**School of Business**

**BIA 652D Multivariate Data Analysis**

**Fall 2016**

Project

Due: Tuesday, Dec 20.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Name \_\_\_Neha Mansinghka\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature \_\_\_Neha Mansinghka\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_12/20/2016\_\_\_\_\_\_\_\_\_\_

Please note that assignments in this class may be submitted to [www.turnitin.com](http://www.turnitin.com/), a web-based anti-plagiarism system, for an evaluation of their originality.

======================================================

Table of Contents

[EXECUTIVE SUMMARY 3](#_Toc470045206)

[INTRODUCTION 4](#_Toc470045207)

[BACKGROUND 4](#_Toc470045208)

[EXAMINING THE DATA 5](#_Toc470045209)

[ANALYSIS AND RESULTS 6](#_Toc470045210)

[FACTOR ANALYSIS 6](#_Toc470045211)

[MULTIPLE REGRESSION ANALYSIS 7](#_Toc470045212)

[CLUSTER ANALYSIS 8](#_Toc470045213)

[CONCLUSION AND RECOMMENDATIONS 9](#_Toc470045214)

[APPENDIX A 11](#_Toc470045215)

[DATASET REFERENCES: 11](#_Toc470045216)

# EXECUTIVE SUMMARY

King County (including Seattle) house sales from May 2014 to May 2015 have been analyzed using factor analysis, multiple regression and cluster analysis. The dataset has around 21000 house sale records and 19 house features. A random sample of 1000 records has been analyzed. Examining the data prior to the analyses showed that assumptions of normality, homoscedasticity and linearity are met except in the case of Price, lot size of the house and average lot size of 15 nearest neighbors. For these variables, there are departures from normality but these are found to be due to outliers at the upper end which are detected and noted for consideration in the analyses. No missing data are found. Factor analysis shows three underlying dimensions or factors that are independent of each other - House Size composed of the size of living area of the house, average living area size of nearest 15 neighbors and bedrooms, Lot Size composed of the size of the lot and average lot size of 15 nearest neighbors and Structural Characteristics composed of floors and age of the house. These summarize the measures represented by these variables and can be used instead of using these 7 variables i.e. for data reduction. Multiple regression shows that the features that best predict and explain the changes in price are bedrooms, bathrooms, living area size of house, floors, view, condition, Grade and number of years since it was built. However, the model shows an inverse relationship between bedrooms and price which is counterintuitive and opposite of the bivariate correlation between bedrooms and price. This reversal of relationship direction in the model is possibly because bedrooms has multicollinearity with one or more features in the model. As a recommendation for further work, a model using factors determined from factor analysis instead of using the variables directly can be tested to see if that better represents the relationship between bedrooms and price. Also, removing influential observations improves the model considerably by reducing the mean squared error however these observations need to be investigated for representativeness to the population to determine whether to include them in this model or analyze as a different population in their own model. Cluster analysis shows 3 clusters of houses. Cluster 3 has higher mean price than Cluster 2 and Cluster 1. Cluster 2 and Cluster 1 mean price are not significantly different. Cluster 3 can be characterized as having most houses of above average grade and condition (compared to Cluster 2 and Cluster 1), most houses with 4 or more bedrooms, over 50% of the houses with 1 floor. It also has the highest mean sqft\_lot15(average lot size of 15 nearest neighbors) and highest mean sqft\_living15 (average interior living area size of 15 nearest neighbors). Cluster 2 can be characterized as having second highest above average Condition houses, second highest number of 4 or more bedroom houses, over half the houses having 1 floor and second highest mean of sqft\_living15 and sqft\_lot15. From Grade perspective, Cluster 2 and Cluster 1 are similar. Cluster 1 can be characterized as having highest number of average or below average condition houses, most houses with 1 or 2 bedrooms, over 50% of the houses with 2 or more floors and having the lowest mean sqft\_living15 and sqft\_lot15. So, for example buyers interested in good condition and good Grade houses which are large (in terms of living area, lot size and number of bedrooms) would be more interested in Cluster 3 houses. Outliers for price and lot size have been excluded from this analysis and like for multiple regression these should be assessed for representativeness to the population and analyzed as a separate population if needed.

# INTRODUCTION

In this project the house sales of King County including Seattle have been analyzed using several multivariate techniques. The aim of the project is to identify any underlying dimensions in the different house features through factor analysis and to predict and explain how the different house features affect the house price through multiple regression. Additionally, cluster analysis is also performed on the dataset to determine if any practically relevant groupings of houses can be identified. The report first provides a background on the dataset itself. Then the results of examining the data are discussed, followed by the analyses and results of the three techniques. It ends with conclusion and management recommendations. Any significant results from these analyses are expected to be useful for both current and potential homeowners and investors interested in King County house market. In addition, commercial entities like real estate agents and house builders interested in this county are also expected to benefit from such analyses.

