NANYANG TECHNOLOGICAL UNIVERSITY

# BC2406 Project Report
# AY20 Semester 1

## Analytics Driven Investment Diversification into Residential Real Estate

## Seminar Group 5
## Team 2
## Professor: Prof Hyeokkoo Eric Kwon


| Names: | Matriculation Numbers: |
|---|---|
| Chong Jie Sheng | U1920968D |
| Leonard Lau Dirong | U1910969B |
| Xavier Soh Jun Jie | U1910126F |
| Brandon Yeow Wei Liang | U1920258E |

**Content Page**

# Executive Summary

White Rock, a global investment management corporation company, is keen on exploring new concepts that will show them how to make faster and more informed decisions by incorporating analytics into their different divisions. This report will identify and explore possible opportunities which will be used to produce a proof-of-concept that can be used to improve White Rock's Investment Management division.

One of the most common problems faced by asset managers these days are the lack of investment opportunities that would add value to their investors. Statistics published by the S&P 500 suggests that 99% of traders underperform as compared to their index, which implies that traders who try to outperform the market either by selectively picking stocks or by diversifying to other stocks end up losing money and wasting their efforts. This, coupled with the management fees charged by the companies, means that investors would be better off buying ETFs that track the S&P 500 which gives better returns in the long run. In order to ensure that the White Rock is able to add value to their investors, there is a need to analyse the possible areas to expand into so that we can provide recommendations on which area and how to best expand into the market.

This report focuses on utilising two-step approaches to explore possible areas of diversification. After conducting research on possible areas of diversification, we found out that Residential Real Estate as an investment avenue is highly viable and not commonly explored by our competitors. To conduct Proof-Of-Concept on our opportunity statement, this report uses the publicly available AirBnb Seattle data to compute possible rental prices as well as Seattle House Sales Prices to compute valuations of properties. For each of the datasets, the two-step approach will be used to sieve out the Key Influencing Factors (KIF) that affect rental prices as well as property valuations. Then, using these KIF, we will input the factors into various analytic models such as Classification And Regression Tree (CART) and Linear Regression to predict the possible prices and rentals.

After obtaining the various analytic models to predict the different factors, evaluation of the different models will be done based on their accuracy as well as highlighting some of the limitations that the models have. Based on the evaluation, recommendations will be given for White Rock to implement this into their Investment Management division in order to achieve the desired outcome of making faster and more informed decisions.

# 1. Introduction to the Opportunity for Diversification

This project is commissioned by White Rock to assist them in exploring potential solutions to make faster and more informed decisions across their various business segments. We will be focusing on Investment Management and a Proof-Of-Concept (POC) using analytics techniques will be developed in an attempt to explore further areas of opportunities.

Traditionally, financial markets are the go-to for most investors, both big and small. Financial markets have various features such as its large consumer base and huge variety, which provides unique features in the form of high liquidity and large diversification possibilities, which in turn, makes it extremely attractive for asset managers. According to an institutional ownership research done in 2017 (Charles, 2017) by Bloomberg, the results show that 80.3% of the shares in S&P 500 are owned by institutions, showing the high correlation between the index's performance and the institution's performance. This results in a problem for the asset managers as they are not providing much value to their customers if their performance correlates closely to the index. Additionally, we foresee two major challenges as White Rock tries to diversify into more stocks outside of the largest companies:

(a) According to the Modern Portfolio Theory, it is implied that there is a soft limit to the amount of diversification done in a portfolio. This is because when considering adding stocks to their portfolio, they have to consider risk level and at a certain number of stocks, increasing the number of stocks no longer decreases the risk level and might even increase the risk level. As such, for asset managers to pursue further diversification in the stock market, it would require them to have acute financial sense to pick out the correct stocks to beat the indexes.

(b) Additionally, statistics released in 2016 by index provider S&P Dow Jones shows that 99% of actively managed US equity firms underperform and fail to beat the indexes for 10 consecutive years (Chris Newlands, 2016). This is further supplemented by an interview with Warren Buffett (CNBC, 2019) where he states that even Berkshire Hathaway cannot beat the S&P 500 index and goes on to say that "the index is still the best way to invest in the stock market for most people". However, he does mention that his best investing gurus only underperformed the index by a bit, further substantiating the point that institutions should look for other forms of diversification to supplement the returns from the index to provide true value to their customers.

## 1.1 Exploring Other Possible Areas of Investments

From the above points, we can see that White Rock should look into other areas of investments other than financial markets in order to gain an edge over its competitors. Out of the various possible areas of diversification, one of the markets that stands out is the Physical Assets market, specifically Real Estate. In the past, before the widespread use of the Internet allowed easy access to the stock markets, physical assets were one of the most popular forms of investments and even up to today, many investors still dabble in real estate to diversify their portfolios.

Real Estate investments differ from investments in the financial markets in a few ways and some of its most attractive features include the tax advantages provided by the governments and also, its ability to hedge against inflation. As Real Estate are physical assets, they are significantly less volatile than stocks and during economic downturns, these assets will still retain their value.


Figure 1: Average Annualised Total Returns over 10 years

The figure above shows the comparison between the real estate market and the stock market, Between 1988 and 2015, REITs have lesser deviation as compared to the stock market, which shows their effectiveness in providing stable gains even during periods of recession and as a form of hedging, allowing institutions to lower their risk. In turn, this shows that real estate might be a good way to strengthen one's portfolio due to its ability to generate steady gains.

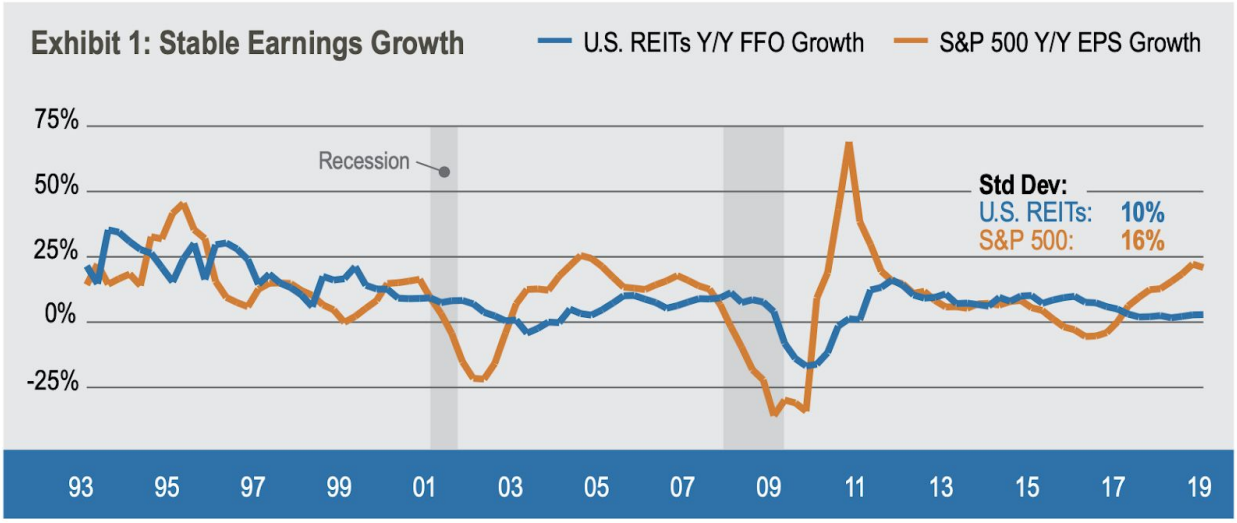
However, some of the reasons why the uptake on Real Estate investing is slow is due to the different laws and regulations in different countries which would require in depth knowledge of property and taxation laws which may require the companies to hire relevant people with the knowledge to proceed with the investments. Additionally, the companies would have to do research over time to find out undervalued areas where they can exploit and make profits out of it.

By looking deeper into the Real Estate industry and analysing major competitors, we can see that many major asset management firms have some stake in the Real Estate market either by their main company or through their subsidiary which further shows the opportunity available to White Rock. Normally, how these firms work is that they purchase commercial properties, construct buildings and sublease them to their other segments or other stakeholders. One popular example of such a firm, in Singapore's context, would be CapitaLand and they have achieved large success in Asia due to their multivariate approach in real estate managements which includes building of commercial properties and renting them out to obtain stable rental income when times are bad or average and flipping them over for a significant margin when property prices are good. This highlights the significant advantage White Rock could potentially gain over its major competitors should it choose to diversify into Real Estate.

## 1.2 Opportunity Statement

Due to the limitations of the traditional investments, we want to identify potential alternative investment avenues. Furthermore, with the portfolio diversification limit as highlighted in the Modern Portfolio Theory, it will be advisable to diversify into not just a wider portfolio but into different markets and products. Our team has identified the Real Estate Market as a potential alternative investment avenue.

In today's world, especially in the events of crises or global pandemics, stocks are becoming less reliable and safe. Looking back at the 2008 financial crisis, most real estate markets fell between 20% - 25%, while stocks fell much more rapidly and recovered much slower. For example, by 2009, the Dow Jones fell 50% which is the largest drop ever in the stock market since the Great Depression (Jan Večerka, 2020). In today's context, Covid-19 has affected many different sectors. In February-March 2020, S&P 500 SPX fell 34% while Case Shiller U.S. National Home Price Index rose 1.0% (Mark Hulbert, 2020). This shows how a global pandemic, such as Covid-19, can have such an adverse impact on the Stock Market whereas the Real Estate Market seems to be relatively more stable in times of crises and global pandemics.




**At December 31, 2018.** Source: UBS, Bloomberg, NBER, Cohen & Steers.

Figure 2: Stable Earnings Growth

The figure above depicts the growth between S&P 500 and U.S. REITs over the years. As seen from the figure, U.S. REITs maintained a relatively constant growth while S&P had much bigger variances. Even during times of recession (marked by the grey column), U.S. REITs maintained relatively constant and growths are constantly higher than S&P 500. This suggests that the Real Estate Market is safer than the typical stock market and will serve as a potential alternative investment avenue.

The Real Estate Market typically consists of both Commercial Property as well as Residential Property. However, with the effects of Covid-19, investment and demand in the Commercial Real Estate Market has dropped rapidly; drops 68% y-o-y in Singapore. Countries such as Hong Kong and mainland China also faced similar declines of at least 60% y-o-y. As a result of the

global pandemic, companies are evolving to maintain operations without the need of a physical office, which leads to a drop in demand for commercial buildings and spaces. This is depicted in the 36% y-o-y fall in investment volume for office assets in Singapore in Q1 of 2020 (Timothy Tay, 2020).

This observation may be due to the trend where more and more companies are enabling employees to work from home for either an extended period of time (until Covid-19 is over), or permanently allowing employees to work from home (Business Insider, 2020). This transition will increase the demand for residential properties and lower the demand for commercial properties further. Furthermore, as the big companies start to introduce these plans, it is likely that the middle sized companies will follow suit, further ingraining the trends and impacting the market further. Specifically, one of the major companies, Recreational Equipment, Inc (REI), have announced they will sell their brand new Seattle headquarters and instead, have multiple headquarters around Seattle, which further substantiates the transition for large scale commercial properties to smaller scale commercial properties or even residential properties.

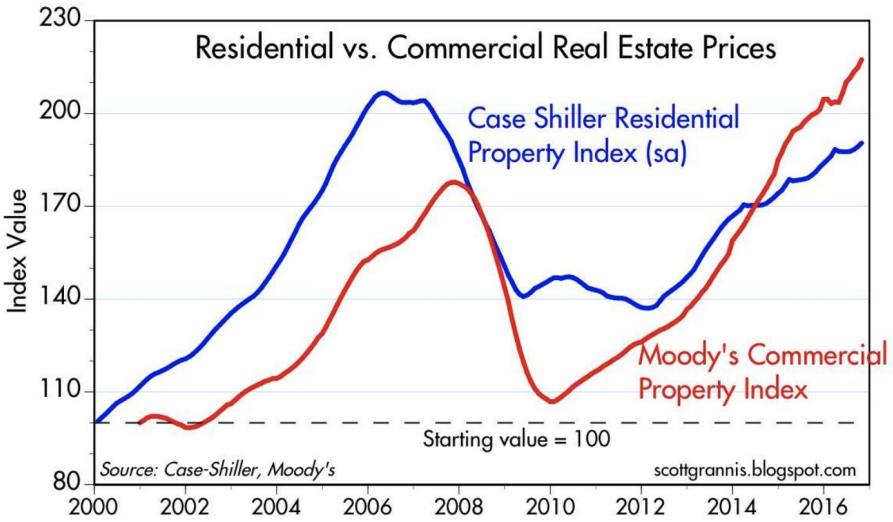



Figure 3: Residential vs Commercial Real Estate Prices

The figure above shows the Index Value for Moody's Commercial Property Index and Case Shiller Residential Property Index. From 2008 onwards, a steep decline was observed for both indexes due to the 2008 Financial Crisis. However, the Residential Index dropped less significantly as compared to the Commercial Index. Although from 2014 onwards, the Commercial Index is seen to be higher than the Residential Index, it is evident that Residential properties are less volatile during crises. This could also be due to the fact that Residential Properties are deemed as a necessity for many and thus may be less susceptible to market conditions.

Thus, in today's globalized world where global pandemics are more prominent, investment in safer avenues are highly sought after which is why our team has identified the Residential Real Estate Market as a Business Opportunity for White Rock.

### 1.3 Feasibility of Business Opportunity

White Rock's entrance into the Residential Real Estate Market is feasible as currently, the Residential Market mostly consists of small firms that only engage in private investment. It is also evident that White Rock's major competitors have yet to enter this market. Together with White Rock's strong reserves, White Rock will be able to have a competitive advantage over its competitors in the Residential Real Estate Market (Dun Bradstreet, 2020). Thus, this market is well-suited for White Rock as the competition level has yet to reach a substantial level.

White Rock being primarily an asset manager, their strong reserves and capital are ideal for investment in the Residential Real Estate Market. As Real Estate is not as liquid and is high-costing, White Rock will have the capability to acquire high-costing housing properties and re-sell at a more desirable time. Furthermore, as Real Estate investments are meant for long-term capital gains, we will be exploring the possibility of using popular online home-stay platforms like AirBnb to obtain rental income which will in turn supplement our overall gain from residential properties. To conclude, the well-established reserves of White Rock will enable them to hold multiple Real Estate portfolios without the need to sacrifice or cut down on their current portfolios.

# 2. Objectives

Since real estate has been identified as a strongly viable segment to diversify into, the objective of this report is to help White Rock develop a predictive model that will improve and speed up decision making in regards to the investment opportunities available in the real estate. Currently, in order to make decisions on real estate purchases, it requires the asset managers to observe the trends and find out undervalued properties to purchase which not only take a lot of technical skills, but also a lot of time.

Our model aims to tackle two problems that asset managers face when making real estate investment decisions.

1) Firstly, one common problem faced by asset managers when making real estate decisions is the amount of research needed to be done on the market before making the investment decision. This involves learning about local law and news to see if there are any external factors which may affect the price of properties. In our model, what we aim to do, is to minimise speculative decisions by producing a model which is grounded on actual statistics, to produce a simple to use model that can be applied to all properties in the area given the right input factors.

2) Secondly, another problem commonly faced by asset managers is that most properties in consideration do not hit the expected returns and as such, significantly lowers the amount of properties shortlisted in the end. Given that home-sharing services like Airbnb are noticing a rapid surge in demands over the past few years (Sherwood, 2019), our model plans to supplement the capital gains of the residential property with stable rental

income which will increase the amount of properties available for consideration as well as accuracy in locating undervalued properties.

To achieve our model, we have devised a 2-step approach - identifying the 'Key Influencing Factors' and using the factors in our 'Prediction Model'.

a) The **Key Influencing Factors** that affect the prices and rentals of housing properties will be determined by analyzing the data sets. These factors will be further analyzed and be used in our modelling to aid in White Rock's decision-making process.

b) The **Prediction Models** will be used to predict the overall profitability of properties through Return On Investments (ROI) of real estate listings. This is achieved by using historical transaction prices as well as factoring in potential revenue from short-term rentals of these housings which will both be determined based on their **Key Influencing Factors.**

By using the **Prediction Models**, asset managers will now be able to better predict ROIs to aid them in their asset acquisition decision making as well as identify more potential investment opportunities.

