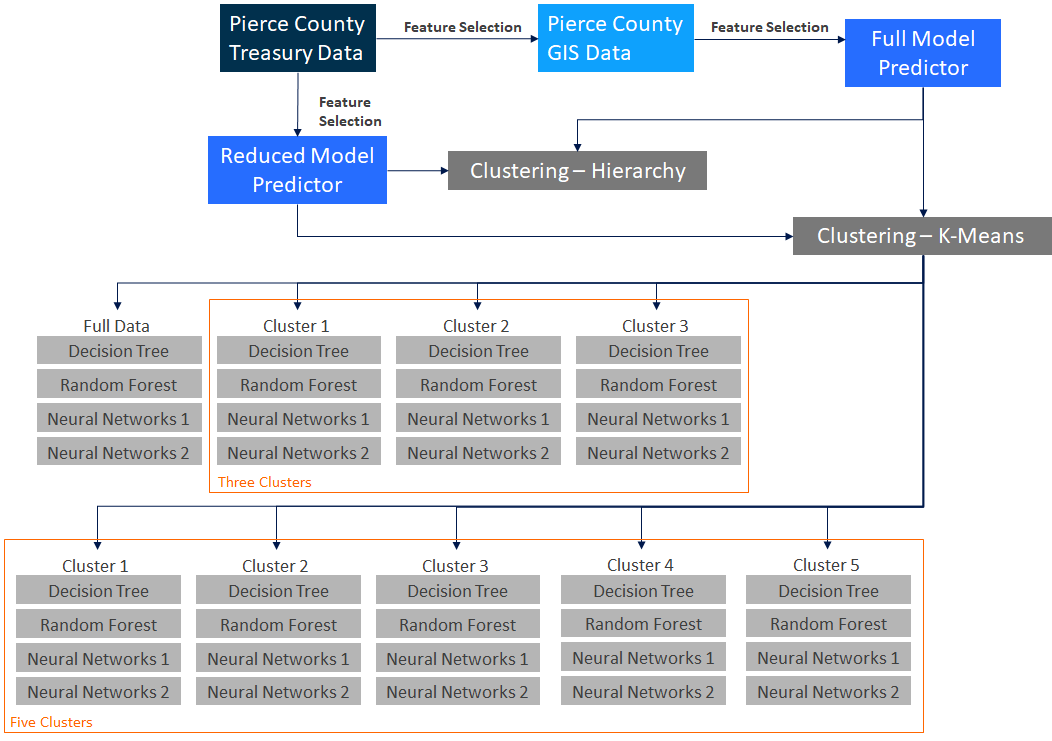


Figure 13: Workflow of Clustering and Modeling

## Clustering

During the clustering process we compared the sum of squares within each cluster to the overall sum of squares as an indicator of cluster homogeneity. Initially, we used three clusters and the comparison indicated the clusters were all very homogeneous with a sum of squares comparison figure of .17. After three clusters we attempted using a different number of centroids, but the best results were yielded using five clusters and a sum of squares comparison of .28. This indication here is that the data is extremely homogeneous and clustering most likely will not improve our models by any discernible amount or could even increase model errors.

## Decision Tree

We adopt multiple decision tree regression models to examine the relationship between housing features and residential property prices in Pierce County, WA (Fan et al. 2006). The decision tree model is able to classify target field by splitting data into subsets that contain similar instances (Woods and Kyral, 1997, p. 42), and address feature importance based on its top-down decision and leaf nodes (Fan et al. 2006). The reason why we adopt decision tree algorithm is that decision tree is easy for interpretation for small-sized tree and robust to the effect of outliers, although it can be very complex and requires many branches for accurate estimation (Moore et al. 2001).

The dependent variable of the decision tree model is the sales prices of residential property in 2018. In order to examine how external attributes influence the impact of intrinsic features on housing prices, we conduct a full model with both intrinsic and external attributes as independent variables (shown in Model (1), Model (2) and Model(3) in Table 1), and a reduced model with only intrinsic attributes as predictors (shown in Model (4), Model (5) and Model(6) in Table 1). Besides, to improve the accuracy of decision tree models, we also apply k-means clustering methods into the model, where we separate the dataset into 3 and 5 groups respectively according to certain similarities (shown in Model (2), Model (3), Model (5) and Model(6) in Table 1). For model validation, we split 80% of the sample dataset as a training dataset for mode construction and 20% of that as a test dataset for model assessment.

Table 4: Decision Tree Models

|  |  |  |
| --- | --- | --- |
| Dependent variable:  The housing sales price in 2018 | Full Model Predictors: Both intrinsic and external features | |
| Model (1) | Full Dataset |
| Model (2) | Full Dataset with 3 clusters |
| Model (3) | Full Dataset with 5 clusters |
| Reduced Model Predictors: Only intrinsic features | |
| Model (4) | Partial Dataset |
| Model (5) | Partial Dataset with 3 clusters |
| Model (6) | Partial Dataset with 5 clusters |

## Random Forest

The second model we used for prediction was a Random Forest Regression. Random Forests operate similarly to decision trees in that they are a collection of decision trees. Model features are split based on information gain indicated by the GINI Index just like the Decision Tree Model, and each individual tree stops once no further information can be gained. Each Decision Tree in the forest has an element of randomness introduced such as subsampling or splitting on varying nodes. The Random Forest model then aggregates the output from each Decision Tree which reduces the variance and increases the accuracy of the model. Further benefits of using a Random Forest are that single Decision Trees have a tendency to overfit the training dataset and perform markedly worse on the test dataset, aggregating the results from a Random Forest helps address this issue. Random Forests do not come without some limitations, they require more computational resources and run time as compared to other algorithms. Additionally, Random Forests are less intuitive and less scrutable compared to other algorithms, for example observing the nodes within each Decision Tree would be impractical. The below figure is a visualization of how a Random Forest is a collection of Decision Trees. The Random Forest Models follow the same organization and training proportion previously indicated for the Decision Tree Models.

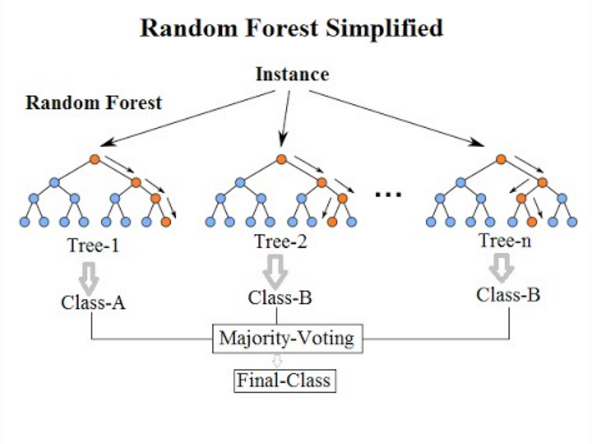


Figure 14: Random Forest Conceptual Diagram

## Artificial Neural Network

In order to predict the housing prices for Pierce County Washington, the third model that we used is Artificial Neural Network (ANN). ANN consists of a connection between numbers of neurons. There are many types of connections, in one type of network which is called the Multi-Layer Perceptron (MLP), the data flows forward to the output without any feedback. Most popular learning rule in MLP is the Error Back Propagation algorithm that we have used for predicting the housing prices. The Back-Propagation training of a network consists of three stages: *Feed forward of the input data*, *calculation of error and back propagation it* and *adjustment of the weights in order to decrease the error*. The algorithm adjusts the weights of the network in order to minimize the Sum Squared Error (SSE). Neural networks are typically organized in layers. Layers are made up of a number of interconnected nodes which contain an activation function. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer. Most ANNs contain some form of 'learning rule' which modifies the weights of the connections according to the input patterns that it is presented with.

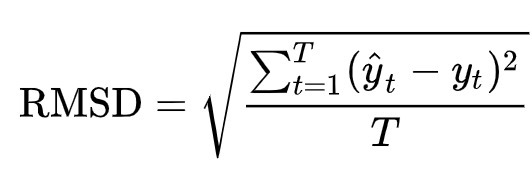
One of the benefits of ANNs is that the information such as in traditional programming is stored on the entire network, not on a database. The disappearance of a few pieces of information in one place does not prevent the network from functioning. However, ANNs require processors with parallel processing power, in accordance with their structure. For this reason, the realization of the equipment is dependent.

# Discussion and Evaluations

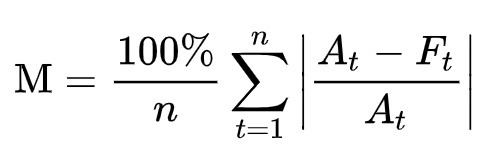
## Measures

We applied two measures of accuracy, mean absolute percentage error (MAPE) and Root Mean Squared Error (RMSE), to assess regression model performance. Hyndman and Koehler (2006) indicated that RMSE is used to compare forecasting residuals of various models for a specific dataset. The smaller the MAPE and RMSE are, the better-fitted a model is. Additionally, when we apply K-means clustering into the models, we adopt the weighted average method to calculate the MAPE and RMSE for the whole dataset.

RMSE Formula:



MAPE Formula:



## Decision Tree

According to the MAPE and RMSE shown in Table 2, we are not surprised to find that k-means clustering doesn’t improve the decision tree model. For both full and reduced decision tree model, using full dataset always produces the best results with lowest MAPE and RMSE (shown in Model (1) and Model (4)). Specifically, in the full model with both intrinsic and external features as independent variables, the MAPE is 18.55%, and the RMSE is $75,311.15. While in the reduced model with only intrinsic features as independent variables, both MAPE and RMSE is slightly worse than that of the full model, 18.68% for MAPE and 76,238 for RMSE. This results also indicate that the performance of a model has been improved after including external features which are extracted from GIS techniques in the decision tree model.

Table 5: Decision Tree Models Assessment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dependent variable:**  The housing sales price in 2018 | **Full Model Predictors: Both intrinsic and external features** | | **MAPE** | **RMSE** |
| Model (1) | Full Dataset | *18.55%* | *75,311.15* |
| Model (2) | Full Dataset with 3 clusters | 20.10% | 78,522.33 |
| Model (3) | Full Dataset with 5 clusters | 20.44% | 77,048.04 |
| **Reduced Model Predictors: Only intrinsic features** | | **MAPE** | **RMSE** |
| Model (4) | Partial Dataset | *18.68%* | *76,238.63* |
| Model (5) | Partial Dataset with 3 clusters | 20.70% | 80,667.04 |
| Model (6) | Partial Dataset with 5 clusters | 21.05% | 80,287.79 |

Figure 15 and Figure 16 exhibit important features from the full model and reduced model respectively, showing a consistent importance among intrinsic features. Both models indicate that the square feet of a house is the most significant determinant of residential property, followed by other intrinsic attributes including quality, basement finished square feet, year built,number of bathrooms and view quality. It’s reasonable that square feet, quality, and basement finished square feet are the top three attributes that influence the housing prices, as a more expensive house is always associated with larger house and basement space, as well as better quality to provide residents a higher living standard. Besides, according to the full model, the number of crimes is the most significant external factors that determine housing prices. Tita et al. (2006) also used a hedonic regression model to find that an increase in numbers of crimes in a neighborhood is more likely to lower housing values, controlling for other characteristics. The proximity to college is also significant through our study, as there might be higher living demand around the college area for students and family.