# BACKGROUND

The dataset for the project has been obtained from Kaggle (<https://www.kaggle.com/harlfoxem/housesalesprediction>). It consists of 19 house features (columns) and approximately 21600 house sale records from May 2014 to May 2015. A random sample of 1000 records has been used for the project. The sample is random ensuring representativeness to the remaining dataset. Working with a large sample size of 21600 records would provide statistically significant results even if they weren’t really significant. The significance tests become quite sensitive as sample size increases. Also, SAS tool has been used for the project. With a large sample size of 21600, SAS would give memory issues as the number of plots in the output increases and environment settings in SAS would need to be modified to overcome such issues. For these reasons, it was decided to use only a random sample of 1000 records for the project.

The following table provides a description and classification of all the house features or variables available in the dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Description | Type | Used in analyses | Comments |
| Id | Unique ID for each home sold | N/A | N |  |
| SellDate | Date of home sale | N/A | N | Used indirectly in Age\_At\_Sale variable. Sale year extracted from this value. |
| Price | Price of each home sold | Metric | Y |  |
| Bedrooms | Number of bedrooms | Metric | Y |  |
| Bathrooms | Number of bathrooms (0.5 means room with toilet but no shower) | Metric | Y |  |
| sqft\_living | Square footage of the apartments interior living space | Metric | Y |  |
| sqft\_lot | Square footage of the land space | Metric | Y |  |
| Floors | Number of floors | Metric | Y |  |
| Waterfront | Overlooking waterfront or not | Non-Metric | N | Binary |
| View | Index of 0 to 4 of the view | Non-Metric | Y | Ordinal |
| Condition | Index of 1 to 5 on condition of house | Non-Metric | Y | Ordinal |
| Grade | An index from 1 to 13 of the construction and design quality | Non-Metric | Y | Ordinal |
| sqft\_above | Square footage of the interior housing space that is above ground level | Metric | N | This is a derived value given in the dataset (sqft\_living – sqft\_basement) |
| sqft\_basement | Square footage of the interior housing space that is below ground level | Metric | Y |  |
| yr\_built | Year house was initially built | Metric | N | Used indirectly in derivation of Age\_At\_Sale variable |
| yr\_renovated | Year of house’s last renovation | Metric | N |  |
| Lat | Latitude | Metric | N |  |
| Long | Longitude | Metric | N |  |
| Zipcode | Zipcode of the area the house is in | Non Metric | Y |  |
| sqft\_living15 | Average square footage of interior housing living space for the nearest 15 neighbors | Metric | Y |  |
| sqft\_lot15 | Average square footage of lot space for the nearest 15 neighbors | Metric | Y |  |
| Age\_At\_Sale | Derived value by subtracting year of sale from yr\_built | Metric | Y |  |
| Multi\_Storey\_House | Binary. Derived value using logic if number of floors greater than 1 then this variable is 1 otherwise it is 0. | Non-Metric | Y |  |
| Above\_Average\_Grade | Binary. Derived value using logic if Grade is greater than 7 then this variable is 1 otherwise it is 0. | Non-Metric | Y |  |

# EXAMINING THE DATA

The dataset was examined before undertaking any analyses for missing data, outliers, normality, homoscedasticity, and linearity. No missing data was found. Empirical test results and visual representations for the metric continuous variables - Price, sqft\_living, sqft\_lot, sqft\_basement, sqft\_living15, sqft\_lot15, Age\_At\_Sale were examined. Although the statistical test for normality (Shapiro Wilk) for all variables has p value less than 0.05 suggesting that the null hypotheses that the data comes from a normally distributed population should be rejected at 0.05 level of significance, the data are not considered non-normal solely on this basis. The sample size of 1000 is still large enough to make these tests sensitive. Therefore, visual representations such as histogram, normal probability plot and boxplot are also examined for each variable along with basic descriptive measures like mean, median, skewness and kurtosis. Data are considered normal or non-normal by using all these in conjunction. Looking at these aspects, Price, sqft\_lot and sqft\_lot15 are the only non-normal variables out of which Price has minor departure from normality due to outliers at the upper end shifting the distribution to the left. The skewness and kurtosis values for Price are 4.58 and 41.50 respectively. sqft\_lot and sqft\_lot15 are most non-normal out of all variables. There are outliers at the upper end for both these variables too making them positively skewed and having a high kurtosis. However, with the removal of outliers both the variables become normally distributed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | With outliers | | Without outliers | |
|  | **Kurtosis** | **Skewness** | **Kurtosis** | **Skewness** |
| sqft\_lot | 80.95 | 8.13 | -0.08 | 0.38 |
| sqft\_lot15 | 61.46 | 7.33 | -0.15 | 0.32 |

As the sample size is large (1000), the non-normality is not considered a major concern. In large sample sizes, the detrimental effects of non-normality on the results are negligible. Furthermore, as departures from normality in this case are due to outliers, these are detected for each of the above variables and noted so results of the various techniques are assessed for how these outliers might impact them.

