Project 2: Analysis of Real Estate Prices in Melbourne, Australia

Timothy Tyree, Matthew Sachs, and Christian Schroeder

School of Data Science, University of Virginia

STAT 6021: Linear Models for Data Science

Dr. Jeffrey Woo

May 13, 2021

Table of Contents

Executive Summary

The Australian real estate market has seen amazing growth over the past decade, especially in the city of Melbourne. As prices continue to increase, the glaring question for investors becomes, “what is a good investment and what is overpriced?” To help inform those investors, we aimed to uncover market inefficiencies that could possibly identify a quality investment vs. an overpriced property. To do this, we needed to accurately predict what a property should be listed at, considering several factors in relation to similar properties in the market. But what are those factors?

We aimed to identify the aspects of a property that drive up price the most, whether that be the building area in square meters, the number of bedrooms, or even the suburb or council area the property is located in. According to Opendoor, a digital platform for residential real estate, adding a bathroom to a home can increase the average resale value of the property by 5.7% (though they also note the likelihood of diminishing returns). These types of real estate heuristics are certainly worth verifying for our market.

During our preliminary analysis of the data, we noticed a difference in prices based on who was selling the property, so we also wanted to explore which sellers are more likely to charge more for a home. Prospective homeowners and investors could then go into a price negotiation with a general understanding of the other agent’s price flexibility.

To assist our efforts, as well as the prospective investors in Melbourne’s real estate market, we developed a web application for inspecting the available real estate data, understanding the factors that affect price, and predicting the price of a property with user-generated inputs.

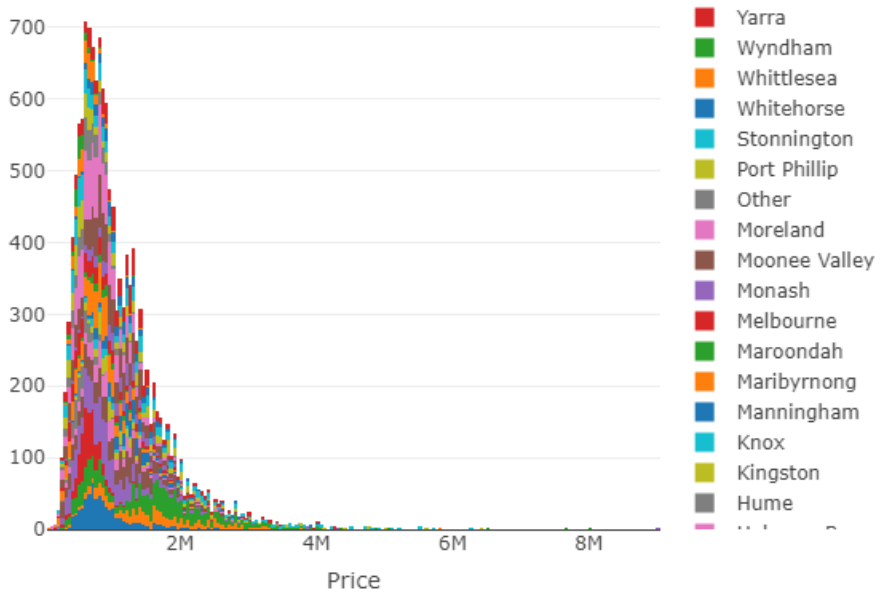
The Data

The data used in this analysis is scraped from multiple Australian real estate websites, and is hosted on Kaggle. The data covers the Melbourne real estate market from January 2016 to November 2017, with a total of 13,581 documented property sales. With this time period in mind, let us operate under the guise that it is currently December 2017; making predictions for the actual current date requires far too much extrapolation. From further investigation into the values that populate this dataset, we determined it was most likely generated through web scraping a select few real estate websites that are popular in Australia,

[www.propertyvalue.com.au] and [www.onthehouse.com.au]. The original data included the following columns, listed by the type of information they provide:

- Property: *Rooms*, *Type*, *Bedroom2*, *Bathroom*, *Car*, *Landsize*, *BuildingArea*, *YearBuilt*
- Sale: *Price*, *Method*, *SellerG*, *Date*
- Region: *Suburb*, *Postcode*, *CouncilArea*, *Regionname*
- Location: *Address*, *Distance*, *Latitude*, *Longitude*

From initial inspection of the data, we can see price follows a right-skewed distribution and has a handful of apparent outliers.

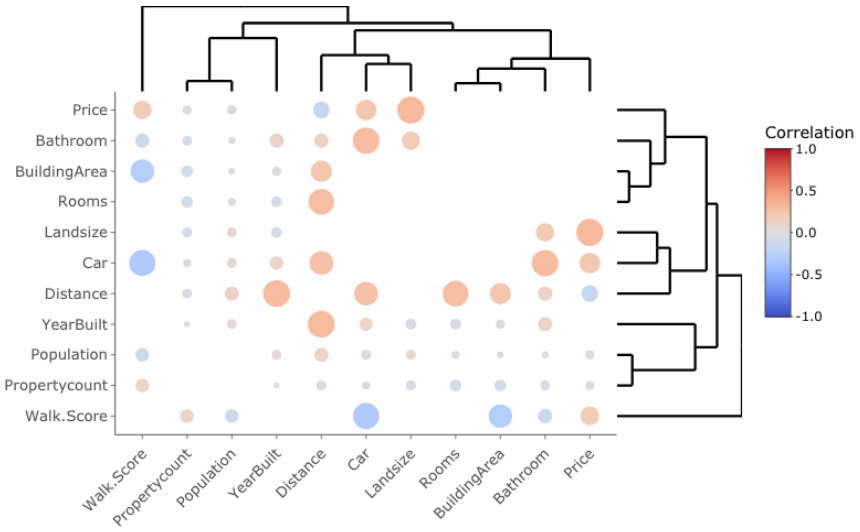


Exploratory Data Analysis

In our exploratory analysis of the dataset we focused on investigating the individual predictors, identifying detriments to the assumptions of the final model, and proper management of outliers and missing values.

