The University of Texas at Austin

Predicting Property Value

SDS 322E

Prof. Arya Farahi

Team 10

Ethan Chen, Justin Chung, Jonathan Debella, Kyle Folker, Joshua Heldt, Austin Yeh, Taehyun Yun

12/06/2022

Introduction:

The goal of our project was to explore the relationships between land size, year, area, and other qualitative and quantitative characteristics of houses with their property values and determine if such correlations exist. We also wanted to use these characteristics to create a model that predicts the value of a property. The motivations behind such analysis were the applicability of such a topic to the economy and an underlying interest in property value fluctuations outside of market health.

Housing markets have experienced considerable dynamic fluctuations over history and have become complex multivariate systems to predict (Duhai et al., 2021). In light of the recent global pandemic, assessing the impacts of Covid-19 on the housing market has been challenging (Nadia et al., 2021), and the impacts of Covid-19 have made the housing market more complex of a system to predict.We believe that specifically focusing on stagnant statistics can give an individual an unbiased understanding of the underlying reasons for property value evaluation. Specifically, we hope to illustrate a relationship between such factors and value which can be held true despite the state of the economy.

In this project we determined which factors have the greatest impact on property value, and which factors can be disregarded. We created and tested models to determine a correlation and determined which model and which factors yielded the most accurate and understandable results. We also hope to gauge an understanding of the fluctuations behind the housing market, and the complex system behind property value prediction.

Data:

The data we utilized for this project was extracted from the Melbourne Housing market, which contained exactly 13,580 houses within the market starting from 2016. Such variables within this dataset include: suburb, address, rooms, area, zip code, seller, etc. We decided to utilize such data by downloading and cleaning the dataset, extracting variables that we believed were useful for the models that we were going to create. To begin the cleaning process of our data set, we first omitted all instances of “NA”, as well as removing variables we deemed insignificant, such as address, postcode, etc. We also omitted all entries that had a Z-score of above 5 as we believed that they were outliers and would skew our models.

We also filtered out data by reducing redundancy within the dataset, specifically longitude and latitude and postcode, as they seem to be representing the same/ similar entities. Lastly, in order to make our dataset more uniform for the models we were going to run, variables were reformatted. Such variables like date were all reformatted from mm/dd/yy into a single numeric quantity, and categorical variables were encoded as dummy variables. The resulting dataset that was untimely used included the following variables: suburb, address, price, seller, region, year, and others. There were 21 variables that were deemed useful and were cleaned.

Exploratory Analysis:

Price Distribution

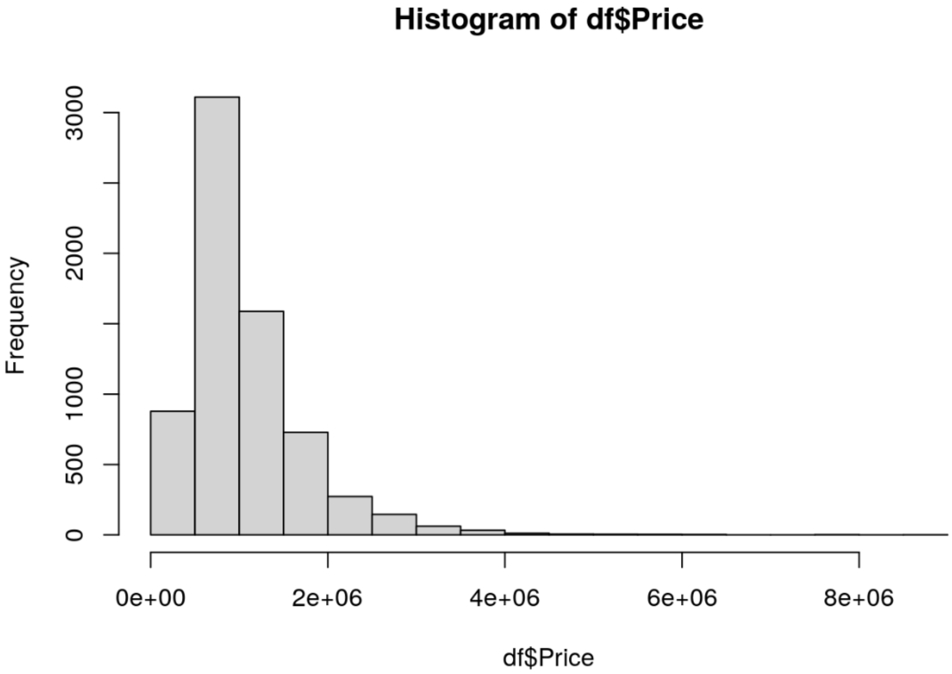


Figure1: Distribution of Price of Properties

In *Figure 1*, we visualized the distribution of prices for the properties in our dataset by plotting a histogram. It appeared that the distribution of property prices was not normal and heavily skewed to the right even despite removing the outliers in our dataset (by removing prices with a Z-score of 5 or greater). Although the dataset did not possess the characteristics of an ideal sample to analyze, we reasoned that the distribution of prices for any property market would be skewed slightly to the right. It would make sense that there would be a greater compact of houses with lower to mid-range prices with a few homes that are in the high-range of prices for any given community. For these reasons, we believed these prices were not outliers and reflected an accurate trend of the property market that determined prices.