Outliers have been identified using the interquartile range (IQR). Values outside the 1.5\*IQR range are deemed as outliers. At the upper end, the outliers for Price are values equal to or above $1M of which there are 74. For sqft\_living, there are 51 such values starting at 3800 sq. ft. For sqft\_living15, there are 49 outliers at the upper end starting at 3300 sq. ft. For sqft\_lot, there are 130 outliers starting at 17000 sq. ft. and for sqft\_lot15 there are 114 outliers starting at 16000 sq. ft. The outliers are not removed from the data for further analyses. These values warrant further investigation to determine whether they are genuinely one-offs or representative of the population.

A scatter matrix was created for all the variables to check for linearity. No non-linear relationships were seen in these scatter plots. The two non-metric variables that are of interest in the analyses are whether the house is a multi-storey house and whether the house has an above average grade. So, the main dependent variable, Price, is tested for homoscedasticity or homogeneity of variance across these variables. The assumption of homogeneity of variance of Price across both groups of each non-metric variable, Multi\_Storey\_House and Above\_Average\_Grade, is satisfied. Again, for this check both the results of the statistical tests (Levene’s, Brown and Forsythe’s and Bartlett’s) and visual representations are utilized. Even though statistical tests conclude heteroscedasticity, the graphical plots are examined and the homogeneity of variance is confirmed.

# ANALYSIS AND RESULTS

## FACTOR ANALYSIS

A factor analysis was performed to detect any underlying dimensions in the house features. The metric variables used for factor analysis are Bedrooms, Bathrooms, sqft\_living, sqft\_lot, Floors, sqft\_basement, sqft\_living15, sqft\_lot15, Age\_At\_Sale. The overall Kaiser’s Measure of Sampling Adequacy (MSA) value initially with all these variables is 0.7 and each variable has MSA above 0.5 suggesting there are some underlying factors. The correlation matrix for these variables also shows some high correlations (above 0.7) suggesting potential underlying factors. Looking at the scree plot, percentage of variance explained plot and the eigenvalues, the first three components or factors are extracted which explain in all 76% of variance and have eigenvalues greater than 1. The unrotated factor loadings are examined and there is some pattern visible but not very clear with a few cross loadings. Therefore, a varimax rotation is performed and the pattern becomes clearer. However, both Bathrooms and sqft\_basement show dual loadings. Conceptually, Bathrooms seems like a more relevant variable to keep as it measures similar dimension as Bedrooms and sqft\_living. Also, the MSA value of Bathrooms (0.84) is higher than MSA value of sqft\_basement (0.52). Therefore, sqft\_basement is dropped and factor analysis is run again with 3 factors as they still explain over 70% variance together and have eigenvalues greater than 1. In the second run, with sqft\_basement dropped, the overall MSA increases to 0.72 and each variable MSA is still over 0.5. The unrotated factor loadings have a few dual loadings. The rotated version shows a much clearer pattern with one dual loading by Bathrooms. The number of factors, 3, is still appropriate. Therefore, the factor analysis is run a third time with Bathrooms dropped and extraction of 3 factors. The 3 factors now account for 80% of variance together and all have eigenvalues greater than 1. The overall MSA drops to 0.63 but each variable still has MSA over 0.5. The unrotated factor loadings show multiple dual loadings but the rotated version shows a clean solution. All loadings are over 0.8 and there are no dual loadings. The groupings of variables into the 3 factors are shown below.



sqft\_living, sqft\_living15 and Bedrooms load highly on Factor 1. sqft\_lot and sqft\_lot15 load highly on Factor 2 and Floors and Age\_At\_Sale load highly on Factor 3. Factor scores and summated scales are calculated using these groupings. The labelling is done for the summated scales and factors based on the similar dimension that the variables measure in each grouping. The first factor is labelled House Size, the second is labelled Lot Size and third is labelled Structural Characteristics.