# 3. Data Preparation

For our data preparation phase we will be cleaning and exploring two sample datasets: Seattle house sales (kc_house_data.csv) and Seattle Airbnb (listings.csv & reviews.csv).

The Seattle Airbnb dataset consists of 3 different parts:

1) 'listings.csv' - This is the primary dataset that we are going to use. It consists of 3818 rows and 92 variables.
2) 'reviews.csv' - This dataset would also be used for us to conduct sentiment analysis on reviews to identify how we could use textual data to supplement our analysis. This data set has 84849 rows and 6 variables
3) 'calendar.csv' - This dataset has been deemed unnecessary and removed from the analysis as we do not need breakdowns of each day and listing.

The Seattle house sales dataset has 21 variables and 21613 rows of data.

The description of the variables from both datasets can be found in the document "Dataset Documentation and Data Dictionary".

Data used are from the same area/region, to reduce unknown factors. Before we proceed with identifying our **Key Influencing Factors** and building our analytics model. We first need to make sure our data set is error-free and cleaned. From the initial data that was recorded. There were multiple variables that had either open-ended textual responses or unrelated content that aren't related to our analysis. These are the variables/factors that we have removed and reasons for removal.

**listings.csv**

| Variables/Factors | Reason for Removal of Variables |
|---|---|
| listing_url, thumbnail_url, medium_url, picture_url, xl_picture_url, host_url, host_thumbnail_url host_picture_url | Links will not provide any insight to the data and is usually for administrative purposes |
| scrape_id, last_scraped | Unnecessary for our analysis as it is information on the dataset instead of the data |
| name, extra_people, minimum_nights, maximum_nights, calendar_updated, calendar_last_scraped, has_availability, first_review, last_review, requires_license, license, jurisdiction_names, cancellation_policy require_guest_profile_picture, require_guest_phone_verification | Unnecessary for our analysis as it is information that is unrelated to the property and will not give us any benefit if we used it to analyse price. Some are also administrative details that Airbnb requires which is also useless in and deemed to have little or no relation to price. |
| host_id, host_name, host_since, host_location, host_about, host_response_time, host_response_rate, host_acceptance_rate, host_neighbourhood, host_listings_count, host_total_listings_count, host_verifications, host_has_profile_pic, host_identity_verified, calculated_host_listings_count | Details of the host are also unnecessary as White Rock will be the host for their investment properties and thus would not affect our analysis. |
| neighbourhood, neighbourhood_cleansed, street, city, state, zipcode, market, smart_location, country_code, country, is_location_exact | Neighbourhood_group_cleansed already contains information on the location of each listing. Therefore the specific information on location (e.g street) gives us little value for analysis. Variables such as city, state can be implied from neighbourhood_group_cleansed. Having these values in would only increase multicollinearity. is_location_exact is unnecessary as exact location is not required |
| weekly_price, monthly_price, availability_90, availability_60, availability_30, reviews_per_month, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value | price, availability_365, review_scores_rating, number_of_reviews covers the variables that we plan to remove. Having these values in would only increase multicollinearity. |

| | |
|---|---|
| summary, space, description, notes, transit, neighbourhood_overview | Textual Data which may be considered for sentiment analysis, but unnecessary for our regression model |

**reviews.csv**

| Variable | Reason for Removal of Variables |
|---|---|
| id, date, reviewer_id, reviewer_name | We are only interested in generating the sentiment score for all listings and map it back to the data from listings.csv |

**kc_house_data.csv**

| Variable | Reason for Removal of Variables |
|---|---|
| date | Unnecessary for our analysis as they provide information about the listings but temporal/forecasting techniques aren't in our toolkit for this analysis. |
| zipcode, lat, long | Provide geographical sensing of the house used for visualization but nothing in our toolkit from his module can effectively use this. |

## 3.1 Data Cleaning

By removing these variables for now, we can see the remaining variables which we will use to proceed with further data exploration and cleaning. We need to correct any data that is to ensure that the data can be more easily used for our analysis later on. Data Cleaning results can be found in Appendix A.
Some possible problems that we identified are :

1) Missing values
2) Check and convert all NA values to a certain value. Drop variable if there are too many NA val
    a) listings.csv
        i) Set median: price, bathroom, bedrooms, beds
        ii) Set to 0: security_deposit, cleaning_fee
        iii) Drop column: square_feet (3721 out of 3818 rows have missing values)
    b) kc_house_data.csv
        i) Drop rows: bedrooms = 0
            (1) These listings have large sqft_lot, low price and no bedrooms. We infer that these are empty plots and will ignore them.
3) Wrong data type
    a) listings.csv
        i) Character to Numeric: price, security_deposit, cleaning_fee
        ii) Character to Factor: neighbourhood_group_cleansed, room_type, property_type, instant_bookable,host_is_superhost

iii) Integer to Numeric: accommodates, bedrooms, beds, guests_included, availability_365, number_of_reviews

b) kc_house_data.csv
   i) Integer to Factor: waterfront, view, condition, grade
   ii) Integer to Numeric: All other variables

4) Insignificant categories
   Group categorical variables into a smaller and manageable number of categories.
   a) listings.csv
      i) property_type - grouping categories into 3 categories: houses, apartments or other.
      ii) Neighbourhood_group_cleansed
5) Wrong values
   a) reviews.csv
      i) Converted encoding to ASCII to remove foreign characters. Stemmed and added german stopwords to reduce occurrences of “die” which is the in German.
         (1) Interpreted as a “die” in english which has negative sentiments.
6) Feature Engineering
   a) kc_house_data.csv
      i) Create a new column, age: 2020 - max(yr_built, yr_renovated)
         (1) We believe that age of the building is a significant factor for the selling price of a house. At the very least, a factor for the category of the house.
   b) K-Means Clustering on kc_house_data.csv and listings.csv (Justification in 3.2)
   c) Sentiment Analysis on reviews.csv to generate sentiment score for listings.csv

## 3.2 Data Exploration

Graphs and summary of variables for both datasets can be found in the Appendix B for reference.

For both datasets, we started off by exploring the relationships between each variable in the dataset. This is done through the use of correlation heatmaps on all numeric variables. Similarly, for both datasets, we seek to predict price given the variables in the dataset. Hence from the heatmap, we can understand the predictive power of each numeric variable.

For variables that are categorical in nature, we would explore the relationship with respect to our output variable, price, by using boxplots.

### 3.2.1 Airbnb Dataset

Leaving out the variables such as id, latitude, longitude which only serves for administrative purposes. We will explore using the rest of the variables together with sentiment seen in Appendix B1.

### Categorical Variables

By plotting boxplots, we can see that room_type does affect the price significantly. Property type does not affect the price as much as room type. However, we have decided to keep this variable as we feel it might still affect listing prices.

As for instant_bookable, host_is_superhost and property_type, there is close to zero difference between prices. Therefore, we have decided to remove these variables from our analysis. Upon further inspection, we also see that there is not much need for us to have instant_bookable and host_is_superhost for short term rental outside of airbnb's context. We realised that room_type is a more important factor to determining price than property_type.

Looking at each neighbourhood group, we can see that the prices do vary from neighbourhoods, hence we will keep this variable for further analysis. But since it has too many categories and that might cause issues with our model, we will cluster it into lesser and more manageable groups of neighbourhoods.

### Sentiment Analysis

Under our feature engineering, we also conducted sentiment analysis on the Airbnb dataset (reviews.csv) to see if sentiments would affect the prices of each listing. Using listing_id from reviews.csv and id from listings.csv as primary keys to join both datasets together, we will use sentiment as a continuous variable for exploration. Visualisation of the sentiment analysis can be seen in Appendix B8 - B11. We explored sentiment score through the Clustering model described later but the distribution of sentiment score across all clusters were identical.

### Continuous Variables


Continuous variable heatmap

To explore the continuous variables, we created a correlation matrix and visualized it using a correlogram.

We will also plot price against each categorical value using bar plots for variables with more distinct values and scatterplots for the rest. We will also add a regression line so we can better see the correlation. Refer to Appendix B7 for the plots. The plots make it easier to visualise the distribution and see each variable's correlation to price (shown on the heatmap)

| price | sentiment | accommodates | bathrooms | bedrooms | beds |
|---|---|---|---|---|---|
| 1.0000 | 0.0050 | 0.6580 | 0.5176 | 0.6370 | 0.5894 |
| **security_deposit** | **cleaning_fee** | **guests_included** | **availability_365** | **number_of_reviews** | **review_scores_rating** |
| 0.2495 | 0.6044 | 0.4257 | -0.0231 | -0.1350 | 0.0543 |

Variables correlation with price

We extracted the correlation of variables with price. From here we can see that accommodates, bathrooms, bedrooms, beds, security_deposit, cleaning_fee and guests_included all have a correlation more than +/-0.2 with respect to price. These will be used for further exploration in our model. sentiments, availability_365, number_of_reviews, review_scores_rating will be dropped.

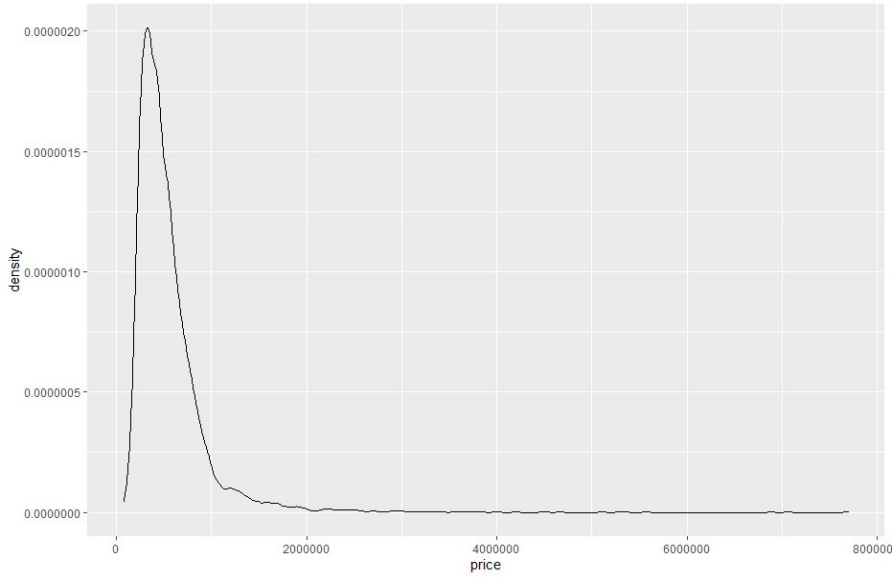
**Clustering**
We wanted to identify if there are any hidden variables/distributions affecting price through clustering our data (using numerical variables indicated above).

We generated clusters with K-Means Clustering. Using the main variable price for visualisation, we can see that clusters identified did not have a distinct price range. Based on further visualisation on the clusters in Appendix B12 we have decided that there are no distinct clusters hence we will be dropping clusters in our model.

## 3.2.2 Seattle Housing Dataset

**Sqft_lot and sqft_living Exploration**
There are two variables, sqft_lot and sqft_living that have similarly named variables, sqft_lot15 and sqft_living15. There was no reference of what each variable meant anywhere on the Internet. We plotted each variable to their respective “15” variation. The plots can be found in the Appendix B16. We observed that the relationship has a very high correlation and suspect that the “15” variation is simply an adjusted version of the variation without the “15”. Our approach to handling this will be to run all 4 variables in the model and use backward elimination algorithm and multi-collinearity scores to decide what to do with them.

**Price Exploration**
As the main variable we seek to predict is the price of a house, we plotted a correlation heatmap to quickly observe the relationship between different variables. This is found in Appendix B14.



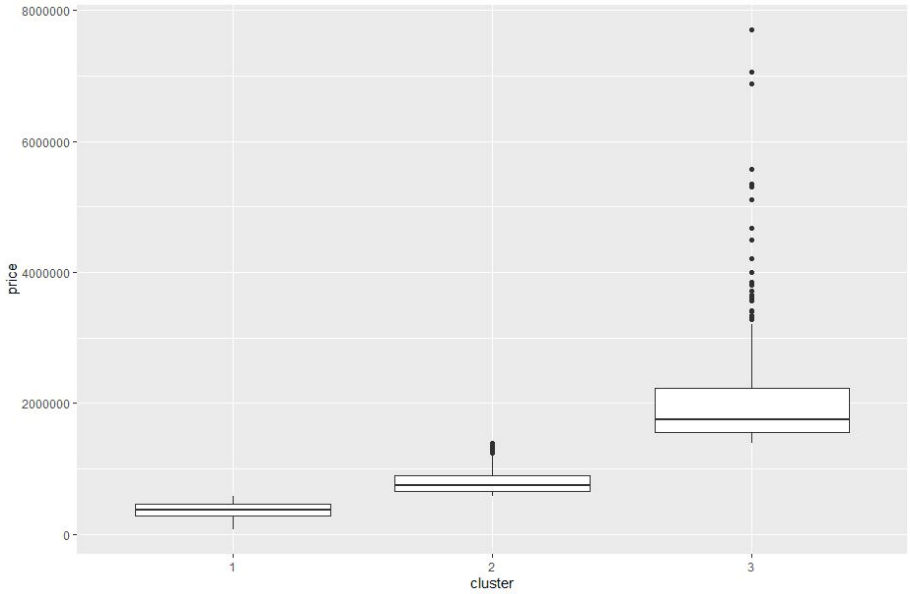

Price Distribution Plot

By observing the price distribution plot and given the size of the dataset, we should observe a normal distribution due to the Central Limit Theorem. However, we can observe minute bumps in the curve which indicates that this is instead a multivariate distribution, and hence there are smaller distribution curves within the dataset. To explore these smaller structures, we utilised clustering to extract the implicit patterns in our data. We have identified 3 distinct clusters and their key characteristics. This is the composition of the clusters.

| 1 | 2 | 3 |
|---|---|---|
| 14957 | 5998 | 642 |

Composition of Cluster by Count

The clusters are generated by K-Means Clustering with an initial cluster count of 5 through the use of the Within Cluster Sum of Squares Plot in the Appendix B15. While the characteristics for upper and middle class housing were distinct, the lower-class housing was very blurred between the three clusters. As such we slowly reduced the cluster count until we could observe a clear distinction between the three clusters as observed in the price-cluster bar plot above. The primary characteristic of each cluster is their price distribution.




Price-Cluster Boxplot

As we can observe, cluster 1 houses are the cheapest, followed by cluster 2 and cluster 3. When plotting the composition of each cluster with every variable, we eventually concluded that cluster 1 describes lower-class housing, cluster 2 describes middle-class housing while cluster 3 describes upper-class housing. The plots can be found in Appendix B17 .

# 4. Forecasting Model

## 4.1 Overall Approach

Given the complicated circumstances and variations that may possibly affect the accuracy of the model, there will be a few assumptions made when developing the model.

a) We will be holding each real estate for at least a year.
b) Median price of real estate does not fluctuate significantly over a year.
    i) This assumption is reflected in our projection of median transaction prices across all 3 clusters. This is done using Python and can be found in Appendix E.
c) There are no significant defects causing the price of the real estate to be low, hence no significant renovation costs.
d) The real estate can be sold at its estimated price with minimal renovation works.
e) The distribution of Airbnb listing price is roughly similar across counties.
f) There are no fees involved with buying and selling real estate across one year.
g) The costs of setting up and maintaining short-term rental in these real estate is accounted for through the cleaning and security deposit fees on Airbnb.

As stated in our objectives, our model comprises two sub-models. In order to gauge the profitability of a house, our first sub-model uses the House Prices dataset to predict the price of a house. By assumption b), this would be the selling price in the future. The difference between the predicted price and the actual price is our profit. Our second sub-model utilises the Airbnb dataset to predict the potential price we would rent out the real estate at. We utilize a fixed occupancy rate to forecast our earnings from short-term rental. Both sub-models are used concurrently to evaluate the potential of a specific listing. In our evaluation, we compared linear regression and Classification and Regression Tree (CART) for both sub-models.