Figure 15: Feature importance of Decision Tree Full Model

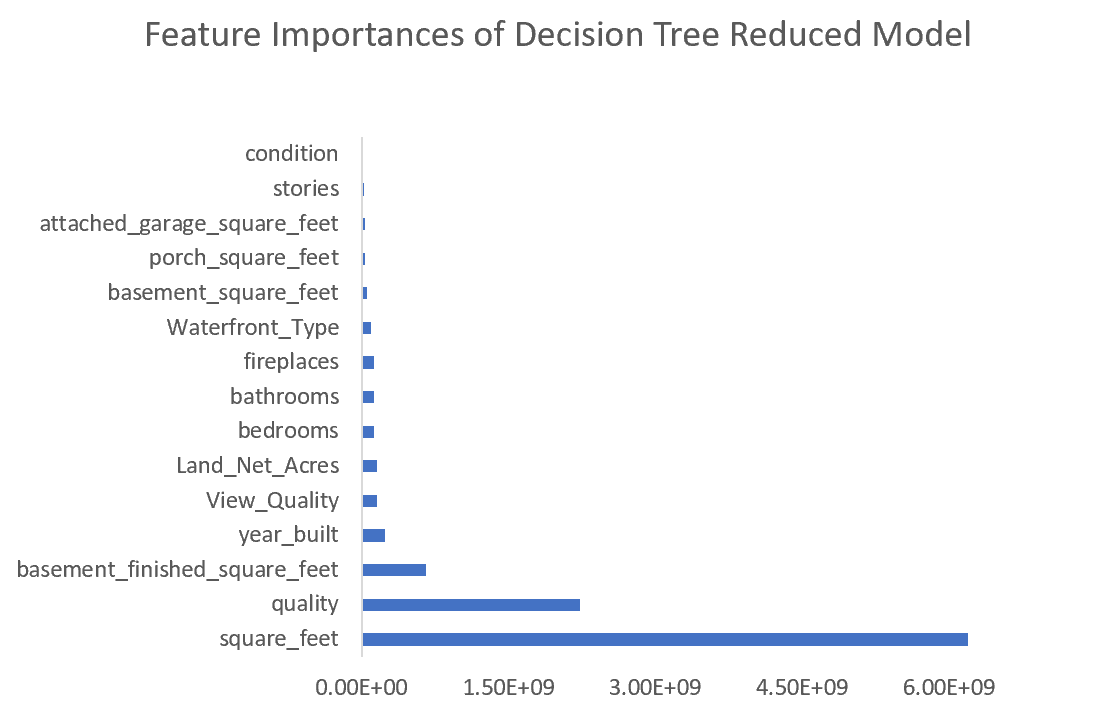


Figure 16: Feature Importance of Decision Tree Reduced Model

## Random Forest

The results from our Random Forest models indicate the external GIS data we included reduced the RMSE and MAPE. Furthermore, the results confirm our earlier inclination that clustering would have little impact on this data, given the homogeneity of our observations. This can be seen given the similar model performance for each dataset, the full data set has under one percentage point of difference between the three models. Combined with the sum of squares comparison during the clustering process we are confident the data is quite homogeneous. The best performance is achieved when using the full data set without clustering, which returns a MAPE of 14.04% and an RMSE of $55,910.26. Each of the full dataset models outperforms its partial data set counterparts by $7,000 to $9,300 RMSE and 2.92% to 3.27% MAPE regardless of the clustering used, definitively indicating the features we added to the Pierce County Assessor-Treasurer data improved the performance of the model.

Table 6: Random Forest Models Assessment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dependent variable:**  Housing sale price in 2018 | **Full Model Predictors: Both intrinsic and external features** | | **MAPE** | **RMSE** |
| Model (1) | Full Dataset | *14.04%* | *55,910.26* |
| Model (2) | Full Dataset with 3 clusters | 15.05% | 59,270.70 |
| Model (3) | Full Dataset with 5 clusters | 14.19% | 58,737.94 |
| **Reduced Model Predictors: Only intrinsic features** | | **MAPE** | **RMSE** |
| Model (4) | Partial Dataset | *16.96%* | *65,209.82* |
| Model (5) | Partial Dataset with 3 clusters | 18.24% | 68,619.58 |
| Model (6) | Partial Dataset with 5 clusters | 17.46% | 65,740.20 |

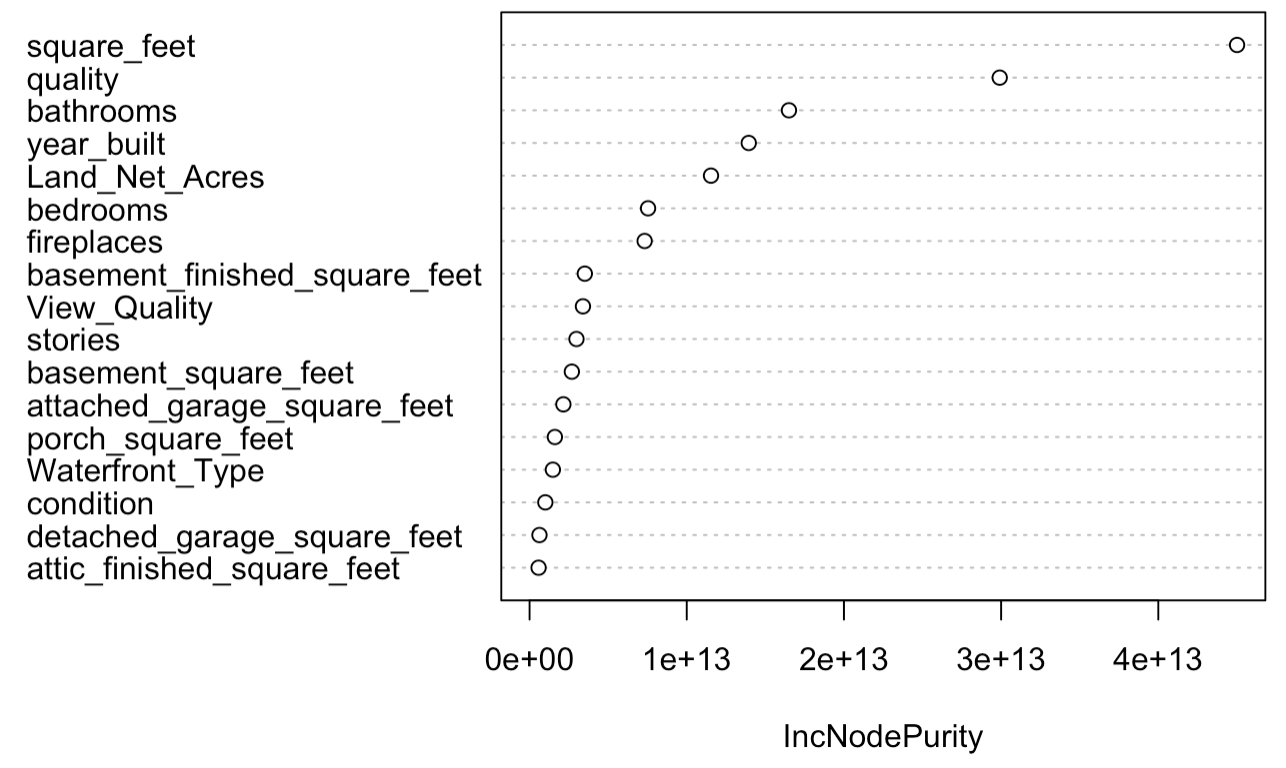
The figures below indicate the most significant features identified by the Random Forest Models. The first is for the partial dataset using only the Assessor-Treasurer Office Data and the second is for the full dataset using the features we extracted from GIS. 

Figure 17: Feature Importance of Random Forest Reduced Model

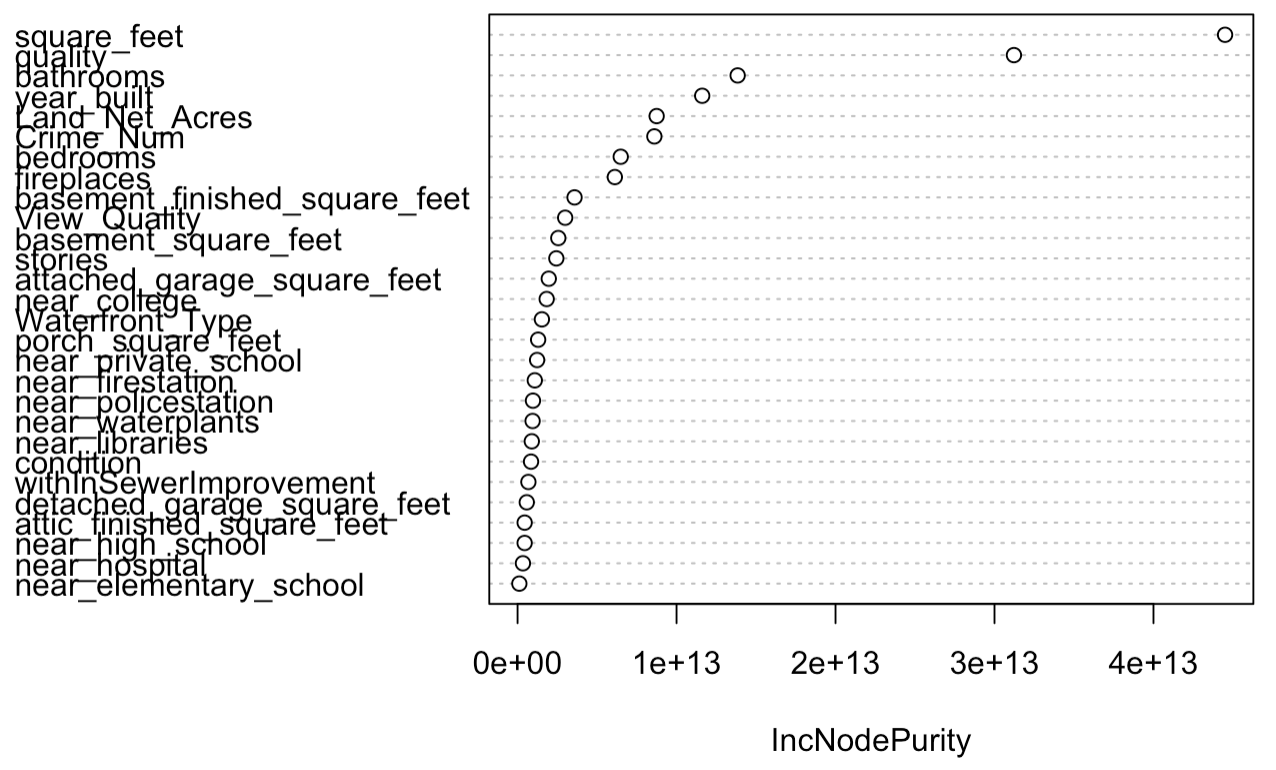


Figure 18: Feature Importance of Random Forest Full Model

Both sets of data indicate the same seven most significant features before diminishing returns, these features include; square footage, quality, bathrooms, year built, land size, bedrooms, and fireplaces. The full dataset models include number of crimes as the fifth most significant feature. The other GIS features we added to create the full dataset are grouped with the remaining features which represent a lower degree of information gain.

## Artificial Neural Network

According to the MAPE and RMSE shown in Table 3, it is important to note that after normalizing the data the k-means clustering improves the artificial neural network model for full model when hidden layer is 1 and cluster size is 5. However, k-means clustering for full model does not improve the results for full model when hidden layer is 2. For reduced artificial neural network model with hidden layer 1, using 3 clusters produces the best results with lowest MAPE and RMSE (shown in Model (5)). Similarly, for reduced model with hidden layer 2, cluster size of 5 produces the best results with lowest MAPE and RMSE (shown in Model (6)). For illustrating the behavior of model, in the full model where both intrinsic and external features are independent variables, the MAPE is 23.07%, and the RMSE is $83,246.44. While in the reduced model with only intrinsic features as independent variables, both MAPE and RMSE is slightly less than that of the full model, 21.78% for MAPE and $117,135.82 for RMSE. These results also indicate that the performance of a model has not improved after including external features which are extracted from GIS techniques in the neural network model.

Table 7: Artificial Neural Network Models Assessment

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dependent variable:** The housing sales price in 2018 | | | | | |
| **Normalized data** | | **Hidden Layer =1** | | **Hidden Layer =2** | |
| **Full Model Predictors: Both intrinsic and external features** | | **MAPE** | **RMSE** | **MAPE** | **RMSE** |
| Model (1) | Full Dataset | 23.32% | 85,203.68 | *24.45%* | *82,428.85* |
| Model (2) | Full Dataset with 3 clusters | 27.84% | 92,273.46 | 28.06% | 98,678.17 |
| Model (3) | Full Dataset with 5 clusters | *23.07%* | *83,246.44* | 30.99% | 118,830.79 |
| **Reduced Model Predictors: Only intrinsic features** | | **MAPE** | **RMSE** | **MAPE** | **RMSE** |
| Model (4) | Partial Dataset | 23.20% | 80,442.80 | 22.66% | 75,099.24 |
| Model (5) | Partial Dataset with 3 clusters | *22.78%* | *90,041.37* | 30.91% | 128,147.74 |
| Model (6) | Partial Dataset with 5 clusters | 26.27% | 94,074.93 | *21.78%* | *117,135.82* |

According to the MAPE and RMSE shown in Table 4, the results with non-normalized data are not convincing either in terms of adding additional external features. The results with reduced model predictors are better in terms of MAPE and RMSE values. Again, the results of clustering only improve the results of partial model with 2 hidden layers and a clustering size of 5. For full model with both intrinsic and external features the results are not the best.