The correlation matrix for the three factors and summated scales is examined. The three factors are orthogonal (i.e. 0 correlation) and the three summated scales also have very low, close to 0, correlations. The correlation between the factor and corresponding summated scale is high (over 0.9) and the correlation between the factor and non-corresponding summated scale is low. This suggests an orthogonal solution i.e. factors are independent of each other.

## MULTIPLE REGRESSION ANALYSIS

Multiple regression analysis is conducted with Price as the dependent variable to determine a model that best predicts the price using the house features (independent variables) and best explains how these independent variables influence Price. First model is created using just one independent variable that has the highest bivariate correlation with Price which in this case is sqft\_living with 0.69 correlation. This model has R2 of 0.48. The square root of mean squared error (Root MSE) is $269062. The residual plots show linearity, independent errors, normality but non-constant variance. Although the overall model and the coefficient for sqft\_living is statistically significant, this is not a good model due to low R2, high Root MSE and the non-constant variance of the errors. A second model is created using stepwise method. 9 independent variables (Bedrooms, Bathrooms, sqft\_living, sqft\_lot, Floors, sqft\_basement, sqft\_living15, sqft\_lot15, Age\_At\_Sale) are provided and the method chose 6 variables (Bedrooms, Bathrooms, sqft\_living, Floors, sqft\_living15, Age\_At\_Sale) for the model.

The R2 and adjusted R2 for this model are both 0.58 i.e. 10% additional variance explained by this model compared to previous model. The Root MSE decreases to $242267. The residual plots, observed versus predicted values plot and the partial regression plots are examined. The non-constant variance of residuals decreases compared to the previous model however there is still some non-constant variance present in the partial regression plots and residual plots suggesting room for improvement. There are still variables influencing the variance of Price that have not been accounted for in the model. It is observed by examining the data again that ordinal non-metric variables View, Condition and Grade also have an impact on Price. They all have 6 or more levels. Therefore, another model is created by including these variables as metric variables and employing the stepwise method. The regression equation from this model is:

**Price = -1035484 – 56457\*Bedrooms + 81692\*Bathrooms + 161.12\*sqft\_living + 38842\*Floors + 81697\*View + 29411\*Condition + 116174\*Grade + 4256.93\*Age\_At\_Sale**