## 4.2 Linear Regression Models

For both models, a 70% train 30% test split will be used to gauge the performance of the model.

### 4.2.1 Variable Selection

The goal of variable selection is to achieve the highest accuracy, while reducing model complexity and thus degree of overfitting. To select the variables used for the model, we will first rely on our knowledge and research on the property market and short term rentals. Following that we have decided to use backward elimination as an automatic selection algorithm and together with analysing the adjusted Variance Inflation Factor (VIF) score to reduce the impact of multicollinearity.

Backward elimination allows us to select variables based on their explainability. VIF on the other hand allows us to manage multicollinearity amongst input variables and thus the correctness of the coefficient of each input variable. Together, they form a robust variable selection methodology for variables that we do not have sufficient knowledge to form an opinion on.

We will also look at the statistical significance of variables for selection.

a. Short-term Rental

i. In addition to variables dropped during the data exploration phase (3.2.1). We used backward elimination to see if we should drop any variables. In this case, security_deposit was dropped

| | Df | Sum of Sq | RSS | AIC |
|---|---|---|---|---|
| - security_deposit | 1 | 5116 | 7131648 | 18068 |
| <none> | | | 7126532 | 18068 |
| - beds | 1 | 9152 | 7135684 | 18069 |
| - guests_included | 1 | 41895 | 7168427 | 18080 |
| - accommodates | 1 | 101754 | 7228287 | 18098 |
| - cleaning_fee | 1 | 169000 | 7295532 | 18119 |
| - bedrooms | 1 | 357195 | 7483727 | 18176 |
| - room_type | 2 | 366160 | 7492692 | 18176 |
| - bathrooms | 1 | 409997 | 7536530 | 18191 |
| - neighbourhood_group_cleansed | 16 | 673240 | 7799772 | 18238 |

ii. We also dropped neighbourhood_group_cleansed as a large proportion of categories within that variable is statistically insignificant. We will not be collapsing the levels to prevent any statistical invalidation.

iii. We then proceeded to analyse adjusted VIF to drop those more than 2. We also expect that the number of beds can be a function of bedrooms and accommodates to be a function of beds, hence we removed both variables to eliminate multicollinearity.

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| accommodates | 5.617869 | 1 | 2.370204 |
| bathrooms | 1.744540 | 1 | 1.320810 |
| bedrooms | 3.149372 | 1 | 1.774647 |
| beds | 4.075514 | 1 | 2.018790 |
| cleaning_fee | 1.900472 | 1 | 1.378576 |
| guests_included | 1.438964 | 1 | 1.199568 |
| room_type | 1.504827 | 2 | 1.107571 |

iv. The summary of the final linear regression model is in Appendix C1.

b. House Prices

i. With our domain knowledge we expect all variables to have an impact on the price of a listing.

ii. We applied backward elimination and removed all variables above of <none>

| | Df | Sum of Sq | RSS | AIC |
|---|---|---|---|---|
| - condition | 4 | 345022524885 | 390070010821891 | 376206 |
| - sqft_lot | 1 | 6388218425 | 389731376515432 | 376221 |
| - sqft_living15 | 1 | 188723974092 | 389913712271099 | 376229 |
| <none> | | | 389724988297007 | 376231 |
| - sqft_lot15 | 1 | 452031301396 | 390177019598403 | 376239 |
| - sqft_above | 1 | 943206002571 | 390668194299578 | 376259 |
| - floors | 1 | 1229646556114 | 390954634853120 | 376270 |
| - bathrooms | 1 | 1541731607881 | 391266719904888 | 376283 |
| - view | 4 | 3063384185450 | 392788372482457 | 376315 |
| - bedrooms | 1 | 2391260612948 | 392116248909955 | 376317 |
| - waterfront | 1 | 7359618772463 | 397084607069470 | 376515 |
| - age | 1 | 11588070564148 | 401313058861155 | 376681 |
| - sqft_living | 1 | 14409704925050 | 404134693222057 | 376791 |
| - grade | 9 | 45562276514806 | 435287264811813 | 377880 |
| - cluster | 2 | 357060812972312 | 746785801269319 | 386426 |

iii. VIF was then applied and sqft_living and sqft_above are analyzed for their impact on the coefficient of other variables.

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| bedrooms | 1.701929 | 1 | 1.304580 |
| bathrooms | 3.446410 | 1 | 1.856451 |
| sqft_living | 8.632815 | 1 | 2.938165 |
| floors | 1.925313 | 1 | 1.387557 |
| waterfront | 1.535802 | 1 | 1.239275 |
| view | 1.812137 | 4 | 1.077144 |
| grade | 5.260956 | 9 | 1.096628 |
| sqft_above | 6.538123 | 1 | 2.556975 |
| sqft_lot15 | 1.078479 | 1 | 1.038498 |
| age | 1.907965 | 1 | 1.381291 |
| cluster | 2.690418 | 2 | 1.280722 |

iv. Sqft_living and sqft_above were removed as they impacted the significance of the coefficient of many other variables and have the highest VIF score.
v. Run backward elimination again to reduce model complexity again.
vi. Summary can be found on Appendix C6

### 4.2.2 Performance

We compute the accuracy, $R^2$ of the model on the train and test sets and compare them. For both models, the deviation between accuracy of train and test set is small and thus the models are not overfitted.

a. Short-term Rental

```
> c(rsql.train.set, rsql.test.set)
[1] 0.5575357 0.5761394
```

b. House Prices

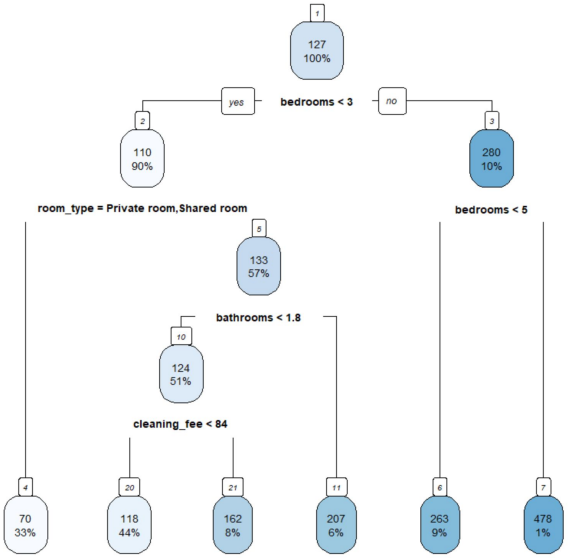
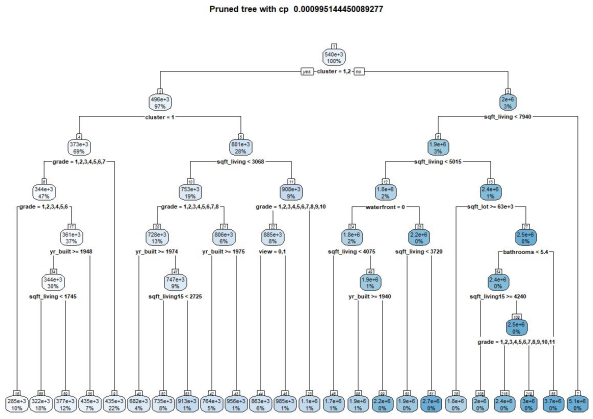
```
> c(rsql.train.set, rsql.test.set)
[1] 0.819876 0.828442
```

## 4.3 CART Models

For CART models, we will be growing each decision tree to its maximum (we will not be including a stopping criteria so that we can get the maximum tree). We will then proceed to prune the tree to try and remove effects of overfitting to finally obtain our model. Hence final model will only have variables from

### 4.3.1 Pruning CART

We will use 10 fold cross validation and 1 SE rule for the pruning of our CART. All trees whose CV error is below the CV error cap are statistically equivalent in terms of error, hence we need to find the simplest tree whose CV error is within the cap. To calculate the CV error cap, we first grow our trees to the max and find the minimum CV error tree and from there add 1 standard error (1SE) to get the error cap. We then proceeded to prune the tree at the error cap.

| a. Short-term Rental | b. House Prices |
|---|---|
|  |  (Refer to Appendix D for a larger version) |





## 4.3.2 Performance

The accuracy of our model R^2 is calculated by taking away the relative error of the last node in the decision tree from 1.

a) Short-term Rental

```
> rsql.val[nrow(rsql.val)]
[1] 0.5364788
```

b) House Prices

```
> rsql.val[nrow(rsql.val)]
[1] 0.858036
```

## 4.4 Overall Evaluation

| Dataset | Linear Regression (Train/Test) | CART |
|---|---|---|
| Airbnb | 0.56/0.58 | 0.54 |
| Seattle Housing Price | 0.82/0.83 | 0.86 |

For the Airbnb dataset, we will be using the Linear Regression model as it performs similarly to CART but is a less complex model, thus it is easier to run on most hardware. For the Seattle Housing Price, we will be using the CART model as it consistently outperforms Linear Regression for random seed values.

## 4.5 Proof of Concept

1. Using the 30% test set from the kc_house_data dataset, we run it through our first model to obtain the predicted profit and profit margin for each house.
2. Additional data is appended to our results to fulfil our assumption g).
    a. Bed is calculated by multiplying the number of bedrooms by 1.5, where some rooms have more than 1 bed. Accommodates is calculated by multiplying the number of beds by 1.5 where some beds are double beds. Guests_included is calculated by multiplying the number of accommodates by 0.5, where 2 occupants can only bring 1 guest.
    b. Security Fee and Cleaning Fee are set at the median of the dataset.
    c. Room_type is set to “Entire home/apt”
3. The results from the previous model are fed into the second model to generate the real estate’s estimated price.
4. We assume that occupancy rate is 60% for any given month to obtain our estimated revenue for the year.
5. Append the result to our dataset for evaluation by the Fund Manager.

Results

```
> head(display.data[which(display.data$price > 1000000)][order(-adjusted_profit_margin)], 20)
      price   profit profit.margin rental_revenue rental_yield adjusted_profit adjusted_profit_margin
 1: 1550000 848810.7     0.5476198       61578.10   0.03972781        910388.8              0.5873476
 2: 1680000 718810.7     0.4278635       69053.69   0.04110339        787864.4              0.4689669
 3: 1488000 510476.9     0.3430624       72700.42   0.04885781        583177.3              0.3919203
 4: 1400000 456550.4     0.3261074       82726.91   0.05909065        539277.3              0.3851981
 5: 1400000 456550.4     0.3261074       80903.55   0.05778825        537453.9              0.3838957
 6: 1500000 498476.9     0.3323179       72700.42   0.04846695        571177.3              0.3807849
 7: 1795000 603810.7     0.3363848       75433.45   0.04202421        679244.2              0.3784090
 8: 1400000 456550.4     0.3261074       65406.96   0.04671926        521957.3              0.3728267
 9: 2230000 718992.9     0.3224183       57931.37   0.02597819        776924.2              0.3483965
10: 1398000 414739.7     0.2966665       68139.99   0.04874105        482879.7              0.3454075
11: 1400000 412739.7     0.2948141       59936.86   0.04281205        472676.6              0.3376261
12: 1450000 406550.4     0.2803796       67230.33   0.04636574        473780.7              0.3267453
13: 1450000 406550.4     0.2803796       57931.37   0.03995267        464481.7              0.3203322
14: 1465000 391550.4     0.2672699       70877.06   0.04838024        462427.4              0.3156501
15: 1430000 382739.7     0.2676501       59936.86   0.04191389        442676.6              0.3095640
16: 1475000 337739.7     0.2289761       52461.27   0.03556696        390201.0              0.2645430
17: 1480000 332739.7     0.2248241       58113.50   0.03926588        390853.2              0.2640900
18: 1490000 322739.7     0.2166038       59936.86   0.04022608        382676.6              0.2568299
19: 1680000 318476.9     0.1895696       94696.77   0.05636713        413173.7              0.2459367
20: 1500000 312739.7     0.2084931       56108.00   0.03740534        368847.7              0.2458985
```

Top 20 Most Profitable Real Estate Listings (>$1,000,000) After Adjustments

To interpret this data, we observe the column of profit.margin and look for rows that are out of sort. When a profit.margin value is smaller than the row below it, then we can observe the effects short-term rental had on its profit margin.

At houses with value > $1,000,000, we observe that rental yields only affect final profit margins at rows 4, 5, 14 and 19. However, for houses valued > $500,000, the impact of short-term rental revenue is visible in rows 5, 8, 9, 12 - 17.

```
> head(display.data[which(display.data$price > 500000)][order(-adjusted_profit_margin)], 20)
      price   profit profit.margin rental_revenue rental_yield adjusted_profit adjusted_profit_margin
 1:  635000 444974.4     0.7007471       77256.82   0.12166428        522231.3              0.8224114
 2:  657500 422474.4     0.6425467       61760.23   0.09393191        484234.7              0.7364786
 3:  609000 376306.5     0.6179089       67230.33   0.11039463        443536.9              0.7283035
 4:  600000 356339.3     0.5938988       68139.99   0.11356665        424479.3              0.7074654
 5:  625000 360306.5     0.5764904       65406.96   0.10465114        425713.5              0.6811416
 6:  685000 401459.8     0.5860728       58113.50   0.08483722        459573.3              0.6709100
 7:  630000 355306.5     0.5639786       56108.00   0.08906032        411414.5              0.6530389
 8:  600000 312555.3     0.5209254       59936.86   0.09989477        372492.1              0.6208202
 9:  649000 336306.5     0.5181919       50637.91   0.07802451        386944.4              0.5962164
10: 1550000 848810.7     0.5476198       61578.10   0.03972781        910388.8              0.5873476
11:  670000 315306.5     0.4706068       77256.82   0.11530868        392563.3              0.5859154
12:  594950 274377.2     0.4611769       69963.35   0.11759535        344340.6              0.5787723
13:  740000 339974.4     0.4594249       88197.01   0.11918515        428171.4              0.5786101
14:  600000 269327.2     0.4488787       77256.82   0.12876136        346584.0              0.5776401
15:  595000 274327.2     0.4610542       56108.00   0.09429916        330435.2              0.5553533
16:  611000 258327.2     0.4227941       78286.48   0.12812845        336613.7              0.5509226
17:  600000 269327.2     0.4488787       59936.86   0.09989477        329264.1              0.5487735
18:  670000 315306.5     0.4706068       50637.91   0.07557896        365944.4              0.5461857
19:  623000 289555.3     0.4647757       45167.81   0.07250050        334723.1              0.5372762
20:  750000 329974.4     0.4399659       69963.35   0.09328447        399937.8              0.5332504
```

Top 20 Most Profitable Real Estate Listings (>$500,000) After Adjustment

# 5. Insights & Analysis

Based on our data exploration, visualization and prediction analysis done above, we have devised models to predict the potential selling price of houses as well as the rental income it can generate by inputting minimal variables. Comparing this to traditional forms of market analysis, managers will now have to spend far less time analysing the past trends of sales and instead, can search for these publicly available data and allow our model to generate the returns. With a prediction accuracy of 86%, we believe that our model can reliably model the actual market value of each real estate. The difference in predicted and actual price would then be our potential profits, allowing for easy identification of potential investments for asset managers to consider. Our short term rental model's prediction accuracy is relatively high as well at 0.58. From our proof of concept, we learnt that for housing valued more than $1,000,000, the impact on rental yield is present, but minimal. On the other hand, for housing valued more than $500,000, the impact of rental yield is quite significant and can reach rental yields of nearly 10%. Reliance on the short-term rental can then be adjusted depending on the characteristics of housings that we acquire in the region.

## 5.1 Limitation of model

Despite our model being largely accurate, we acknowledge that there are certain factors that we were unable to include which may affect the accuracy of the model if included.

Firstly, one of our model's largest limitations is the non-inclusion of external factors such as location and utilities into the model. Our model was only based on datasets from Seattle, whereas the model may be subjective depending on the geographical location. Furthermore, the variables included were only the physical aspects of the houses such as the size and amenities/facilities available. However, valuation of houses also takes into account external factors, such as the surroundings of the property in determining the price of a property. Our model is unable to take into account such factors as it is difficult to quantify such variables due to the subjective nature of these variables and the impact of these external factors may vary greatly depending on the user. In order to circumvent this limitation, what this means for White Rock is that manual intervention will still be needed after sieving out the profitable properties, and experts should be brought in to evaluate the external factors before recomputing the ROIs.

