Real Estate Market Analysis Sydney & Melbourne

STAT 31631 – Statistical Modelling

Department of Statistics & Computer Science

University of Kelaniya

Academic Year 2022/2023

By

GROUP 11

- PS/2020/002 – A.S.N. Perera
- PS/2020/202 - M.A.S.C. Siriwardhana
- PS/2020/096 – D.A.G. Ranthilina
- PS/2020/155 - G.K.M. Pathum
- PS/2020/249 - R.H.R.S. Nishshanka
- PS/2020/074 - N.C.R. Gunarathna
- PS/2020/153 - D.M.S. Indrashan
- PS/2020/156 - C.K. Kariyawasam
- PS/2020/134 – M.R.R.C. Gunarathne

INTRODUCTION

The real estate markets in Sydney and Melbourne are among the most vibrant and dynamic in the world, attracting significant attention from both local and international investors. These cities are known for their unique blend of culture, economic opportunity, and lifestyle, making them prime locations for residential and commercial real estate investment. Understanding the factors that influence property prices in these markets is essential for stakeholders looking to make informed investment decisions. Real estate is a complex asset class influenced by a multitude of factors, including economic conditions, social trends, and specific property attributes. For investors, grasping the nuances of these influences can mean the difference between profitable and poor investment decisions. This study aims to dissect these influences through a comprehensive regression analysis, focusing on a variety of variables such as property size, condition, location, and market trends over time. By examining data from recent property sales, we hope to illuminate the underlying patterns that drive price fluctuations in Sydney and Melbourne. The significance of this analysis extends beyond immediate financial gains. By providing a detailed understanding of property price determinants, this study can aid in urban planning, housing policy formulation, and sustainable development efforts. Moreover, it offers a data-driven foundation for future research in real estate economics, contributing to the broader body of knowledge in this field.

PROBLEM STATEMENT

The real estate markets in Sydney and Melbourne are characterized by significant volatility and complexity, driven by a myriad of economic, social, and property-specific factors. Investors, policymakers, and urban planners often face challenges in accurately predicting property prices due to the intricate interplay of these variables. Despite the availability of extensive property data, there remains a gap in effectively analyzing and understanding the key determinants that most significantly influence property values.

Specifically, the problem is twofold:

1. Lack of Predictive Insight: Existing models and analyses often fail to provide reliable predictions of property prices due to their inability to adequately account for the diverse range of factors at play.

2. Uncertainty in Decision-Making: Without a clear understanding of what drives property prices, stakeholders are left to make investment, policy, and planning decisions based on incomplete or outdated information, potentially leading to suboptimal outcomes.

To address this problem, there is a need for a comprehensive regression analysis that can integrate multiple variables, including economic indicators, property attributes, and temporal trends, to accurately predict property prices in Sydney and Melbourne. This study aims to fill this gap by developing a robust predictive model that identifies and quantifies the impact of key factors influencing property values, thereby providing valuable insights for informed decision-making.

OBJECTIVES

- Identify Key Factors:
  Determine the most significant variables influencing property prices in Sydney and Melbourne.

- Predict Property Prices:
  Develop a regression model to predict property prices based on identified factors.

- Analyze Trends:
  Examine temporal trends in property prices to understand market fluctuations over time.

- Provide Recommendations:
  Offer actionable insights for investors and policymakers based on the analysis.

SIGNIFICANCE OF THE STUDY

The real estate sector is a cornerstone of the Australian economy, with Sydney and Melbourne serving as pivotal hubs. The significance of this study lies in its potential to provide deep insights into the complex mechanisms governing property prices in these cities.

Economic Impact

Real estate prices are a critical indicator of economic health. Fluctuations in property values can have wide-ranging effects on the economy, influencing consumer spending, investment flows, and overall financial stability. By identifying the key factors that drive property prices, this study helps investors and policymakers anticipate market movements and make strategic decisions to foster economic stability and growth.

Investment Strategies

For investors, both individual and institutional, understanding the determinants of property prices is crucial for optimizing portfolio performance. This study aims to offer actionable insights into which factors most significantly impact property values, enabling investors to tailor their strategies effectively. Whether considering residential properties or commercial real estate, these insights can guide investment decisions to maximize returns and mitigate risks.

Urban Planning and Policy Development

City planners and policymakers can greatly benefit from this study's findings. By understanding how various factors such as infrastructure development, population growth, and zoning regulations affect property prices, planners can make more informed decisions that promote balanced and sustainable urban development. The insights from this study can also inform housing policies aimed at improving affordability and accessibility, ensuring that urban growth benefits all residents.

