Statistical Analysis of Airbnbs in Sydney, Australia

17th October 2023

James Huvenaars, Ethan Scott, Julia Cornejo, Archangelo Ouano, Divine Ahuchogu
Data 602, L01

Purpose

The main goal of this project was to gain more insight into the types of Airbnb rentals in Sydney, Australia. When planning a trip, information about price, location, property type, and amenities are significant factors to consider when deciding where to stay. Our analysis's main practical implications are a better understanding of the neighbourhood and room type dispersions and the factors that affect pricing.

Data

The dataset used in this project contains over 24,000 listings for Airbnb in Sydney, Australia. The sample of listings was collected from June 4, 2023, until September 4, 2023.

This dataset is available under the Creative Commons Attribution 4.0 International License, so we must give appropriate credit, link to the licence, and indicate if changes were made. We may do so in any reasonable manner but not in any way that suggests the licensor endorses us or our use. We acquired the dataset from http://insideairbnb.com/get-the-data/ and made no changes to this dataset. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit https://creativecommons.org/licenses/by/4.0/.

Guiding Questions

The main focus of our statistical investigation will be the rental price per night between different property types and neighbourhoods.

The statistical methods used are as follows:

1. **Guiding question 1** - *Do different neighbourhoods and room types affect price?*
    - We analyzed the distribution of room types and neighbourhoods in the dataset. We then created and compared confidence intervals of their mean price values to find which have a statistical difference from one another. We completed this test through ANOVA testing.

- We then completed a Fisher LSD test to measure the statistically significant difference between specific neighbourhoods and room types.

2. **Guiding Question 2 -** *Can pricing be modelled as a linear regression of multiple factors such as the number of beds, bathrooms, neighbourhood, etc.?*
    - Five linear regression models were used to test which factors most affect Airbnb pricing and create a formula for the regression. The regression models used were Simple Linear Regression, Stepwise Selection, LASSO, Ridge, and Elastic Net.
    - A comparison of Mean Squared Error (MSE), Root Mean Squared Error (RSME), and Mean Absolute Error (MAE) was then used to find the best-fit model.

Guiding Question 1 - *Do different neighbourhoods and room types affect price?*

Anova testing 1 - Measuring the difference in price by neighbourhood.

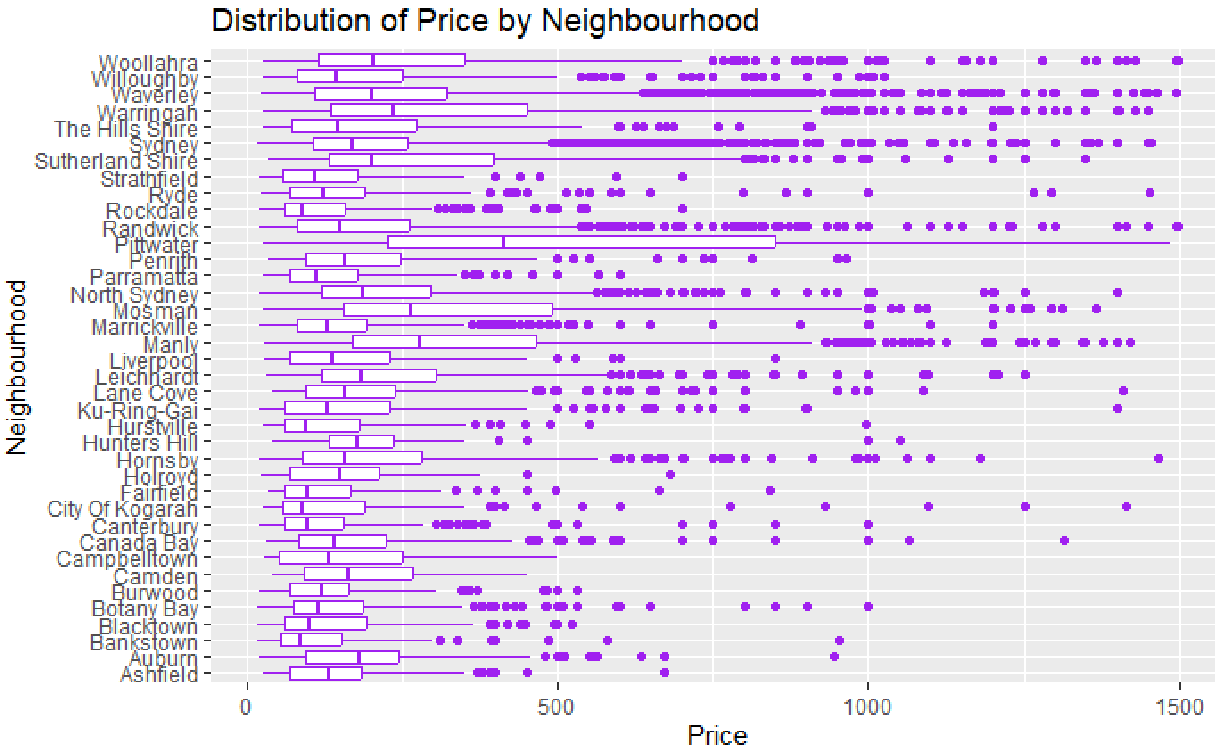
Hypothesis creation

H_0: There is no statistical difference in mean price from neighbourhood to neighbourhood in Sydney.