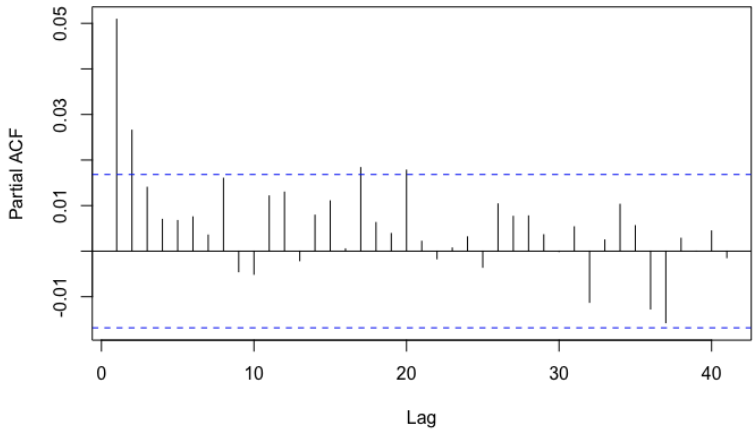
Checking Correlation, Autocorrelation, and Multicollinearity

We began by visualizing the correlation between quantitative predictors and the response. Our correlation plot uses a color bar to visualize the correlation coefficient, and the size of the dot represents the log p-value.



As the plot shows, many of the predictors describing a home's amenities have a correlation coefficient of around 0.3. Intuition tells us that a house with more than 2 bedrooms is likely to have a second bathroom, so this was expected. We feared there would be multicollinearity issues due to this, but the p-values showed us that all of the quantitative predictors are significant in the presence of the others. Furthermore, after fitting a model, the variance inflation factors were all below 10.

Since housing markets fluctuate through time, we needed to check for autocorrelation. We created a partial correlation plot and saw that many of the lags were above the threshold.



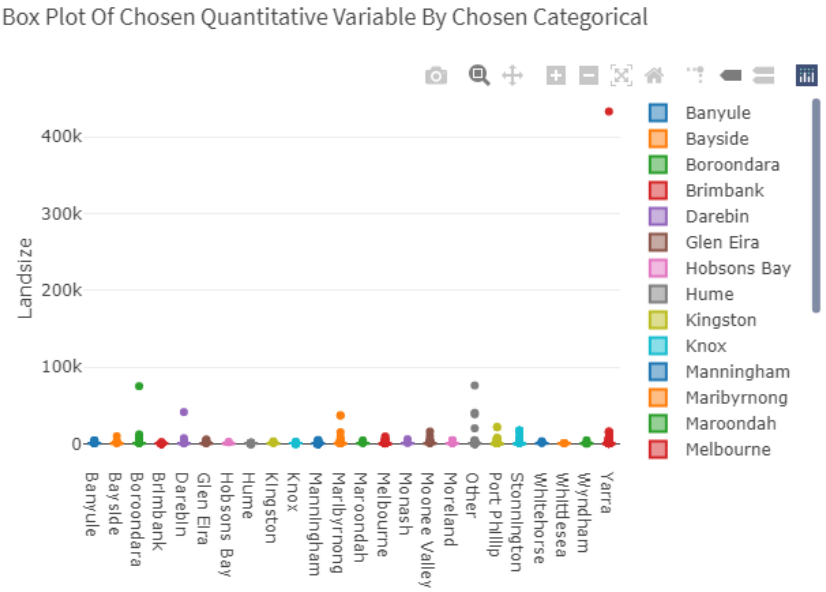
As you can see, the errors follow an AR(20) structure, which would require an autoregressive moving average model (ARMA). Since ARMA is beyond the scope of this class, we decided (and Dr. Woo agreed) that we should not transform the errors. To give the time

columns predictive power, we created *Year* and *Month* columns and treated them as categorical predictors.

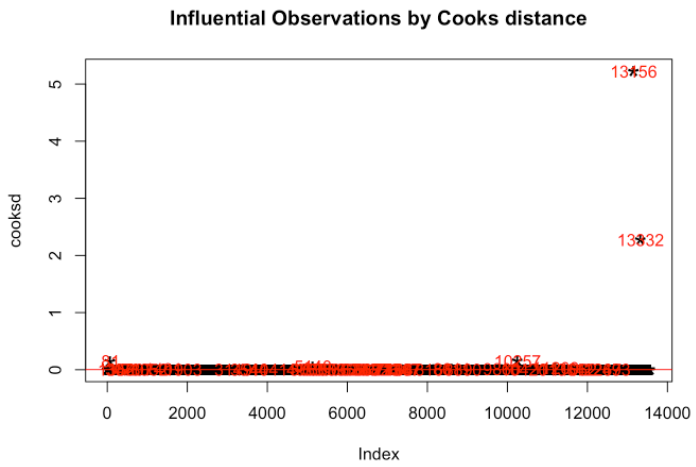
Identifying Outliers, and Missing and Inaccurate Data

An issue we noticed early on in our analysis was the substantial amount of missing values throughout the data. When checking for NA values, we found that there were 6,417 missing values for *BuildingArea*, 5,344 missing values for *YearBuilt*, 1,369 missing values for *CouncilArea*, and 62 missing values for *Car*.