Social and Environmental Considerations

Beyond economic and investment perspectives, this study addresses important social and environmental factors. The impact of property prices on social equity, community stability, and environmental sustainability is profound. By analyzing the role of features such as property condition, renovation status, and environmental amenities (like waterfront views), this study contributes to a holistic understanding of real estate dynamics. It underscores the importance of integrating social and environmental considerations into real estate valuation and urban development strategies

In summary, this study's significance is multifaceted, impacting economic stability, investment strategies, urban planning, social equity, environmental sustainability, and academic research. By shedding light on the factors influencing property prices in Sydney and Melbourne, it aims to contribute valuable knowledge and practical insights to a wide range of stakeholders.

METHODOLOGY

1. Data Collection

Objective: To gather comprehensive data on property sales in Sydney and Melbourne.

- Data Sources: The data will be collected from reputable sources, including government databases (e.g., Australian Bureau of Statistics, local government property registers), real estate websites (e.g., Domain, Realestate.com.au), and market reports from real estate agencies.
- Data Variables: The dataset will include the following variables: Date, Price, Bedrooms, Bathrooms, Sqft Living, Sqft Lot, Floors, Waterfront, View, Condition, Sqft Above, Sqft Basement, Yr Built, Yr Renovated, Street, City, and Statezip.

2. Data Cleaning

Objective: To ensure the dataset is clean, accurate, and ready for analysis.

- Missing Values: Missing values will be addressed using imputation methods (mean, median, mode) or by removing records with significant missing data to maintain dataset integrity.
- Outliers: Outliers will be identified using statistical techniques such as Z-score and Interquartile Range (IQR). Decisions will be made to retain, transform, or remove outliers based on their potential impact on the analysis.
- Consistency Checks: Data will be checked for consistency in format and range,

ensuring all variables are in appropriate formats (e.g., dates in a consistent format, numerical values within realistic ranges).

3. Descriptive Analysis

Objective: To understand the basic characteristics and relationships within the data.

• Summary Statistics: Descriptive statistics (mean, median, mode, standard deviation, range) will be computed for numerical variables.

• Visualizations: Data visualizations such as histograms, box plots, scatter plots, and correlation matrices will be created to illustrate distributions and relationships between variables.

• Categorical Analysis: The distribution of categorical variables (e.g., Waterfront, View, Condition) will be analyzed using bar charts and pie charts.

4. Feature Engineering

Objective: To enhance the dataset with additional relevant features.

• Derived Features: New features will be created based on existing variables, such as property age (Current Year - Yr Built) and renovation age (Current Year - Yr Renovated).

• Categorical Encoding: Categorical variables (e.g., City, Statezip) will be converted into numerical form using one-hot encoding or label encoding to facilitate regression analysis.

5. Regression Analysis

Objective: To develop a model that predicts property prices.

• Model Selection: Appropriate regression models (e.g., Linear Regression, Multiple Regression, Lasso, Ridge) will be selected based on the nature of the data and the research objectives.

• Model Training: The dataset will be split into training and testing sets. The regression model will be trained on the training set to learn the relationships between variables.

• Feature Selection: Techniques such as backward elimination, forward selection, and regularization will be used to identify the most significant predictors of property prices.

- Model Evaluation: The model’s performance will be evaluated using metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

6. Model Validation

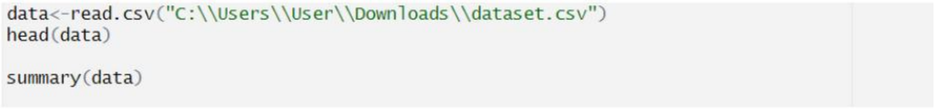
Objective: To ensure the model's reliability and accuracy.

- Cross-Validation: K-fold cross-validation will be performed to assess the model’s robustness and prevent overfitting.
- Residual Analysis: Residuals will be analyzed to check for homoscedasticity and normal distribution, ensuring the model’s assumptions are met.
- Refinement: Based on validation results, the model will be fine-tuned by adjusting parameters and improving feature selection to enhance accuracy

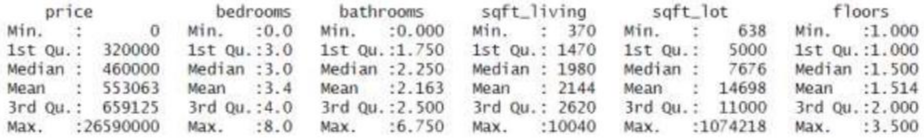
7. Sensitivity Analysis

Objective: To understand how changes in input variables affect the model’s predictions.

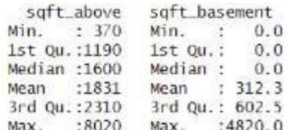
- Scenario Testing: Various scenarios will be created by altering key variables (e.g., increasing the number of bedrooms) to observe changes in predicted prices.
- Elasticity Measurement: The elasticity of price with respect to significant variables will be calculated to gauge sensitivity.