Table 8: Artificial Neural Network Models Assessment

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dependent variable:** The housing sales price in 2018 | | | | | |
| **Non-Normalized data** | | **Hidden Layer =1** | | **Hidden Layer =2** | |
| **Full Model Predictors: Both intrinsic and external features** | | **MAPE** | **RMSE** | **MAPE** | **RMSE** |
| Model (1) | Full Dataset | *29.84%* | *123,443.38* | *29.12%* | *133,580.48* |
| Model (2) | Full Dataset with 3 clusters | 31.64% | 130,187.92 | 32.91% | 127,028.47 |
| Model (3) | Full Dataset with 5 clusters | 36.96% | 129,602.44 | 30.67% | 128,415.53 |
| **Reduced Model Predictors: Only intrinsic features** | | **MAPE** | **RMSE** | **MAPE** | **RMSE** |
| Model (4) | Partial Dataset | *29.84%* | *123,443.38* | 29.84% | 123,443.38 |
| Model (5) | Partial Dataset with 3 clusters | 32.91% | 127,028.47 | 35.41% | 129,366.13 |
| Model (6) | Partial Dataset with 5 clusters | 31.99% | 117,005.19 | *23.98%* | *136,434.81* |

## Model Comparison

After computing the MAPE and RMSE for each of our 32 models we compared the best performers between each of the algorithms used. The table below indicates the best model for each Decision Tree, Random Forest, and Neural Network for both datasets and indicates the corresponding MAPE and RMSE of that model. Across all modeling techniques the full dataset produces the best results, confirming that our GIS feature creation contributes to the accuracy of the models. Additionally, for each dataset the un-clustered model (1 or 4) produces the best or equivalent result as the clustered models due to the homogeneous nature of the data, as previously stated. Overall, the Random Forest proved to be the highest performing model with a MAPE of 14.04% and $55,910.26 RMSE. This model outperformed the Decision Tree and Neural Network by 5% and 15% respectively for MAPE. Additionally, the Random Forests’ RMSE was $20,000 to $70,000 less than the Decision Trees and Neural Networks.

Table 9: Comparison of Model Performance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dependent variable:**  Housing sale price in 2018 | **Full Model Predictors: Both intrinsic and external features** | | **MAPE** | **RMSE** |
| Decision Tree Model | Model 1 | *18.55%* | *75,311.15* |
| Random Forest Model | Model 1 | *14.04%* | *55,910.26* |
| Neural Network | Model 1 (One Hidden Layer) | *29.84%* | *123,443.38* |
| **Reduced Model Predictors: Only intrinsic features** | | **MAPE** | **RMSE** |
| Decision Tree | Model 4 | *18.68%* | *76,238.63* |
| Random Forest | Model 4 | *16.96%* | *65,209.82* |
| Neural Network | Model 4 (One Hidden Layer) | *29.12%* | *133,580.48* |

# Conclusion

## Influential Features

Overall the random forest full model without any clustering produces the best performance in predicting the determinants of residential property values in our study. Although clustering doesn’t improve the model performance, the inclusion of external characteristics from GIS lead to a better-fitted result.

(1) Space and Rooms

Based on the best model, we found that square feet is the most important features that determines house prices. It is intuitive that the higher square feet lead to a higher prices, which shows the similar tendency from our study (figure 19).

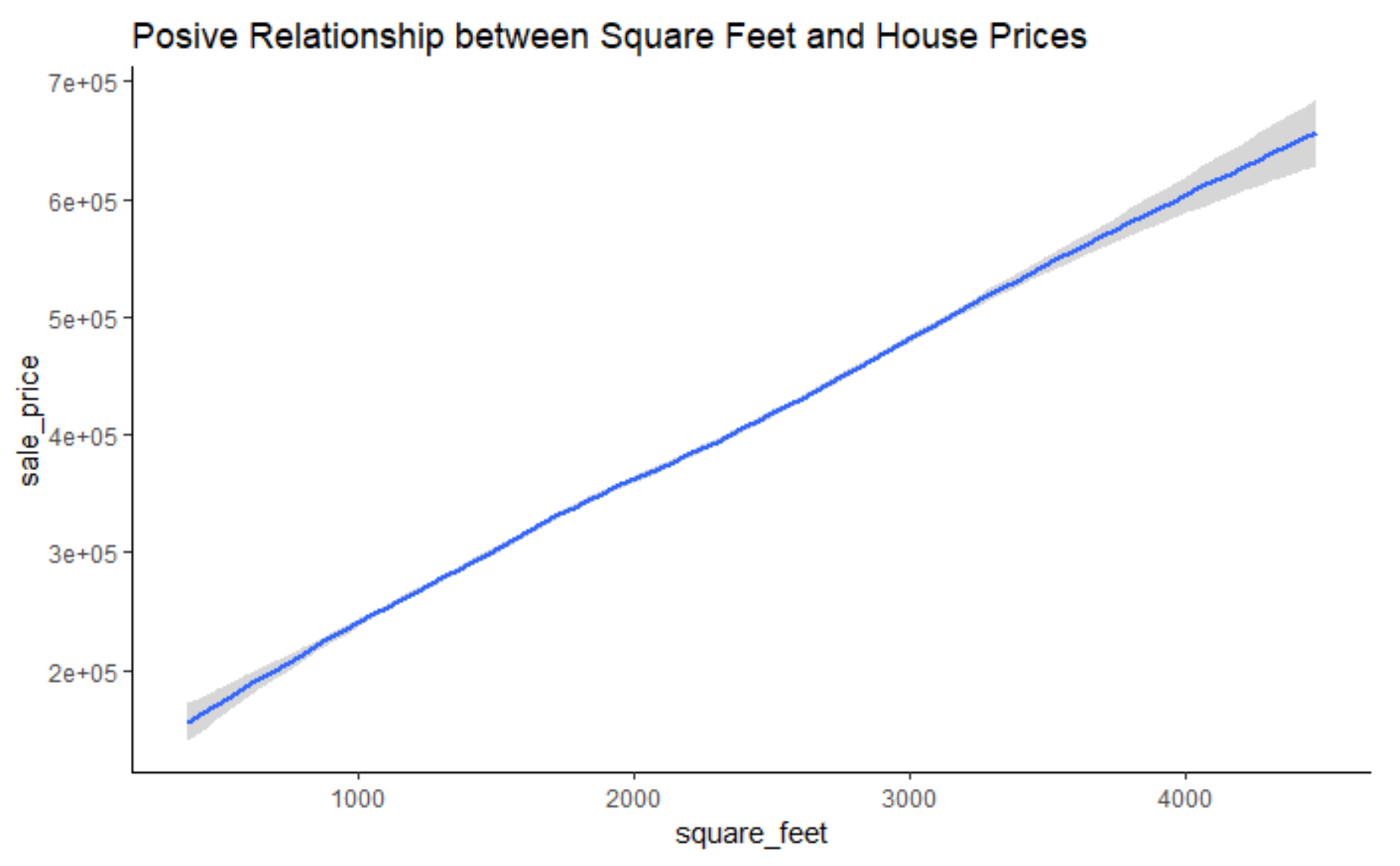


Figure 19: The relationship between square feet and house prices

As increased space of a house is always associated with more bathrooms and bedrooms and able to accommodate more household members, house prices also increase with additional number of bathrooms and bedrooms (Ebru and Eban 2011). As assumed, similar results are gained from our study (Figure 20 and Figure 21). From figure 20, when the number of bathrooms reach to around 2.25, the house prices increase faster than before. In figure 21, when there’s under 4 bedrooms, the house prices increase dramatically with an additional bedroom, but such growth of prices slows down if there’s more than 4 rooms.