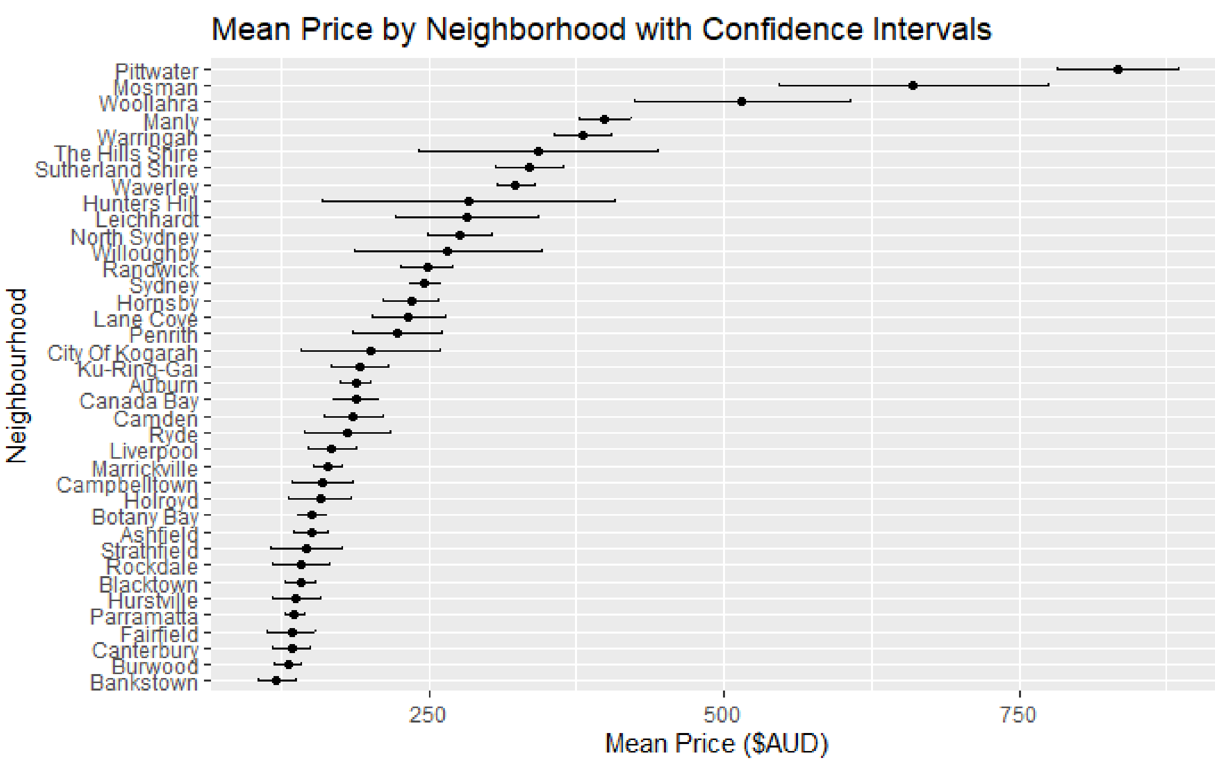
H_A: There is a statistical difference in mean price from neighbourhood to neighbourhood in Sydney.

Data Visualization

The graph below is a box plot of the mean prices per neighbourhood. This graph only shows the listings priced less than $1,500 to improve visibility.



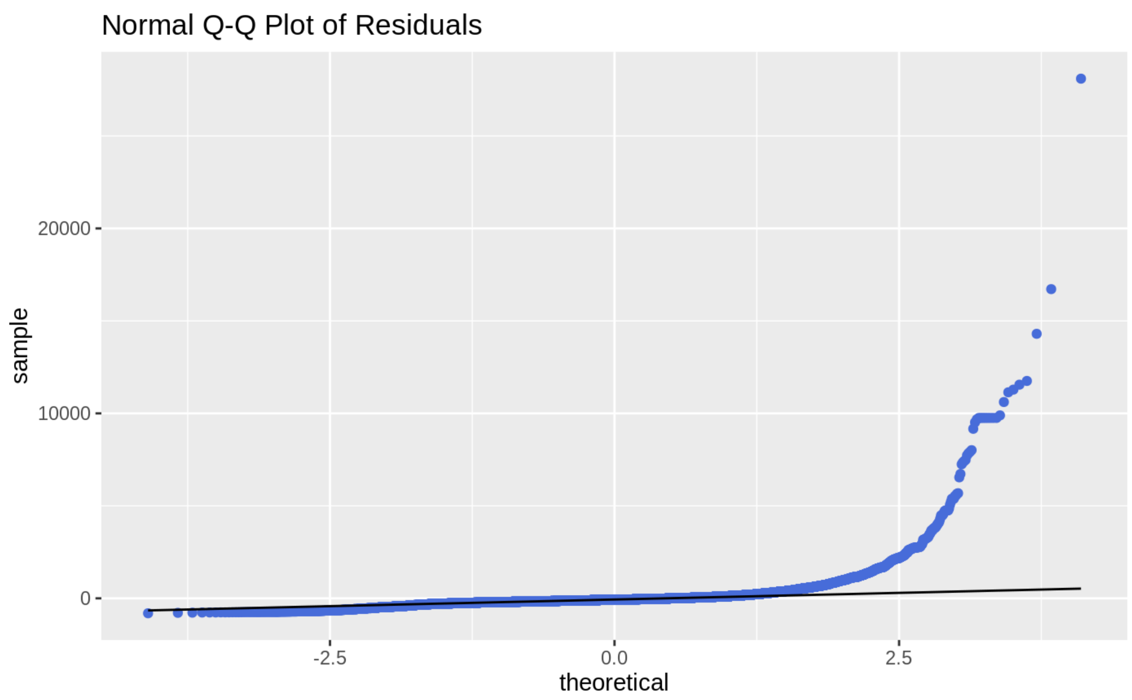
The graph below is similar but just displays the mean price and 95% confidence intervals for all neighbourhoods.