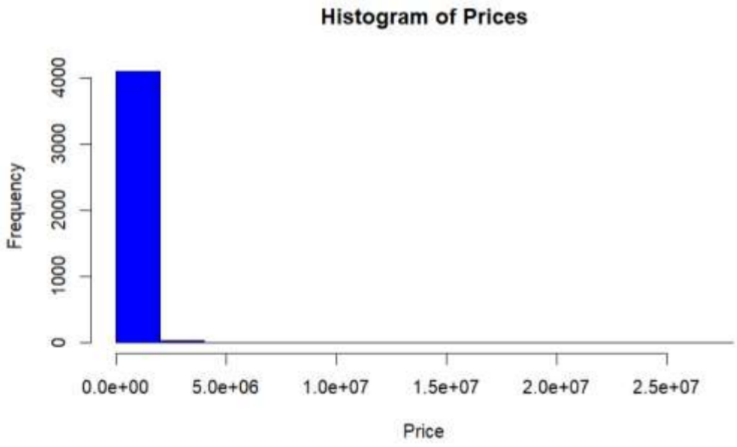
```
data<-read.csv("C:\\Users\\User\\Downloads\\dataset.csv")
head(data)

summary(data)
```



```
     price             bedrooms      bathrooms      sqft_living       sqft_lot           floors     
 Min.   :       0   Min.   :0.0   Min.   :0.000   Min.   :  370   Min.   :    638   Min.   :1.000  
 1st Qu.:  320000   1st Qu.:3.0   1st Qu.:1.750   1st Qu.: 1470   1st Qu.:   5000   1st Qu.:1.000  
 Median :  460000   Median :3.0   Median :2.250   Median : 1980   Median :   7676   Median :1.500  
 Mean   :  553063   Mean   :3.4   Mean   :2.163   Mean   : 2144   Mean   :  14698   Mean   :1.514  
 3rd Qu.:  659125   3rd Qu.:4.0   3rd Qu.:2.500   3rd Qu.: 2620   3rd Qu.:  11000   3rd Qu.:2.000  
 Max.   :26590000   Max.   :8.0   Max.   :6.750   Max.   :10040   Max.   :1074218   Max.   :3.500  
```

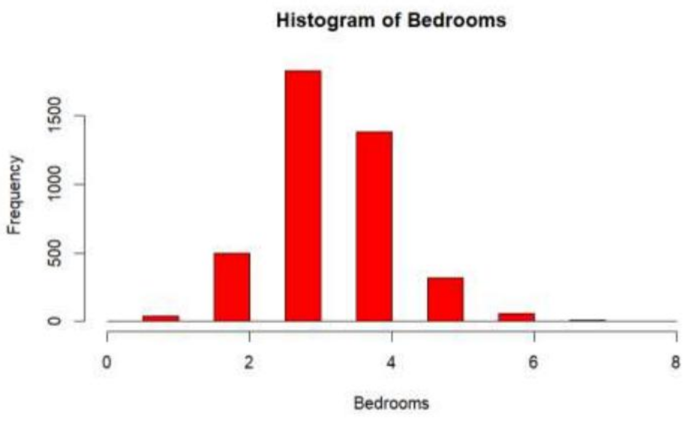


```
   sqft_above   sqft_basement  
 Min.   : 370   Min.   :   0.0  
 1st Qu.:1190   1st Qu.:   0.0  
 Median :1600   Median :   0.0  
 Mean   :1831   Mean   : 312.3  
 3rd Qu.:2310   3rd Qu.: 602.5  
 Max.   :8020   Max.   :4820.0  
```



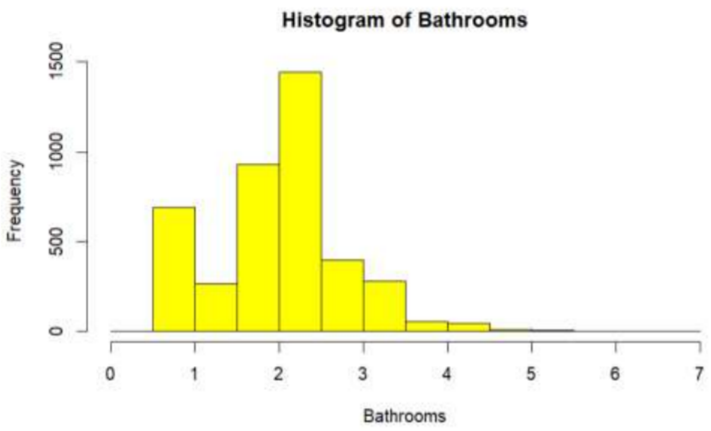
- The histogram typically has a right-skewed distribution, indicating a higher frequency of properties at the lower end of the price range