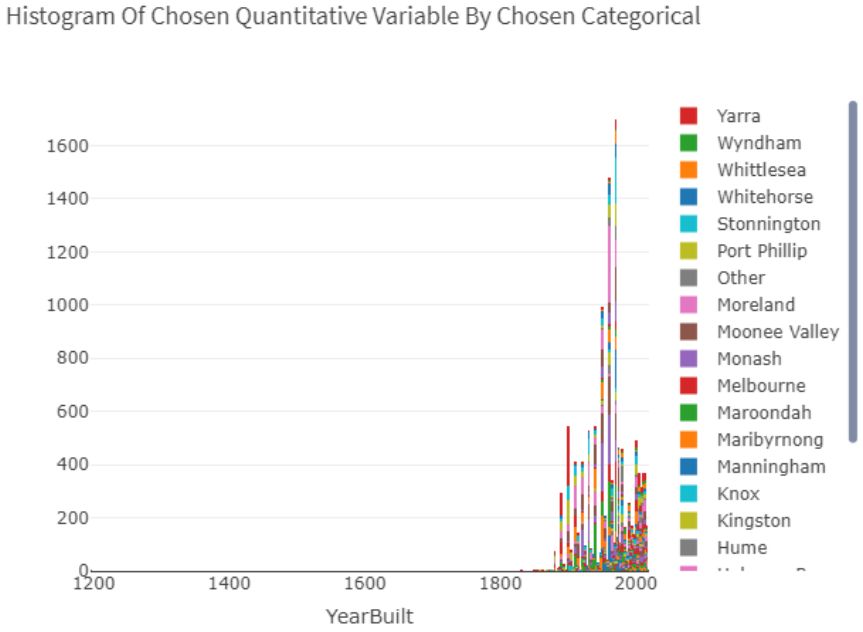
Our efforts to identify outliers in the data coincided with our efforts to find inaccurate values. A probable side-effect of gathering the data through web-scraping websites with data entry errors, we found inaccuracies in the reporting of *Landsize*, and *BuildingArea* for several properties. For instance, a property with a listed *Landsize* of 433014 square meters that when investigated online is found to only be around 107.



The above mentioned outliers were initially identified using Cook's Distance to find influential observations. A full regression model was fit and then cook's distance calculated and plotted. Two majorly influential points were visually identified, but the extremity of these points made it difficult to visually identify any other outlying points.



Other inaccuracies in data entry were evident throughout the dataset. More unreliable values were found in the *YearBuilt* column, with several properties reportedly being built in the 12th century.

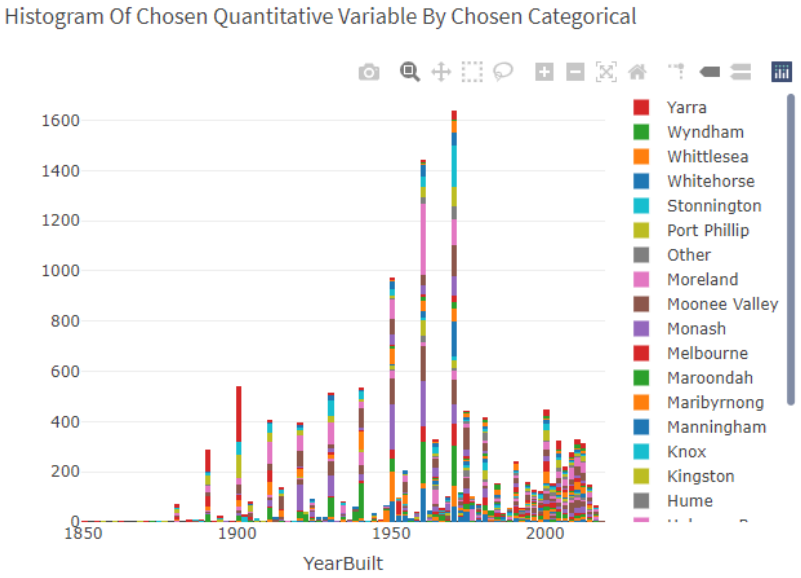


Furthermore, there were differences in date formatting in the *Date* column (e.g. 04/03/2017 vs. 4/3/2017), and in many of the categorical variables such as *Suburb, CouncilArea, Postcode,* and *Seller* there were large imbalances between the frequency of classes within their respective category.

Data Cleaning

Due to the variety of inconsistencies and errors found in the data, we needed to employ several steps of data cleaning before we could fit our model. First, we decided to impute the missing values in the *Landsize, BuildingArea,* and *Car* columns. Then, we wanted to bin together the categorical classes with a low frequency. Lastly, we used the lubridate package to standardize our datetime formatting.