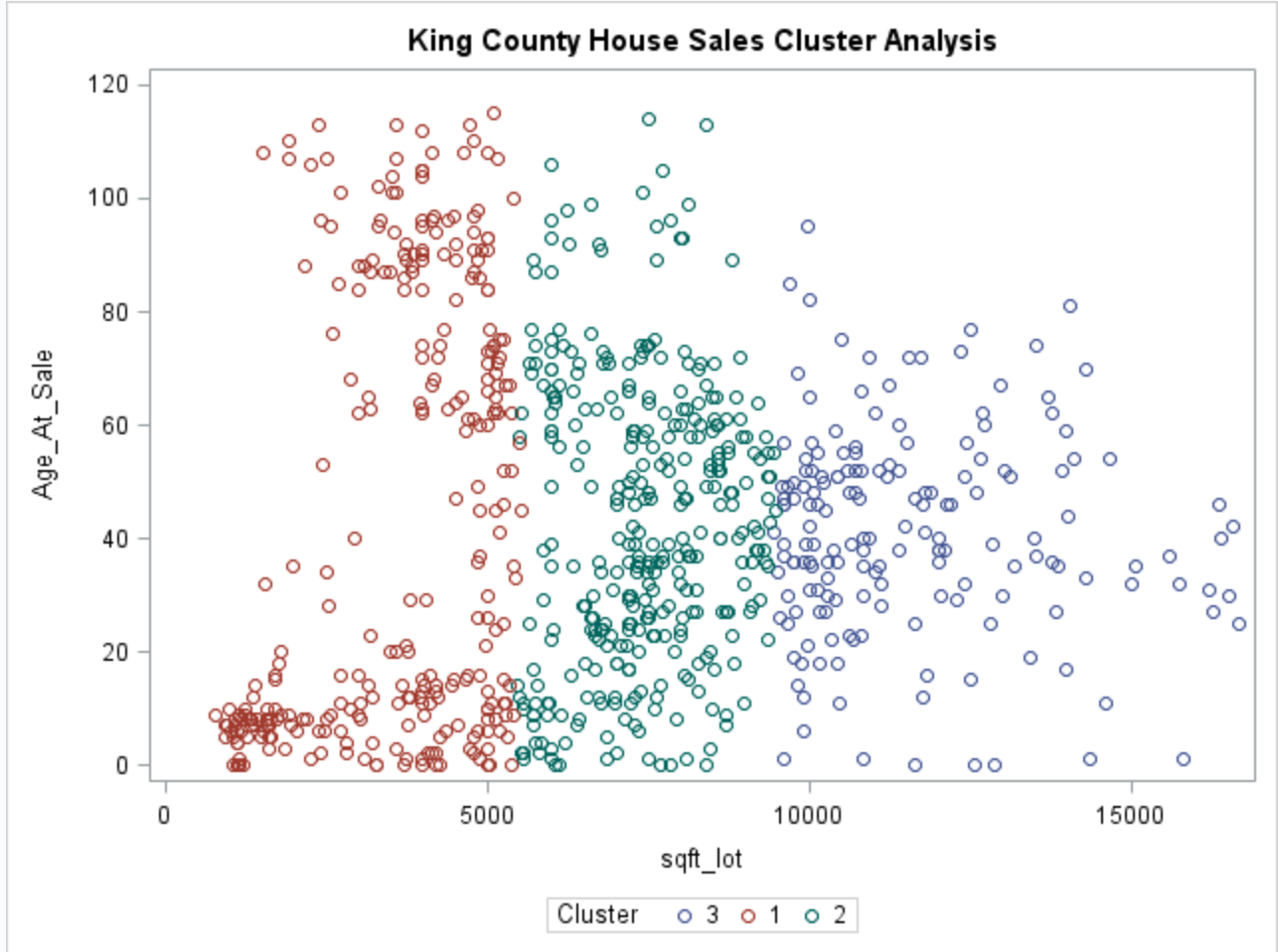
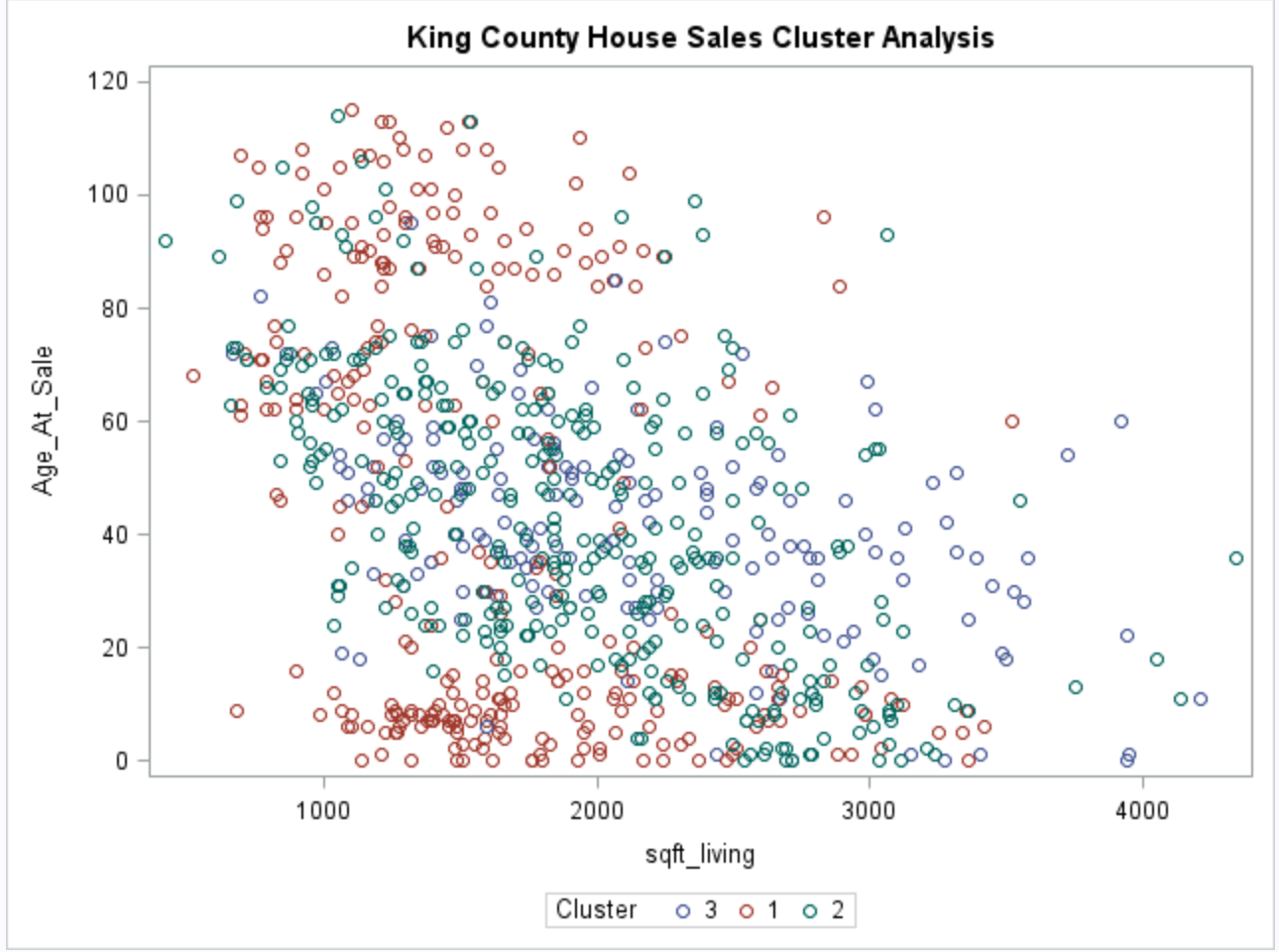
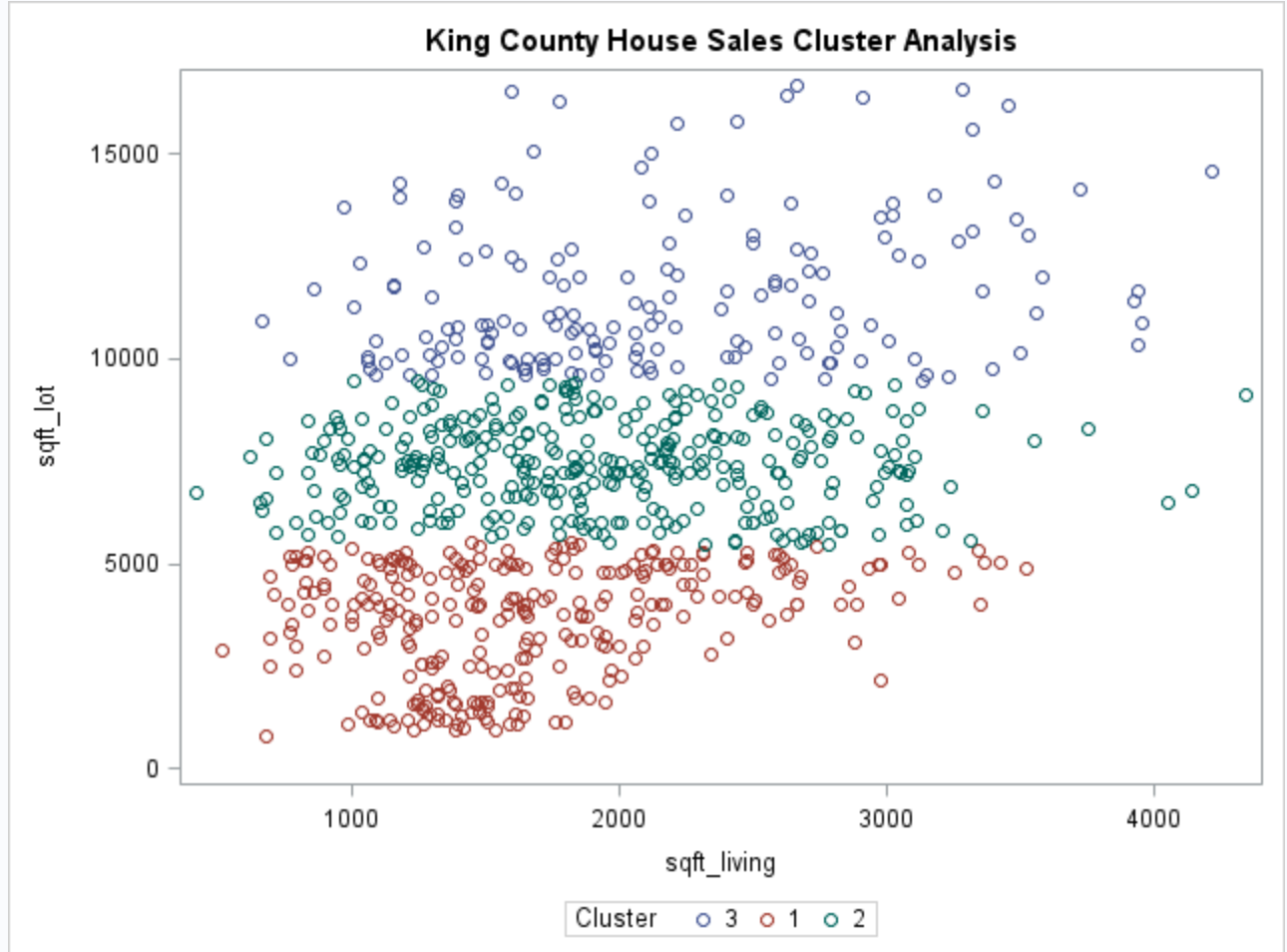
The overall model and each of the coefficients are statistically significant. The R2 and adjusted R2 for this model are both 0.64 i.e. 8% additional variance explained by this model compared to previous model. The Root MSE decreases to $223658. The residual plots, observed versus predicted values plot and the partial regression plots are examined too. All the assumptions (normality, linearity and independence of errors) are met and the non-constant variance is fixed to a large extent too. This model is considered as the best fit. To further improve the model as the Root MSE is considered on the higher side, the influential observations (outliers and high leverage observations) are identified using Cook’s D, Hat Matrix, CovRatio, DFFITS and studentised deleted residuals measures. 17 observations are identified as influential observations across all of these measures. The previous model is re-run without these influential observations. The R2 increases slightly to 0.65 and adjusted R2 stays the same. The Root MSE decreases too $157893. The plots look good too. However these influential observations cannot simply be removed. They need to be investigated further. Therefore, the previous model with the influential observations is considered as the final model.