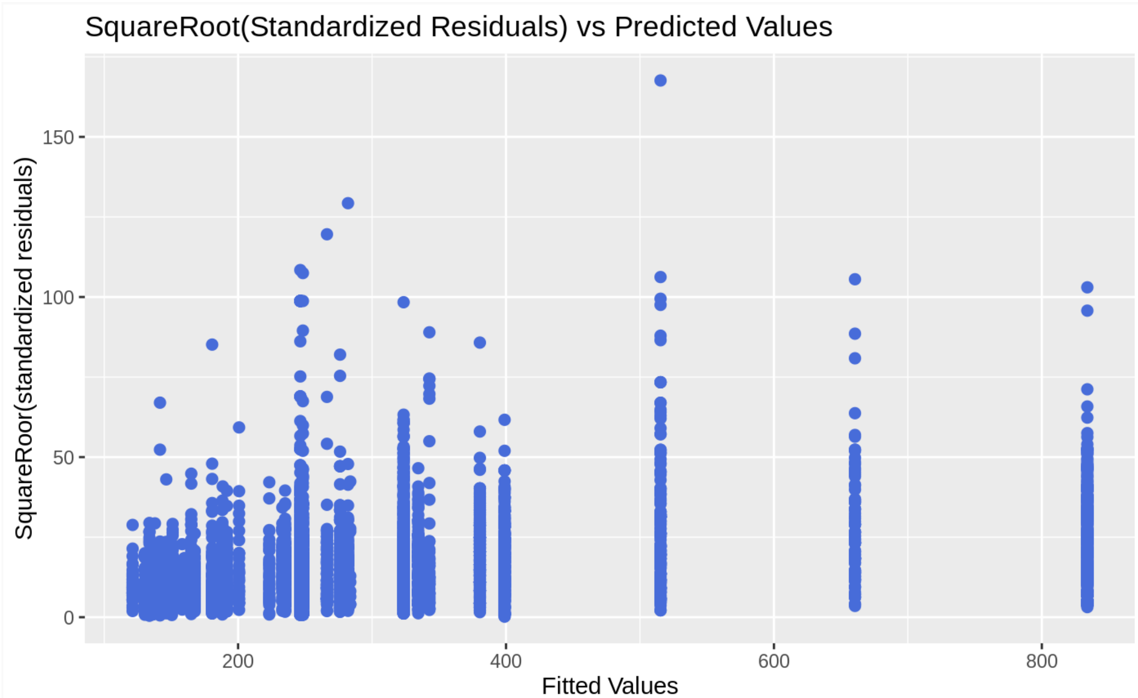
Condition Checking

Below, the conditions of ANOVA are being checked for normality and homoscedasticity.

First, we are checking the normality of residuals. The distribution is roughly normal, with some values trailing off on the right end, signifying that there are some outliers slightly skewing the data. We can consider the condition of normality met.



In the second plot, we are checking for homoscedasticity. The distribution is roughly equivalent across values of X. The residuals show homoscedasticity through an overall rectangular distribution instead of a wedge or patterned shape.