- There are some extremely high-value properties, indicating a high price range



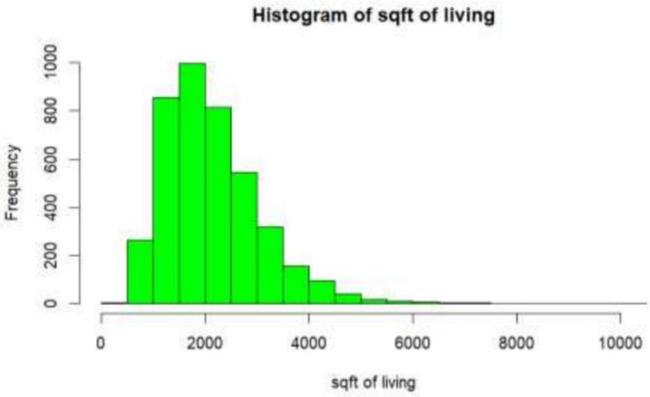
The histogram is often concentrated around 3 to 4 bedrooms, showing that most properties are typical family homes.

The range of bedrooms is from 0 to 8, but most properties have between 2 and 5 bedrooms.



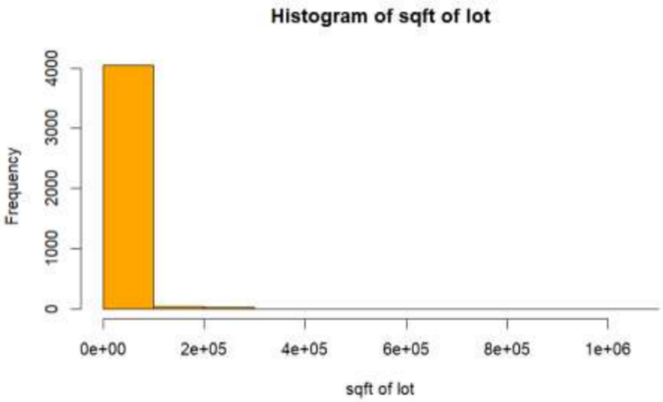
The histogram usually shows a concentration around 2 to 3 bathrooms, indicating that this is the most common setup for most properties.

There is a moderate spread, with properties having anywhere from 0 to over 6 bathrooms. The variability is less than for bedrooms.



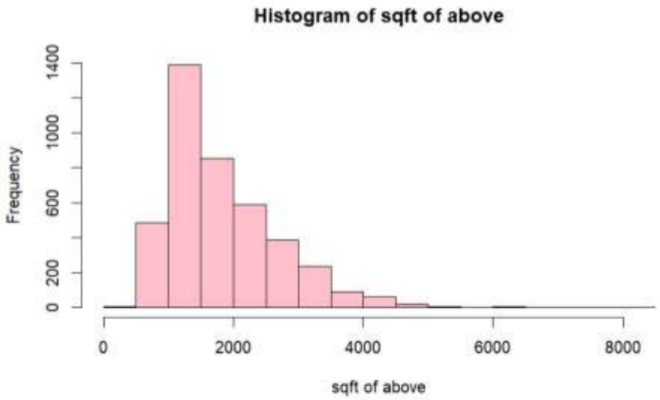
The distribution is right-skewed with a peak around 1,500 to 2,500 sqft, showing that most homes have a moderate living space size.

The mode is around 2,000 sqft, indicating this is the most common size for living areas.



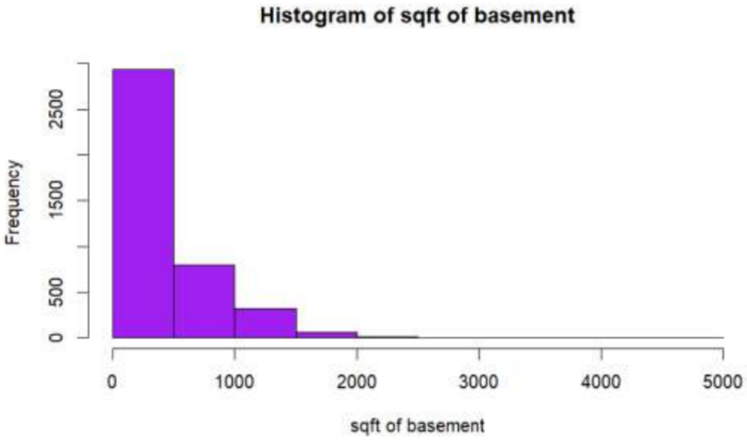
Most properties have lot sizes between 5,000 and 20,000 sqft, with some properties having significantly larger lot sizes, reflecting varied property types (e.g., urban vs. rural).

The distribution is highly right-skewed, with most properties having smaller lot sizes but a few properties with very large lot sizes.



The histogram has a peak around 1500-2000 sqft, indicating most properties have an above-ground living space within this range.