Analysis of Correlation Matrix

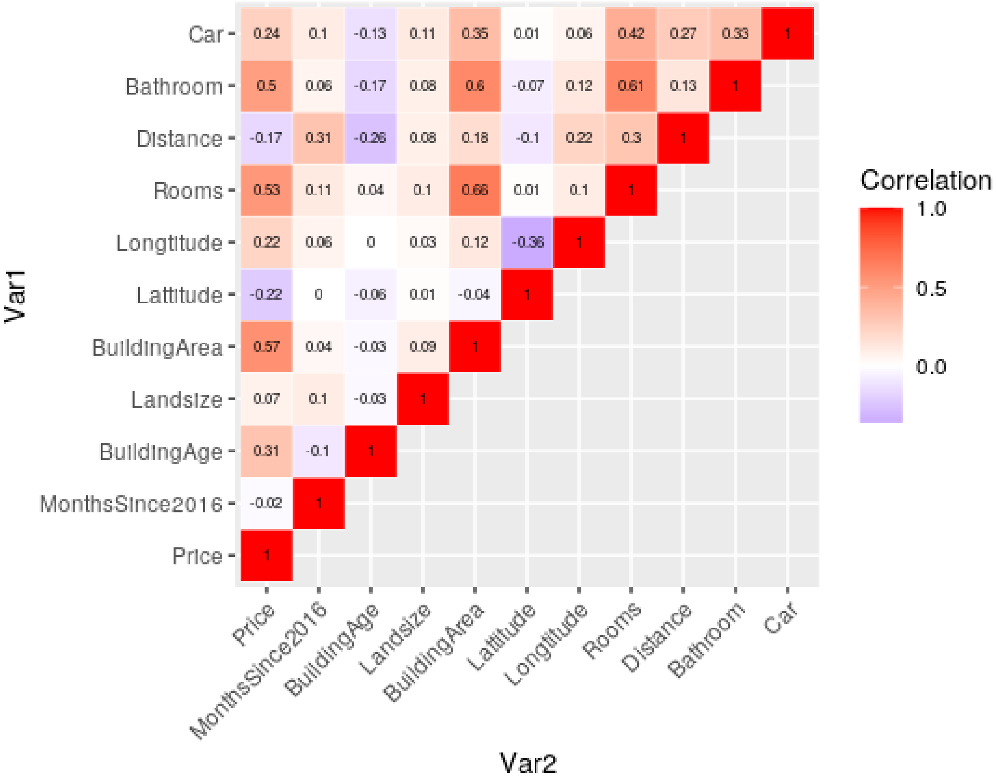


Figure 2: Correlation Matrix

In *Figure 2*, we visualized the correlations between the quantitative characteristics of the properties in our data set with the prices by creating a correlation matrix. We found that the quantitative variables were most correlated to the prices of the properties. Building area, room count, and bathroom count had the highest correlations to price with |correlation values| greater than 0.49. Land size and months since the property was purchased in 2016 had the smallest correlations to price with |correlation values| less than 0.1.