Having met the conditions of the ANOVA model (residual normality and homoscedasticity), the hypothesis was calculated with the following output:

```
                          Df     Sum Sq  Mean Sq F value                Pr(>F)
neighbourhood_cleansed    37  561798346 15183739   52.71 <0.0000000000000002 ***
Residuals              23984 6908755144   288057
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P-value = <0.0000000000000002

F value = 52.71.

Since the p-value is less than the critical value of 0.05, we can reject the null hypothesis and infer from these data that there is a statistically significant difference in mean price from neighbourhood to neighbourhood in Sydney.

Model Application

We completed a Fisher LSD Test to compare the means of each neighbourhood. Below is the sum of outputs, outlining how many neighbourhoods have a statistically significant difference in means and how many do not.

The statistically significant difference column indicates the number of neighbourhoods where the confidence interval for the difference in mean price doesn't cross 0. In contrast, the not statistically significant difference column shows the number of neighbourhoods where the confidence interval for the difference in mean price crosses 0.

Neighbourhood <chr>	Statistically_significant_difference <int>	Not_statistically_significant_difference <dbl>
Sydney	4	33
Manly	15	22
Randwick	8	29
Waverley	2	35
Mosman	15	22
Marrickville	11	26
Warringah	3	34
Leichhardt	11	26
Hornsby	11	26
Woollahra	0	37
Canterbury	15	22
Sutherland Shire	2	35
Ryde	7	30
Ku-Ring-Gai	10	27
Pittwater	11	26
North Sydney	8	29
Willoughby	1	36
Rockdale	7	30
The Hills Shire	1	36
Penrith	6	31
Ashfield	13	24
Parramatta	9	28
Lane Cove	10	27
Hurstville	13	24
Hunters Hill	3	34
Auburn	13	24
Burwood	15	22
Camden	8	29
Blacktown	14	23
Liverpool	10	27
City Of Kogarah	8	29
Bankstown	15	22
Canada Bay	11	26
Botany Bay	14	23
Holroyd	8	29
Strathfield	7	30
Campbelltown	8	29
Fairfield	13	24

Anova testing 2 - Measuring the difference in price by room type.

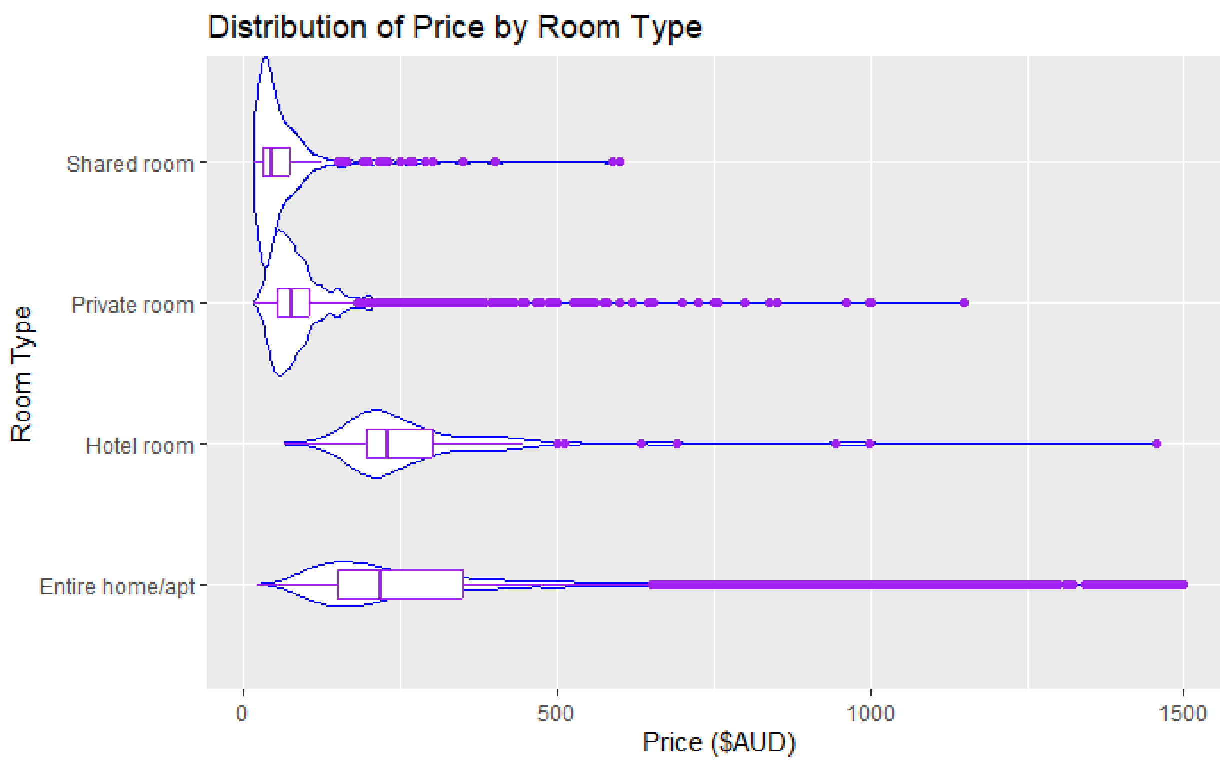
Hypothesis creation:

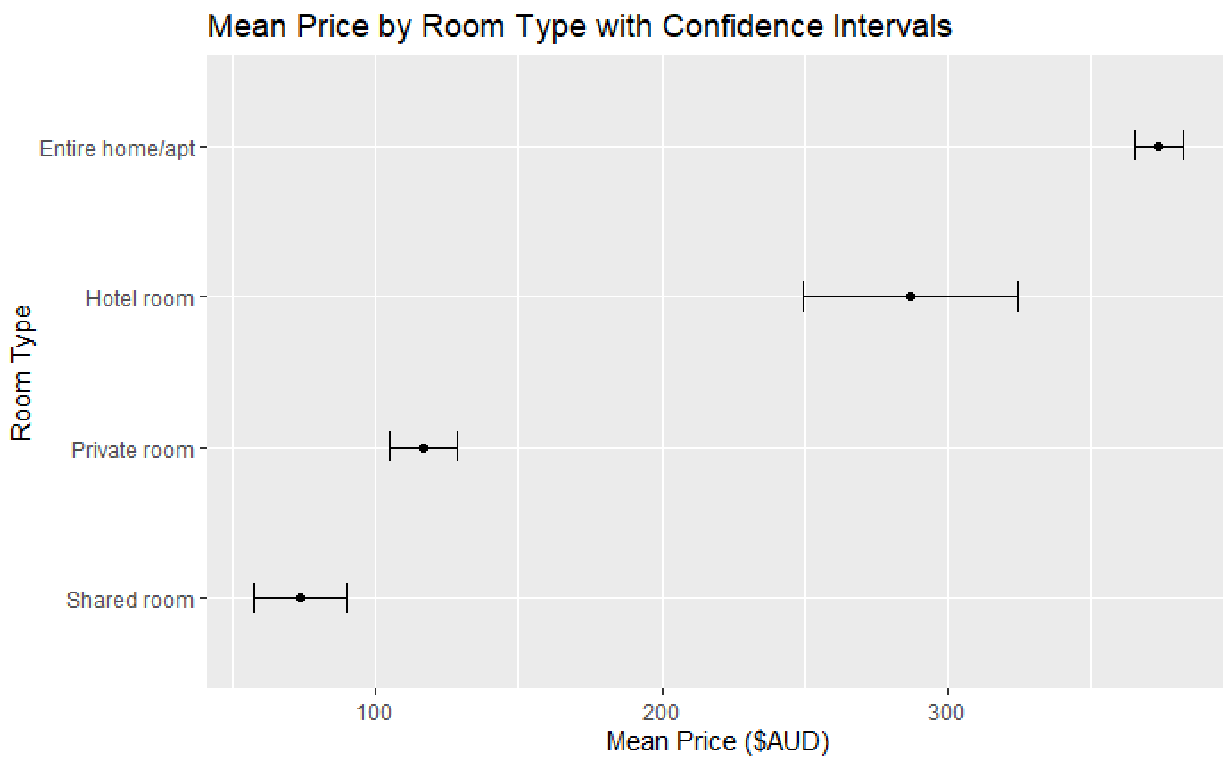
H_0: There is no statistical difference in mean price between different room types.

H_A: There is a statistical difference in mean price between different room types.

Data Visualization

Below are visualizations of the distribution of mean prices by room type displayed just as violin plots and confidence intervals.

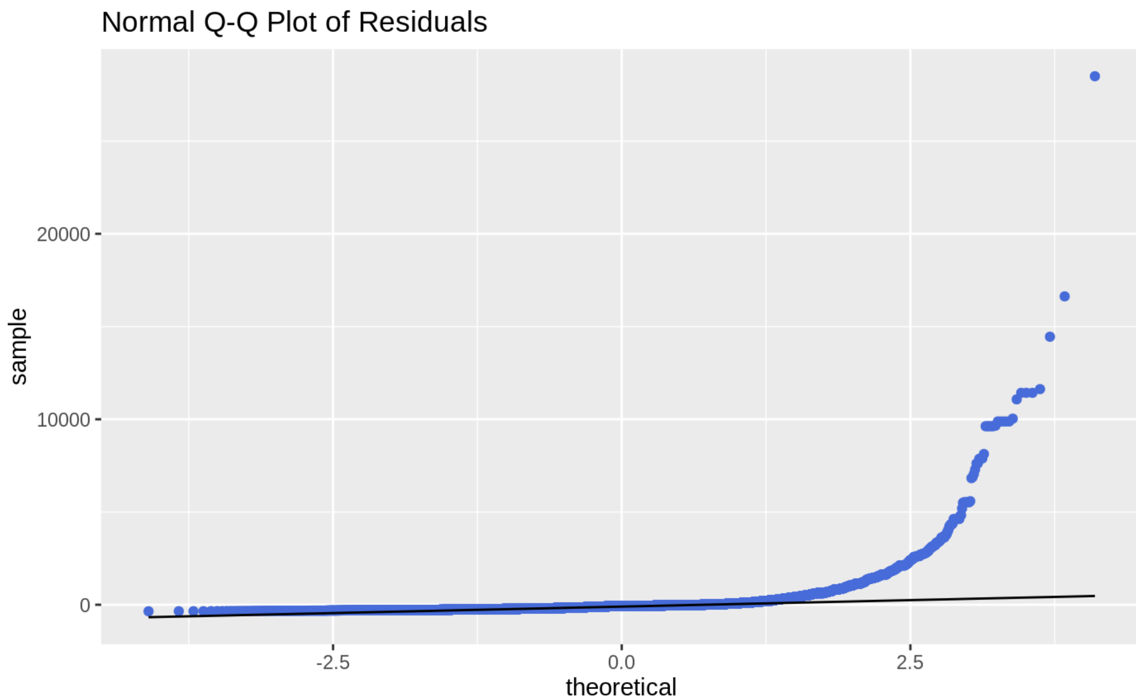




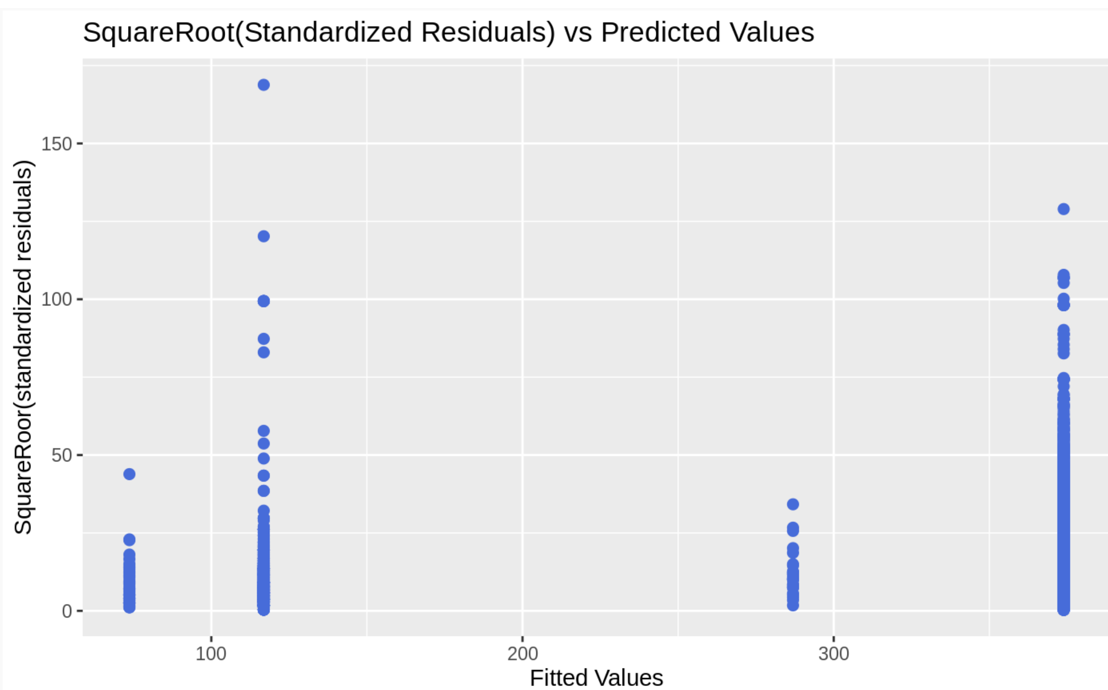
Condition Checking

Below, the conditions of ANOVA are being checked for normality and homoscedasticity.

In the first plot, checking the normality of residuals, we can see that the distribution is roughly normal, with some values trailing off on the right end, signifying that there are some outliers slightly skewing the data. We can consider the condition of normality being met.



In the second plot, checking for homoscedasticity, we can see that the distribution is roughly equivalent across all values of X. The residuals show homoscedasticity through an overall rectangular distribution instead of a wedge or patterned shape.



Having met the conditions of the ANOVA model (residual normality and homoscedasticity), the hypothesis was calculated with the following output:

```
                Df     Sum Sq   Mean Sq F value               Pr(>F)    
room_type        3  333355076 111118359   373.9 <0.0000000000000002 ***
Residuals    24018 7137198414    297160                                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P-value = <0.0000000000000002

F value = 373.9.

Since the p-value is less than the critical value of 0.05, we reject the null hypothesis and can conclude that there is a statistical difference in mean rental price between room types.

Model Application

Below are the outputs of a Fisher LSD Test that specifies the difference in means for all combinations of different room types to showcase which have a statistically significant difference in means.

```
  Posthoc multiple comparisons of means : Fisher LSD 
    95% family-wise confidence level

$room_type
                                  diff     lwr.ci     upr.ci                pval    
Hotel room-Entire home/apt    -86.98339 -194.6840   20.71720             0.11343    
Private room-Entire home/apt -257.05807 -272.4280 -241.68810 < 0.0000000000000002 ***
Shared room-Entire home/apt  -300.21417 -365.0528 -235.37556 < 0.0000000000000002 ***
Private room-Hotel room      -170.07469 -278.2428  -61.90655             0.00206 ** 
Shared room-Hotel room       -213.23079 -338.4033  -88.05828             0.00084 ***
Shared room-Private room      -43.15610 -108.7684   22.45622             0.19733    

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The room types with statistically significant differences were:

- Private room and Entire home/apt
- Shared room and Entire home/apt
- Private room and Hotel room

- Shared room and Hotel room.

The room types with statistically insignificant differences were:

- Shared room and Private room
- Hotel room, and Entire home/apt.

From these data, we can infer that Private room/Entire home, Shared room/Entire home, Private room/ Hotel room and Shared room/Hotel room have statistically different means. We can also infer that Hotel room/Entire home and Shared room/private room do not have statistically different means.

Guiding Question 2 - *Can pricing be modelled as a linear regression of multiple factors such as the number of beds, bathrooms, neighbourhood, etc.?*

We checked all variables that may be relevant to the price of an Airbnb to measure if they had a statistically significant effect on the overall rental price. We removed some entirely irrelevant variables (ex, listing_url) to lessen the burden of the calculations. In our model creation, we used five different modelling methods to find the best overall fit, prioritizing the model that best predicts the price value. The regression models used were Simple Linear Regression, Stepwise Selection, LASSO, Ridge, and Elastic Net.

Regression Model 1 - (Simple) Linear Regression

Model Creation