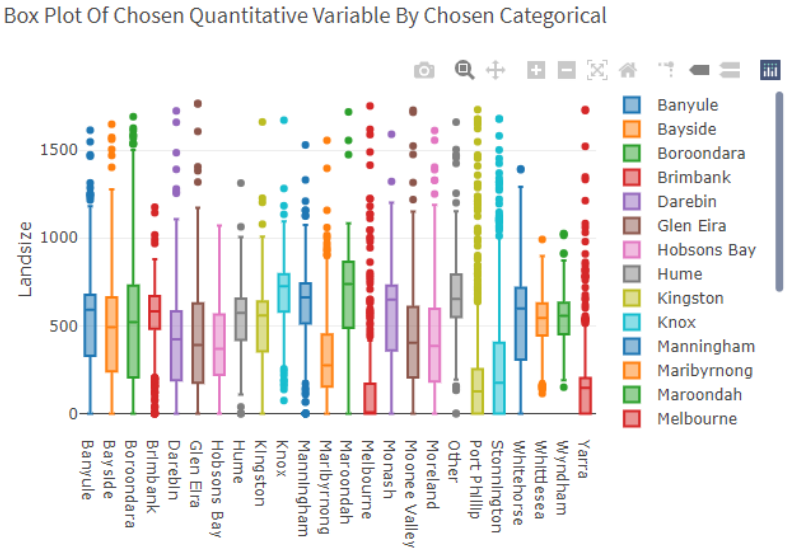
To impute the missing values we calculated the median value for each column. To ensure the most accurate results, we calculated multiple medians based on how we grouped the data. The medians for *BuildingArea* were determined by *Rooms* and *Type (EX. What is the median building area for a townhome with 3 rooms?)*, the medians for *YearBuilt* were determined by *Suburb* and *Type (EX. What is the median year built for houses in Carlton?)*, and the medians for *Car* were based on *Type*. *Type* was used in each column because of the expected differences between the variables for apartments, townhouses, and houses. In total, 11,823 missing values between the three columns were imputed with a median value relative to the other predictors.



In addition to missing values, we needed to consolidate the values of several of the categorical variables to more manageable ranges to hopefully reduce the complexity of the final model. The categorical variables we found the most important to the final model, and thus in need of consolidation, were *CouncilArea*, *Postcode*, *Suburb*, and *Seller*. To reduce the number of unique values, we set a frequency threshold for each variable. If a value did not occur in the data more times than the threshold, we changed the value to "Other." The threshold for *CouncilArea*

was 100, 25 for *Postcode*, 20 for *Suburb*, and 20 for *Seller*. This process significantly reduced the number of unique values for these variables and in turn would benefit the model building process through reduced complexity.

In removing outliers we focused on *Landsize*, *BuildingArea*, and *YearBuilt*. For *Landsize* and *BuildingArea* a threshold was used to filter out all values above the 98th percentile. Based on the distribution of values in *YearBuilt*, we believed it necessary to filter out any values below the year 1750. The effects of removing these outliers are obvious when comparing the distribution of *Landsize* to the figure of the same variable earlier in the paper.



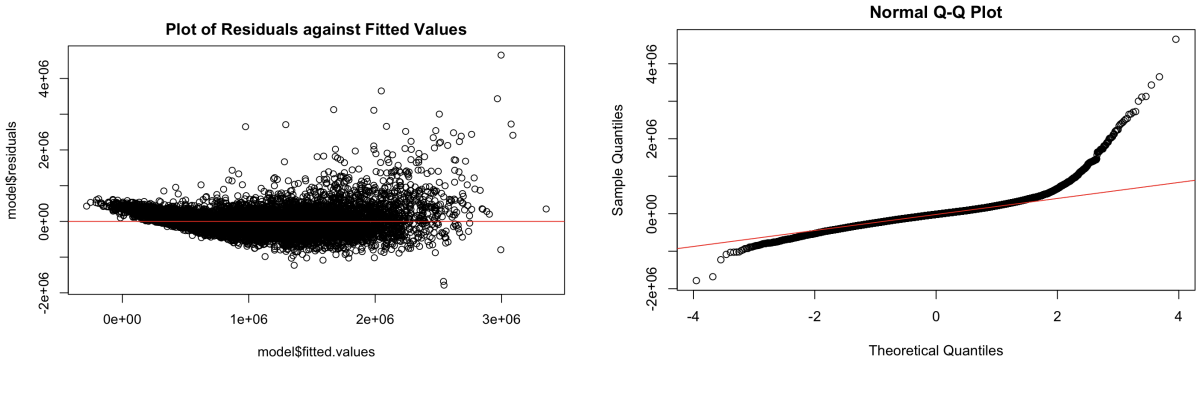
Model Building Process

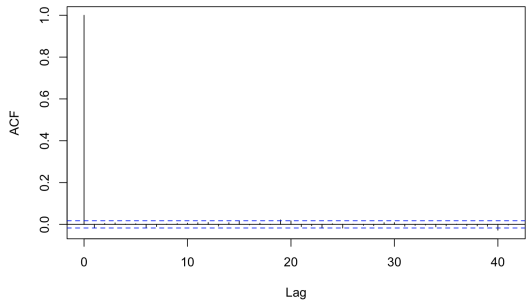
After our exploratory analysis and cleaning the dataset we were ready to start building the final model for predicting *Price*. Due to the large number of classes within each categorical predictor, we wanted to use automatic search procedures to find a baseline model to evaluate. Exhaustive search proved to be too expensive, so we created a full model and an intercept only model, and found a baseline model using forward selection. Forward selection adds predictors until the AIC no longer improves, and the optimal model via forward selection is as follows;

```
Step:  AIC=326410.2
Price ~ BuildingArea + Suburb + Type + Landsize + CouncilArea +
    Bathroom + YearBuilt + SellerG + Year + Rooms + Method +
    Car + Walk.Score + Month + Lattitude + Distance
```

```
Residual standard error: 310300 on 12629 degrees of freedom
Multiple R-squared:  0.7349,    Adjusted R-squared:  0.7293
F-statistic: 131.6 on 266 and 12629 DF,  p-value: < 2.2e-16
```