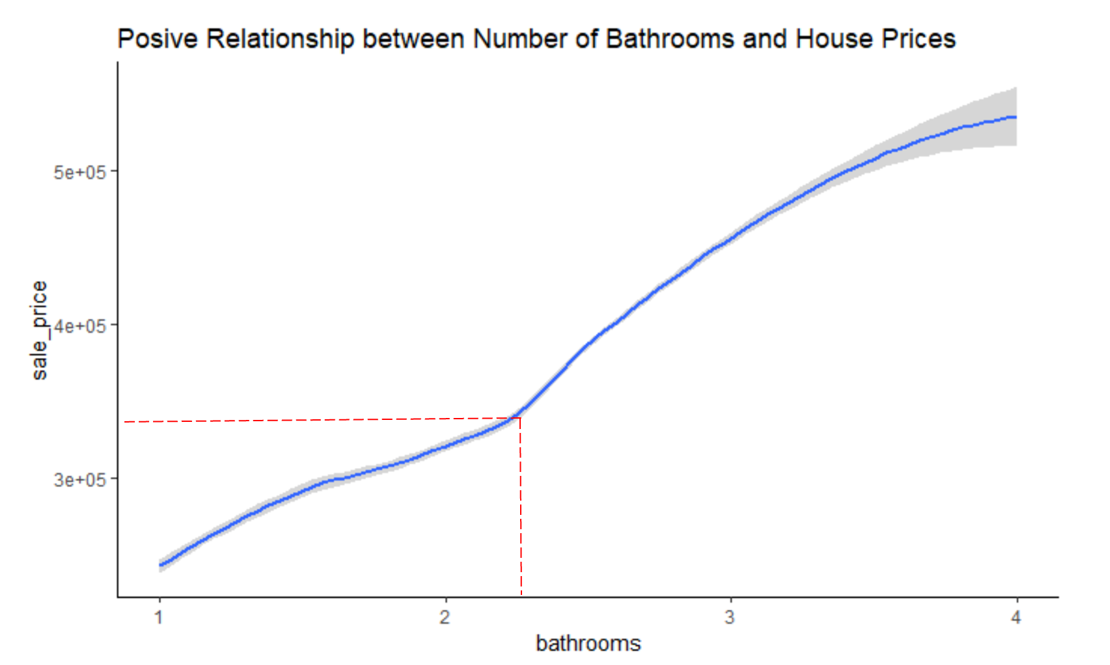


Figure 20: The relationship between bathroom number and house price

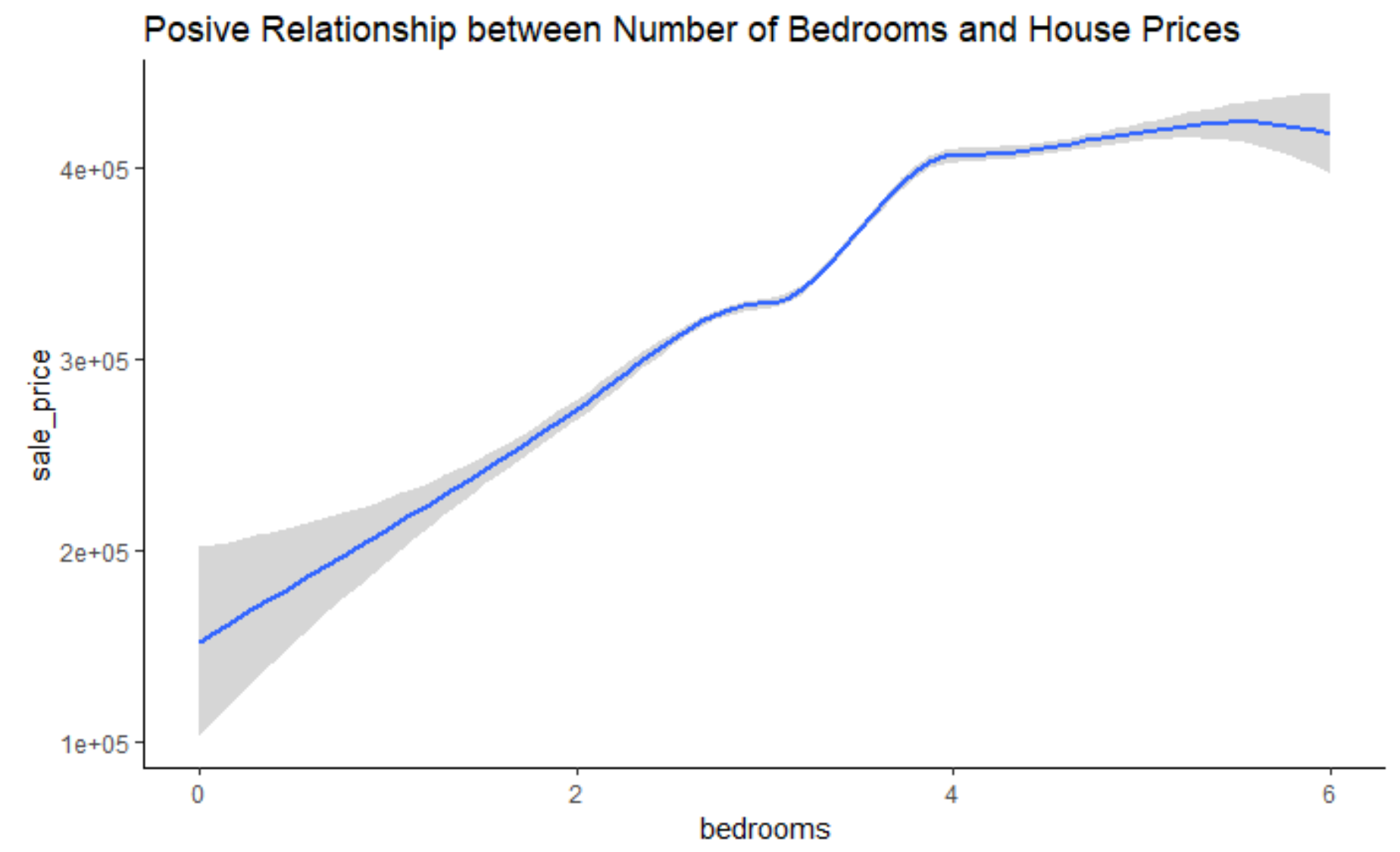


Figure 21: The relationship between bedroom number and house price

(2) Quality

Another important feature is quality, which refers to the quality of the materials used, workmanship, architectural attractiveness and functional design. A house with better quality is always associated with a higher construction cost to afford better material used, more excellent designer, etc. From our study, the better quality always leads to higher housing values. On average, when the quality is regarded as good or higher, the sales price is no less than $50,000 (Figure 22).

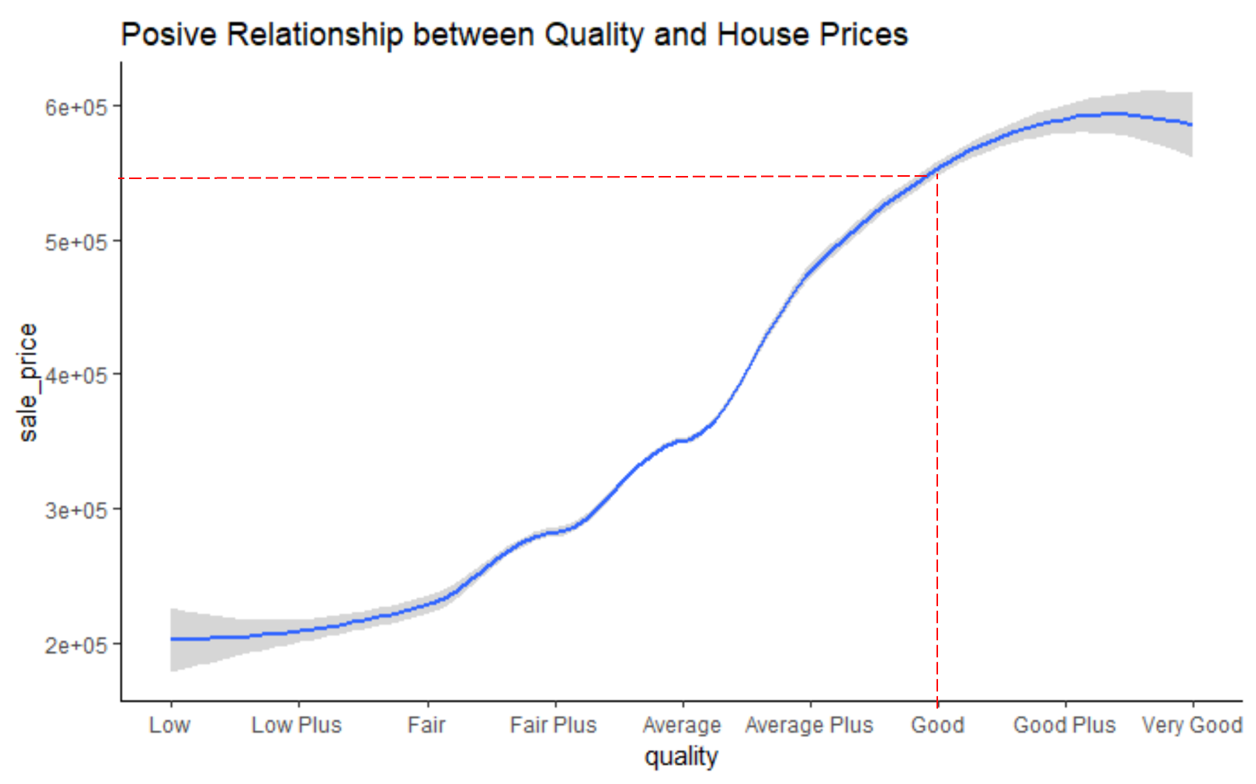


Figure 22: The relationship between quality and house price

(3) Number of Crime

Crime number is the fifth significant determinant of residential property value. As we expected, the number of crimes has a negative effect on the housing prices in Pierce County. From Figure 7, we are surprised to find that the crime number didn’t lead to an obvious decrease in sales prices until the number of crimes reach to 4,000 in the past 12 months, and there’s an apparent negative impact of crime number on house prices.

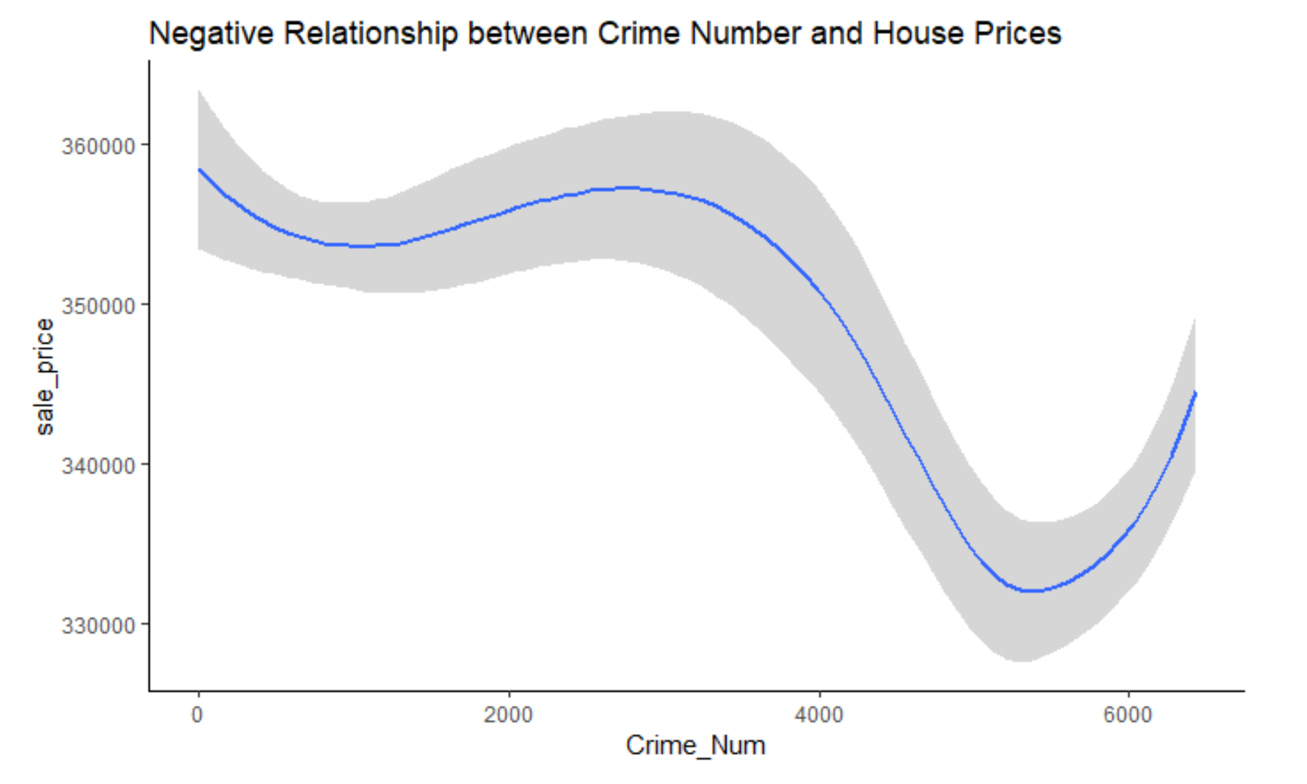


Figure 23: The relationship between crime numbers and house price

(3) Functional Space

Fireplace and basement are both significant factors that affect housing prices in our findings. In Figure 24, we also found that the house prices grow with additional fireplaces. Besides, in Figure 25, on average, a basement is significant to improve a home's value. It is intuitive that a functional space is important for residents, where they are able to have more utility.

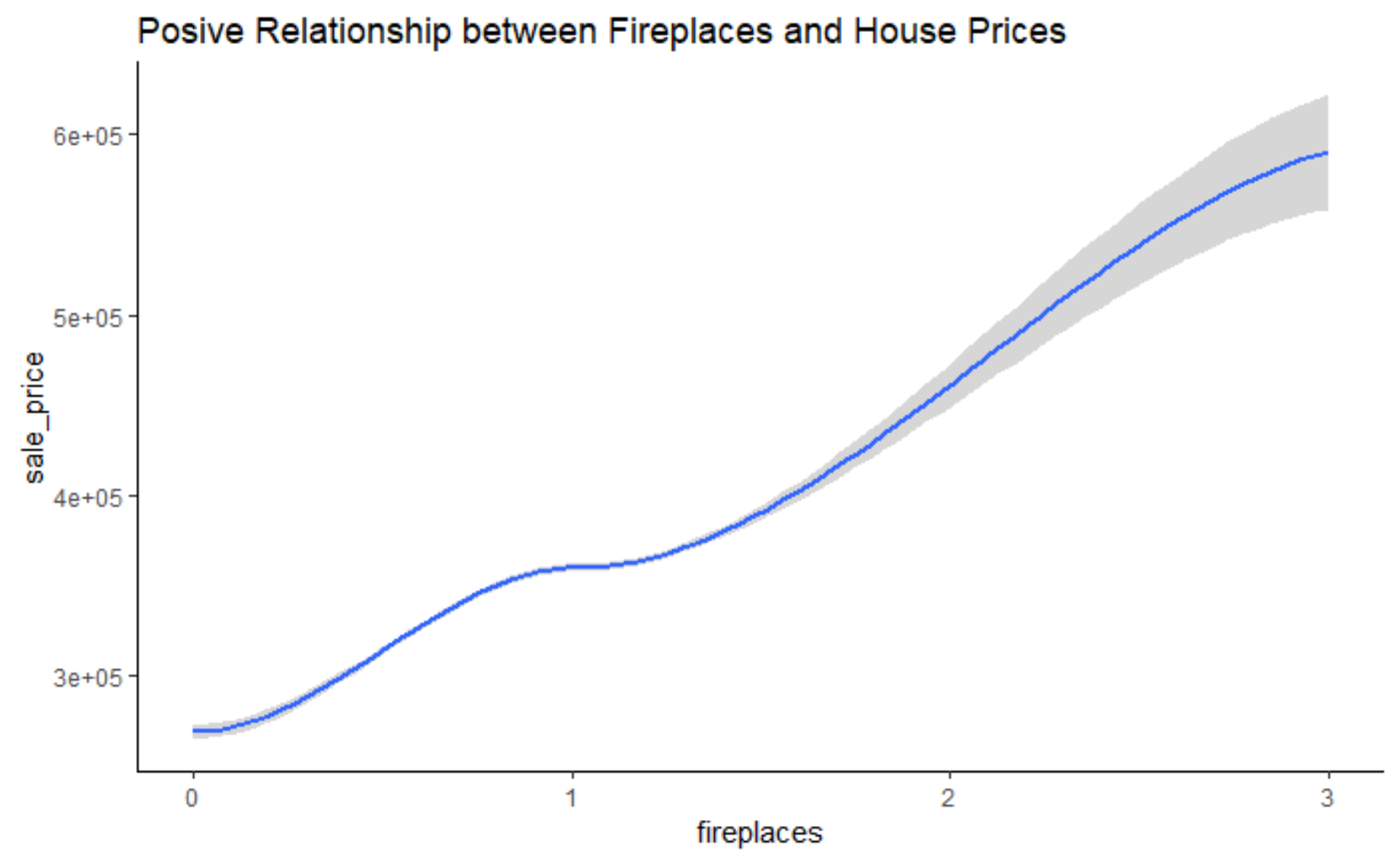
Figure 24: The relationship between fireplaces and house price



Figure 25: The relationship between basement and house price

(4) Garage

Furthermore, by comparing the significance of attached garage square feet and detached garage square feet, we found that attached garage is a more important factor on housing prices than detached garage, because residents as a buyer might be more concerned about the attached garage.