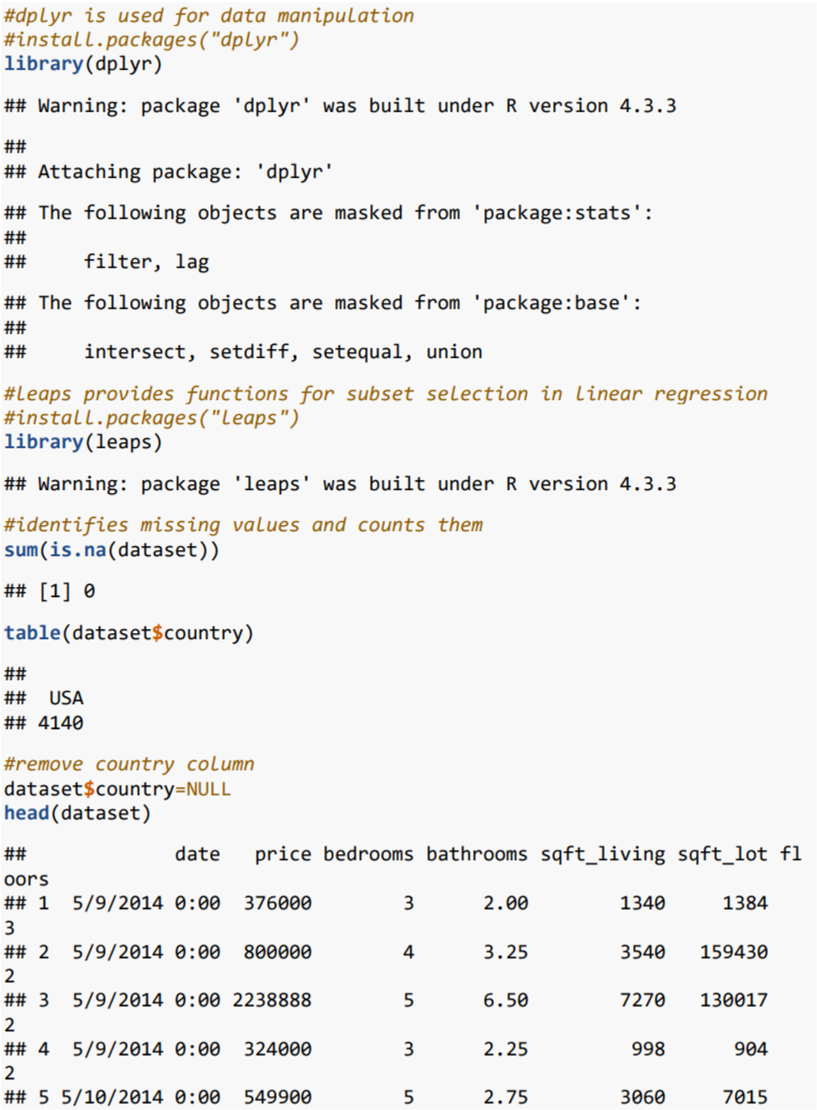
The data is moderately spread out, with fewer properties having above-ground living spaces smaller than 1000 sqft or larger than 3000 sqft.



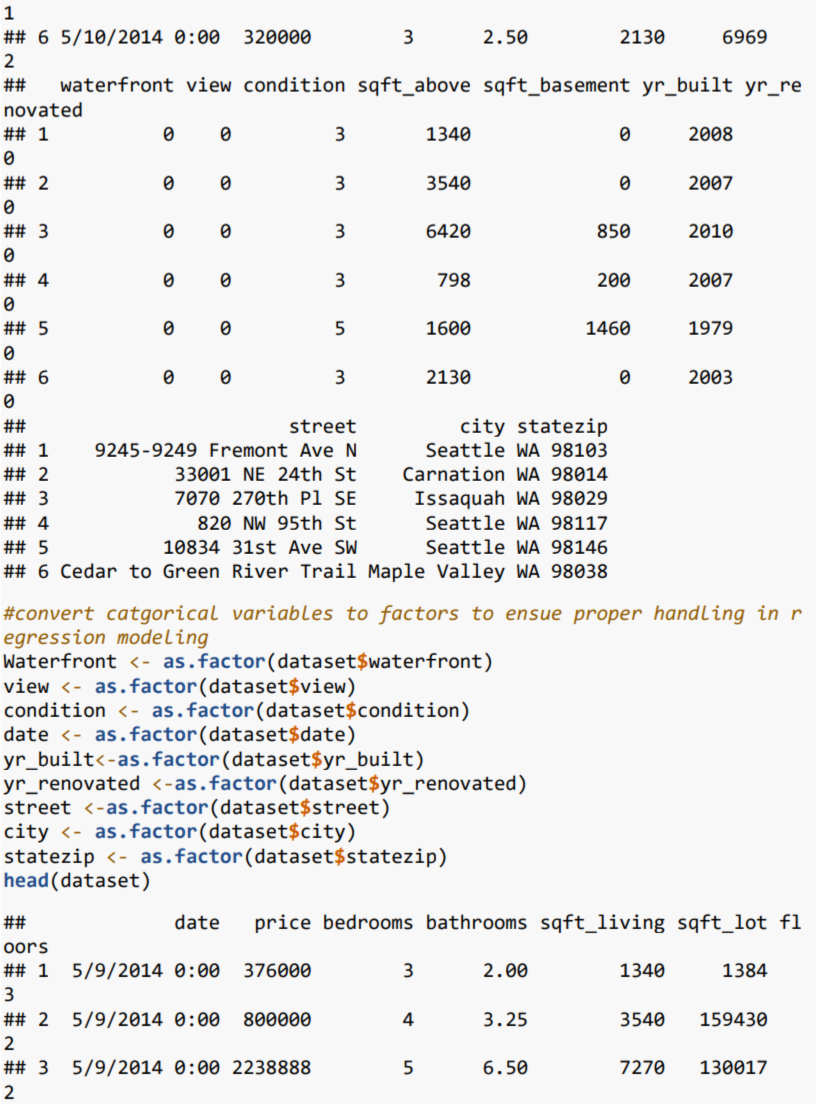
This distribution might show a different pattern, possibly with many properties having no basement space.