```
Call:
lm(formula = price ~ ., data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1738.2  -131.3   -29.4    63.6 16549.3
```

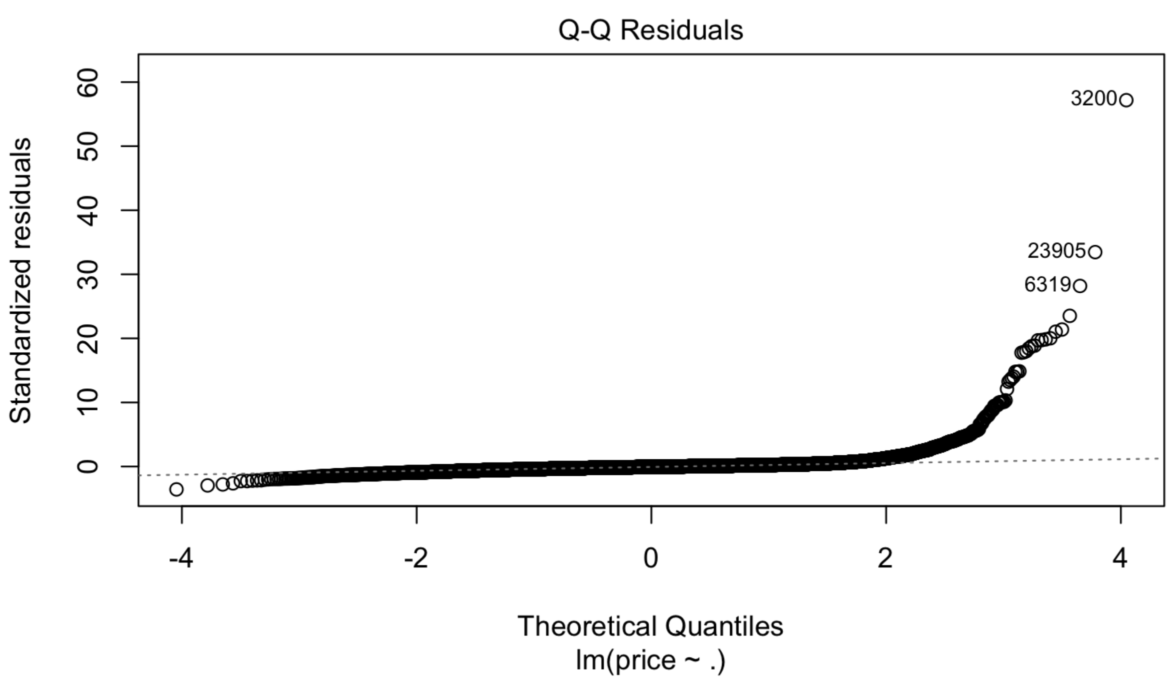
```
Coefficients:
                                          Estimate       Std. Error t value           Pr(>|t|)    
(Intercept)                        -138.64573324059  51.89093843840  -2.672           0.007549 ** 
host_is_superhostt                   61.07731567559  10.90378920198   5.601  0.0000000215451896 ***
host_has_profile_pict               -45.74332889019  33.58333092835  -1.362           0.173187    
host_identity_verifiedt               5.17289706607  10.63650301568   0.486           0.626736    
neighbourhood_cleansedAuburn         10.12950559605  46.33341417106   0.219           0.826947    
neighbourhood_cleansedBankstown     -63.24278898678  56.26792588657  -1.124           0.261045    
neighbourhood_cleansedBlacktown    -137.92984170182  51.32196469695  -2.688           0.007204 ** 
neighbourhood_cleansedBotany Bay     -4.36509516586  47.04178239341  -0.093           0.926070    
neighbourhood_cleansedBurwood       -24.74776035217  52.70599531798  -0.470           0.638687    
neighbourhood_cleansedCamden       -196.08671501428  72.60074663526  -2.701           0.006921 ** 
neighbourhood_cleansedCampbelltown -134.22694883046  70.62081958393  -1.901           0.057360 .  
neighbourhood_cleansedCanada Bay      4.13100214784  48.96990507848   0.084           0.932773    
neighbourhood_cleansedCanterbury    -41.35787014833  51.51154401650  -0.803           0.422051    
neighbourhood_cleansedCity Of Kogarah -34.58730858392 57.88924981976 -0.597           0.550198    
neighbourhood_cleansedFairfield    -140.88590831642  60.91174242188  -2.313           0.020736 *  
neighbourhood_cleansedHolroyd      -155.70301629103  77.08014063084  -2.020           0.043396 *  
neighbourhood_cleansedHornsby       -36.33945503112  47.68134308292  -0.762           0.445991    
neighbourhood_cleansedHunters Hill    0.49895487540 101.19254592150   0.005           0.996066    
neighbourhood_cleansedHurstville    -54.19708586138  59.36287768872  -0.913           0.361265    
neighbourhood_cleansedKu-Ring-Gai   -36.40065591838  50.18690660367  -0.725           0.468276    
neighbourhood_cleansedLane Cove      19.43679519036  54.90986574371   0.354           0.723360    
neighbourhood_cleansedLeichhardt     68.05385576827  45.00264601642   1.512           0.130495    
neighbourhood_cleansedLiverpool    -199.98112603612  59.06608659536  -3.386           0.000711 ***
neighbourhood_cleansedManly         122.81853001721  41.70173100924   2.945           0.003232 ** 
neighbourhood_cleansedMarrickville   16.42495547273  43.55748115026   0.377           0.706113    
neighbourhood_cleansedMosman        336.01238618374  48.85450650336   6.878  0.0000000000062650 ***
neighbourhood_cleansedNorth Sydney   97.42610974998  42.86683103317   2.273           0.023052 *  
neighbourhood_cleansedParramatta    -70.26674841020  46.89121340168  -1.499           0.134018    
neighbourhood_cleansedPenrith      -137.63225770071  57.29264050826  -2.402           0.016303 *  
neighbourhood_cleansedPittwater     372.69166056870  42.32369305918   8.806 < 0.0000000000000002 ***
neighbourhood_cleansedRandwick       53.71911768927  40.42862608928   1.329           0.183950    
neighbourhood_cleansedRockdale        3.44546271039  46.56086052413   0.074           0.941012    
neighbourhood_cleansedRyde            7.91969018326  46.43017900232   0.171           0.864562    
neighbourhood_cleansedStrathfield   -14.50100043653  59.76815609130  -0.243           0.808302    
neighbourhood_cleansedSutherland Shire 25.39664312257 45.77379963373  0.555           0.579018    
neighbourhood_cleansedSydney         83.44864984943  39.11813430147   2.133           0.032917 *  
neighbourhood_cleansedThe Hills Shire -38.73224978830 50.85525733451 -0.762           0.446298    
neighbourhood_cleansedWarringah      64.81017673347  41.48524927424   1.562           0.118246    
neighbourhood_cleansedWaverley      130.06218157653  39.74742805410   3.272           0.001069 ** 
neighbourhood_cleansedWilloughby     36.63315513169  47.20346735519   0.776           0.437718    
neighbourhood_cleansedWoollahra     271.25492152163  42.71598236170   6.350  0.0000000002198709 ***
room_typeHotel room                  92.91240799490  55.06303436894   1.687           0.091546 .  
room_typePrivate room                 0.15759694090   9.66182952098   0.016           0.986986    
room_typeShared room                -23.37148584825  34.42341565076  -0.679           0.497183    
accommodates                         80.81257063876   3.46875962587  23.297 < 0.0000000000000002 ***
beds                                  5.55638081038   4.93586608392   1.126           0.260300    
minimum_nights_avg_ntm                0.70380208491   0.07297214535   9.645 < 0.0000000000000002 ***
maximum_nights_avg_ntm                0.00000003256   0.00000011707   0.278           0.780909    
availability_365                      0.46070165763   0.03222330349  14.297 < 0.0000000000000002 ***
number_of_reviews                    -0.59907547262   0.07650187280  -7.831  0.0000000000000051 ***
instant_bookablet                   -48.08768551222   8.34215425693  -5.764  0.0000000083197345 ***
hostfor                              -0.00490347121   0.00398891422  -1.229           0.218984    
host_about_word_count                 0.51820799903   0.07306047990   7.093  0.0000000000013594 ***
bathrooms_count                       8.02512324794   0.54934700496  14.608 < 0.0000000000000002 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 495.1 on 19165 degrees of freedom
Multiple R-squared:  0.239,	Adjusted R-squared:  0.2369 
F-statistic: 113.6 on 53 and 19165 DF,  p-value: < 0.00000000000000022