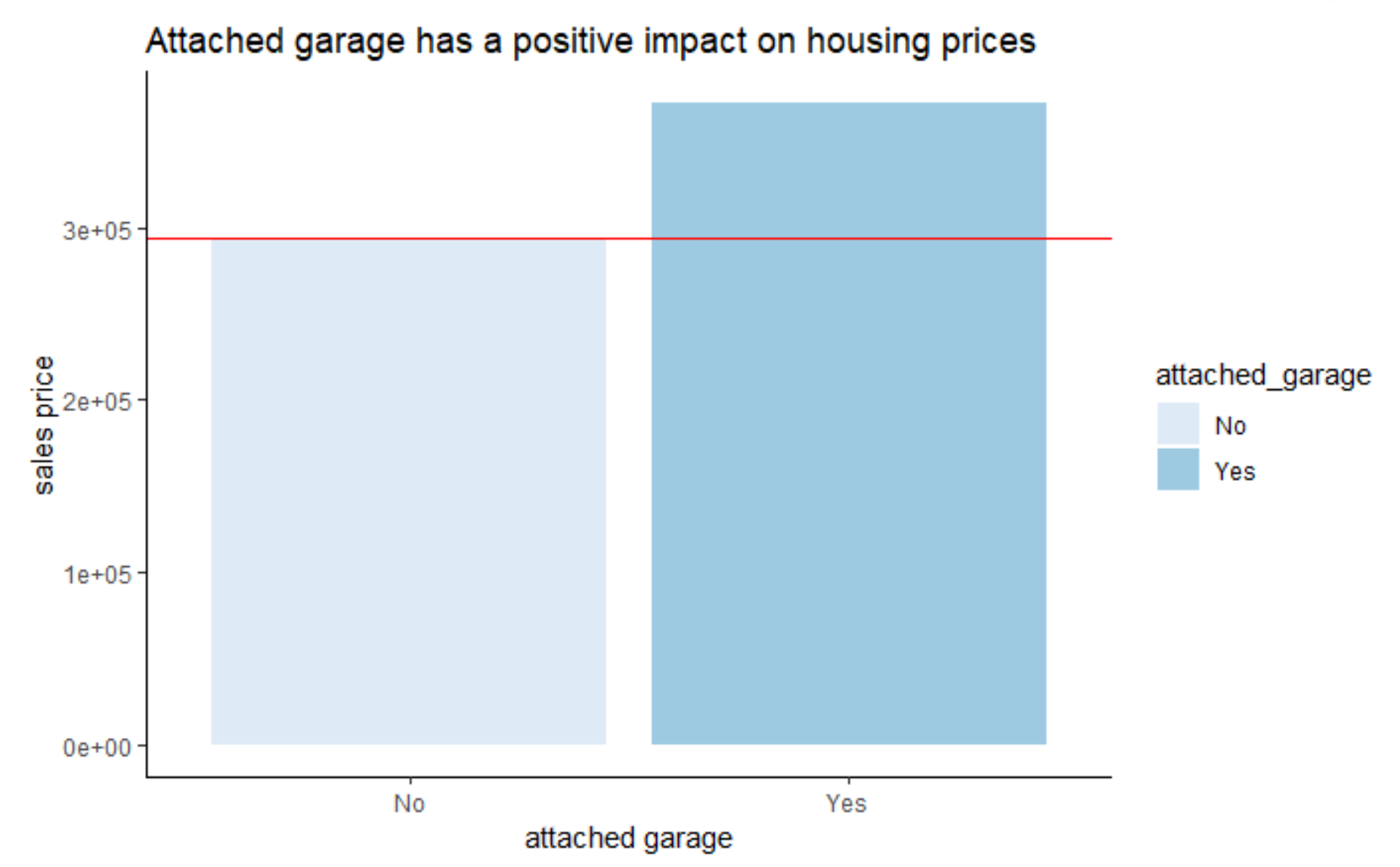


Figure 26: The relationship between the attached garage and house price

(5) Locational Attributes

Besides number of crimes, we also notice that the proximity to college has a slightly more significant effect on house prices relative to other locational characteristics, such as the proximity to private schools, fire station, police station, libraries, high schools, hospitals and elementary school and whether the house has sewer improvement or not. When a house is located near college, on average, there is a slight increase with less than $5,000 in house price according to our findings (figure 27). Overall, the impact of all locational characteristics on housing prices isn’t significant in our study.

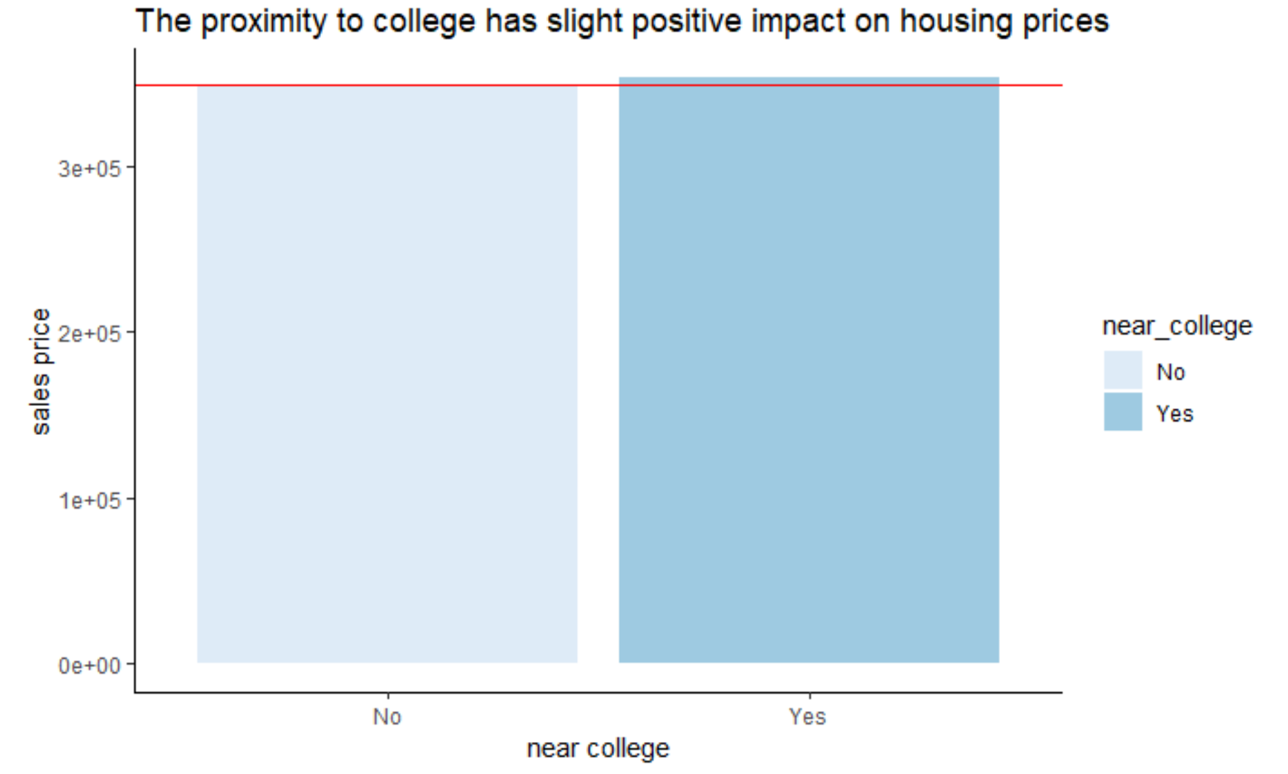


Figure 27: The relationship between proximity to college and house price

## Business Implication

(1) Innovation will improve the property’s selling price.

This study takes the residential property’s original construction year and the reconstruction year as the built year. The final result from the model shows that the new built properties hold a higher price than the old buildings do. In this case, for real estate developers, after acquiring the parcel, it makes sense to invest in the reconstruction and structure improvement work to allow this building a new look. And for local government, it also helps to improve the local houses market value simply by doing the façade innovation.

(2) Allow more spaces for bedroom and bathroom.

The number of bedrooms and bathrooms helps increase the properties’ value. For property developers, who are arranging space between lobby, living room, kitchen, bedroom, bathroom, allowing more space for bedroom and bathroom can help a lot to increase the property value, not the space of basement, living room, or playroom.

(3) Design a fireplace if possible.

Fireplace usually is the central of the conventional house. Today, it is a symbol of a warm family to many people. This study’s result shows that the fireplace has a positive impact on properties’ price. For developers, if it is difficult to build a real fireplace in the house, designing an electronic fireplace screen will also help.

(4) Design a basement and attached garage.

The proprieties with basement area and the attached garage have a greater value. People love to use the basement as the private gym, playroom, video room, personal lab for geeks. The extra space in the base allows more functions to the property which has a positive impact on properties’ value. The attached garage will help to increase the properties’ value, but not the detached garage. People prefer parking in the garage then walking into the house directly.

(5) Prosperities closing to the school do not have a higher price.

Prosperities that are close to the educational facilities do not hold a higher price. The possible reason was that people go to the school by school bus, public transportation, or using the private car, the short spatial distance does not help to increase the properties’ price.

(6) Sewer Improvement Area does not help to increase properties’ value.

Many areas in the Pierce County have improved the sewer system. However, the improved system does not help to improve the price of properties which are benefited from the better system.

## Limitation

From the Pierce County database, we include many influential features which impact the housing prices, out of those there are some features that have more than 87 percent null values. However, we include these features in the model evaluations because we consider them important for our study. These features include: View\_Quality, Basement Square Feet, Basement Finished Square Feet, Bedrooms and Bathrooms.

Hence, one of the limitations of our study include existing potential bias in the model because of these null values present in the original dataset obtained from Pierce county database. Furthermore, another limitation of our study is not including any additional features related to walkability factors which influences the housing prices as discussed in the literature review.

The data limitations of this study include the potential problem of omitted variable bias (OVB). If OVB is present in the statistical modeling above, it indicates that there is an unaccounted component in the error term that is correlated with both the dependent variable and one or more of the independent variables. Therefore, there is potential biases included in the coefficient on those variables either positively or negatively.  To correct the potential limitations which might occur due to OVB, more data would be needed on housing characteristics or other potentially important variables. For example, we have not included the income characteristics of household which could potentially create OVB.

## Further Direction

(1) Neighborhood

The neighborhood characteristics such as general population demographics can have significant impact on housing prices. In our study, we have not included these numbers however, we consider this as an important factor which influences the housing prices.

(2) Education

Another housing characteristic which is important for future enhancements is education. For example, a household which includes people with higher education prefers to live in a community where there are better facilities available such as good neighboring hospitals, better schools and area with less crime rate. A higher number of such features could have significant impact on the sale price.