Relationship Between BuildingArea and Distance with Price

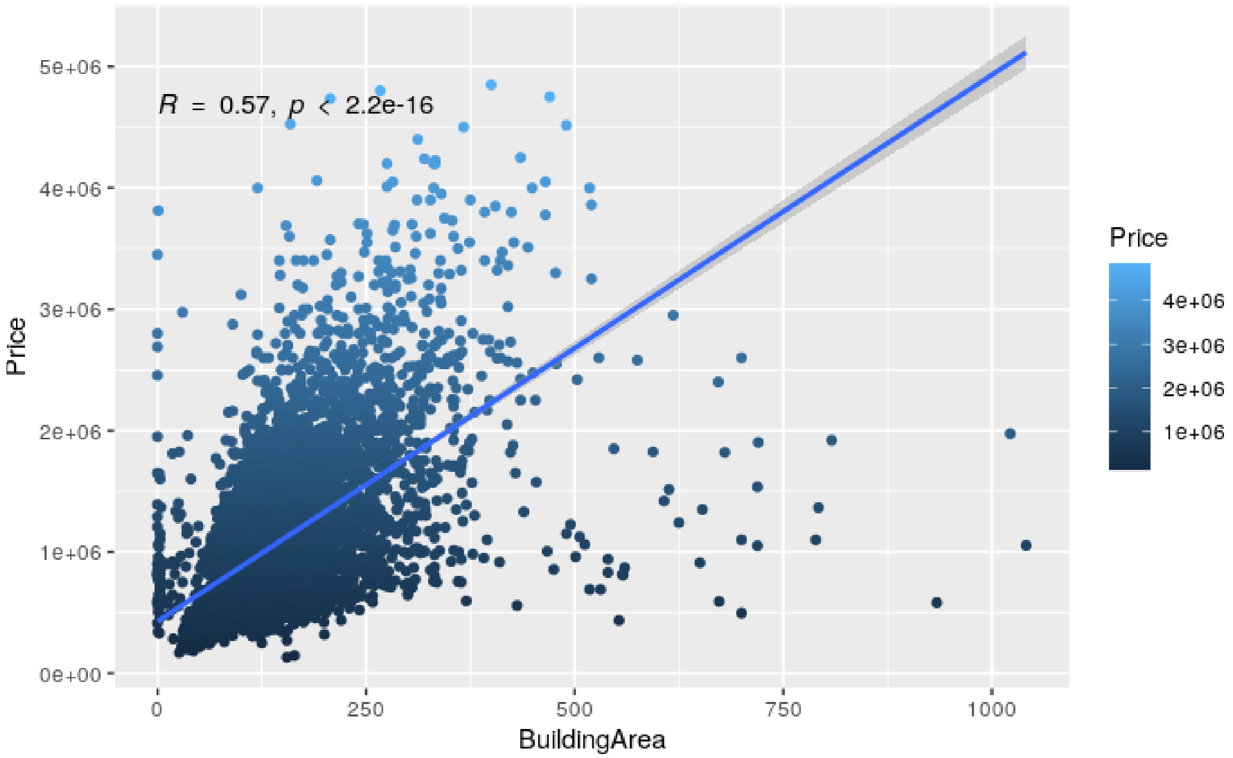


Figure 3: Relationship Between BuildingArea and Price

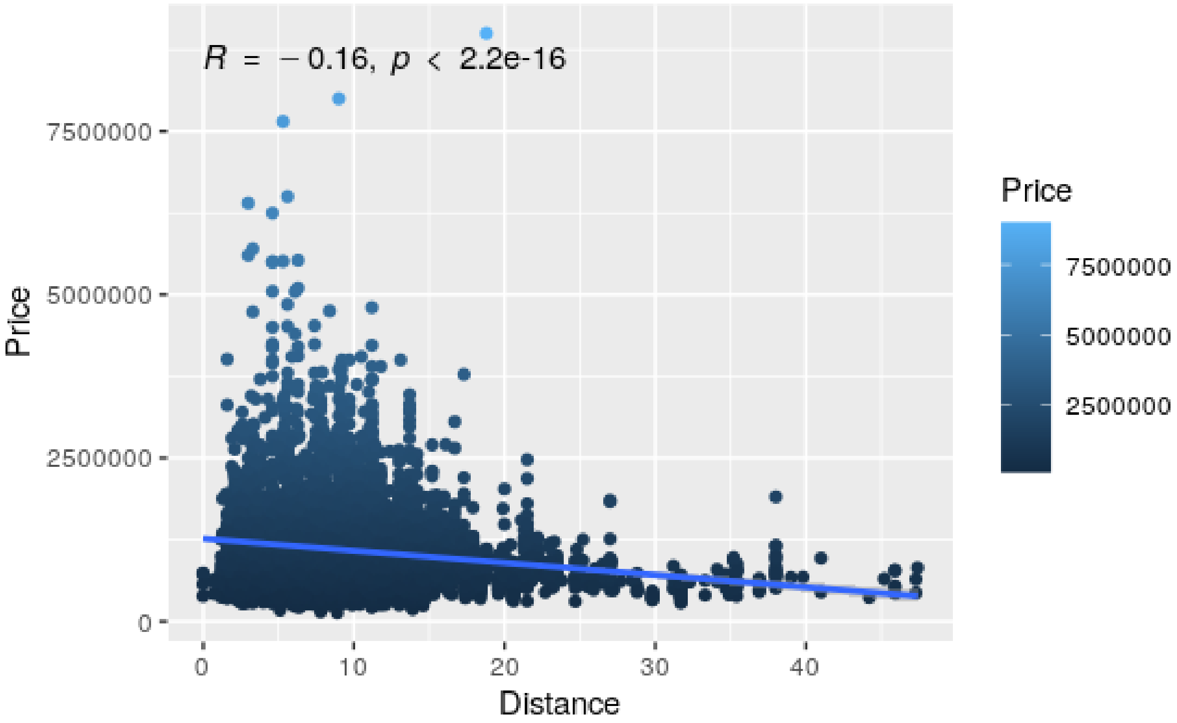


Figure 4: Relationship Between Distance and Price

In *Figure 3*, we observed that the price of property increases as the building area of the property increases. This can be explained with its correlation coefficient of R = 0.57, and this relationship is intuitive since a property with larger building area would have expended more costs in its manufacturing process.

In *Figure 4*, we observed that the price of property slightly decreases as the distance of the property from the central business district increases. This can be explained with its correlation coefficient of R = -0.16, and this relationship is intuitive since a property that is further away from the central district of economic activity would have a lower value.

Hypothesis

In *Figures 3* and *4*, we observed noticeable relationships between the quantitative characteristics of a property with price. However, we did not observe or explore any noticeable relationships between the qualitative characteristics of the property in our data set with price. Hence we created two hypotheses on two qualitative characteristics we expected would have a significant impact on property price.

- **Hypothesis 1**: The number of rooms a property has will impact its price
  - H_o: The mean prices of properties are not different based on the number of rooms
  - H_a:The mean prices of properties are different based on number of rooms

- **Hypothesis 2**: The region a property is located in will impact its price
  - H_o: The mean prices of properties are not different based on regions
  - H_a: The mean prices of properties are different based on regions