Secondly, the dates recorded for the Seattle Airbnb dataset (2016 - 2017) and the Seattle house sale price dataset (2014 - 2015) are different and not up-to-date. This may result in a timing difference as the value of houses may differ between the years. This may suggest that the values in the 2 datasets are not comparable due to the timing difference. However, as there were no global major events happening over the past few years, there is no reason to believe that there would be a large deviation in price between 2014 to 2018. Additionally, we chose to use data that were slightly backdated as up-to-date data (during 2019 - 2020) might not be a good representative due to COVID which had a negative impact on housing prices and rentals. From our data exploration of the datasets (Appendix E), we found out that the dataset itself further validated our assumptions of a constant selling price where clusters of similar amenities are grouped together and analysed. However, we do acknowledge that using more data that are of the same year range would increase the accuracy of the model and may highlight specific trends which are specific to the area of exploration and that may cause slight deviations to accuracy.

# 6. Overall Recommendations to White Rock

After our analysis, our overall recommendation for White Rock is to enter the residential real estate market. As evident from our model, we have found properties that are below transaction prices and White Rock should seize this opportunity to enter the market before its other major competitors do. In order to improve and smoothen White Rock's entry, we have a few sub recommendations that White Rock could pursue in the midst of preparing.

## 6.1 Automation of processes in the asset acquisition process

By implementing our model into the decision making process, there can be an increase in efficiency due to the reduced research time on property trends in the area. This is achieved through the use of predictive data analytics techniques to automate the identification of profitable investments.

Typically, asset managers will identify undervalued assets from the pool of available investment options and forecast the predicted price of the assets identified. They will research and estimate a ROI to determine if the asset is profitable to invest in. However, with our model, it can be implemented into White Rock's system to automatically test all assets in the pool of available investment options and generate the ROI of each option. Upon testing, each asset's ROI can be tabulated and the system will then filter out the options with a positive ROI for further decision making and evaluation. This will help to automate many of the manual processes involved in the asset acquisition process and reduce the research time of identifying undervalued assets. Furthermore, as asset managers usually have discretionary trading powers, which enable them to acquire assets on their own authority (James Chen, 2019), this automated process will reduce the occurrences of errors in human judgements as well as serve as a viable internal control regarding asset acquisition. Using the model, the groundwork for viable investments will be largely reduced and White Rock can instead, direct their efforts to focus on external factors such as market trends etc.

## 6.2 Partnering with a well-established company to smoothen entry into new markets

Our second recommendation to White Rock is to partner up with an well-established company such as Airbnb to smoothen entry into the residential real estate market. As mentioned earlier in

our introduction, online homestay platforms are becoming more and more popular amongst travellers due to its low cost and more homely feel. Airbnb is one of the largest vacation rental online marketplace companies and is known globally for its services which will improve our ability to obtain consistent rental. Given White Rock's background as a global asset manager, the existing connections that Airbnb have in their various markets will help White Rock in terms of data collection as well as the ability to connect to the rental customer base. Furthermore, as stated in our assumptions, we assume that there are no property taxes and stamp duties which may not hold true in real life. Thus, collaborating with big companies who are already well-established in the residential market like Airbnb allows us to find out the countries with lower taxes and stamp duties for White Rock to shortlist specific countries and conduct pilot studies. This further helps to improve predictive accuracy by lowering the difference between predicted and actual values as well as stabilising rental income by tapping on their large consumer base and relevant data mined.

### 6.3 Tapping on White Rock resources to better improve model accuracy

Our third recommendation is to integrate White Rock's resources and data collection ability with the model to improve the model accuracy. One of the limitations of our model mentioned above is that our model does not factor in external factors such as the location and the surroundings of the housing properties. White Rock collects a lot of data such as activities monitoring and intends to further mine the value of such data to organisations in Singapore or any market. For countries like Singapore where land is limited, value of houses do not only depend on the physical aspects of the house provided, but also depends on the utilities available in the surrounding environment such as availability of shopping malls, distance from MRT stations etc (Si Jie, 2019). As White Rock By tapping on White Rock's activities monitoring database, our model will be able to better determine the value of housing by factoring in the utilities in the surrounding area. As such, by integrating both White Rock's internal data and our model, we can achieve better accuracy in our price predictions and better estimate the ROI of the assets in Singapore.

Additionally, with the vast historical data in White Rock's database, forecasting can also be achieved by comparing trends over time and factoring in data such as Government Development Plans and Population Plans. With forecasting, ROI can be better determined and may reduce the need for human judgement to evaluate the ROI based on external factors.

# 7. Conclusion

This report highlights the use of analytic models in order to explore residential real estate which is a possible area of diversification. As mentioned earlier, White Rock is looking for the usage of analytics in their various segments to make faster and more accurate decisions. By looking at sample datas of property and rental prices in an area, we are able to determine the key influencing factors which allows White Rock to narrow down the data they need to feed into their predictive models to obtain an accurate output. This in turn helps to speed up and allow them to make faster and more accurate decisions.

# References


1) Aaron Treloar. (2019, February 11). *The top 5 factors affecting rental rates for rental properties*. Access Property Management Group - West Michigan Professional Rental Property Management Company. https://www.accesspmgroup.com/the-top-5-factors-affecting-rental-rates-for-rental-properties/

2) Bembridge, R. (2020, May 7). *Why property investments are a safe choice during the current pandemic*. PropertyWire. https://www.propertywire.com/blog/why-property-investments-are-a-safe-choice-during-the-current-pandemic/

3) Business Insider. (2020, September 2). *18 major companies that have announced employees can work remotely long-term*. Business Insider Nederland. https://www.businessinsider.nl/companies-asking-employees-to-work-from-home-due-to-coronavirus-2020?international=true&r=US#earlier-this-summer-outdoor-retailer-rei-announced-that-it-is-selling-its-brand-new-unused-8-acre-corporate-campus-in-bellevue-washington-in-an-august-12-statement-ceo-eric-artz-said-the-company-will-lean-into-remote-working-as-an-engrained-supported-and-normalized-model-for-employees-6

4) Charles Mcgrath. (2017, April 25). *80% of equity market cap held by institutions*. Pensions & Investments. https://www.pionline.com/article/20170425/INTERACTIVE/170429926/80-of-equity-market-cap-held-by-institutions

5) Chris Newlands. (2016, October 24). *99% of actively managed US equity funds underperform*. Financial Times. https://www.ft.com/content/e139d940-977d-11e6-a1dc-bdf38d484582

6) Dun & Bradstreet. (2020). *Residential Property Investment Industry Insights From D&B Hoovers*. https://www.dnb.com/business-directory/industry-analysis.residential-property-investment.html

7) Hulbert, M. (2020, July 25). *Look what happened to home prices when the coronavirus sent stocks into a bear market*. MarketWatch.

https://www.marketwatch.com/story/look-what-happened-to-home-prices-when-the-coronavirus-sent-stocks-into-a-bear-market-2020-07-21

8) James Chen. (2019, June 5). *Asset management company (AMC)*. Investopedia. https://www.investopedia.com/terms/a/asset_management_company.asp

9) KZB Real Estate. (2020, April 23). *Commercial vs residential properties in a recession*. https://www.kzbrealestate.com/commercial-vs-residential-properties-in-a-recession/

10) Ryan Boykin. (2020, October 19). *The advantages of real estate vs. stocks*. Investopedia. https://www.investopedia.com/investing/reasons-invest-real-estate-vs-stock-market/

11) Sherwood, H. (2019, May 5). *How Airbnb took over the world*. the Guardian. https://www.theguardian.com/technology/2019/may/05/airbnb-homelessness-renting-housing-accommodation-social-policy-cities-travel-leisure

12) Si Jie. (2019, September 9). *Property valuation Singapore - Three methods that you need to know*. iCompareLoan Resources. https://www.icompareloan.com/resources/property-valuation-singapore/

13) Tay, T. (2020, May 11). *Singapore commercial real estate investment volume drops 68% Y-o-Y in 1Q2020*. EdgeProp.sg: Singapore Property for Sale & Rent, Latest Property News. https://www.edgeprop.sg/property-news/singapore-commercial-real-estate-investment-volume-drops-68-y-o-y-1q2020

14) Tom Bohjalian. (2019, February). *A REIT Defense for the Late Cycle*. https://assets.cohenandsteers.com/assets/content/resources/insight/MP866_REIT_Late_Cycle.pdf

# Appendices

## Appendix A - Variable Cleaning

listings.csv

1. Missing Values

Before cleaning

```
sapply(cleanlistings.dt, function(x) sum(is.na(x)))
                      id            host_is_superhost neighbourhood_group_cleansed
                       0                            0                            0
                latitude                    longitude                property_type
                       0                            0                            0
               room_type                 accommodates                    bathrooms
                       0                            0                           16
                bedrooms                         beds                  square_feet
                       6                            1                         3721
                   price             security_deposit                 cleaning_fee
                       1                         2005                         1030
          guests_included             availability_365            number_of_reviews
                       0                            0                            0
         instant_bookable          review_scores_rating
                       0                          647
```

After cleaning

```
sapply(cleanlistings.dt, function(x) sum(is.na(x)))
                      id            host_is_superhost neighbourhood_group_cleansed
                       0                            0                            0
                latitude                    longitude                property_type
                       0                            0                            0
               room_type                 accommodates                    bathrooms
                       0                            0                            0
                bedrooms                         beds                        price
                       0                            0                            0
          security_deposit                cleaning_fee              guests_included
                       0                            0                            0
          availability_365             number_of_reviews             instant_bookable
                       0                            0                            0
      review_scores_rating
                       0
```

2. Wrong Datatype - Data cleaning to factor and numeric types

Before cleaning

```
                      id            host_is_superhost neighbourhood_group_cleansed
               "integer"                    "factor"                    "factor"
                latitude                    longitude                property_type
               "numeric"                    "numeric"                    "factor"
               room_type                 accommodates                    bathrooms
                "factor"                    "integer"                    "numeric"
                bedrooms                         beds                  square_feet
               "integer"                    "integer"                    "integer"
                   price             security_deposit                 cleaning_fee
               "numeric"                    "numeric"                    "numeric"
          guests_included             availability_365            number_of_reviews
               "integer"                    "integer"                    "integer"
         instant_bookable          review_scores_rating
                "factor"                    "integer"
```

After cleaning

```
                id            host_is_superhost neighbourhood_group_cleansed
         "integer"                    "factor"                      "factor"
          latitude                   longitude                 property_type
         "numeric"                    "numeric"                      "factor"
         room_type                accommodates                     bathrooms
          "factor"                    "numeric"                     "numeric"
          bedrooms                        beds                   square_feet
         "numeric"                    "numeric"                     "integer"
             price            security_deposit                  cleaning_fee
         "numeric"                    "numeric"                     "numeric"
   guests_included            availability_365            number_of_reviews
         "numeric"                    "numeric"                     "numeric"
instant_bookable          review_scores_rating
          "factor"                    "numeric"
```

3. Insignificant categories
   a. Before cleaning

```
             Apartment Bed & Breakfast           Boat       Bungalow          Cabin
             1          1708            37              8             13             21
Camper/RV          Chalet     Condominium           Dorm          House           Loft
            13             2             91              2           1733             40
         Other           Tent       Townhouse      Treehouse          Yurt
            22             5             118             3              1
```

   b. After cleaning

```
Apartment     House     Other
     1839      1924        55
```

reviews.csv

1. Wrong Values

   Converted encoding to ASCII to remove foreign characters.

```
texts(data.corpus) <- iconv(texts(data.corpus), from="UTF-8", to="ASCII", sub = "")
```

   Stemmed and added german stopwords to reduce occurrences of “die” which is the in German.

```
> sum(ntoken(data.tokens))
[1] 5797164
> # Stopword removal
> data.tokens <- tokens_remove(data.tokens, pattern = stopwords("en"))
> data.tokens <- tokens_remove(data.tokens, pattern = stopwords("german"))
> sum(ntoken(data.tokens))
[1] 2991007
```

house_sales.csv

1. Missing Values

   Before cleaning

```
> nrow(cleaned.house.sale.data[bathrooms == 0 | bedrooms == 0])
[1] 16
```

   After cleaning

```
> nrow(cleaned.house.sale.data[bathrooms == 0 | bedrooms == 0])
[1] 0
```

2. Wrong Datatype - Data cleaning to factor and numeric types

Before cleaning

```
            id          price       bedrooms      bathrooms    sqft_living       sqft_lot         floors
 "integer64"      "integer"      "integer"      "numeric"      "integer"      "integer"      "numeric"
    waterfront           view      condition          grade    sqft_above sqft_basement       yr_built
     "integer"      "integer"      "integer"      "integer"      "integer"      "integer"      "integer"
 yr_renovated sqft_living15     sqft_lot15
     "integer"      "integer"      "integer"
```

After cleaning

```
            id          price       bedrooms      bathrooms    sqft_living       sqft_lot         floors
 "integer64"      "numeric"      "numeric"      "numeric"      "numeric"      "numeric"      "numeric"
    waterfront           view      condition          grade    sqft_above sqft_basement       yr_built
     "integer"      "integer"      "integer"      "integer"      "numeric"      "numeric"      "numeric"
 yr_renovated sqft_living15     sqft_lot15
     "numeric"      "numeric"      "numeric"
```

3. Feature engineering

```
> # Feature Engineering
> cleaned.house.sale.data$age <- 2020 - pmax(cleaned.house.sale.data$yr_built,
+                                            cleaned.house.sale.data$yr_renovated)
> summary(cleaned.house.sale.data$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   21.00   43.00   46.62   66.00  120.00
```

# Appendix B - Visualization Graphs and Variable Summary

Figure 1: Seattle Airbnb Listing Dataset Variable Summary

```
> #Summary of cleaned dataset
> summary(cleanlistings.dt)
       id            host_is_superhost      neighbourhood_group_cleansed    latitude       longitude        property_type
 Min.   :    3335   f:3040            Other neighborhoods: 794          Min.   :47.51   Min.   :-122.4   Apartment:1839
 1st Qu.: 3258256   t: 778            Capitol Hill       : 567          1st Qu.:47.61   1st Qu.:-122.4   House    :1924
 Median : 6118244                     Downtown           : 530          Median :47.62   Median :-122.3   Other    :  55
 Mean   : 5550111                     Central Area       : 369          Mean   :47.63   Mean   :-122.3
 3rd Qu.: 8035127                     Queen Anne         : 295          3rd Qu.:47.66   3rd Qu.:-122.3
 Max.   :10340165                     Ballard            : 230          Max.   :47.73   Max.   :-122.2
                                      (Other)            :1033
          room_type     accommodates      bathrooms       bedrooms          beds            price       security_deposit  cleaning_fee
 Entire home/apt:2541   Min.   : 1.000   Min.   :0.000   Min.   :0.000   Min.   : 1.000   Min.   : 20.0   Min.   :  0.0    Min.   :  0.00
 Private room   :1160   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.: 75.0   1st Qu.:  0.0    1st Qu.:  0.00
 Shared room    : 117   Median : 3.000   Median :1.000   Median :1.000   Median : 1.000   Median :100.0   Median :  0.0    Median : 30.00
                        Mean   : 3.349   Mean   :1.258   Mean   :1.307   Mean   : 1.735   Mean   :127.7   Mean   :120.8    Mean   : 45.06
                        3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.: 2.000   3rd Qu.:150.0   3rd Qu.:200.0    3rd Qu.: 65.00
                        Max.   :16.000   Max.   :8.000   Max.   :7.000   Max.   :15.000   Max.   :999.0   Max.   :995.0    Max.   :300.00

 guests_included  availability_365 number_of_reviews instant_bookable review_scores_rating
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00    f:3227           Min.   : 20.00
 1st Qu.: 1.000   1st Qu.:124.0   1st Qu.:  2.00    t: 591           1st Qu.: 94.00
 Median : 1.000   Median :308.0   Median :  9.00                     Median : 96.00
 Mean   : 1.673   Mean   :244.8   Mean   : 22.22                     Mean   : 94.79
 3rd Qu.: 2.000   3rd Qu.:360.0   3rd Qu.: 26.00                     3rd Qu.: 98.00
 Max.   :15.000   Max.   :365.0   Max.   :474.00                     Max.   :100.00
```