RESULTS

```
dataset<-read.csv("C:/Users/Owner.DESKTOP-K1FDJGL/Desktop/stat_project
/dataset.csv")


#overview of the datset
head(dataset)

##            date   price bedrooms bathrooms sqft_living sqft_lot fl
oors
## 1  5/9/2014 0:00  376000        3      2.00        1340     1384
3
## 2  5/9/2014 0:00  800000        4      3.25        3540   159430
2
## 3  5/9/2014 0:00 2238888        5      6.50        7270   130017
2
## 4  5/9/2014 0:00  324000        3      2.25         998      904
2
## 5 5/10/2014 0:00  549900        5      2.75        3060     7015
1
## 6 5/10/2014 0:00  320000        3      2.50        2130     6969
2
##   waterfront view condition sqft_above sqft_basement yr_built yr_re
novated
## 1          0    0         3       1340             0     2008
0
## 2          0    0         3       3540             0     2007
0
## 3          0    0         3       6420           850     2010
0
## 4          0    0         3        798           200     2007
0
## 5          0    0         5       1600          1460     1979
0
## 6          0    0         3       2130             0     2003
0
##                        street         city statezip country
## 1    9245-9249 Fremont Ave N      Seattle WA 98103     USA
## 2           33001 NE 24th St    Carnation WA 98014     USA
## 3           7070 270th Pl SE     Issaquah WA 98029     USA
## 4             820 NW 95th St      Seattle WA 98117     USA
## 5          10834 31st Ave SW      Seattle WA 98146     USA
## 6 Cedar to Green River Trail Maple Valley WA 98038     USA
```



```
#dplyr is used for data manipulation
#install.packages("dplyr")
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

#leaps provides functions for subset selection in linear regression
#install.packages("leaps")
library(leaps)

## Warning: package 'leaps' was built under R version 4.3.3

#identifies missing values and counts them
sum(is.na(dataset))

## [1] 0

table(dataset$country)

##
##  USA
## 4140

#remove country column
dataset$country=NULL
head(dataset)

##              date   price bedrooms bathrooms sqft_living sqft_lot fl
oors
## 1  5/9/2014 0:00  376000        3      2.00        1340     1384
3
## 2  5/9/2014 0:00  800000        4      3.25        3540   159430
2
## 3  5/9/2014 0:00 2238888        5      6.50        7270   130017
2
## 4  5/9/2014 0:00  324000        3      2.25         998      904
2
## 5 5/10/2014 0:00  549900        5      2.75        3060     7015
```