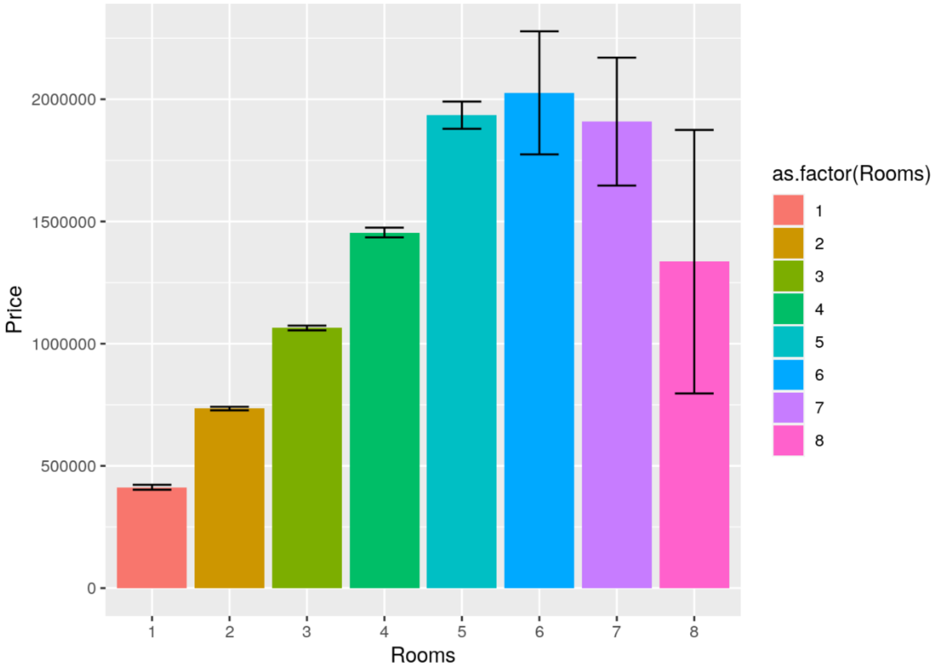


Figure 5: Visualization on One-Factor ANOVA between Room and Price

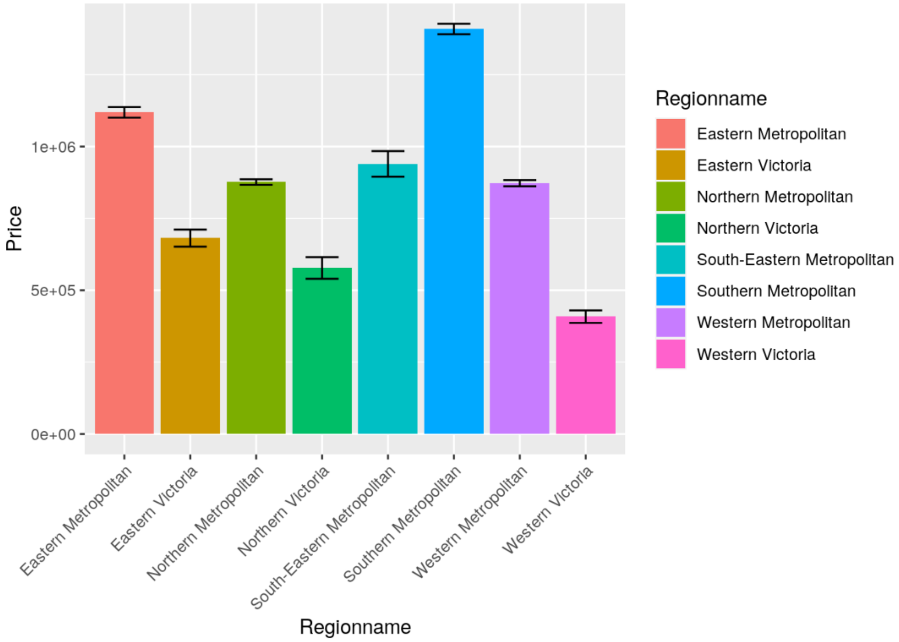
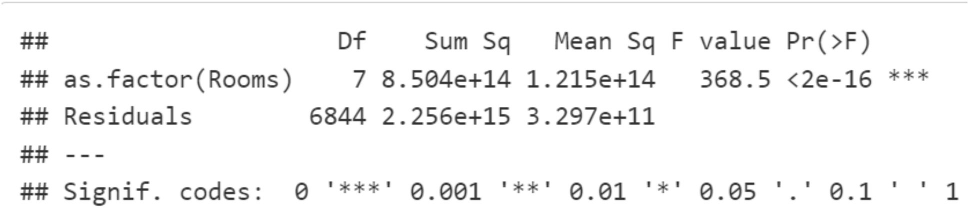


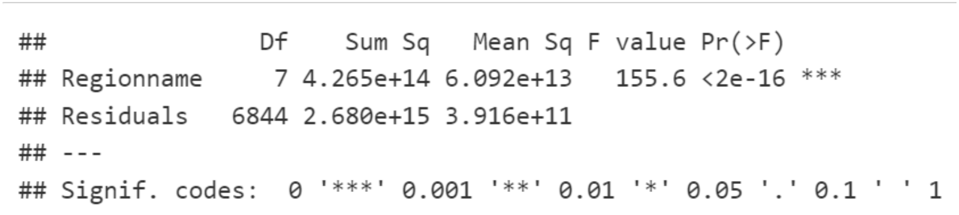
Figure 6: Visualization of One-Factor ANOVA between Regionname and Price

To have statistically significant conclusions on these two hypotheses, we ran two respective one-factor ANOVA tests on room count and region name with property price. The ANOVA tests were visualized in *Figure 5* for room count and *Figure 6* for region name. The summary of the test statistics can be seen below:



```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)    
## as.factor(Rooms)    7 8.504e+14 1.215e+14   368.5 <2e-16 ***
## Residuals        6844 2.256e+15 3.297e+11                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 1: Summary of ANOVA Test Statistics for Room Count