Although, Grade and Floors are included as metric variables in the final model, their non-metric versions (Above\_Average\_Grade and Multi\_Storey\_House) are tested too and models with their metric versions are marginally better.

## CLUSTER ANALYSIS

Cluster Analysis is performed on the data to identify any natural groupings of houses based on certain features. The features or variables used to perform the cluster analysis are sqft\_living, sqft\_lot and Age\_At\_Sale. To prevent multicollinearity, one variable from each of the factors identified in the factor analysis (with an orthogonal factor solution) is used here. Also, from previous analyses, these variables are identified as significant measures of this dataset. 9 outliers are removed from the data before running cluster analysis. These are records that are outliers across all these 3 variables.

First hierarchical clustering is performed using Ward’s method. The cubic clustering criterion (CCC), Pseudo F and Pseudo T-Squared plots are examined to get the appropriate number of clusters. However, in this instance, this cannot be obtained from these plots. Therefore, non-hierarchical clustering is performed with 3, 4 and 5 clusters. It is observed that the Price and sqft\_lot outliers still in the dataset are distorting the clusters somewhat. There are no significant distinguishing characteristics in the resulting clusters. So, the outliers identified in examining the data step were removed for Price and sqft\_lot and non-hierarchical clustering was performed again with 3, 4 and 5 clusters. The 3-cluster solution is best out of all three. The following plots show the division of three clusters based on sqft\_lot, sqft\_living and Age\_At\_Sale:



The plots show that the algorithm has formed Cluster 3 mostly of houses with sqft\_lot above 10000 sq.

ft., sqft\_living between 1000 and 3500 sq. ft. and Age\_At\_Sale between 15 years and 70 years. Cluster 2 is formed mostly of houses with sqft\_lot between 5000 sq. ft. and 10000 sq. ft., sqft\_living starting below 1000 sq. ft and going up to 3000 sq. ft. and Age\_At\_Sale between 0 and 80 years. Cluster 1 consists mostly of houses with sqft\_lot between 0 and 5000 sq. ft, sqft\_living between 500 and 3000 sq. ft. and Age\_At\_Sale between 0 and 120 years. However, majority of Cluster 1 houses fall between 0 and 30 years old and between 60 and 120 years. There are relatively fewer houses between 30 and 60 years in Cluster 1, Cluster 2 and Cluster 3 have more houses in this age range. These are not the distinguishing characteristics of the clusters though. As these variables were provided in the algorithm, it would divide up the houses in clusters based on these variables. To validate these groupings, a MANOVA analysis is performed for Price, sqft\_living15 and sqft\_lot15 to see whether these features differ significantly across the three clusters. If they do, they would be the distinguishing characteristics of the clusters.

The results of MANOVA show that these differ significantly across the 3 clusters. The means comparison tests show that mean Price of Cluster 3 ($489958) is significantly different from that of Cluster 2 and 1 but mean Price of Cluster 2 ($420796) and Cluster 1 ($452393) is not significantly different. sqft\_living15 mean is significantly different for all 3 clusters. Cluster 1 has mean of 1664 sq. ft., Cluster 2 has mean of 1877 sq. ft. and Cluster 3 has mean of 2129 sq. ft. The mean of sqft\_lot15 is also significantly different for all 3 clusters with Cluster 1 having a mean of 3995 sq. ft., Cluster 2 having a mean of 7530 sq. ft. and Cluster 3 having a mean of 10647 sq. ft.