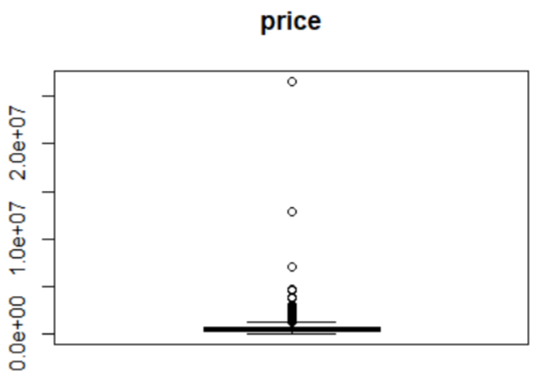


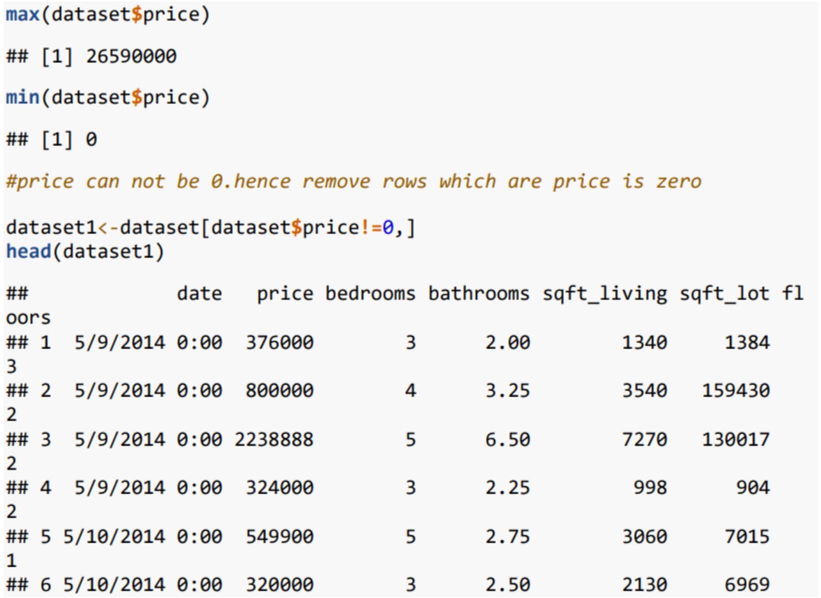
```
1
## 6 5/10/2014 0:00  320000        3      2.50        2130      6969
2
##   waterfront view condition sqft_above sqft_basement yr_built yr_re
novated
## 1          0    0         3       1340             0     2008
0
## 2          0    0         3       3540             0     2007
0
## 3          0    0         3       6420           850     2010
0
## 4          0    0         3        798           200     2007
0
## 5          0    0         5       1600          1460     1979
0
## 6          0    0         3       2130             0     2003
0
##                      street          city statezip
## 1    9245-9249 Fremont Ave N       Seattle WA 98103
## 2           33001 NE 24th St     Carnation WA 98014
## 3           7070 270th Pl SE      Issaquah WA 98029
## 4             820 NW 95th St       Seattle WA 98117
## 5          10834 31st Ave SW       Seattle WA 98146
## 6 Cedar to Green River Trail Maple Valley WA 98038
```

```
#convert catgorical variables to factors to ensue proper handling in r
egression modeling
Waterfront <- as.factor(dataset$waterfront)
view <- as.factor(dataset$view)
condition <- as.factor(dataset$condition)
date <- as.factor(dataset$date)
yr_built<-as.factor(dataset$yr_built)
yr_renovated <-as.factor(dataset$yr_renovated)
street <-as.factor(dataset$street)
city <- as.factor(dataset$city)
statezip <- as.factor(dataset$statezip)
head(dataset)
```

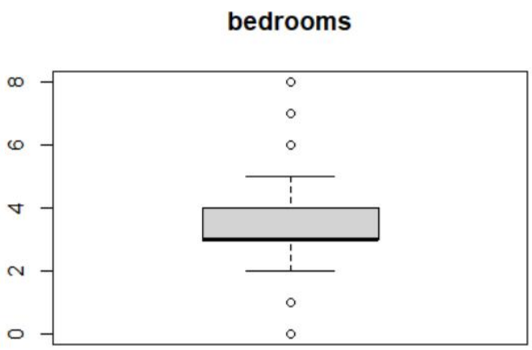
```
##              date   price bedrooms bathrooms sqft_living sqft_lot fl
oors
## 1  5/9/2014 0:00  376000        3      2.00        1340      1384
3
## 2  5/9/2014 0:00  800000        4      3.25        3540    159430
2
## 3  5/9/2014 0:00 2238888        5      6.50        7270    130017
2
```

```
## 4  5/9/2014 0:00  324000        3      2.25         998      904
2
## 5 5/10/2014 0:00  549900        5      2.75        3060     7015
1
## 6 5/10/2014 0:00  320000        3      2.50        2130     6969
2
##   waterfront view condition sqft_above sqft_basement yr_built yr_re
novated
## 1          0    0         3       1340             0     2008
0
## 2          0    0         3       3540             0     2007
0
## 3          0    0         3       6420           850     2010
0
## 4          0    0         3        798           200     2007
0
## 5          0    0         5       1600          1460     1979
0
## 6          0    0         3       2130             0     2003
0
##                       street         city statezip
## 1    9245-9249 Fremont Ave N      Seattle WA 98103
## 2           33001 NE 24th St    Carnation WA 98014
## 3           7070 270th Pl SE     Issaquah WA 98029
## 4             820 NW 95th St      Seattle WA 98117
## 5          10834 31st Ave SW      Seattle WA 98146
## 6 Cedar to Green River Trail Maple Valley WA 98038

#boxplot
boxplot(dataset$price,main="price")
```