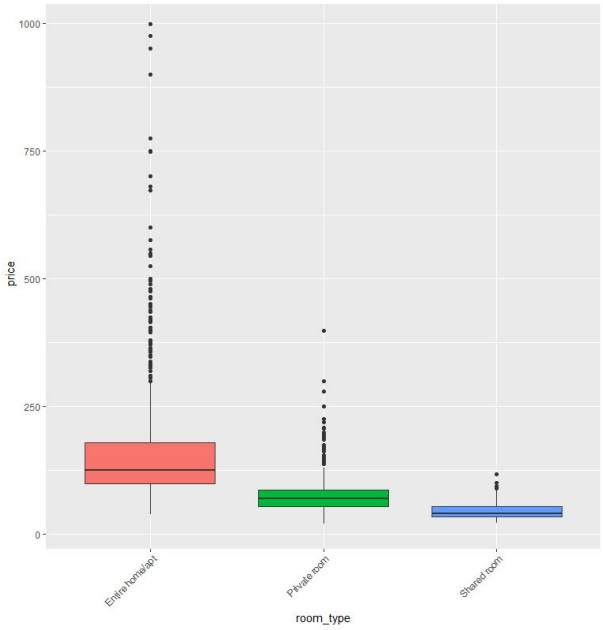
Figure 2: Seattle Airbnb Review Dataset Variable Summary

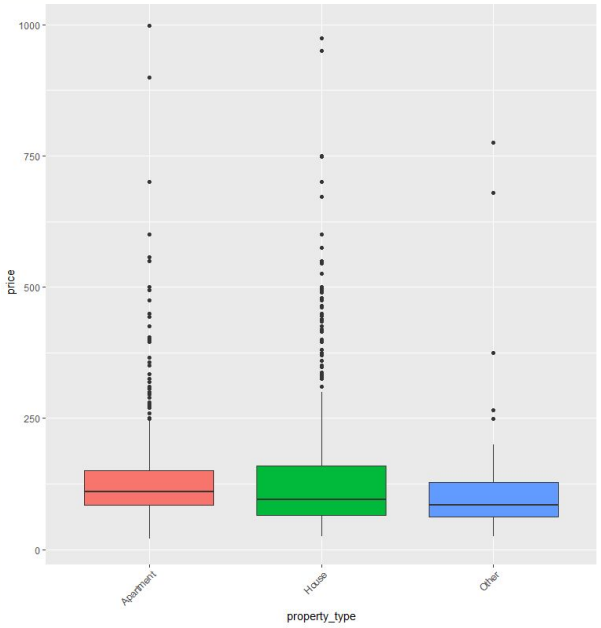
```
> summary(review.data)
    doc_id             text             listing_id             id                date             reviewer_id
 Length:84849       Length:84849       Min.   :    4291   Min.   :    3721   Length:84849       Min.   :      15
 Class :character   Class :character   1st Qu.:  794633   1st Qu.:17251274   Class :character   1st Qu.: 5053141
 Mode  :character   Mode  :character   Median : 2488228   Median :32288093   Mode  :character   Median :14134759
                                       Mean   : 3005067   Mean   :30587645                      Mean   :17013007
                                       3rd Qu.: 4694479   3rd Qu.:44576477                      3rd Qu.:27624023
                                       Max.   :10248139   Max.   :58736511                      Max.   :52812740

 reviewer_name
 Length:84849
 Class :character
 Mode  :character
```

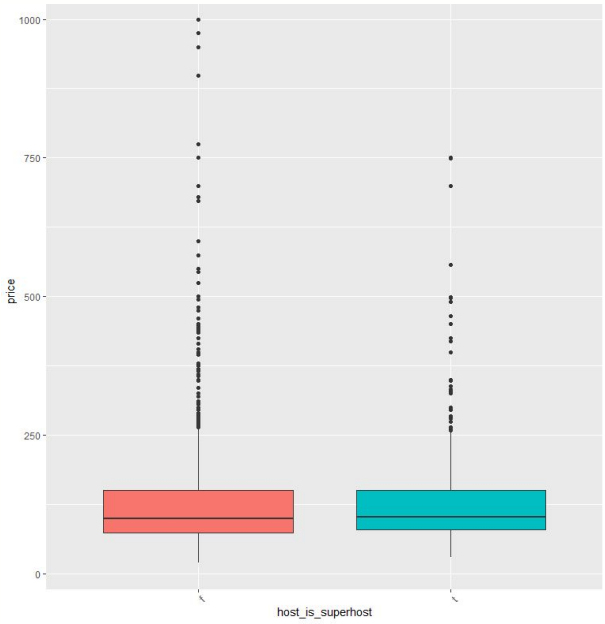
Figure 3: Seattle House Prices Dataset Variable Summary

```
> summary(cleaned.house.sale.data)
     price            bedrooms        bathrooms      sqft_living       sqft_lot          floors
 Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000
 1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000
 Median : 450000   Median : 3.000   Median :2.250   Median : 1910   Median :   7618   Median :1.500
 Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494
 3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000
 Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500
   waterfront           view          condition        grade          sqft_above   sqft_basement       yr_built
 Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000   Min.   : 290   Min.   :   0.0   Min.   :1900
 1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951
 Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000   Median :1560   Median :   0.0   Median :1975
 Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657   Mean   :1788   Mean   : 291.5   Mean   :1971
 3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997
 Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000   Max.   :9410   Max.   :4820.0   Max.   :2015
  yr_renovated    sqft_living15    sqft_lot15
 Min.   :   0.0   Min.   : 399   Min.   :   651
 1st Qu.:   0.0   1st Qu.:1490   1st Qu.:  5100
 Median :   0.0   Median :1840   Median :  7620
 Mean   :  84.4   Mean   :1987   Mean   : 12768
 3rd Qu.:   0.0   3rd Qu.:2360   3rd Qu.: 10083
 Max.   :2015.0   Max.   :6210   Max.   :871200
>
```

Table 4: Figures illustrating relationship of price to individual categorical variables in the listings dataset



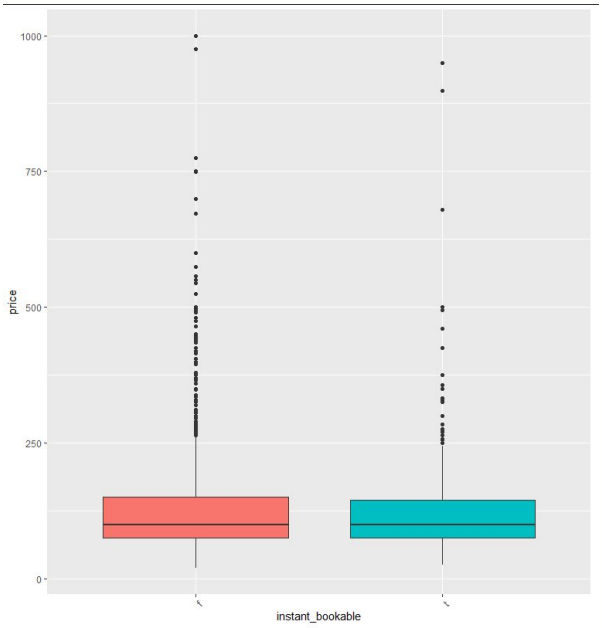



Figure 5: Illustration of relationship between price and neighbourhood_group

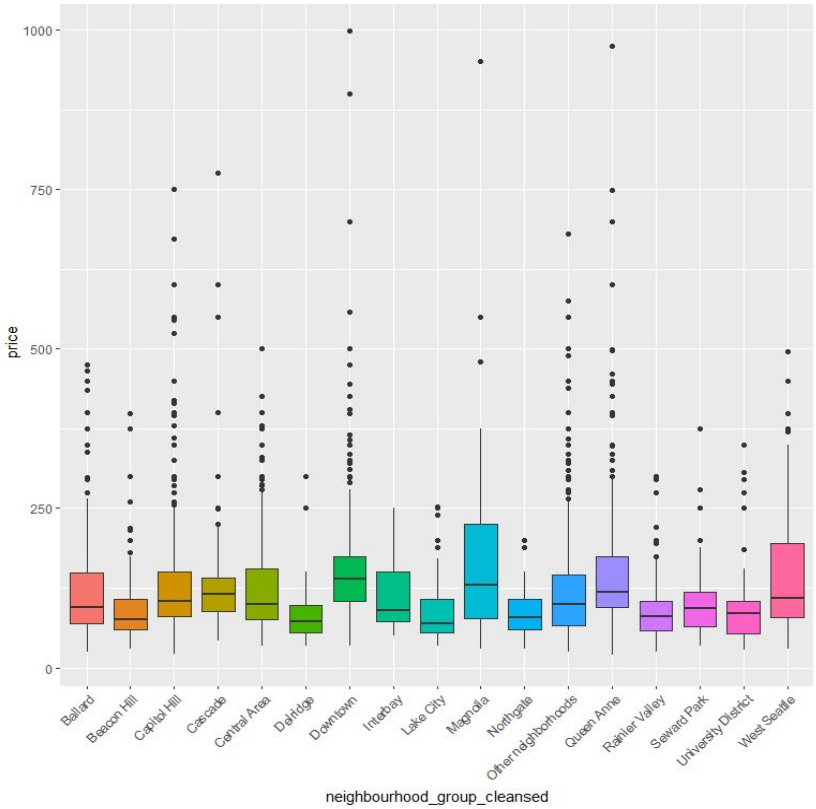



Figure 6: Correlation heatmap of all numerical variables

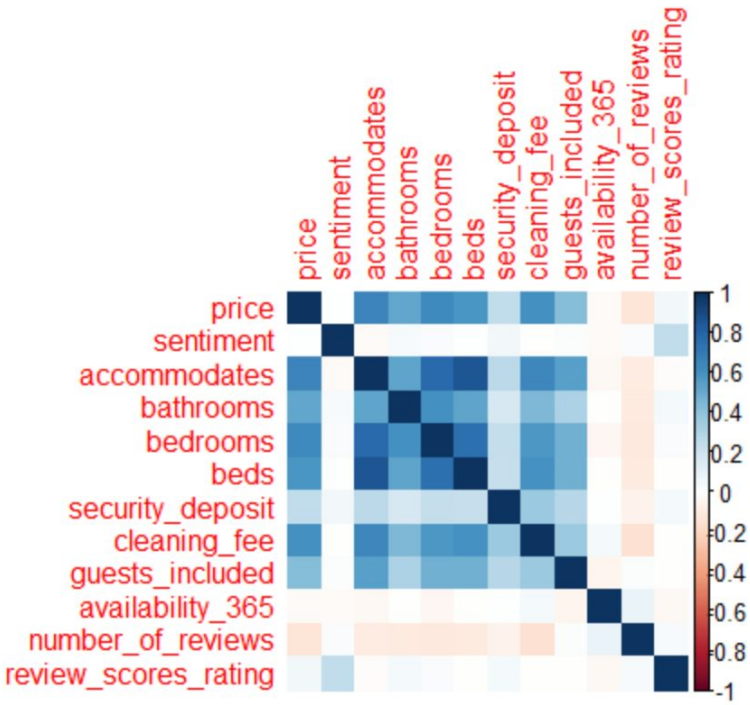


Table 7: Figures illustrating the relationship between price/median price and the numeric variables in the listings dataset

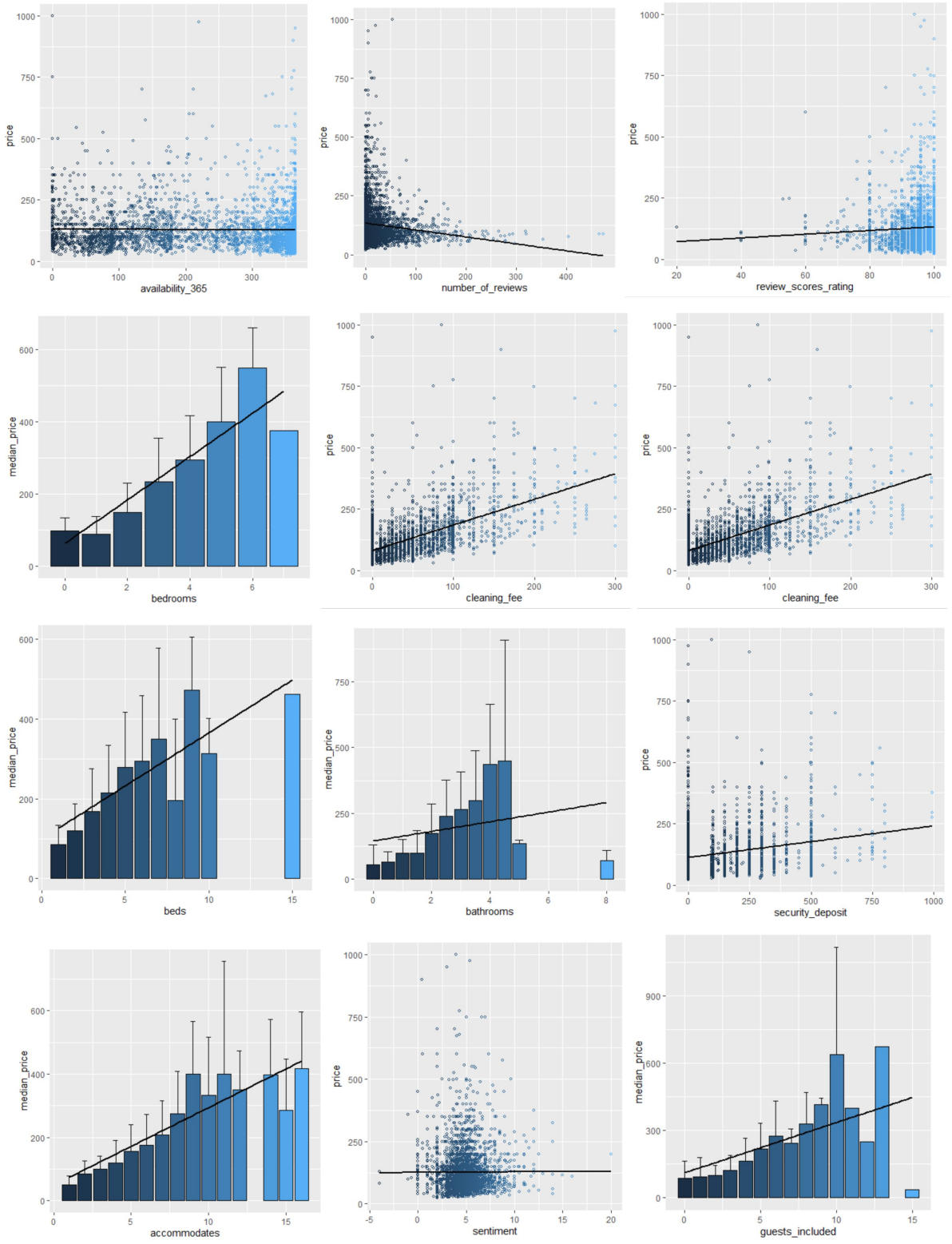


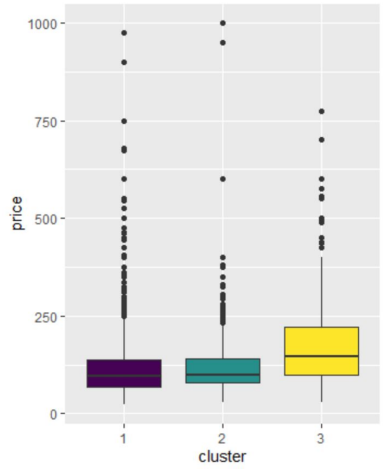
Figure 8: Top keywords amongst all reviews