The below table shows how the three clusters vary on features such as View, Condition, Grade, Bedrooms and Floors:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Composition of each cluster (in % of houses)** | | |
| **House Features** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
| View rating of 0 | 95 | 95 | 94 |
| View rating above 0 | 5 | 5 | 6 |
| **Total** | **100** | **100** | **100** |
| Average and below average Condition | 74 | 65 | 61 |
| Above average Condition | 26 | 35 | 39 |
| **Total** | **100** | **100** | **100** |
| Average or below average Grade | 58 | 61 | 49 |
| Above average Grade | 42 | 39 | 51 |
| **Total** | **100** | **100** | **100** |
| 1 Bedroom | 2 | 2 | 0 |
| 2 Bedrooms | 24 | 10 | 5 |
| 3 Bedrooms | 52 | 48 | 45 |
| 4 Bedrooms | 19 | 31 | 43 |
| More than 4 bedrooms | 3 | 9 | 7 |
| **Total** | **100** | **100** | **100** |
| 1 Floor | 37 | 59 | 61 |
| 2 Floors | 54 | 41 | 38 |
| 3 Floors | 9 | 0 | 1 |
| **Total** | **100** | **100** | **100** |

# CONCLUSION AND RECOMMENDATIONS

Factor analysis shows that the three underlying dimensions or factors are House Size composed of the size of living area of the house, average living area size of nearest 15 neighbors and Bedrooms, Lot Size composed of the size of the lot and average lot size of 15 nearest neighbors and Structural Characteristics composed of Floors and age of the house. Multiple regression analysis provides an equation that can be used for Price prediction as well as knowing how different house features influence price. For example, according to the model, for every unit increase or decrease in Bathrooms the price will increase or decrease by $81692, for every unit increase or decrease in sqft\_living the Price will increase or decrease by $161, for every unit increase or decrease in Floors the Price will increase or decrease by $38842, for every unit increase or decrease in View the Price will increase or decrease by $81697, for every unit increase or decrease in Condition the Price will increase or decrease by $29411, for every unit increase or decrease in Grade the Price will increase or decrease by $116174 and for every unit increase or decrease in Age\_At\_Sale the Price will increase or decrease by $4257. This explains the influence each feature has on the house price. However, the relationship between Bedrooms and Price is not represented correctly by this model. Per the model, these two have an inverse relationship i.e. for every unit increase in Bedrooms the Price will decrease and increase with decrease in number of Bedrooms. This goes against the relationship shown in the bivariate correlations where Bedrooms and Price have a positive correlation meaning as number of Bedrooms increases Price increases. It is also illogical that as the number of Bedrooms increases, the Price decreases. One reason for this reversal of relationship direction in the multiple regression equation is multicollinearity among the variables. It is recommended to create a model with the factors instead of individual variables and see if that better explains the relationship between Bedrooms and Price. Also, the influential observations need to be analyzed for representativeness to the population to determine whether they should be included in the model or perhaps should be analyzed as a separate population with their own model. Cluster Analysis shows houses can be divided in three clusters based on certain features. Cluster 3 has the highest percentage of above average Condition and Grade houses at 39% and 51% respectively. Cluster 2 has the second highest above average Condition houses while Cluster 1 has the lowest. Only a quarter of Cluster 1 houses are above average Condition. Cluster 2 and Cluster 1 have about the same percentage of houses of above average Grade. Cluster 3 has the highest percentage of 4 bedrooms or more houses at 50%, Cluster 2 has the second highest and Cluster 1 has the lowest at 22%. The percentage of 3 bedroom houses is comparable in all three clusters. The percentage of 1 and 2 bedroom houses is highest in Cluster 1 at 26%, about half of this in Cluster 2 and only 5% in Cluster 3. Cluster 1 has majority of multi-storey houses i.e. 2 or more floors (around 63%) and Cluster 2 and 3 have majority of houses with 1 floor (over 50% in each cluster). Cluster 3 mean price is higher than that of Cluster 2 and 1. Cluster 2 and 1 mean price is not significantly different. The average living area size of 15 nearest neighbors is highest for Cluster 3 houses, second highest for Cluster 2 houses and lowest for Cluster 1 houses. The average lot size of 15 nearest neighbors is also highest for Cluster 3, second highest for Cluster 2 houses and lowest for Cluster 1 houses. The three clusters do not differ distinctly on View. The outliers for Price and Lot size have been excluded from this Cluster Analysis but like the recommendation for multiple regression, these need to assessed for representativeness to the population and analyzed as a separate cluster if needed.

# APPENDIX A

## DATASET REFERENCES:

<https://www.kaggle.com/harlfoxem/housesalesprediction>

<https://rstudio-pubs-static.s3.amazonaws.com/155304_cc51f448116744069664b35e7762999f.html>

<http://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r#c>