```
##               Df    Sum Sq   Mean Sq F value Pr(>F)    
## Regionname     7 4.265e+14 6.092e+13   155.6 <2e-16 ***
## Residuals   6844 2.680e+15 3.916e+11                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2: Summary of ANOVA Test Statistics for Region Name

As seen in *Table 1*, for the ANOVA test conducted between the different number of room counts and price, we saw that the p-value was less than 0.001. Therefore we rejected our null hypothesis in *hypothesis 1* that the mean prices of properties are not different based on the number of rooms. Furthermore, we had significant evidence to conclude that the group mean price of properties are different based on different numbers of rooms. Therefore, we included the number of rooms in our model as it can significantly impact the property price prediction.

As seen in *table 2*, for the ANOVA test conducted between the different regions and price, we saw that the p-value was less than 0.001. Therefore we rejected our null hypothesis in *hypothesis 2* that the mean prices of properties are not different based on the region the property is located. Furthermore, we had significant evidence to conclude that the group mean price of properties are different based on different regions. Therefore, we included the region names in our model as it can significantly impact the property price prediction.

For our third hypothesis, we went back to the relationship between the distance of a property from the central business district area and price explored in *Figure 4*. As mentioned earlier, we observed that there was a relatively small correlation between distance and price with a correlation coefficient of r = |0.16|. We wanted to further analyze why a property's distance from the central business district would not affect price as much as we anticipated. In *hypothesis 2,* we found that there was statistically significant evidence to conclude that the group mean price of properties are different based on different regions. Hence, we were led to believe that the region a property is located in may affect the relationship distance has with price.

- **Hypothesis 3**: The correlation between Distance and Price depends on Region

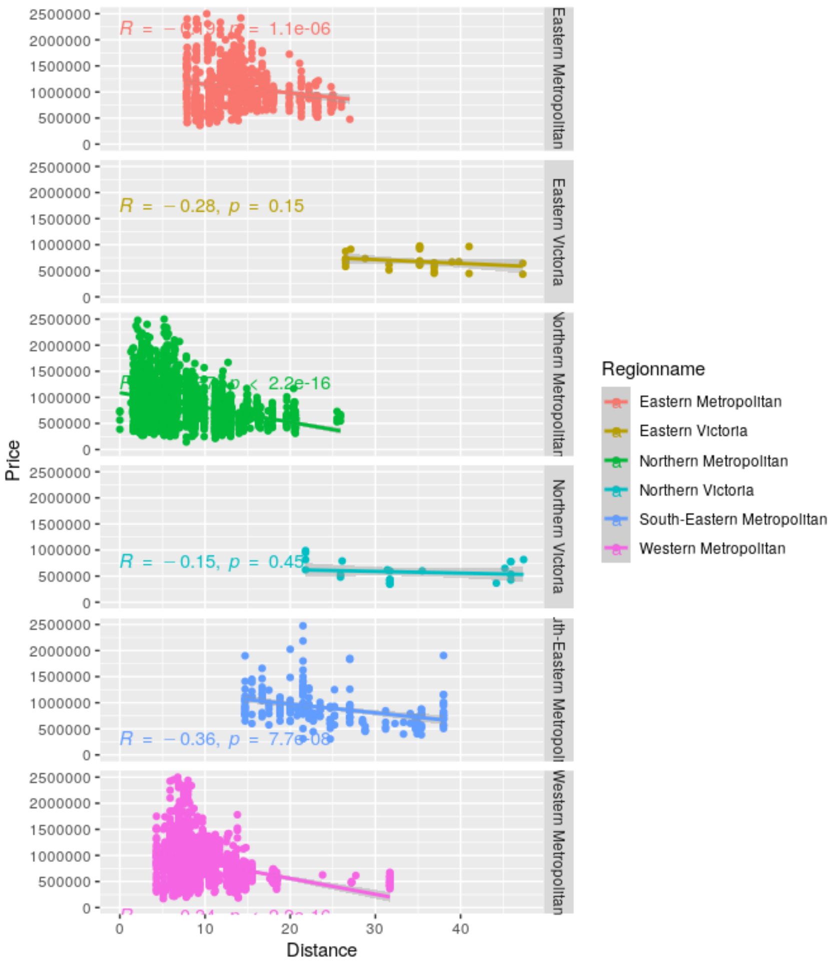


Figure 7: Relationship Between Price and Distance Based on Regionname

In *Figure 7*, we visualized the relationship between distance and price based on the different regions the properties in our data set were located in. In this graphic, we found that there appeared to be no correlation between distance and price in Eastern and Northern Victoria whereas in the other regions, there seemed to be a correlation between distance and price.

With that, we were able to conclude that the correlational relationship between distance and price is affected by the region a property is located in. We found this interesting as it may imply that based on the region an individual chooses to live in, their reason to travel to the central business district may be affected.

For our fourth hypothesis, we went back to the distribution of property prices in our data set that was seen in *Figure 4*. As mentioned earlier, we observed that the distribution of prices appeared to be skewed to the left and not normal. We wanted to further analyze this distribution by logarithmically transforming our prices.