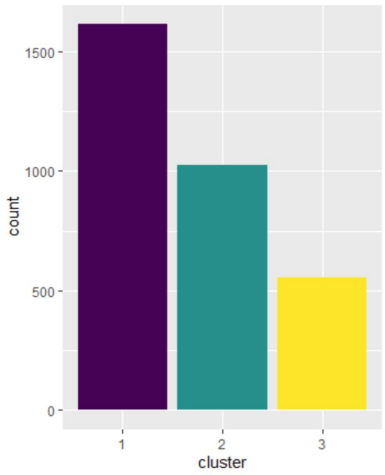

Figure 9: Top keywords amongst reviews with positive sentiments

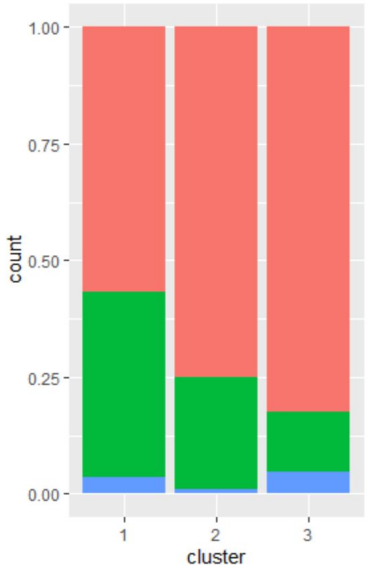

Figure 10: Top keywords amongst reviews with negative sentiments

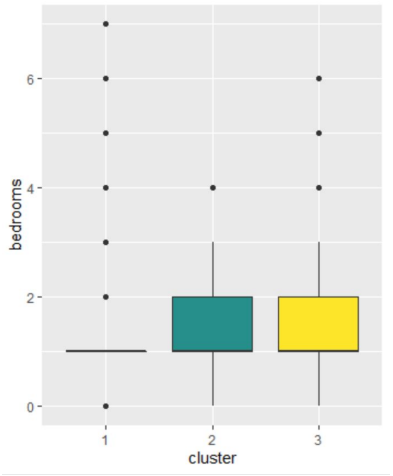

Figure 11: Top keywords amongst reviews with neutral sentiment

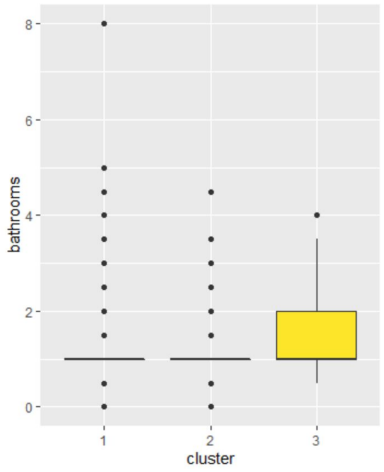
Table 12: Figures illustrating the characteristics in each cluster for listings.csv

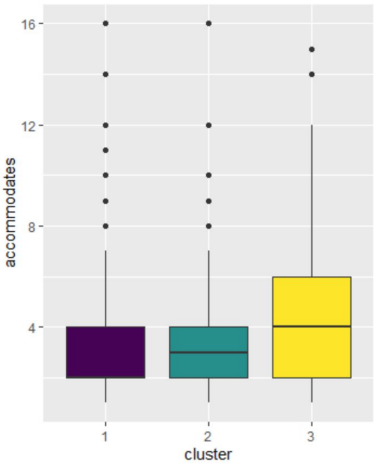


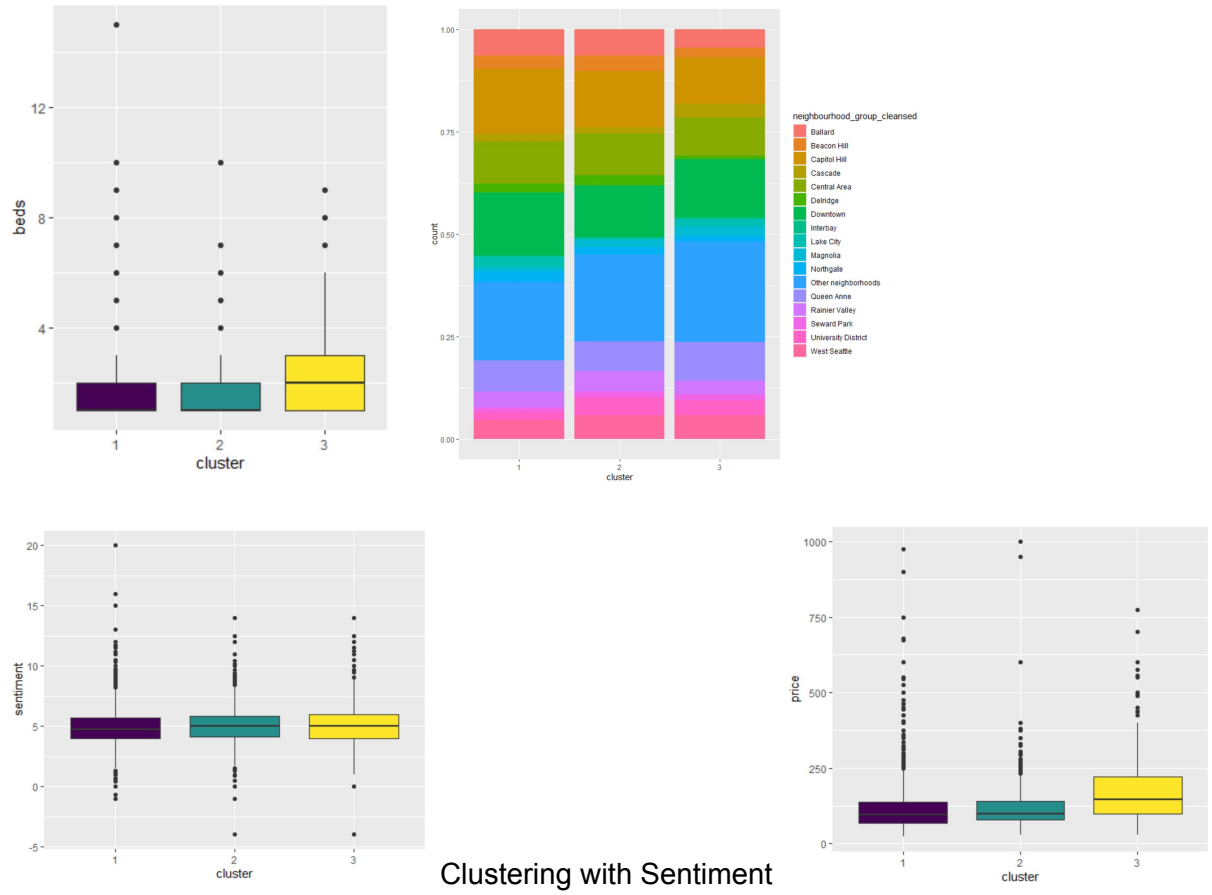
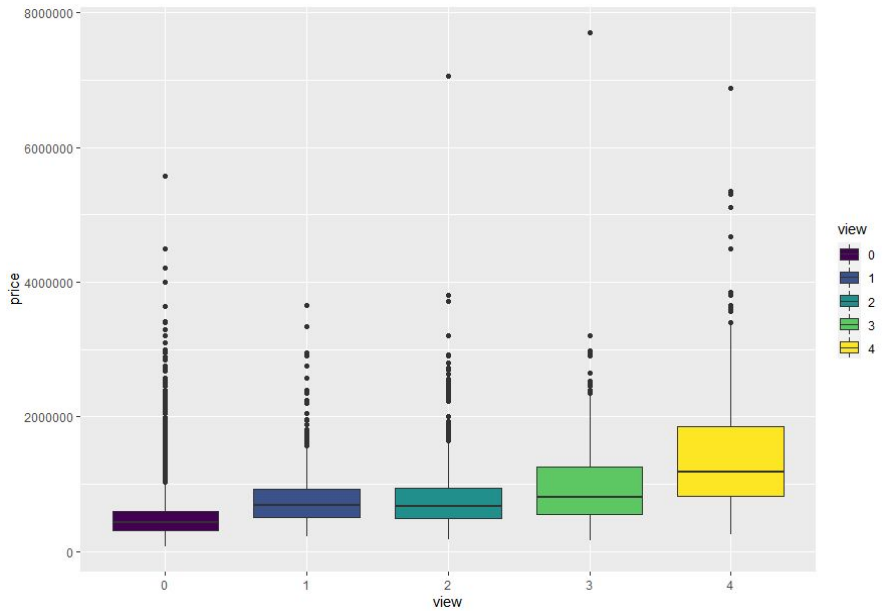
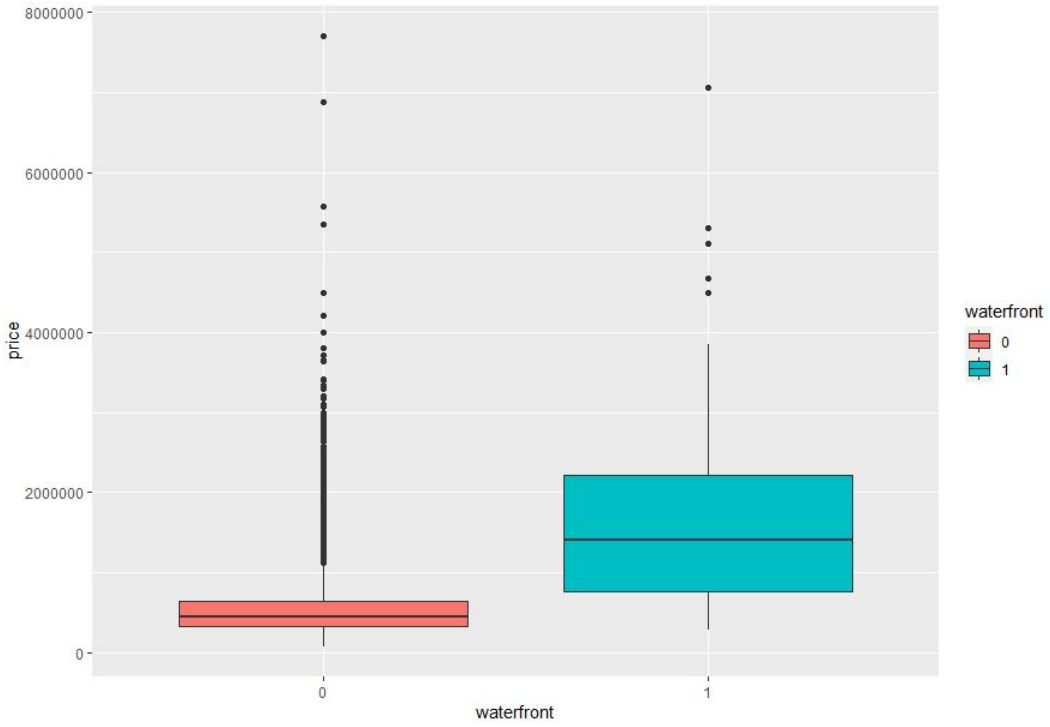
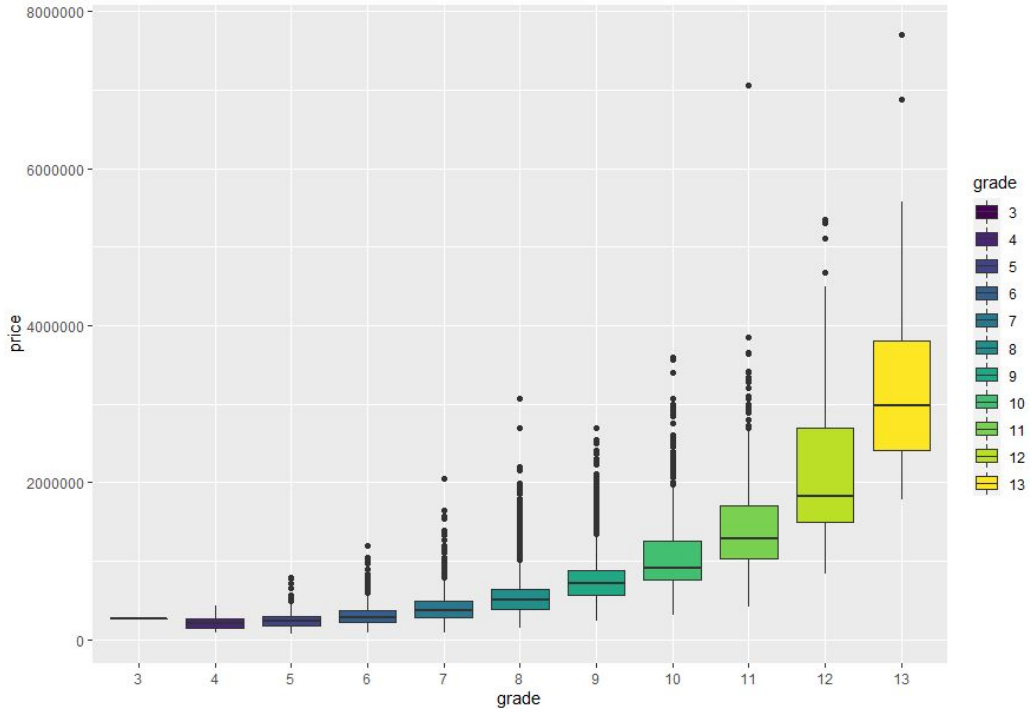



Table 13: Figures illustrating relationship of price to individual categorical variables in the kc_house_data dataset




8000000
6000000
4000000
2000000
0
price
0
1
waterfront
waterfront
0
1


8000000
6000000
4000000
2000000
0
price
3
4
5
6
7
8
9
10
11
12
13
grade
grade
3
4
5
6
7
8
9
10
11
12
13