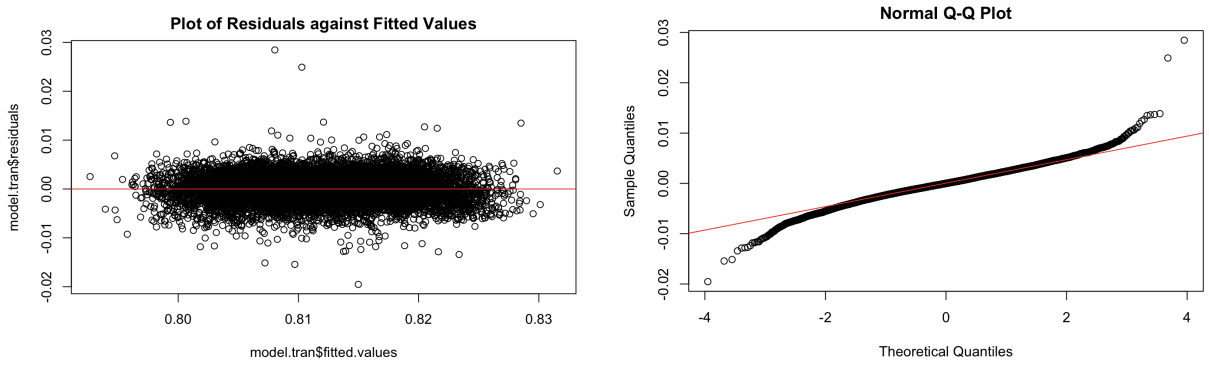
Automatic search procedures still need to be validated, so we then tested the regression assumptions for the baseline model. We found that the model does not have constant variance, the normality is skewed, and there is independence.



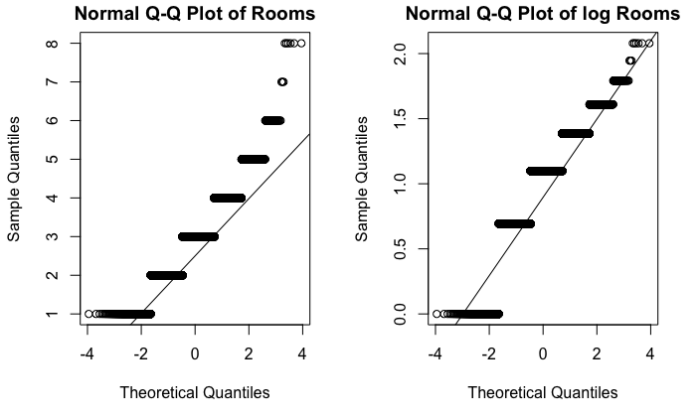


Transformations

We used a log-likelihood plot to find the optimal lambda range for a box-cox transformation on the response variable. The range contained zero, so we used a log transformation on *Price*. After applying the transformation, our residual plot and QQ plot showed better normality, and constant variance.



We created normality plots for quantitative predictors side by side with their respective log plot and found that *Rooms* would benefit from a log transformation.

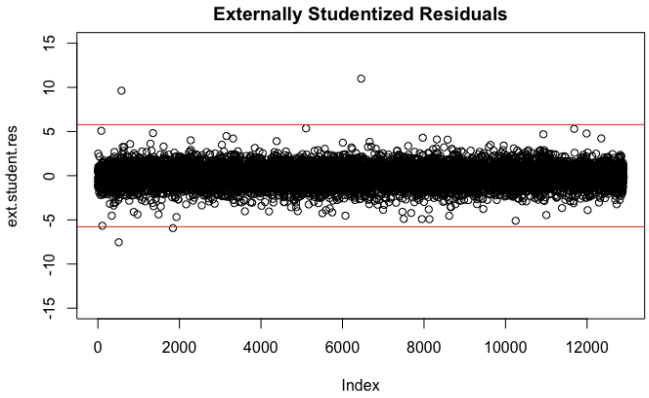


After transforming the response variable and *Rooms*, we saw a substantial increase in the predictive performance of our transformed model with the adjusted r-squared increasing from 0.73 to 0.82.

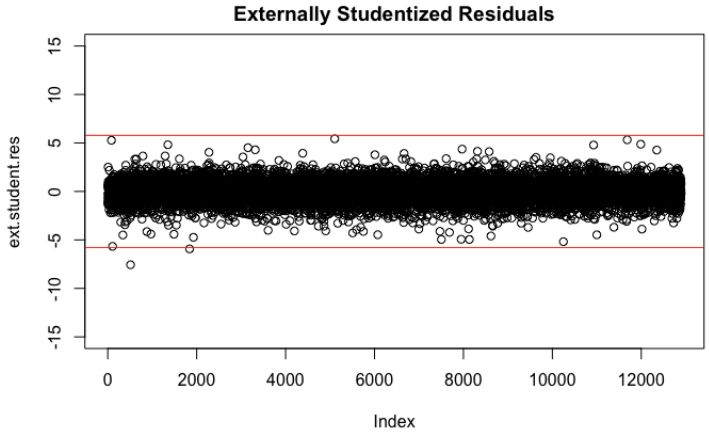
```
Residual standard error: 0.002637 on 12629 degrees of freedom
Multiple R-squared:  0.8282,    Adjusted R-squared:  0.8246
F-statistic: 228.9 on 266 and 12629 DF,  p-value: < 2.2e-16
```