Warning: longer object length is not a multiple of shorter object lengthWarning: longer objec
length is not a multiple of shorter object lengthTesting Mean Squared Error (MSE): 391432.3 
Testing Root Mean Squared Error (RMSE): 625.6455 
Testing Mean Absolute Error (MAE): 307.9846 
Training Mean Squared Error (MSE): 210070.6 
Training Root Mean Squared Error (RMSE): 458.3346 
Training Mean Absolute Error (MAE): 176.0348
```

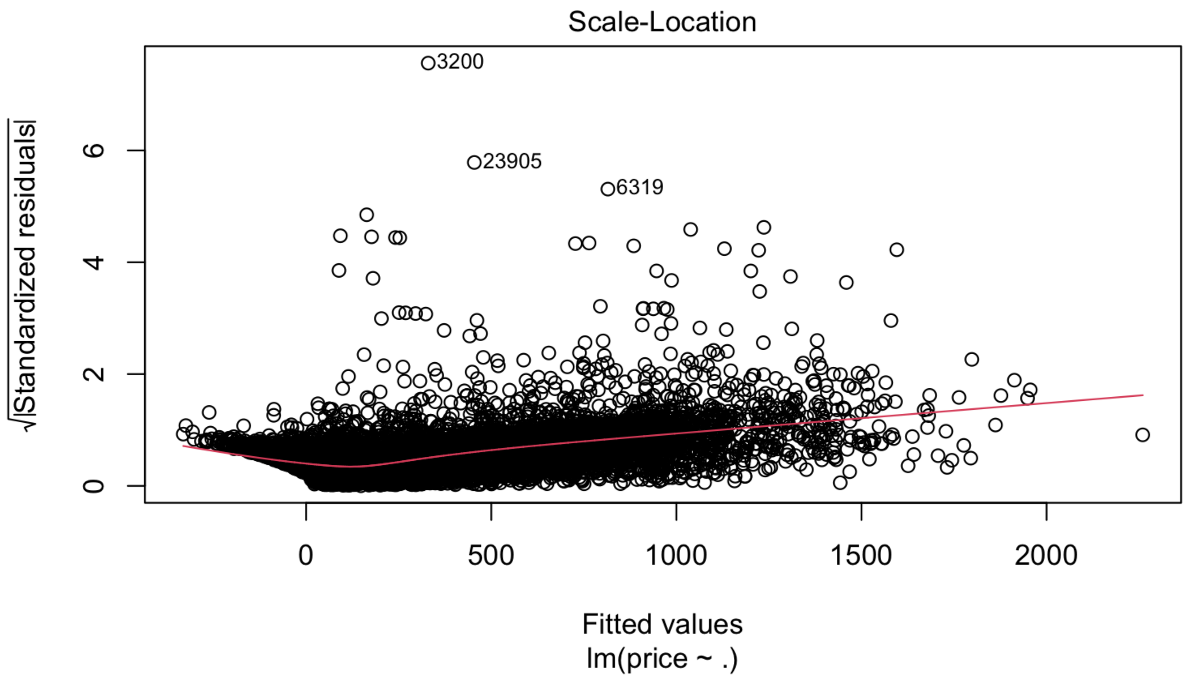
Condition Checking

Below are the conditions of the model being checked for normality and homoscedasticity.

In the first plot, checking the normality of residuals, we can see that the distribution is roughly normal, with some values trailing off on the right end, signifying that there are some outliers slightly skewing the data. We can consider the condition of normality met.



In the second plot, checking for homoscedasticity, we can see that the distribution is roughly equivalent across values of X. The residuals show homoscedasticity through an overall rectangular distribution as opposed to a wedge or patterned shape:



Regression Model 2 - Stepwise Selection

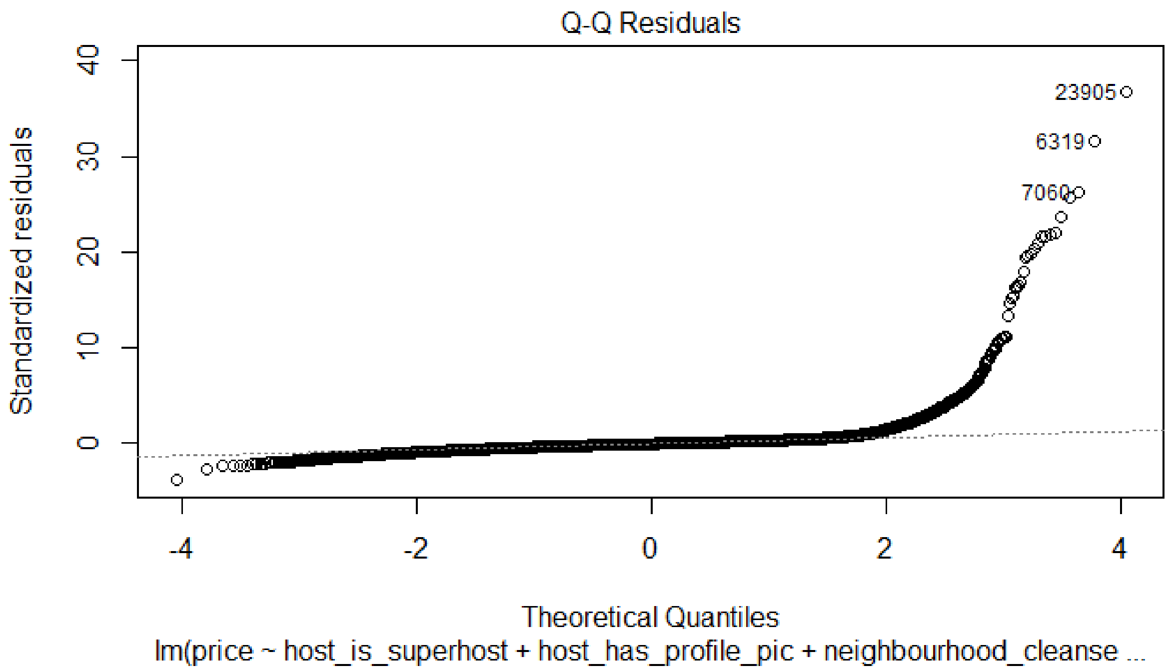
```{r}
#Checking the model
summary(model_stepwise)
```