(3) Income Level

At last, we consider income level of household as one of the most significant factors in determining the housing sale price.  A higher income level households might be interested in purchasing a property which has higher sale price. Higher median income could have potential impact on the housing sale price.

At the end, the final and obvious question to ask from the above analysis is: does including the external features and clustering improves the overall model and is this approach superior to the one where the model does not include any external features and clusters?  While this study has some limitations and potential for better outcomes, the results on adding external influential characteristics offer a contribution to better understand the current state of housing prices in the Pierce County region of Washington state.

# References

Thaler, Richard. 1978 "A Note on the Value of Crime Control Evidence from the Property Market " Journal of Urban Economics, 5(1) 137-45

Gibbons, Steve. 2004 "The Costs of Urban Property Crime " Economic Journal, 114(499) F441-63

Tita, George & L. Petras, Tricia & Greenbaum, Robert. (2006). Crime and Residential Choice: A Neighborhood Level Analysis of the Impact of Crime on Housing Prices. Journal of Quantitative Criminology. 22. 299-317. 10.1007/s10940-006-9013-z.

Weicher J.C. and Hartzell D. (1982). Hedonic analysis of home prices: results for 59 metropolitan areas. Research in Real Estate 2 267-91.

Ceccato, V., Wilhelmsson, M. (2018) Does crime impact real estate prices? An assessment of accessibility and location1 In: Gerben J.N. Bruinsma and Shane D. Johnson (ed.), Oxford Handbook on Environmental Criminology Oxford University Press

C. M. Hui, Eddie & K. Chau, C & Pun, Lilian & Law, M.Y.. (2007). Measuring the neighboring and environmental effects on residential property value: Using spatial weighting matrix. Building and Environment. 42. 2333-2343. 10.1016/j.buildenv.2006.05.004.

Black, S. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education. The Quarterly Journal of Economics,114(2), 577-599.

Boyle, A., Barrilleaux, C., & Scheller, D. (2014). Does Walkability Influence Housing Prices? Social Science Quarterly (Wiley-Blackwell), 95(3), 852–867.

Tan Teck-Hong, (2011),"Neighborhood preferences of house buyers: the case of Klang Valley, Malaysia", International Journal of Housing Markets and Analysis, Vol. 4 Iss 1 pp. 58 - 69

Oxford S. (2002). Valuing locational externalities: a GIS and multilevel modeling approach. Environment and Planning B: Planning and Design 29(1) 105-127.

Des Rosiers F., Lagana A, Thériault M. and Beaudoin M. (1996). Shopping centres and house values: an empirical investigation. Journal of Property Valuation and Investment 14(4) 41-62.

Kendree J.M. and Rauch D.A. (1990). Toward a theory on the effects of view and size on the price of real estate. the ARES annual meeting, Lake Tahoe, NV.

McMillan M.L., Reid B.G. and Gillen D.W. (1980). An extension of the hedonic approach for estimating the value of quiet. Land Economics 56(3) 315-28.

Fan, G., Ong, Z. S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach.Urban Studies, 43(12), 2301–2315.

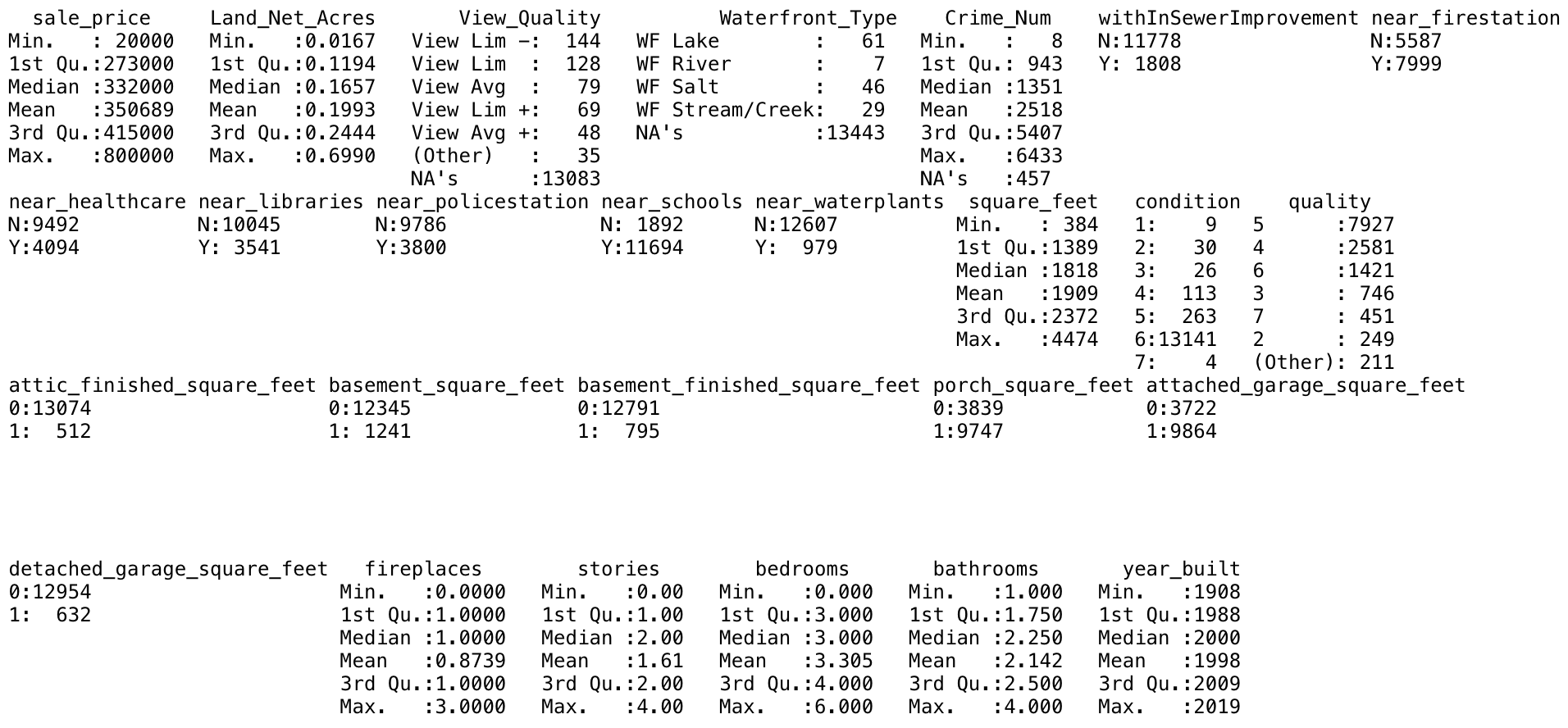
Moore, T., Jesse, C., & Kittler, R. (2001, May). An overview and evaluation of decision tree methodology. In American Statistical Association quality and productivity conference papers. University of Texas, Austin.

Ebru, Ç., & Eban, A. (2011). Determinants of house prices in Istanbul: a quantile regression approach. Quality & Quantity, 45(2), 305-317.

# Appendices

## Statistical summary – Before the outlier treatment

## Statistical summary – After the outlier treatment



## Sample Codes

(1) Data integration - create view to select variables from raw table in SQL.

create view v\_appraisal\_account

as

select Parcel\_Number as aa\_Parcel\_Number, Appraisal\_Account\_Type, Land\_Economic\_Area, Land\_Net\_Acres, Appraisal\_Date, Utility\_Sewer, Utility\_Water, Utility\_Electric, View\_Quality, Waterfront\_Type

from dbo.appraisal\_account;

(2) Data integration - create view to select variables from raw table in SQL.

create view v\_appraisal\_account

as

select Parcel\_Number as aa\_Parcel\_Number, Appraisal\_Account\_Type, Land\_Economic\_Area, Land\_Net\_Acres, Appraisal\_Date, Utility\_Sewer, Utility\_Water, Utility\_Electric, View\_Quality, Waterfront\_Type

from dbo.appraisal\_account;

(3) Data integration - joining the partial master table in SQL.

create view v\_master\_table\_mid\_product\_nodup2

as

select vs.s\_Parcel\_Number as Parcel\_Number,vs.sale\_price, vs.Sale\_Date, vraa.Land\_Economic\_Area, vraa.Land\_Net\_Acres, vraa.Appraisal\_Date,vraa.Utility\_Sewer,

vraa.Utility\_Water, vraa.Utility\_Electric, vraa.View\_Quality, vraa.Waterfront\_Type,

cd.Crime\_Num, ISNULL(si.withInSewerImprovement, 'N') AS withInSewerImprovement, ISNULL(nf.near\_firestation, 'N') AS near\_firestation,

ISNULL(nh.near\_healthcare, 'N') AS near\_healthcare, ISNULL(nl.near\_libraries, 'N') AS near\_libraries, ISNULL(np.near\_policestation, 'N') AS near\_policestation,

ISNULL(ns.near\_schools, 'N') as near\_schools, ISNULL(nw.near\_waterplants, 'N') as near\_waterplants

from dbo.v\_sale\_2018\_nodup as vs

left join dbo.v\_resi\_appraisal\_account as vraa on vs.s\_Parcel\_Number = vraa.aa\_Parcel\_Number

left join dbo.v\_tax\_account as vta on vs.s\_Parcel\_Number = vta.ta\_Parcel\_Number

left join dbo.crime\_data as cd on (vta.Range = cd.Range and vta.Township = cd.Township)

left join dbo.IN\_SEWER\_IMPROVEMENT\_DISTRICT as si on vs.s\_Parcel\_Number = si.taxparceln

left join dbo.NEAR\_FIRESTATION as nf on vs.s\_Parcel\_Number = nf.taxparceln

left join dbo.NEAR\_HEALTHCARE as nh on vs.s\_Parcel\_Number = nh.taxparceln

left join dbo.NEAR\_LIBRARIES as nl on vs.s\_Parcel\_Number = nl.taxparceln

left join dbo.NEAR\_POLICESTATION as np on vs.s\_Parcel\_Number = np.taxparceln

left join dbo.NEAR\_SCHOOLS as ns on vs.s\_Parcel\_Number = ns.taxparceln

left join dbo.NEAR\_WATERPLANTS as nw on vs.s\_Parcel\_Number = nw.taxparceln;

(4) Data integration - joining the final master table in R.

master <-

master %>%

inner\_join(full\_improvement\_update, by = c("Parcel\_Number"))

(5) Outlier Treatment

boxplot(master$bedrooms)

boxplot(master$bathrooms)

boxplot(master$square\_feet)

boxplot(master$Land\_Net\_Acres)

boxplot(master$stories)

boxplot(master$sale\_price)

hist(master$bedrooms)

hist(master$bathrooms)

hist(master$square\_feet)

hist(master$Land\_Net\_Acres)

hist(master$stories)

hist(master$sale\_price)

master <- master %>% filter(between(sale\_price, 20000, 800000),

Land\_Net\_Acres < .7,

between(square\_feet, 120, 4500),

bedrooms <= 6,

between(bathrooms, 1, 4),

between(stories, 0, 4),

fireplaces < 4)

(6) Clustering and Sum or Squares Comparison

m1\_scaled <- scale(m1\_full)

km1 <- kmeans(m1\_scaled, 3)

bt1 <- km1$betweenss/km1$totss

bt1

(7) Random Forest Model Sample Code

set.seed(1234) #to generate the same random numbers

rs <- sample(nrow(m1\_full), .8\*nrow(m1\_full))

training <- m1\_full[rs,]

testing <- m1\_full[-rs,]

dim(training)

dim(testing)

attach(training)

rfm <- randomForest(sale\_price ~., data = training, ntree=1000)

print(rfm)

mse <- sum((rfm$predicted - training$sale\_price)^2)/nrow(training)

mse

p1 <- predict(rfm, testing[,-1])

mse2 <- sum((p1 - testing$sale\_price)^2)/nrow(testing)

mse2

mape <- MAPE(p1, testing$sale\_price)

mape

varImpPlot(rfm)

importance(rfm)