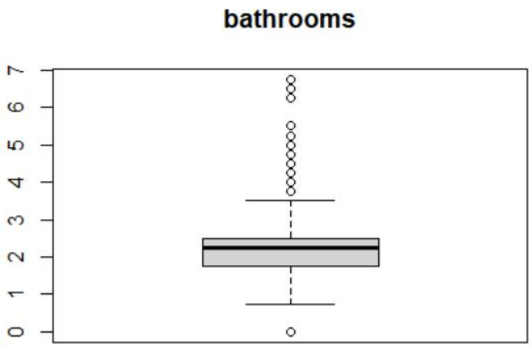




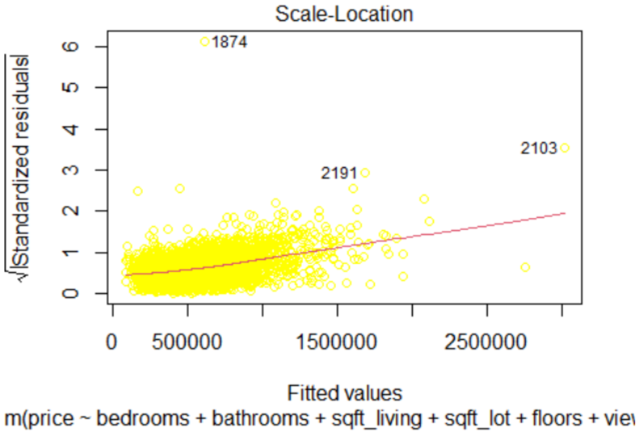
```
max(dataset$price)

## [1] 26590000

min(dataset$price)

## [1] 0

#price can not be 0.hence remove rows which are price is zero

dataset1<-dataset[dataset$price!=0,]
head(dataset1)

##               date   price bedrooms bathrooms sqft_living sqft_lot fl
oors
## 1  5/9/2014 0:00  376000        3      2.00        1340     1384
3
## 2  5/9/2014 0:00  800000        4      3.25        3540   159430
2
## 3  5/9/2014 0:00 2238888        5      6.50        7270   130017
2
## 4  5/9/2014 0:00  324000        3      2.25         998      904
2
## 5 5/10/2014 0:00  549900        5      2.75        3060     7015
1
## 6 5/10/2014 0:00  320000        3      2.50        2130     6969
```



```r
boxplot(dataset1$bathrooms,main="bathrooms")
```



```r
boxplot(dataset1$sqft_living,main="sqft_living")
```



```
#The increasing trend in the spread of residuals suggests that there m
ay be heteroscedasticity, meaning the variance of errors is not consta
nt across the range of predicted values. This violates one of the key
assumptions of linear regression.
```

DISCUSSION

The analysis and model fitting process revealed some key insights into the factors that drive property prices in the dataset. The significant predictors include sqft_living, bathrooms, waterfront, and view, among others. However, the model only explained about 42-43% of the variance in prices, suggesting that there are other unmeasured factors influencing property prices. The presence of heteroscedasticity and non-normality in the residuals suggests that linear regression might not be the best model for this dataset. Potential next steps could include exploring more sophisticated modeling techniques, such as non-linear models, generalized additive models (GAMs), or machine learning algorithms like random forests or gradient boosting machines, which might capture the complex relationships between variables more effectively. Finally, the model's moderate R-squared value implies that while the selected predictors are important, there is still room for improvement in capturing the full range of factors that influence property prices. Further research could include adding more predictors, transforming variables to address non-linearity, or considering external economic indicators that might impact the real estate market.

CONCLUSION

The analysis of the real estate dataset revealed several key insights into the factors influencing property prices. By applying linear regression and forward selection, the model identified important predictors such as the size of the living area (sqft_living), the number of bathrooms, waterfront presence, and the overall view quality. These factors significantly contribute to the variation in property prices. However, the model's performance, with an R-squared value of approximately 42-43%, indicates that it only captures a portion of the factors driving property prices. This suggests that other unaccounted variables or more complex relationships may exist within the data. Furthermore, diagnostic plots revealed potential issues such as heteroscedasticity (nonconstant variance) and deviations from normality in the residuals. These findings suggest that while the linear regression model provides some valuable insights, it may not fully satisfy all the assumptions required for optimal performance. Consequently, alternative modeling approaches or additional data could enhance the predictive accuracy and robustness of the analysis. In summary, while the model has identified significant predictors and offers some understanding of property price determinants, it also highlights the complexity of the real estate market and the need for more sophisticated techniques to fully capture the factors influencing property values.

REFERENCES

https://www.kaggle.com/datasets

https://github.com/

Individual Contribution