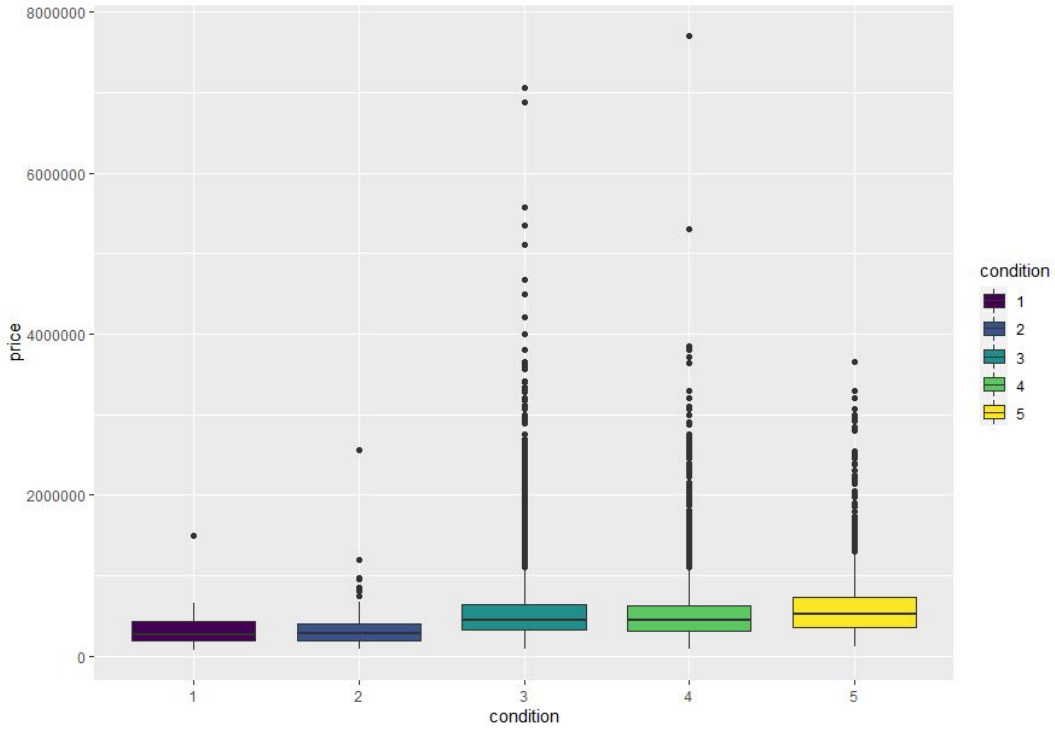



Figure 14: Correlation Heat Map of kc_house_data numerical data

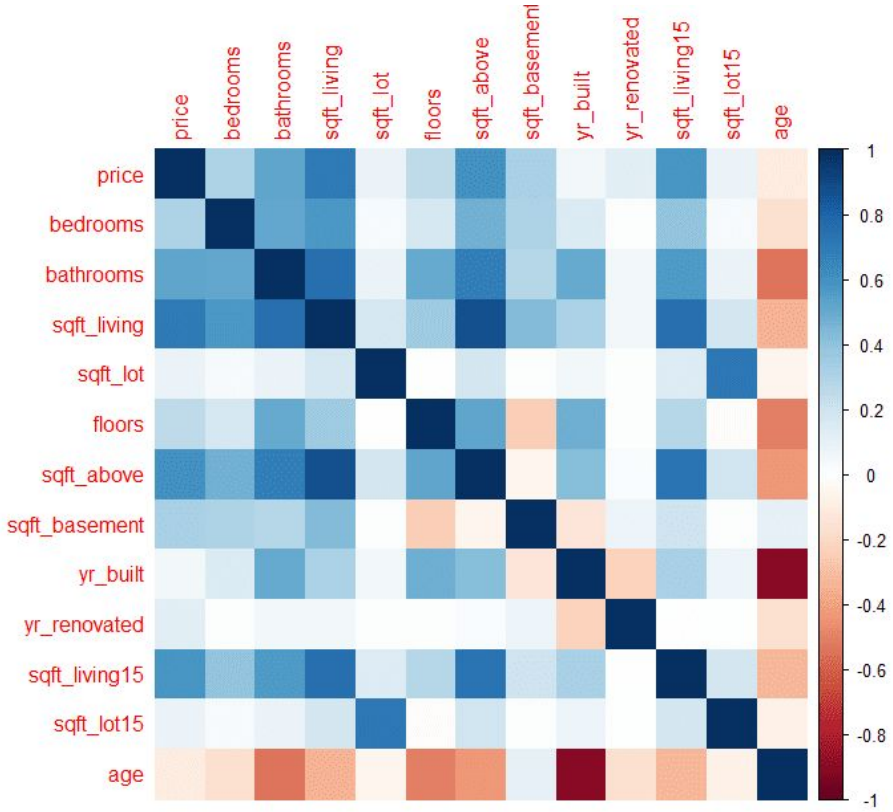


Figure 15: Within Cluster Sum of Squares Plot in kc_house_data dataset

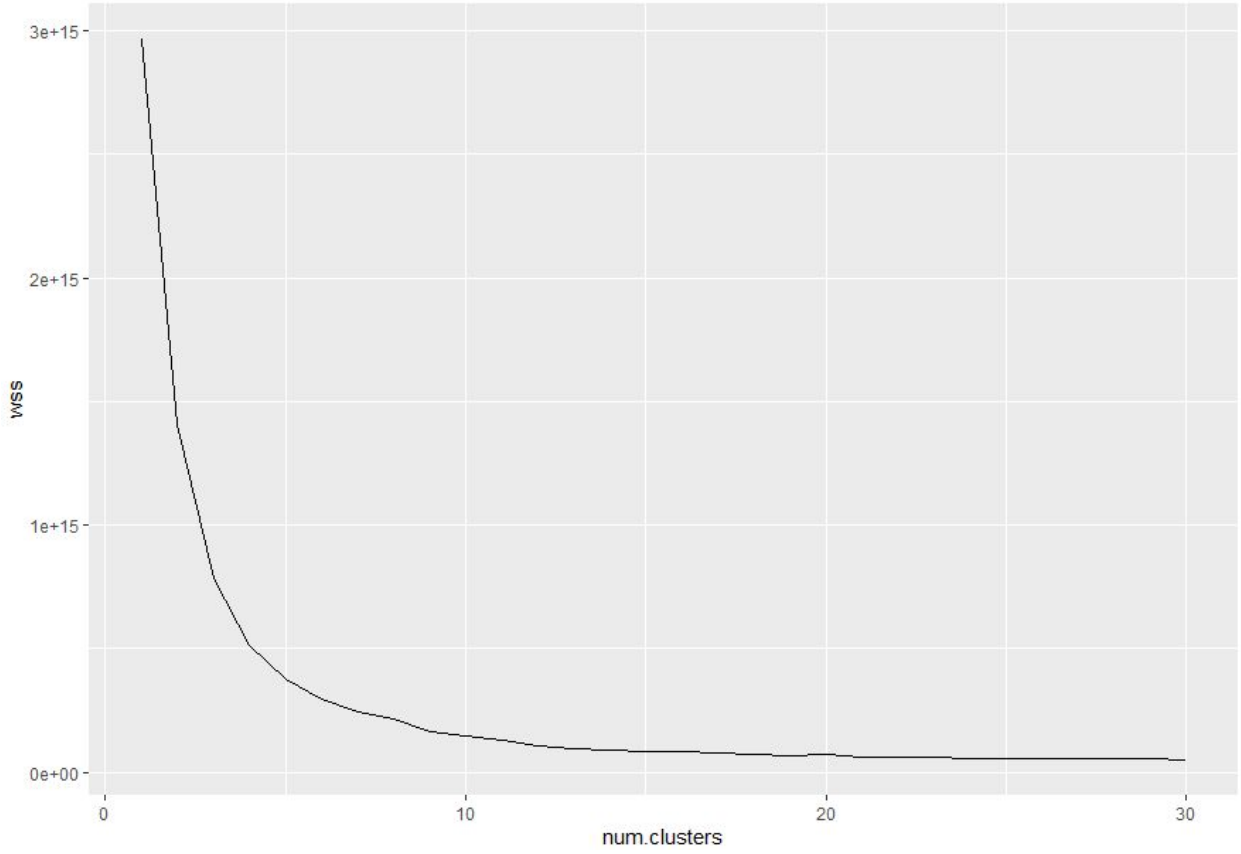
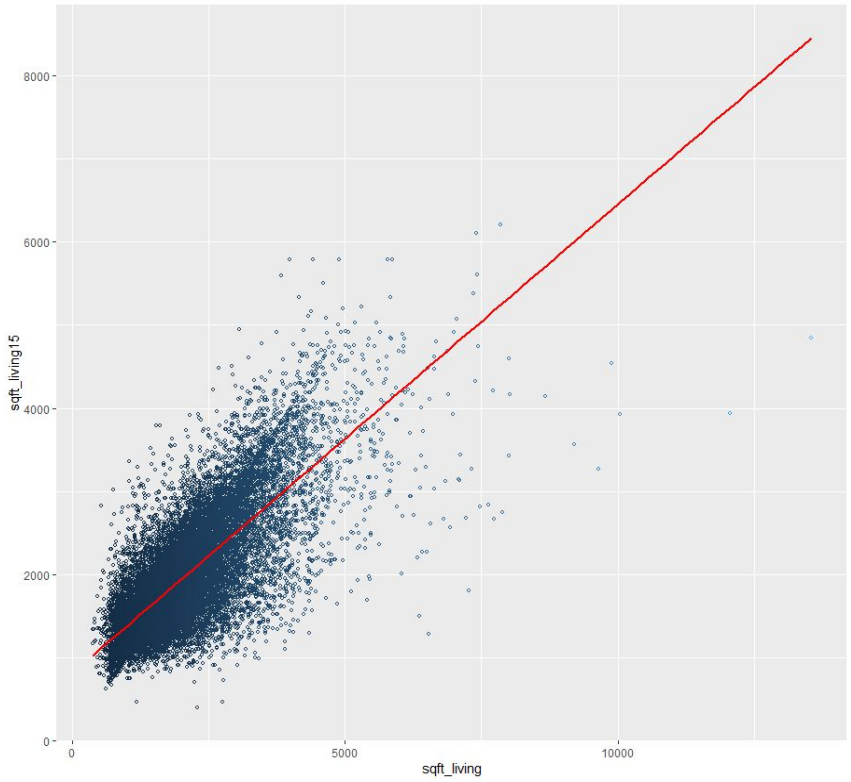


Table 16: Figures illustrating the high correlation between the unknown variables in dataset



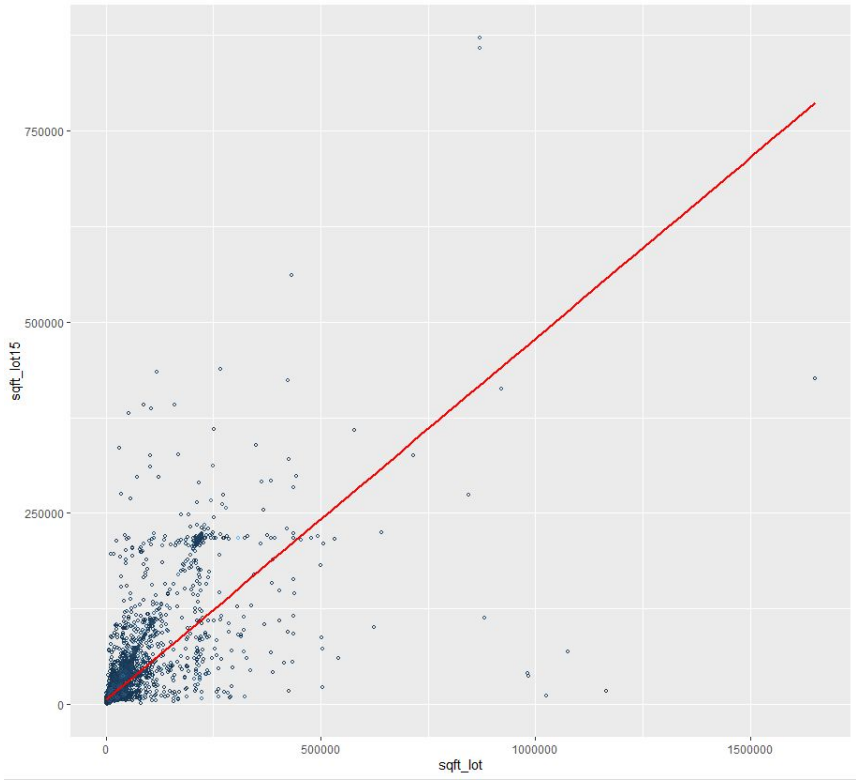
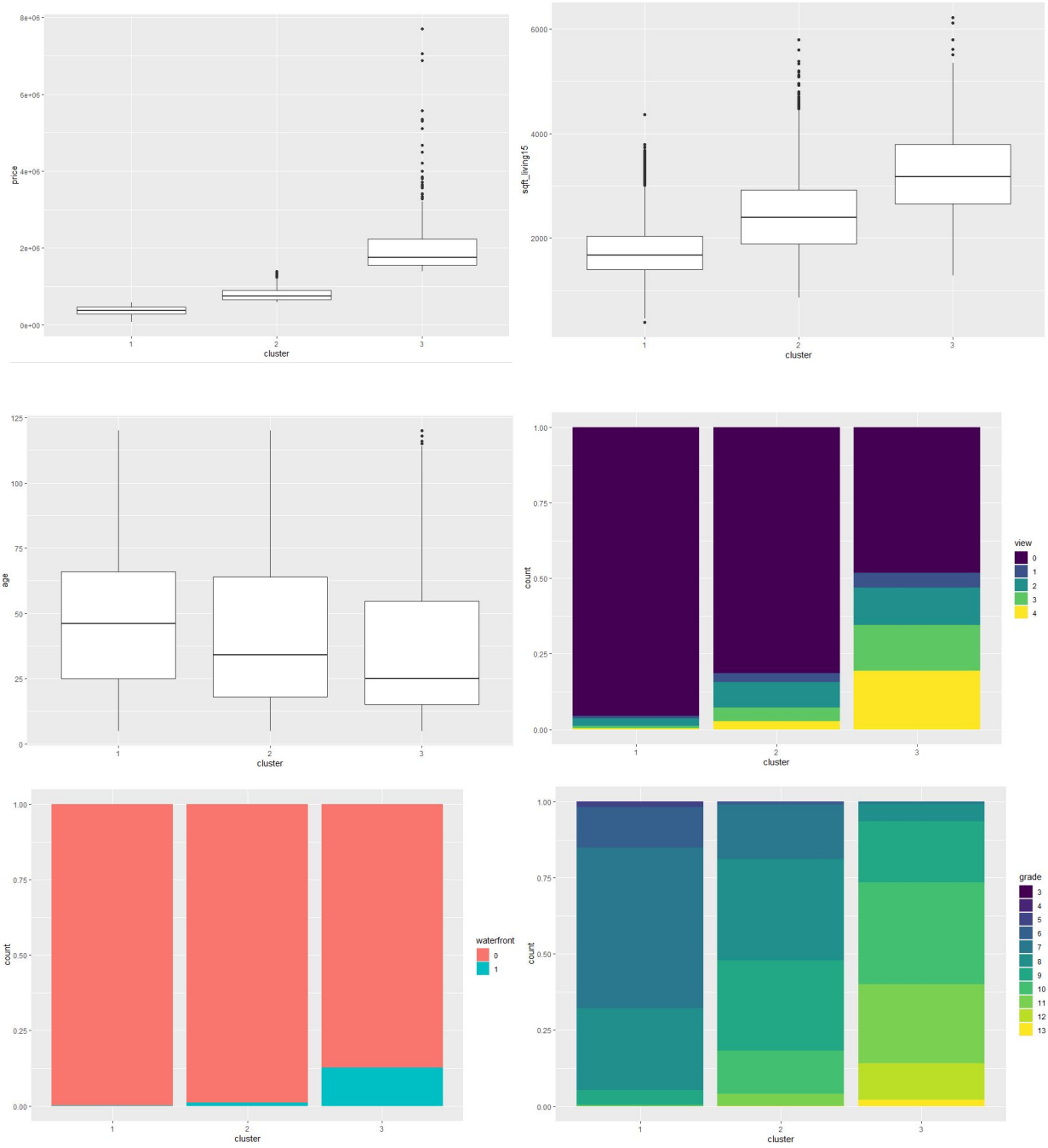
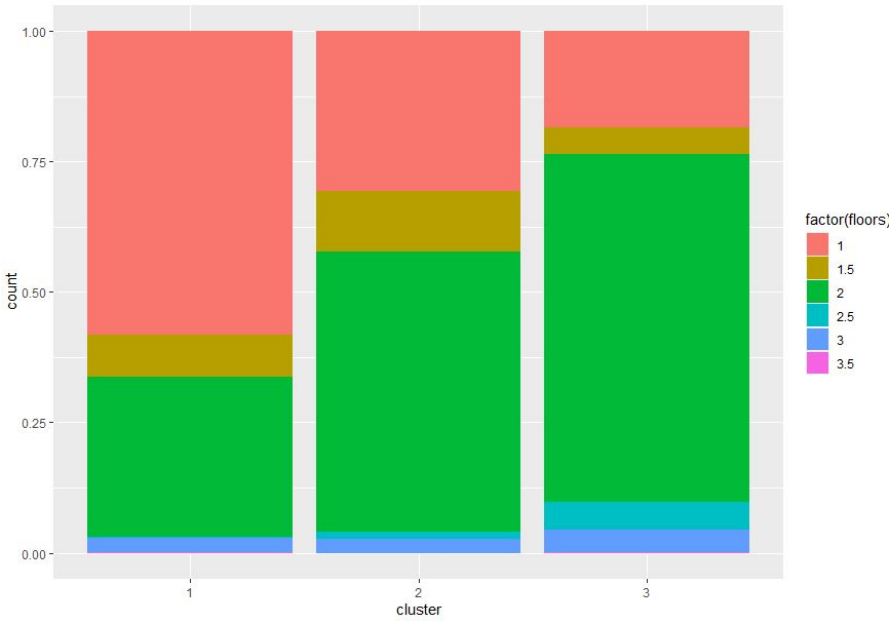



Table 17: Figures illustrating the key characteristics in each cluster for kc_house_data.csv