Model Diagnostics and Remedial Measures in MLR

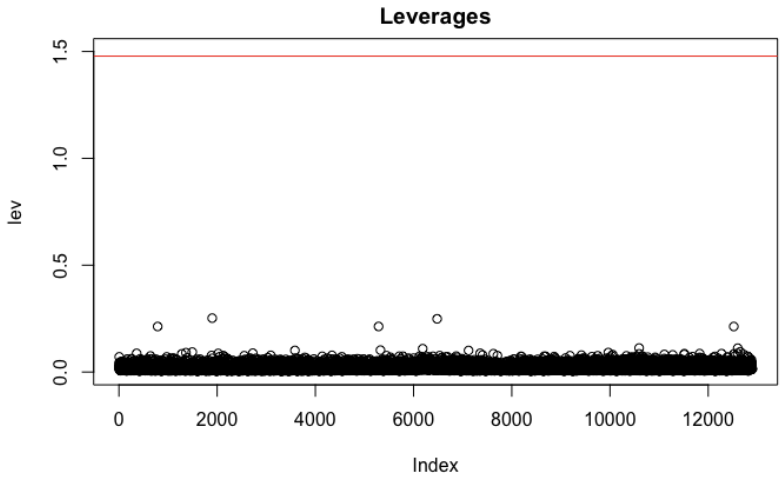
Next we needed to confirm our regression model was appropriate by ensuring that the externally studentized residuals followed a t-distribution with *n-p-1* degrees of freedom, and we used those residuals to detect outliers in the response variable.



We found 4 outliers in *Price*, 2 of them were verified, but the other 2 were data entry errors. The first, 171 Moreland Rd, had *Price* listed as $145,000, but the actual price is $1,550,000, [www.developmentready.com.au/properties/171-moreland-road-coburg-vic-3058]. The second was a similar data entry error. After running the models with the corrected data entry errors, we were left with only 2 outliers in the response. Both outliers were accurate data, so we left them in.



Then, we needed to look for outlying values in the predictor variables; Using *2*p/n* as the cutoff threshold, we were able to show that there are none.



Lastly, we wanted to find observations that if removed, would drastically change the fitted value or the estimated coefficients. Using Cook's Distance, we were able to show that there are no influential observations present within the model.

Finalized Model

After obtaining a baseline model via forward selection, applying the appropriate transformations, and performing outlier detection, we had obtained a solid model that meets all of the regression assumptions. The model summary is as follows;

```
Residual standard error: 0.003982 on 12629 degrees of freedom
Multiple R-squared:  0.8308,    Adjusted R-squared:  0.8272
F-statistic: 233.1 on 266 and 12629 DF,  p-value: < 2.2e-16
```

While we were able to prove this model is appropriate, we were left wondering if it's the best one. With so many parameters, we decided that a graphical user interface would be an interesting tool to allow us to play with different models without repeating large amounts of code.

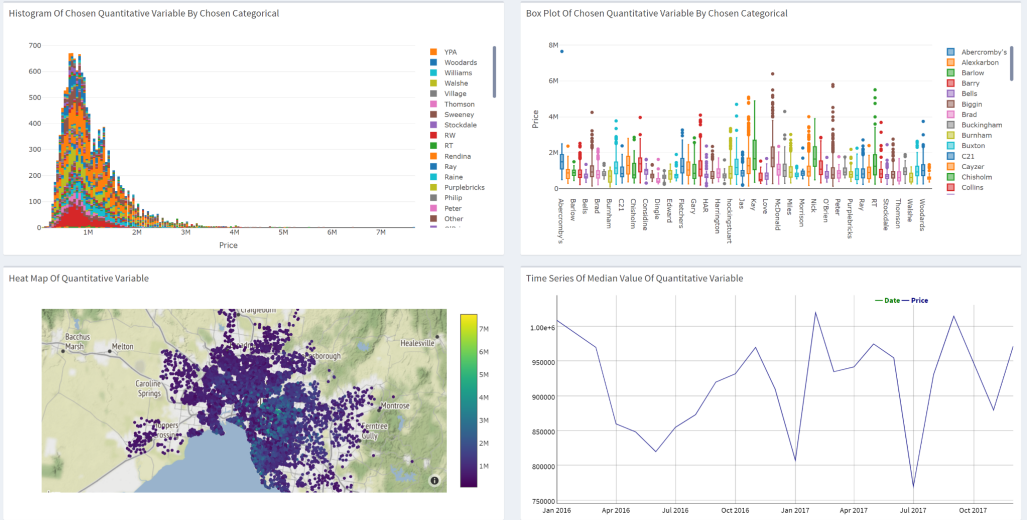
Dashboard

To assist in data analysis and iterative modelling, we created a web application to interact with the dataset. The dashboard is split into three tabs, Exploratory, Simple Linear Regression and Multivariate Linear Regression.

The Exploratory tab allows for easy visualization of filtered data. In the sidebar, the user can select a date range, quantitative variables to visualize, categorical variables to group the data by, and also choose which levels of the selected categorical variable to filter the data to (multiple

levels can be compared simultaneously). When a filter option is changed, the plots are automatically updated to represent the new selected data. Several plots are included for visualization, as well as an attribute table of the data at the bottom.