- **Hypothesis 4**: Price is not normally distributed

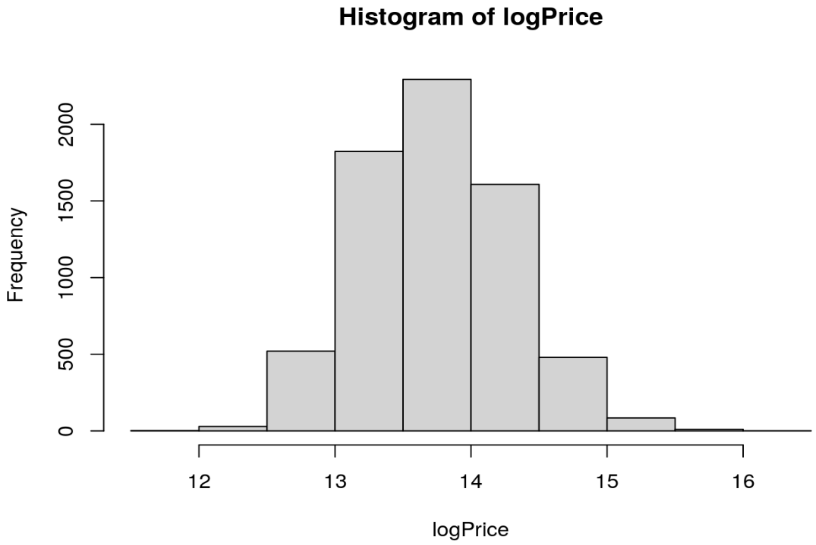


Figure 8: Distribution of Price after Logarithmic Transformation

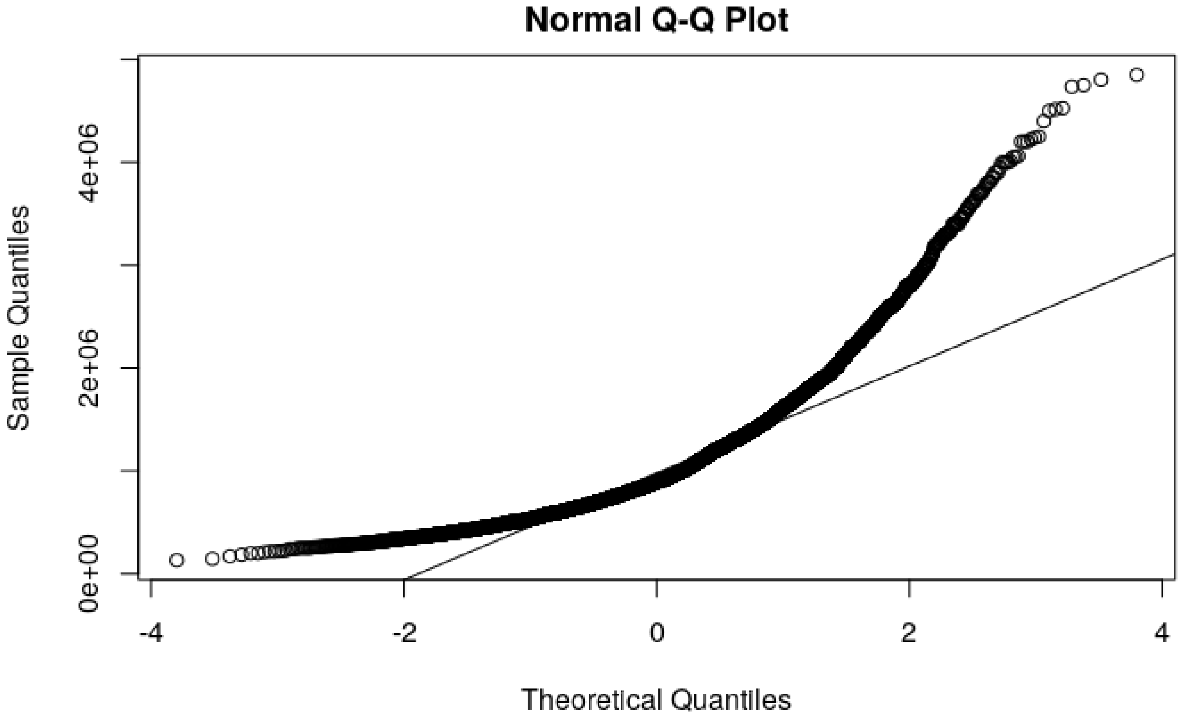


Figure 9: Normal Q-Q Plot of Price before Log Transform

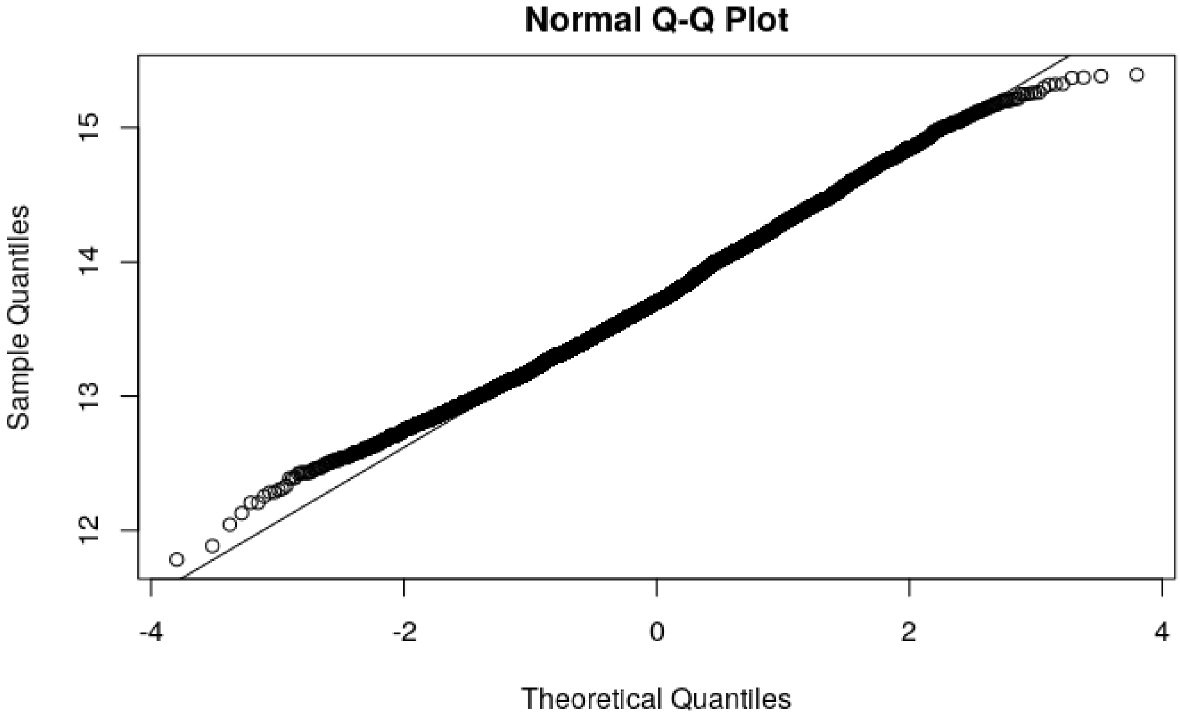


Figure 10: Normal Q-Q Plot of Price After Log Transform

In *Figure 1* and *Figure 9,* the theoretical and sample quantiles of the price distribution show that the price is not normally distributed. This would cause complicated problems in our model, resulting in inaccurate prediction, so we need to transform it. After logarithmically transforming the prices, we visualized the distribution of prices post transformation in a histogram as seen in *Figure 8* and *Figure 10*. In the visualizations, we observed that the post transformation price has a bell-shaped histogram and matching theoretical and sample quantiles, indicating normal distribution, This gave us significant evidence to believe that the prices in our data set was not normally distributed as we supposed, and that we should log transform the prices for a better prediction in our model.

Modeling:

We believed a regression analysis would be appropriate as one of the goals of this project was to forecast the value of property based on its relationships with multiple independent variable characteristics of a property. For our regression analysis, we utilized the RandomForestRegressor and LinearRegression models from the Scikit-learn package to predict the value of property by using the price of the property in our train set.

	Random Forest Regression	Linear Regression
MAE	$148,529.94	$207,100.19
R^2	0.8805	0.7721

****MAE = Mean Average Error**
**** R^2 = Variance**

Table 3: Summary of Analytical Statistics for Models

The linear regression model fit our prices to linear trends based on the relationships it had with the multiple variables we included in our study. These relationships would be averaged down to a single linear regression that would best forecast the price of the property.

The random forest regression grows decision trees to search for the most ideal feature amongst a random subset of features instead of looking for the most important feature. Thus this model may better predict prices than the linear regression model, but it may be more prone to overfitting more easily than the linear regression model.

We utilized mean average error (**MAE**) and variance (**R^2**) as analytical statistics to determine the accuracy of the models we created. As seen in *Table 3*, we found that the Random Forest Regression had a mean average error of **MAE = $148,529.94** and a variance of **R^2 = 0.8805.** We also found that the Linear Regression had a mean average error of **MAE = $207,100.19** and a variance of **R^2 = 0.7721.**

Discussion:

Model Analysis

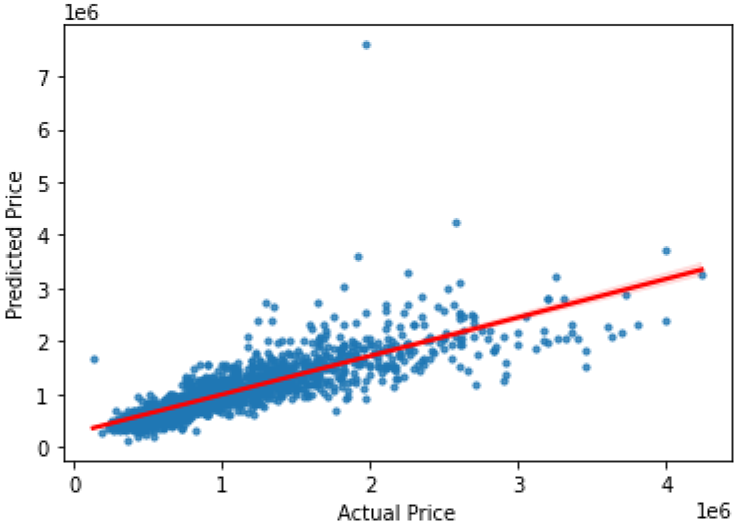


Figure 11: Random Forest Regression Diagnosis Plot

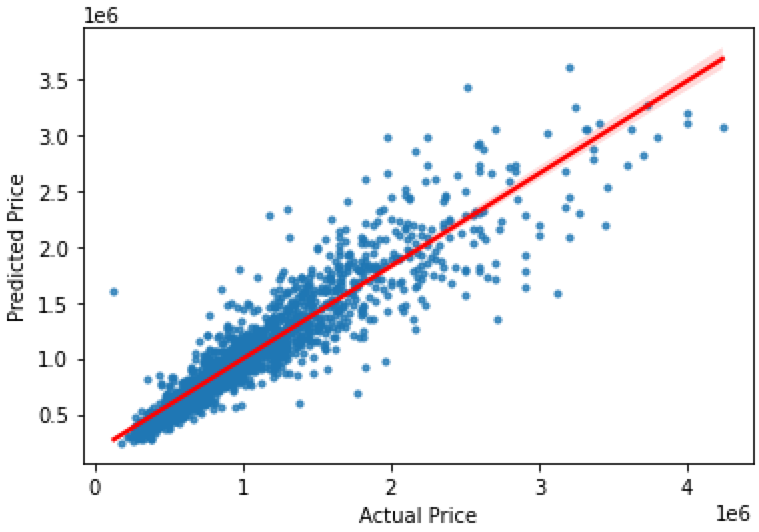


Figure 12: Linear Regression Diagnosis Plot

Figures 11 and 12 were visualizations of the Random Forest Regression and Linear Regression model we computed and how well the models predicted prices of properties based on the characteristic of the property that we included. We observed that the price points were better fit around the regression line created by the Random Forest Regression model than the Linear Regression model. We also saw a more positive trend with the Linear Regression than the Random Forest Regression.

Hence we were led to assume that the Random Forest Regression model was a better forecaster of property value than was the Linear Regression model we generated. The summary statistics confirmed this assumption as seen in *Table 3*, the Random Forest Regression model had a significantly lower mean average error and a higher variance R^2 scores. This led us to conclude that the Random Forest Regression model had lower margins of error in the prices they predicted and the model explained the variance of the predicted prices much better than the Linear Regression model. Furthermore, we found that the Random Forest Regression model is a much better forecaster of property value than the Linear Regression model.

Limitations

- Went from 13580 entries to 6830 entries as a result of omitting NAs
- Some of the nonessential variables may have had value
- Data represents the housing market of Melbourne and results may be limited to the conditions in Melbourne.
- Possibility of overfitting model from not properly cleaning out variables that are insignificant to predicting price

Conclusion:

Findings

- Hypothesis I - The number of rooms statistically significantly impacts the property price prediction
- Hypothesis II - The region a property is located statistically significantly impacts the property price prediction
- Hypothesis III - The value of property is not normally distributed in property markets
- Hypothesis IV - The distance from the central business district has little-to-no effect on predicting the value of properties depending on what region they are in
- Modeling - Random Forest Regression model can forecast the value of a property within a margin of error of $148,529.94.

Future Iterations and Moving Forward

During the data cleaning portion of our project, we dropped 6,750 entries from our data set because they were NA. We believe these values could have significant impact on the model and believe for future iterations of this project, we can find an alternative to dropping the 6,750 entries lost from omitting NAs. Perhaps we can encode them as dummy variables or other statistical substitutes that would enable us to use these NA values.

We also believe that the results and findings of our project were not generalizable to global property markets since our data was only on the Melbourne housing market. For future iterations of this project, we could include data sets from different cities and countries to make the analysis more generalizable to a global property market.

We believe that our project is on the right track towards finding a model that predicts the value of properties in the market. The property market is a complex multivariate system with many characteristics that will impact the price of a property. Moving forward, we can improve our model by finding more characteristics that will better improve its forecasting abilities.

Acknowledgement:

- Austin Yeh: data cleaning, modeling, and coding of such dataset, hypothesis testing, documenting and reporting.
- Taehyun Yun: Modeling, heavy coding, documentation, reporting, debugging and programming and Github for computing and exploratory analysis.
- Ethan Chen: Report writing, data cleaning and exploratory analysis, programming, editing and debugging code, researching market, variables and housing fluctuations
- Jonathan Debella: data cleaning, modeling, and programming. Editing and creating presentation, hypothesis tests, and debugging models and visualizations
- Justin Chung: Cleaning and visualizations of data for hypothesis testing, debugging and reporting code and model errors, interpreting and understanding of the data and variables within such models.
- Kyle Folker: Cleaning and programming dataset, reformatting and remodeling data for models and creating visualizations.
- Joshua Heldt: Creating visualizations, cluster analysis, exploratory data analysis, reporting and documenting, debugging and implementation of cod, models, and visualizations

Name:	Contribution Score
Austin Yeh	100%
Taehyun Yun	100%
Ethan Chen	100%
Jonathan Debella	95%
Justin Chung	100%
Kyle Folker	90%
Joshua Heldt	90%

Bibliography:

1.) Yue, Dahai, and Ninez A Ponce. “Booms and Busts in Housing Market and Health Outcomes for Older Americans.” *Innovation in aging* vol. 5,2 igab012. 10 Apr. 2021, doi:10.1093/geroni/igab012

2.) Balemi, Nadia et al. “COVID-19's impact on real estate markets: review and outlook.” *Financial markets and portfolio management* vol. 35,4 (2021): 495-513. doi:10.1007/s11408-021-00384-6