1.00
0.75
0.50
0.25
0.00
count
1
2
3
cluster
factor(floors)
1
1.5
2
2.5
3
3.5

## Appendix C - Prediction Model Summaries

Figure 1: Linear Regression Model Summary on listings.csv

```
Call:
lm(formula = price ~ bathrooms + bedrooms + cleaning_fee + guests_included +
    room_type, data = train.listings)

Residuals:
    Min      1Q  Median      3Q     Max
-241.46  -26.73   -6.00   19.10  864.87

Coefficients:
                        Estimate Std. Error t value             Pr(>|t|)
(Intercept)             35.16684    3.39883  10.347 < 0.0000000000000002 ***
bathrooms               35.44857    2.78219  12.741 < 0.0000000000000002 ***
bedrooms                28.46542    2.13181  13.353 < 0.0000000000000002 ***
cleaning_fee             0.35378    0.03453  10.245 < 0.0000000000000002 ***
guests_included          4.97871    1.11081   4.482            0.00000777 ***
room_typePrivate room -44.52319    3.11132 -14.310 < 0.0000000000000002 ***
room_typeShared room  -69.74165    8.12825  -8.580 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.68 on 2226 degrees of freedom
Multiple R-squared:  0.5575,    Adjusted R-squared:  0.5563
F-statistic: 467.5 on 6 and 2226 DF,  p-value: < 0.00000000000000022
```

Figure 2: CART model before pruning on listings.csv

```
Regression tree:
rpart(formula = price ~ accommodates + bathrooms + bedrooms +
    beds + security_deposit + cleaning_fee + guests_included +
    room_type + property_type, data = cleanlistings.dt, method = "anova",
    control = rpart.control(cp = 0))

Variables actually used in tree construction:
[1] accommodates     bathrooms        bedrooms         beds             cleaning_fee
[6] guests_included  property_type    room_type        security_deposit

Root node error: 25166010/3191 = 7886.6

n= 3191
```

Figure 3: CP plot of CART model before pruning on listings.csv

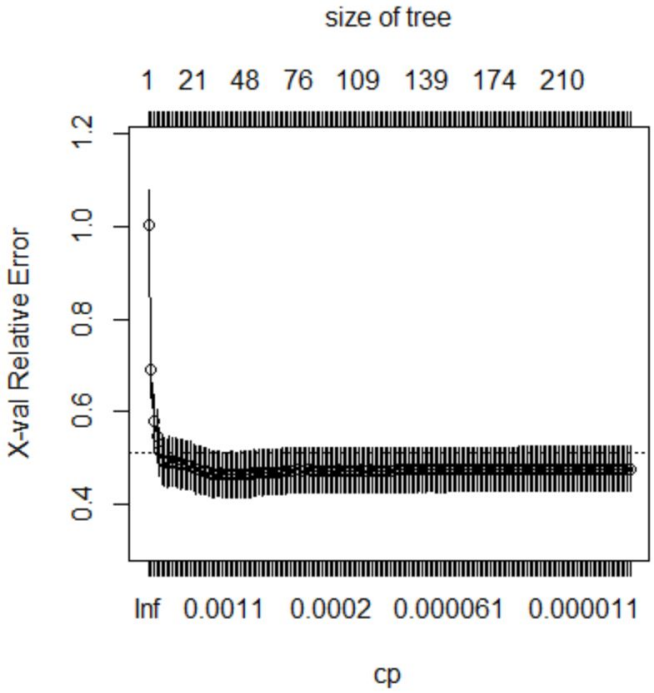


Figure 4: CART model after pruning on listings.csv

```
Call:
rpart(formula = price ~ accommodates + bathrooms + bedrooms +
    beds + security_deposit + cleaning_fee + guests_included +
    room_type + property_type, data = cleanlistings.dt, method = "anova",
    control = rpart.control(cp = 0))
  n= 3191

          CP nsplit rel error    xerror       xstd
1 0.32833018      0 1.0000000 1.0013663 0.07827699
2 0.10385939      1 0.6716698 0.6903766 0.06000621
3 0.04527254      2 0.5678104 0.5799935 0.05706949
4 0.04239764      3 0.5225379 0.5464851 0.05692885
5 0.01661904      4 0.4801403 0.5127015 0.05331818
6 0.01563211      5 0.4635212 0.4942100 0.05228714

Variable importance
       bedrooms    accommodates       bathrooms            beds    cleaning_fee
             33              16              12              10               9
      room_type guests_included   property_type
              9               8               2
```

Figure 5: CP plot of CART model after pruning on listings.csv

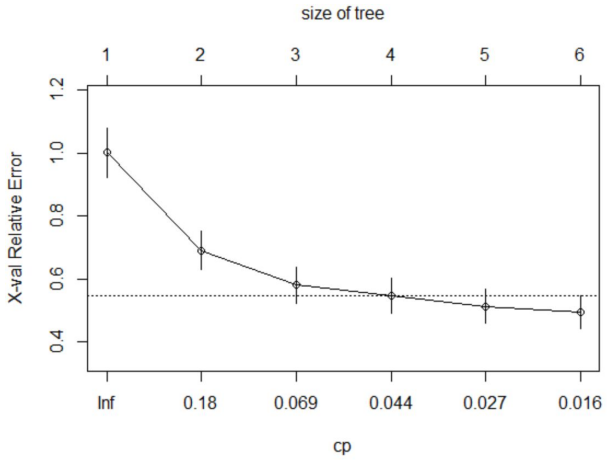



Figure 6: Linear Regression Model Summary on kc_house_data.csv

```
Call:
lm(formula = price ~ bathrooms + waterfront + view + grade +
    sqft_lot15 + age + cluster, data = cleaned.house.sale.data.clusters.train)

Residuals:
     Min       1Q   Median       3Q      Max
-1772731   -86415   -11315    74156  4686637

Coefficients:
                Estimate    Std. Error t value             Pr(>|t|)
(Intercept)  985930.48914   18377.78055  53.648 < 0.0000000000000002 ***
bathrooms     65425.85227    2478.94840  26.393 < 0.0000000000000002 ***
waterfront1  286597.80002   17916.51400  15.996 < 0.0000000000000002 ***
view.L        84601.09688    9567.22490   8.843 < 0.0000000000000002 ***
view.Q        18861.25418    8369.15701   2.254              0.024231 *
view.C        55901.41282    9351.77315   5.978    0.000000002313186 ***
view^4        -3644.86455    7954.16201  -0.458              0.646791
grade.L     1503819.22685   83915.84143  17.921 < 0.0000000000000002 ***
grade.Q      970965.60071   87846.94103  11.053 < 0.0000000000000002 ***
grade.C      584031.40209   78346.67821   7.454    0.000000000000095 ***
grade^4      434077.83936   62413.96496   6.955    0.000000000003670 ***
grade^5      201370.16913   45513.07018   4.424    0.000009734104957 ***
grade^6      116290.70294   30838.67341   3.771              0.000163 ***
grade^7       65002.87355   19662.94972   3.306              0.000949 ***
grade^8       30363.22286   11874.47628   2.557              0.010567 *
grade^9        8102.68231    6809.18907   1.190              0.234079
grade^10       4475.20091    3714.49652   1.205              0.228301
sqft_lot15       -0.15759       0.05102  -3.089              0.002013 **
age            1501.82775      59.07814  25.421 < 0.0000000000000002 ***
cluster.L    833298.45659    7032.93174 118.485 < 0.0000000000000002 ***
cluster.Q    241073.05012    4068.62911  59.252 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 163100 on 15686 degrees of freedom
Multiple R-squared:  0.8199,    Adjusted R-squared:  0.8196
F-statistic:  3570 on 20 and 15686 DF,  p-value: < 0.00000000000000022
```

Figure 7: CART model before pruning on kc_house_data.csv

```
Regression tree:
rpart(formula = price ~ . - id, data = cleaned.house.sale.data.clusters,
    method = "anova", control = rpart.control(cp = 0))

Variables actually used in tree construction:
 [1] age           bathrooms     bedrooms      cluster       condition     floors        grade         sqft_above
 [9] sqft_basement sqft_living   sqft_living15 sqft_lot      sqft_lot15    view          waterfront    yr_built
[17] yr_renovated

Root node error: 2910863855476463/21597 = 134780935106

n= 21597
```

Figure 8: CP plot of CART model before pruning on kc_house_data.csv

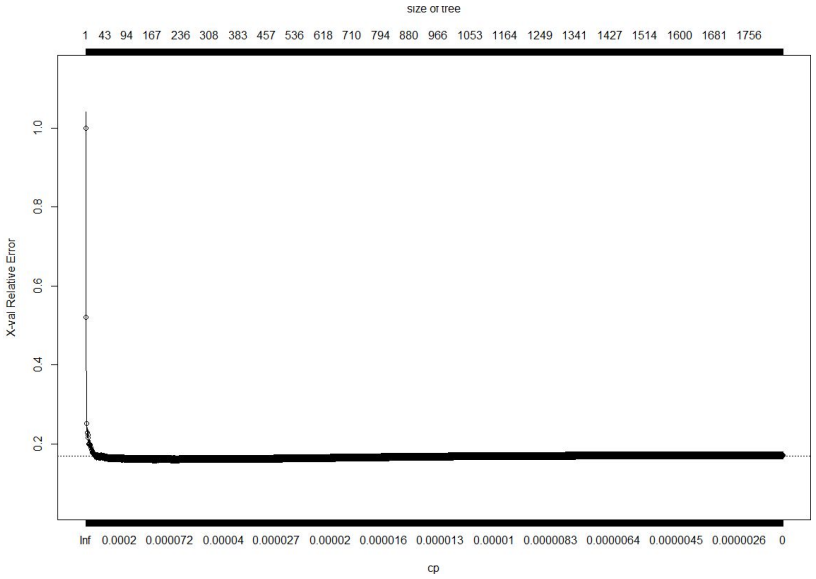
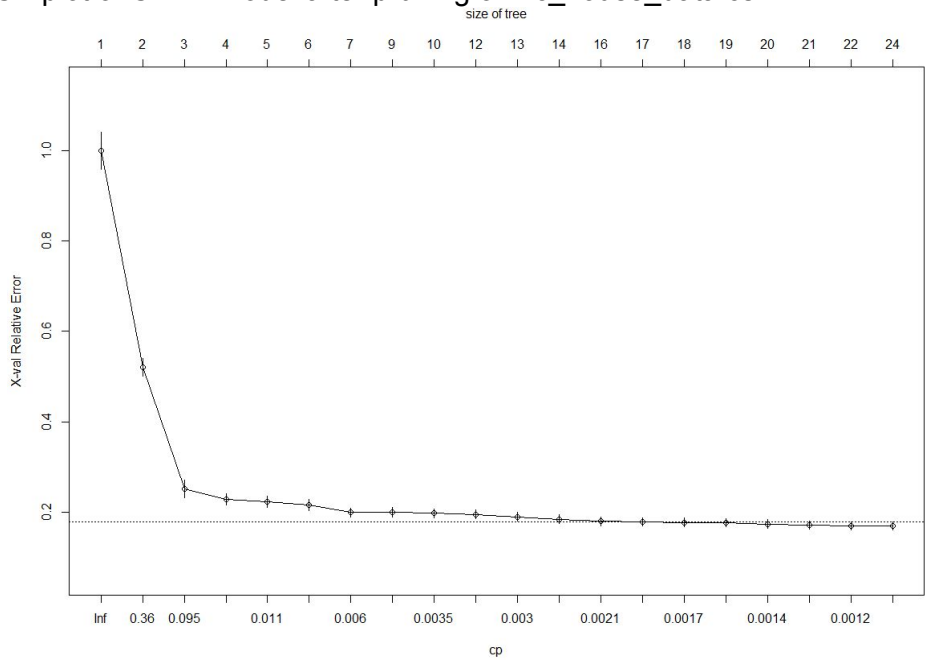


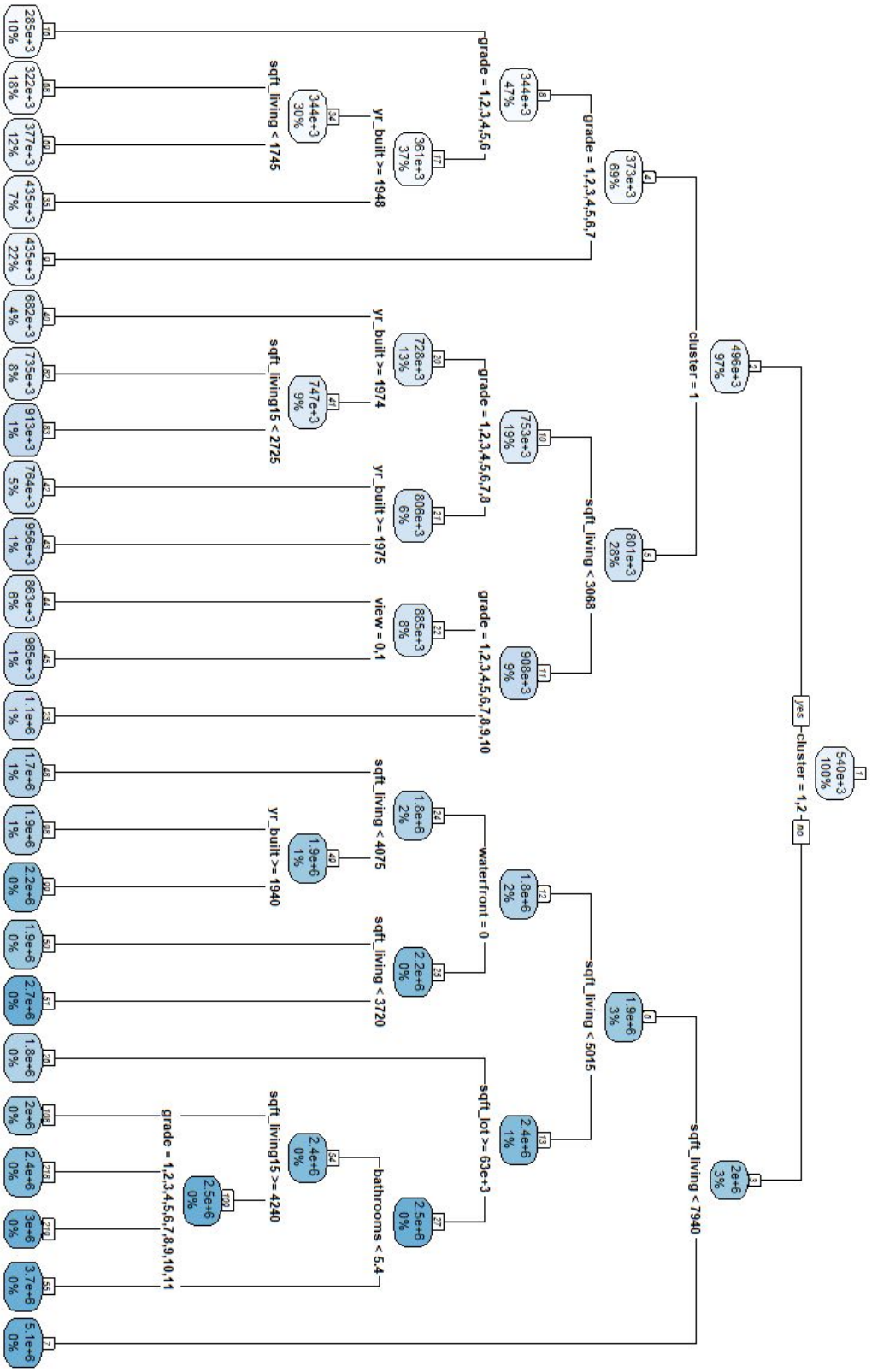
Figure 9: CART model after pruning on kc_house_data.csv

```
Regression tree:
rpart(formula = price ~ . - id, data = cleaned.house.sale.data.clusters,
    method = "anova", control = rpart.control(cp = 0))

Variables actually used in tree construction:
[1] bathrooms     cluster       grade         sqft_living   sqft_living15
[6] sqft_lot      view          waterfront    yr_built

Root node error: 2910863855476463/21597 = 134780935106

n= 21597
```

Figure 10: CP plot of CART model after pruning on kc_house_data.csv



## Appendix D - Visualization of CART model for house sale dataset



# Appendix E - Reliability of Assumptions