(8) Decision Tree Model

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn import metrics

from sklearn import tree

import matplotlib as plt

# import the regressor

from sklearn.tree import DecisionTreeRegressor

# import export\_graphviz

from sklearn.tree import export\_graphviz

from sklearn.metrics import mean\_squared\_error

from sklearn.ensemble import RandomForestRegressor

from sklearn.datasets import make\_regression

from sklearn.utils import check\_array

def mean\_absolute\_percentage\_error(y\_true, y\_pred):

y\_true = y\_true.values

y\_true = y\_true.reshape(-1,1)

y\_pred = y\_pred.reshape(-1,1)

y\_true = check\_array(y\_true)

y\_pred = check\_array(y\_pred)

return round(np.mean(np.abs((y\_true - y\_pred) / y\_true)) \* 100,2)

########## load data ########

m1\_c1 = pd.read\_csv("m2\_c5\_kmean-5clusters\m1\_c1\_kmean.csv", sep=',', header=0)

m1\_c2 = pd.read\_csv("m2\_c5\_kmean-5clusters\m1\_c2\_kmean.csv", sep=',', header=0)

m1\_c3 = pd.read\_csv("m2\_c5\_kmean-5clusters\m1\_c3\_kmean.csv", sep=',', header=0)

m1\_c4 = pd.read\_csv("m2\_c5\_kmean-5clusters\m1\_c4\_kmean.csv", sep=',', header=0)

m1\_c5 = pd.read\_csv("m2\_c5\_kmean-5clusters\m1\_c5\_kmean.csv", sep=',', header=0)

m1\_full = pd.read\_csv("m1\_full.csv", sep=',', header=0)

m2\_c1 = pd.read\_csv("m2\_c5\_kmean-5clusters\m2\_c1\_kmean.csv", sep=',', header=0)

m2\_c2 = pd.read\_csv("m2\_c5\_kmean-5clusters\m2\_c2\_kmean.csv", sep=',', header=0)

m2\_c3 = pd.read\_csv("m2\_c5\_kmean-5clusters\m2\_c3\_kmean.csv", sep=',', header=0)

m2\_c4 = pd.read\_csv("m2\_c5\_kmean-5clusters\m2\_c4\_kmean.csv", sep=',', header=0)

m2\_c5 = pd.read\_csv("m2\_c5\_kmean-5clusters\m2\_c5\_kmean.csv", sep=',', header=0)

m2\_full = pd.read\_csv("m2\_full.csv", sep=',', header=0)

def get\_df\_name(df):

name =[x for x in globals() if globals()[x] is df][0]

return name

########## function for full data set #############

def fullModel(m1\_c1):

y = m1\_c1['sale\_price']

m1\_c1['View\_Quality'] = m1\_c1['View\_Quality'] .astype('category')

m1\_c1['Waterfront\_Type'] = m1\_c1['Waterfront\_Type'] .astype('category')

m1\_c1['withInSewerImprovement'] = m1\_c1['withInSewerImprovement'] .astype('category')

m1\_c1['near\_firestation'] = m1\_c1['near\_firestation'] .astype('category')

m1\_c1['near\_hospital'] = m1\_c1['near\_hospital'] .astype('category')

m1\_c1['near\_libraries'] = m1\_c1['near\_libraries'] .astype('category')

m1\_c1['near\_policestation'] = m1\_c1['near\_policestation'] .astype('category')

m1\_c1['near\_waterplants'] = m1\_c1['near\_waterplants'] .astype('category')

m1\_c1['condition'] = m1\_c1['condition'] .astype('category')

m1\_c1['quality'] = m1\_c1['quality'] .astype('category')

m1\_c1['attic\_finished\_square\_feet'] = m1\_c1['attic\_finished\_square\_feet'] .astype('category')

m1\_c1['basement\_square\_feet'] = m1\_c1['basement\_square\_feet'] .astype('category')

m1\_c1['basement\_finished\_square\_feet'] = m1\_c1['basement\_finished\_square\_feet'] .astype('category')

m1\_c1['porch\_square\_feet'] = m1\_c1['porch\_square\_feet'] .astype('category')

m1\_c1['attached\_garage\_square\_feet'] = m1\_c1['attached\_garage\_square\_feet'] .astype('category')

m1\_c1['detached\_garage\_square\_feet'] = m1\_c1['detached\_garage\_square\_feet'] .astype('category')

m1\_c1['fireplaces'] = m1\_c1['fireplaces'] .astype('category')

m1\_c1['near\_private\_school'] = m1\_c1['near\_private\_school'] .astype('category')

m1\_c1['near\_elementary\_school'] = m1\_c1['near\_elementary\_school'] .astype('category')

m1\_c1['near\_high\_school'] = m1\_c1['near\_high\_school'] .astype('category')

m1\_c1['near\_college'] = m1\_c1['near\_college'] .astype('category')

m1\_c1['Crime\_Num'] = m1\_c1['Crime\_Num'].fillna(0)

X = m1\_c1[['Land\_Net\_Acres','View\_Quality',

'Waterfront\_Type', 'Crime\_Num', 'withInSewerImprovement',

'near\_firestation', 'near\_hospital', 'near\_libraries',

'near\_policestation', 'near\_waterplants', 'square\_feet',

'condition', 'quality', 'attic\_finished\_square\_feet',

'basement\_square\_feet', 'basement\_finished\_square\_feet',

'porch\_square\_feet', 'attached\_garage\_square\_feet',

'detached\_garage\_square\_feet', 'fireplaces', 'stories', 'bedrooms',

'bathrooms', 'year\_built', 'near\_private\_school', 'near\_elementary\_school', 'near\_college', 'near\_high\_school']]

########### decision tree regression ##############

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.20,random\_state=109)

# create a regressor object

regressor = DecisionTreeRegressor(criterion='mse', splitter='best', max\_depth=5, min\_samples\_split=2, min\_samples\_leaf=1, min\_weight\_fraction\_leaf=0.0, max\_features=None,

random\_state=0, max\_leaf\_nodes=None, min\_impurity\_decrease=0.0, min\_impurity\_split=None, presort=False)

# fit the regressor with X and Y data

reg = regressor.fit(X\_train, y\_train)

y\_pred = reg.predict(X\_test)

MSE = mean\_squared\_error(y\_test, y\_pred)

print("MSE of "+ get\_df\_name(m1\_c1) + '\t' ,MSE)

print('MAPE of '+ get\_df\_name(m1\_c1) + '\t',mean\_absolute\_percentage\_error(y\_test, y\_pred),'%')

# export the decision tree to a tree.dot file

# for visualizing the plot easily anywhere

feat\_importance = regressor.tree\_.compute\_feature\_importances(normalize=False)

indices = np.argsort(feat\_importance)[::-1]

# Print the feature ranking

print("Feature ranking:")

for f in range(X.shape[1]):

print("%d. feature %d (%f)" % (f + 1, indices[f], feat\_importance[indices[f]]))

export\_graphviz(regressor, out\_file =get\_df\_name(m1\_c1) + '\_5\_cluster\_tree.dot',feature\_names=X.columns)

fullModel(m1\_c1)

fullModel(m1\_c2)

fullModel(m1\_c3)

fullModel(m1\_c4)

fullModel(m1\_c5)

fullModel(m1\_full)

########## function for original data set #############

def orignalModel(m2\_c1):

y = m2\_c1['sale\_price']

m2\_c1['View\_Quality'] = m2\_c1['View\_Quality'] .astype('category')

m2\_c1['Waterfront\_Type'] = m2\_c1['Waterfront\_Type'] .astype('category')

m2\_c1['condition'] = m2\_c1['condition'] .astype('category')

m2\_c1['quality'] = m2\_c1['quality'] .astype('category')

m2\_c1['attic\_finished\_square\_feet'] = m2\_c1['attic\_finished\_square\_feet'] .astype('category')

m2\_c1['basement\_square\_feet'] = m2\_c1['basement\_square\_feet'] .astype('category')

m2\_c1['basement\_finished\_square\_feet'] = m2\_c1['basement\_finished\_square\_feet'] .astype('category')

m2\_c1['porch\_square\_feet'] = m2\_c1['porch\_square\_feet'] .astype('category')

m2\_c1['attached\_garage\_square\_feet'] = m2\_c1['attached\_garage\_square\_feet'] .astype('category')

m2\_c1['detached\_garage\_square\_feet'] = m2\_c1['detached\_garage\_square\_feet'] .astype('category')

m2\_c1['fireplaces'] = m2\_c1['fireplaces'] .astype('category')

X = m2\_c1[['Land\_Net\_Acres','View\_Quality',

'Waterfront\_Type', 'square\_feet',

'condition', 'quality', 'attic\_finished\_square\_feet',

'basement\_square\_feet', 'basement\_finished\_square\_feet',

'porch\_square\_feet', 'attached\_garage\_square\_feet',

'detached\_garage\_square\_feet', 'fireplaces', 'stories', 'bedrooms',

'bathrooms', 'year\_built']]

########### decision tree regression ##############

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.20)

# create a regressor object

regressor = DecisionTreeRegressor(criterion='mse', splitter='best', max\_depth=7, min\_samples\_split=2, min\_samples\_leaf=1, min\_weight\_fraction\_leaf=0.0, max\_features=None,

random\_state=0, max\_leaf\_nodes=None, min\_impurity\_decrease=0.0, min\_impurity\_split=None, presort=False)

# fit the regressor with X and Y data

reg = regressor.fit(X\_train, y\_train)

y\_pred = reg.predict(X\_test)

MSE = mean\_squared\_error(y\_test, y\_pred)

print("MSE of "+ get\_df\_name(m2\_c1) + '\t',MSE)

print('MAPE of '+ get\_df\_name(m2\_c1) + '\t',mean\_absolute\_percentage\_error(y\_test, y\_pred),'%')

# export the decision tree to a tree\_5cluster.dot file

# for visualizing the plot easily anywhere

export\_graphviz(regressor, out\_file =get\_df\_name(m2\_c1) + '\_5\_cluster\_tree.dot',feature\_names=X.columns)

feat\_importance = regressor.tree\_.compute\_feature\_importances(normalize=False)

print(feat\_importance)

orignalModel(m2\_c1)

orignalModel(m2\_c2)

orignalModel(m2\_c3)

orignalModel(m2\_c4)

orignalModel(m2\_c5)

orignalModel(m2\_full)

(9) Artificial Neural Network Model

import pandas as pd

import patsy

import numpy as np

import statsmodels.api as sm

import statsmodels.formula.api as smf

import statsmodels.tools as sm\_tools

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split, GridSearchCV

from sklearn.svm import SVC

from sklearn.neural\_network import MLPClassifier

from sklearn.metrics import mean\_squared\_error

import graphviz

import os

import matplotlib.pyplot as plt

from sklearn.utils import check\_array

df\_pierce\_county = pd.read\_csv("m1\_full.csv", sep=',', header=0)

print(df\_pierce\_county.describe())

print(df\_pierce\_county.info())

def mean\_absolute\_percentage\_error(y\_test, y\_pred):

y\_test = y\_test.values

y\_test = y\_test.reshape(-1,1)

y\_pred = y\_pred.reshape(-1,1)

y\_test = check\_array(y\_test)

y\_pred = check\_array(y\_pred)

return np.mean(np.abs((y\_test - y\_pred)/y\_test)) \* 100

y = df\_pierce\_county['sale\_price']

print(df\_pierce\_county.dtypes)

df\_pierce\_county['View\_Quality'] = df\_pierce\_county['View\_Quality'] .astype('category')

df\_pierce\_county['Waterfront\_Type'] = df\_pierce\_county['Waterfront\_Type'] .astype('category')

df\_pierce\_county['withInSewerImprovement'] = df\_pierce\_county['withInSewerImprovement'] .astype('category')

df\_pierce\_county['near\_firestation'] = df\_pierce\_county['near\_firestation'] .astype('category')

df\_pierce\_county['near\_hospital'] = df\_pierce\_county['near\_hospital'] .astype('category')

df\_pierce\_county['near\_libraries'] = df\_pierce\_county['near\_libraries'] .astype('category')

df\_pierce\_county['near\_policestation'] = df\_pierce\_county['near\_policestation'] .astype('category')

df\_pierce\_county['near\_waterplants'] = df\_pierce\_county['near\_waterplants'] .astype('category')

df\_pierce\_county['condition'] = df\_pierce\_county['condition'] .astype('category')

df\_pierce\_county['quality'] = df\_pierce\_county['quality'] .astype('category')

df\_pierce\_county['attic\_finished\_square\_feet'] = df\_pierce\_county['attic\_finished\_square\_feet'] .astype('category')

df\_pierce\_county['basement\_square\_feet'] = df\_pierce\_county['basement\_square\_feet'] .astype('category')

df\_pierce\_county['basement\_finished\_square\_feet'] = df\_pierce\_county['basement\_finished\_square\_feet'] .astype('category')

df\_pierce\_county['porch\_square\_feet'] = df\_pierce\_county['porch\_square\_feet'] .astype('category')

df\_pierce\_county['attached\_garage\_square\_feet'] = df\_pierce\_county['attached\_garage\_square\_feet'] .astype('category')

df\_pierce\_county['detached\_garage\_square\_feet'] = df\_pierce\_county['detached\_garage\_square\_feet'] .astype('category')

df\_pierce\_county['fireplaces'] = df\_pierce\_county['fireplaces'] .astype('category')

df\_pierce\_county['near\_private\_school'] = df\_pierce\_county['near\_private\_school'] .astype('category')

df\_pierce\_county['near\_elementary\_school'] = df\_pierce\_county['near\_elementary\_school'] .astype('category')

df\_pierce\_county['near\_high\_school'] = df\_pierce\_county['near\_high\_school'] .astype('category')

df\_pierce\_county['near\_college'] = df\_pierce\_county['near\_college'] .astype('category')

df\_pierce\_county['Crime\_Num'] = df\_pierce\_county['Crime\_Num'].fillna(0)

print(df\_pierce\_county.columns)