The plots available include a histogram of the chosen quantitative variable grouped by the chosen categorical variable, a box plot of the same parameters, a heat map showing the geographic location of each property colored by the chosen quantitative variable, and a time series of the median value of the chosen quantitative variable.

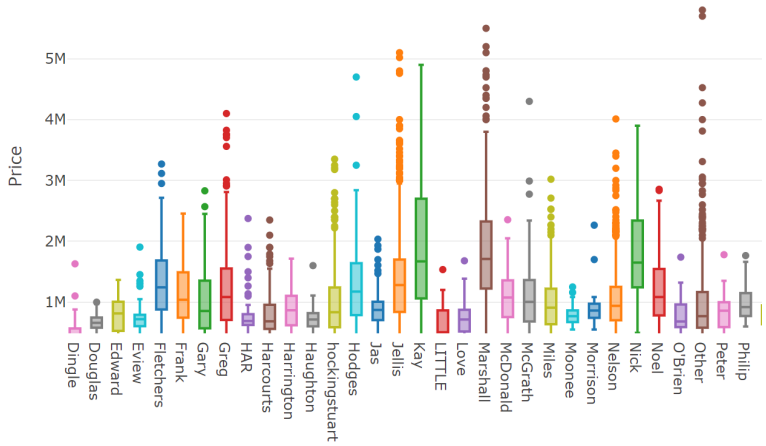


The histogram allowed us to quickly and easily isolate abnormalities in our selected quantitative variable. Looking at it by *YearBuilt*, for example, the extreme kurtosis of our distribution showed us that we had one property dated back to year 1196. This was very clearly incorrect data that we were able to identify.

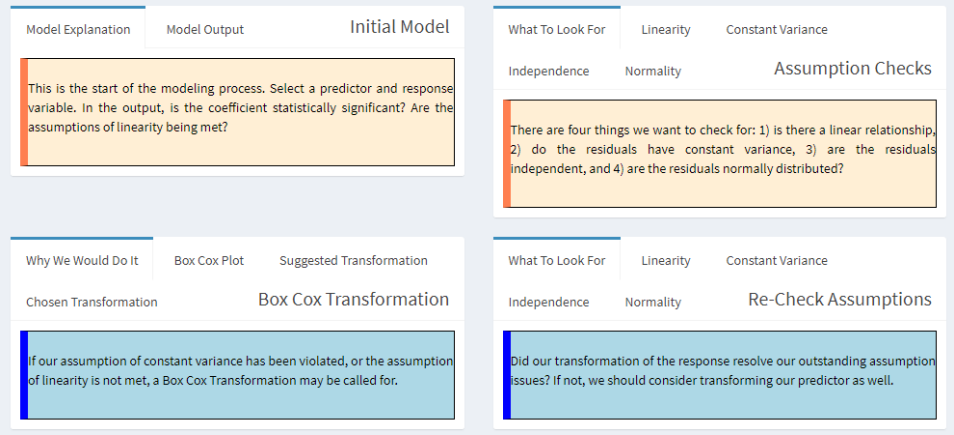
Using the box-cox plot to view *Price* by *SellerG* we were able to compare median price and visually identify some of the real estate brokerages most likely to sell more expensive properties. After zooming in to get a closer look, we can see that the two highest median-price sellers are Kay and Marshall, with Marshall's median *Price* being slightly higher at 1.711 million AUD.

The heat map allows us to inspect a number of different trends. We were able to identify that properties further from downtown are typically larger in size, on plots larger in area. Similarly, higher priced properties clustered in downtown Melbourne and near the southern beach. This would be telling for our future modeling attempts, as we would discover that latitude had a strong negative coefficient.

The time series plot portended the PACF plot we would later run into. Having this chart in the exploratory analysis alerted us that we would have to either deseasonalize the data using decomposition or use month/year as categorical variables.

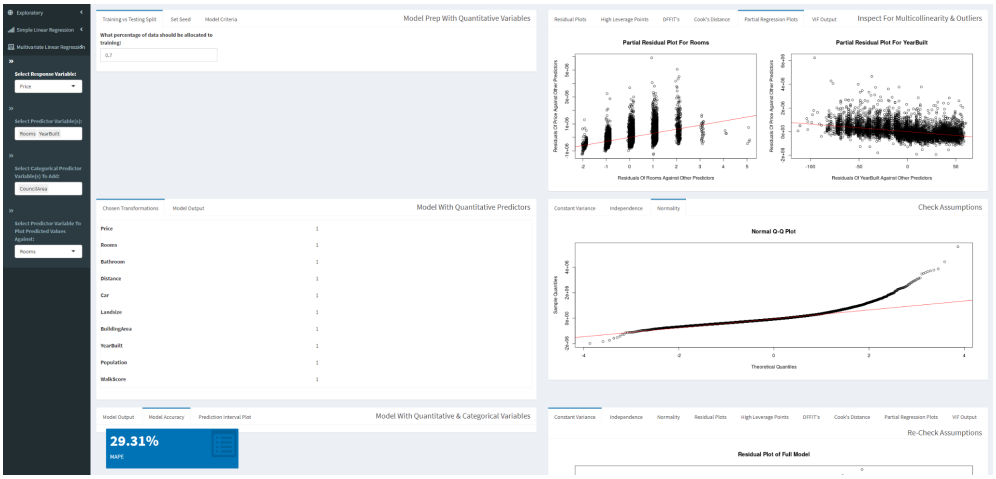


The Simple Linear Regression tab is a great tool for quickly building a regression model of a chosen response variable against a chosen predictor variable. The user can select the response and predictor as well as additional categorical variables to add to the model. The tool will then build the model and produce the right plots for checking the assumptions. Through this tool, the user is walked step by step through the model building and variable transformation process until they reach a well fit model. At each step, tips and explanations are provided for why the steps are taken.