X = df\_pierce\_county[['Land\_Net\_Acres','View\_Quality',

'Waterfront\_Type', 'Crime\_Num', 'withInSewerImprovement',

'near\_firestation', 'near\_hospital', 'near\_libraries',

'near\_policestation', 'near\_waterplants', 'square\_feet',

'condition', 'quality', 'attic\_finished\_square\_feet',

'basement\_square\_feet', 'basement\_finished\_square\_feet',

'porch\_square\_feet', 'attached\_garage\_square\_feet',

'detached\_garage\_square\_feet', 'fireplaces', 'stories', 'bedrooms',

'bathrooms', 'year\_built', 'near\_private\_school', 'near\_elementary\_school', 'near\_college', 'near\_high\_school']]

# ------------------------------------------------------------- #

# --------------------- Neural Network ------------------------ #

# ------------------------------------------------------------- #

# -------------------- Pierce County Model ------------------------##

# Convert all categorical variables to a matrix of zeros and ones

df\_pierce\_county = pd.get\_dummies(df\_pierce\_county)

print(df\_pierce\_county.head())

## Standardizing data improves computations and makes sure all features are weighted equally for NNs

scaler = StandardScaler()

df\_pierce\_county = scaler.fit\_transform(df\_pierce\_county)

## Split the dataset into training and testing data

X\_train\_PC, X\_test\_PC, y\_train\_PC, y\_test\_PC = train\_test\_split(df\_pierce\_county,y,test\_size =0.2,random\_state=109)

nn1\_m1\_full = MLPClassifier(hidden\_layer\_sizes = (3), activation='logistic', random\_state=109)

nn1\_m1\_full.fit(X\_train\_PC, y\_train\_PC)

y\_pred\_nn1\_m1\_full = nn1\_m1\_full.predict(X\_test\_PC)

print("Neural Network Classifier Model Identity: PierceCounty")

MSE = mean\_squared\_error(y\_test\_PC, y\_pred\_nn1\_m1\_full)

print("MSE for NN with 2 hidden layers and identity activation :",MSE)

MAPE = mean\_absolute\_percentage\_error(y\_test\_PC, y\_pred\_nn1\_m1\_full)

print("MAPE: ", MAPE)

## Data Dictionary Details

(1) appraisal\_account

Table 10: Data dictionary - Appraisal Account

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **valid values** | **Data Type** | **Length** | **Example Values** | **Measure** | **Null Ratio** |
| **Parcel\_Number** | Unique 10-digit number assigned to each property. | 9-digits tax parcel number | varchar | 10 | 0019012000 | Nominal | 0% |
| Appraisal\_Account\_Type | How a parcel is classified. (Com Condo, Com Leasehold, Com Multi Unit, Commercial, Industrial, Mobile Home, Reference, Res Com Condo, Res Leasehold, Residential, Trended invest) | Com Condo/ Com Leasehold/ Com Multi Unit/ Commercial/ Condominium/ Industrial/ Mobile Home/ Reference/ Res Leasehold/ Residential/ Trended Invest | varchar | 15 | Residential | Nominal | 0.00% |
| Land\_Net  \_Acres | Size of the individual parcel in acres. | Positive number | decimal | (15,4) | 18.0800 | Numerical | 0.00% |
| Waterfront\_Type | Describes the type of waterfront property adjoins or has legal access to. | Text describing the waterfront type | varchar | 30 | WF Salt | Nominal | 0.39% |
| View\_Quality | Assigned to reflect the market appeal of the overall view available from the dwelling or property. | Text describing the view quality | varchar | 30 | View Lim | Nominal | 95.69% |

Note: Parcel\_Number is the key value

(2) improvement

Table 11: Data dictionary - Improvement

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **valid values** | **Data Type** | **Length** | **Example Values** | **Measure** | **Null Ratio** |
| **Parcel Number** | Unique 10 digit number assigned to each property. | 10-digit number | varchar | 10 | 19121002 | Nominal | 0.00% |
| Property Type | Links the buildings on the individual records to the proper cost and depreciation tables. | Options include Commercial, Duplex, Industrial, Mobile Home, Multiple Unit, Out Building, Residential, Townhouse and Triplex. | varchar | 15 | Residential | Nominal | 0.00% |
| Square Feet | Sum of the square feet for the 'built-as' types identified for the building. Each building may have one or more 'built as' type (see IMPROVEMENT BUILTAS table), though usually only one. The exceptions are 'built as' # 124 and 126 (Add on only Res & Com) or any type of Storage Tank, where it represents the count of those 'built as' as opposed to the sum of their square feet. | Up to 7-digit number | Integer |  | 1673 | Numerical | 0.00% |
| Percent Complete | Describes the stage of new construction. A completed structure is listed at 100% complete. Structures less than 100% complete are listed at the appropriate percent complete based on the appraisers observation. Depending on the duration of the construction, in some cases the appraiser may move the structure to 100% complete with an adjustment to the building for any unfinished work. | Decimal from .00 to 1.00 | Decimal | 15,4 | 1.0000 | Numerical | 0.1% |
| Condition | Captures the overall depreciation of a structure. Condition is a reflection of the maintenance and upkeep of the structure. | Average  Good  Bad | varchar | 15 | Average | Ordinal | 13.82% |
| Quality | Indication of the quality of the materials used, workmanship, architectural attractiveness, and functional design. | Average  Good  Bad | varchar | 15 | Good | Ordinal | 23.81% |
| Attic Finished Square Feet | Finished living area in the attic. | Up to 4-digit number | Integer |  | 441 | Numerical | 0.00% |
| Basement Square Feet | Total square footage of the basement. | Up to 4-digit number | Integer |  | 501 | Numerical | 87.39% |
| Basement Finished Square Feet | Finished square footage of the basement. | Up to 4-digit number | Integer |  | 240 | Numerical | 87.39% |
| Porch Square Feet | Total number of square feet associated with all porches. | Up to 4-digit number | Integer |  | 275 | Numerical | 0.00% |
| Attached Garage Square Feet | Total square footage of the attached or built in garage(s). | Up to 4-digit number | Integer |  | 400 | Numerical | 0.00% |
| Detached Garage Square Feet | Total detached garage(s) square footage. | Up to 4-digit number | Integer |  | 650 | Numerical | 0.25% |
| Fireplaces | Total count of single, double or PreFab stoves. | 1-digit number | Integer |  | 1 | Numerical | 0.25% |

Note: Parcel\_Number is the key value

(3) improvement\_builtas

Table 12: Data dictionary - Improvement Builtas

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | DescriptionValid Values | Valid Values | Data Type | Length | Example | Data Measurement | Null Ratio |
| Parcel Number | Unique 10 digit number assigned to each property. | 10-digit number | Text | 10 | 19121002 | Nominal | 0.00% |
| Stories | Number of floors/building levels above grade. Stories do not include attic or basement areas. | Up to 2-digit number | Decimal | 13,2 | 1.00 | Numerical | 0.61% |
| Bedrooms | Number of bedrooms listed for a residential property. (Collected for informational purposes only.) | Up to 3-digit number | Integer |  | 2 | Numerical | 93.39% |
| Bathrooms | Number of baths listed for a residential property. The number is listed as a decimal, i.e. 2.75 = two full and one three-quarter baths. A tub/sink/toilet combination (plus any additional fixtures) is considered 1.0 bath. A shower/sink/toilet combination (plus any additional fixtures) is 0.75 bath. A sink/toilet combination is .5 bath. | Up to 3-digit number | Decimal | 7,2 | 1.5 | Numerical | 94.79% |
| Year Built | Year the building was built, as stated by the building permit or a historical record. | YYYY | Integer |  | 1994 | Numerical | 99.59% |

Note: Parcel\_Number is the key value

(4) sale

Table 13: Data dictionary - Sale

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **Valid Values** | **Data Type** | **Length** | **Example Values** | **Measure** | **Null Ratio** |
| **Parcel\_Number** | Unique identifier for a property. | 10-digit number | varchar | 10 | 19121002 | nominal | 0.00% |
| Sale\_Date | Date the legal document (deed) was executed. | YYYY-MM-DD | date |  | 1998-10-01 | nominal | 0.00% |
| Price | Dollar amount recorded on the ETN. | Dollar amount in decimal format | decimal | 15,2 | 175000.00 | numerical | 0.00% |

Note: Parcel\_Number is the key value

(5) tax\_account

Table 14: Data dictionary - Tax Account

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **valid values** | **Data Type** | **Length** | **Example Values** | **Measure** | **Null Ratio** |
| **Parcel\_Number** | Unique 10 digit number assigned to each property. | 10 digit number | varchar | 10 | 2805020810 | nominal | 0.00% |
| Range | Range | 2-digit numerical code from 00 to 11 | varchar | 10 | 03 | nominal | 7.55% |
| Township | Township | 2-digit numerical code from 15 to 22 | varchar | 10 | 21 | nominal | 7.55% |

Note: Parcel\_Number is the key value

(6) crime

Table 15: Data dictionary - Crime

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **valid values** | **Data Type** | **Length** | **Example Values** | **Measure** | **Null Ratio** |
| **Range** | Range | 2 digit numerical code from 00 to 11 | varchar | 10 | 0 | nominal | 0.00% |
| **Township** | Township | 2 digit numerical code from 15 to 22 | varchar | 10 | 20 | nominal | 0.00% |
| Crime\_Num | Number of crimes in the past 12 months | Numbers between 3 and 6433 | int | NULL | 91 | numerical | 0.00% |

Notes: Range and Township is the composite primary key

(6) GIS Data

Table 16: Data dictionary - GIS Data

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Description** | **valid values** | **Data Type** | **Length** | **Example Values** | **Measure** | **Null Ratio** |
| **Parcel\_Number** | Unique identifier for a property. | 10-digit number | varchar | 10 | 19121002 | nominal | 0.00% |
| withInSewerImprovement | The parcel located in the Pierce County Sewer Improvement Area | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_firestation | The parcel within one mile distance to a fire station | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_libraries | The parcel within one mile distance to a library | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_policestation | The parcel within one mile distance to a police station | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_waterplants | The parcel within one mile distance to a water treatment plant | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_hospital | The parcel within ten mile distance to a hospital | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_private\_school | The parcel within three mile distance to a private school | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_elementary\_school | The parcel within three mile distance to an elementary school | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_high\_school | The parcel within three mile distance to a high school | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| near\_college | The parcel within three mile distance to a college | 1-digit number | varchar | 1 | 1 or 0 | nominal | 0.00% |
| Crime\_Num | Number of crimes in the past 12 months | Numbers between 3 and 6433 | int | NULL | 91 | numerical | 0.00% |

Note: Parcel\_Number is the key value

1. (Pierce County, n.d.) [↑](#footnote-ref-1)
2. (Pierce County, n.d.) [↑](#footnote-ref-2)