The Multivariate Linear Regression tab is expectedly similar to the Simple tab, but allows for more in-depth model building. The tool allows the user to select a response variable and multiple predictor variables, quantitative and categorical, to start the model building with. Numerous outputs are provided to inform the model building process, including regression

subsets, residual plots, Cook's Distance, partial regression plots, VIF tests, plots for checking regression assumptions, and mean absolute percentage error for checking the overall model accuracy. Most significantly, we have given the user the ability to a) build a model for whatever information they have on hand regarding a property of interest, b) input these variables into a dynamic matrix, and c) produce an estimated price with a corresponding prediction interval. The user can also download the predicted data from the testing dataset for further analysis.



The combination of the three tabs available in the dashboard will allow any Melbourne real estate investor to easily analyze and understand the dataset, and gain a better understanding of the many variables that influence the price of a property.

Does an Additional Bathroom Increase Price?

Now that we had our final model, we were able to run a few tests to answer our questions about the market. One of the questions we had was, "does adding an additional bathroom increase the average home value by 5.7%?" We checked this with a t-test. We compared the coefficient for *Bathroom*, 0.06268, to the log value of 5.7% of the average *Price*, 11.01228.

H0: Bathroom Coefficient = 11.01228 Ha: Bathroom Coefficient != 11.01228

```{r}
bathroomCOEF <- 0.06268
bathroomSE <- 0.003868

mean <- mean(melbourne$Price)
pctmean <- mean*0.057
logmeanprice <- log(pctmean)
t <- (bathroomCOEF-logmeanprice)/bathroomSE
abs(t)

critical <- qt(0.025, 12895)
critical
```

```
[1] 2833.786
[1] -1.960148
```

Because the absolute value of the test statistic is greater than the critical value, we reject the null hypothesis that the coefficient for *Bathroom* is 11.01208. In other words, we fail to prove that for each additional bathroom the average value of a home will increase by 5.7%, while holding the other predictors constant.

Conclusions

We began this project with the goal of uncovering market inefficiencies and leveraging these for strong property investments. Using our optimized model, we were able to establish prediction intervals on our testing dataset: 3,869 properties if we use a 70%-30% split. We downloaded this data from our dashboard and subsequently identified 201 properties that fell outside of our prediction intervals. 91 of these 201 properties were classified as "undervalued"--meaning the sell price was lower than the lower range of our intervals.

Since our categorical variables of interest, like *SellerG*, failed Levene's test, we could not confidently perform post-hoc testing and ANOVA. We proceeded to interpret our results like an "end user" of our dashboard would; we explored the prediction data in Excel and looked at items of interest. By isolating the real estate agents/agencies associated with our outliers, we compared the "undervalued" (and "overvalued") frequency with the total number of properties sold under that name. A few agencies jumped out as entities to avoid in negotiations, and conversely there were a handful of smaller agents who might prove to be weaker price hagglers.

We also found Opendoor's prediction that an additional bathroom would increase the average value of a property by 5.7% to be false. This could be because of the diminishing returns that Opendoor also noted in their article. An additional bathroom may greatly increase the value

of a smaller home, but may not add any significant boost in price for larger ones; similarly, there may be a significant price increase when going from 0 bathrooms to 1 bathroom, or 1 to 2, but there may be no marginal value whatsoever beyond that. Log transformation of the bathroom predictor did not change the contextual dollar value impact, but we could have explored various other transformations and even utilized bathrooms as a categorical variable. For example, we could have set bathroom levels: 0, 1 to 2, and 3 plus bathrooms. Such factorization could provide a more intuitive understanding of bathrooms' impact on price and actually be supportive of Open Door's assertion. On the other hand, this may be a great example of sample data not generalizing to the characteristics of a population set. Our sample, Melbourne, has experienced a drastic rise in housing prices over the past few years. It seems analogous to San Francisco, where the home's amenities don't matter as much as the land value the home sits on, and with that in mind, it would make sense that our hypothesis test failed.

One of the other conclusions we were able to draw was that time heavily factored into selling price: whether that be the date it was sold, or whether that be the year it was built. Selling price was heavily affected by month, suggesting surprising seasonality over the course of the year. The month coefficients encompassed a broad range of values. As for year built, older homes, counter to popular opinion, actually sold for more than newer ones. It may be worth exploring the interaction of year built with home type; an inordinate proportion of newer properties are apartments, which, as a level of housing type, have a smaller coefficient than that corresponding to the "house" level. We opted out of interactions in favor of a more intuitive model, but they would certainly be helpful in extricating the true drivers of some of these relationships.

Lastly, we developed some hypotheses regarding the 17% of unexplained variance in our response. The most glaring omission from our dataset was the listing date. The number of days between listing and close is primarily a function of supply and demand, and appropriate pricing along these curves. If a property's list price outpaces the market's demand, this property is likely to find itself on the market considerably longer. Having the list date would allow us to not only explore appropriate low-ball offers for properties that have been listed for a while, but it would allow us to isolate properties that require quick action--giving us an opportunity to make an offer before other prospective buyers make a decision.