```
Call:
lm(formula = price ~ host_is_superhost + neighbourhood_cleansed +
    accommodates + minimum_nights_avg_ntm + availability_365 +
    number_of_reviews + instant_bookable + hostfor + host_about_word_count +
    bathrooms_count, data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1775.4  -134.1   -29.7    67.3 28261.9

Coefficients:
                                          Estimate  Std. Error t value           Pr(>|t|)
(Intercept)                            -174.533427   42.244889  -4.131  0.000036198220704 ***
host_is_superhostt                       60.617856   11.014490   5.503  0.000000037716983 ***
neighbourhood_cleansedAuburn              9.163941   48.647435   0.188            0.85059
neighbourhood_cleansedBankstown         -66.463245   58.470862  -1.137            0.25568
neighbourhood_cleansedBlacktown        -143.186799   53.612365  -2.671            0.00757 **
neighbourhood_cleansedBotany Bay        -10.654430   48.763417  -0.218            0.82705
neighbourhood_cleansedBurwood           -39.099702   53.960155  -0.725            0.46870
neighbourhood_cleansedCamden           -206.910857   78.769054  -2.627            0.00863 **
neighbourhood_cleansedCampbelltown      -140.386568   75.800555  -1.852            0.06403 .
neighbourhood_cleansedCanada Bay         11.928442   50.612585   0.236            0.81368
neighbourhood_cleansedCanterbury        -52.317455   53.552294  -0.977            0.32861
neighbourhood_cleansedCity Of Kogarah   -45.459216   60.995349  -0.745            0.45611
neighbourhood_cleansedFairfield        -148.105936   61.593931  -2.405            0.01620 *
neighbourhood_cleansedHolroyd          -147.201421   82.034735  -1.794            0.07277 .
neighbourhood_cleansedHornsby           -34.335852   49.601028  -0.692            0.48879
neighbourhood_cleansedHunters Hill      -51.654286  106.853854  -0.483            0.62881
neighbourhood_cleansedHurstville        -63.163565   60.853179  -1.038            0.29930
neighbourhood_cleansedKu-Ring-Gai       -29.032534   52.153131  -0.557            0.57775
neighbourhood_cleansedLane Cove          22.782647   55.642812   0.409            0.68222
neighbourhood_cleansedLeichhardt         64.935169   46.998675   1.382            0.16710
neighbourhood_cleansedLiverpool        -193.116413   59.698743  -3.235            0.00122 **
neighbourhood_cleansedManly             130.947427   43.667775   2.999            0.00271 **
neighbourhood_cleansedMarrickville       23.137460   45.687593   0.506            0.61256
neighbourhood_cleansedMosman            348.486253   50.782192   6.862  0.000000000006979 ***
neighbourhood_cleansedNorth Sydney       84.607142   44.715053   1.892            0.05849 .
neighbourhood_cleansedParramatta        -69.346589   48.896637  -1.418            0.15614
neighbourhood_cleansedPenrith          -138.218828   59.629455  -2.318            0.02046 *
neighbourhood_cleansedPittwater         378.452733   44.284183   8.546 < 0.0000000000000002 ***
neighbourhood_cleansedRandwick           46.647696   42.505991   1.097            0.27246
neighbourhood_cleansedRockdale           -9.604078   48.239606  -0.199            0.84219
neighbourhood_cleansedRyde              -30.389393   48.326800  -0.629            0.52947
neighbourhood_cleansedStrathfield       -28.276693   61.125344  -0.463            0.64366
neighbourhood_cleansedSutherland Shire   34.172470   47.643924   0.717            0.47323
neighbourhood_cleansedSydney             91.794275   41.157730   2.230            0.02574 *
neighbourhood_cleansedThe Hills Shire    -3.143500   52.789723  -0.060            0.95252
neighbourhood_cleansedWarringah          68.481451   43.462679   1.576            0.11513
neighbourhood_cleansedWaverley          127.534243   41.732566   3.056            0.00225 **
neighbourhood_cleansedWilloughby         27.268381   49.564022   0.550            0.58221
neighbourhood_cleansedWoollahra         295.352232   44.701377   6.607  0.000000000040187 ***
accommodates                             84.218657    1.769386  47.598 < 0.0000000000000002 ***
minimum_nights_avg_ntm                    0.746125    0.070647  10.561 < 0.0000000000000002 ***
availability_365                          0.455180    0.032150  14.158 < 0.0000000000000002 ***
number_of_reviews                        -0.573604    0.076686  -7.480  0.000000000000077 ***
instant_bookablet                       -43.909642    8.460918  -5.190  0.000000212778370 ***
hostfor                                  -0.008825    0.004007  -2.202            0.02767 *
host_about_word_count                     0.549174    0.074232   7.398  0.000000000000143 ***
bathrooms_count                           7.247513    0.553212  13.101 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 504.3 on 19172 degrees of freedom
Multiple R-squared:  0.2291,	Adjusted R-squared:  0.2272
F-statistic: 123.8 on 46 and 19172 DF,  p-value: < 0.00000000000000022
```

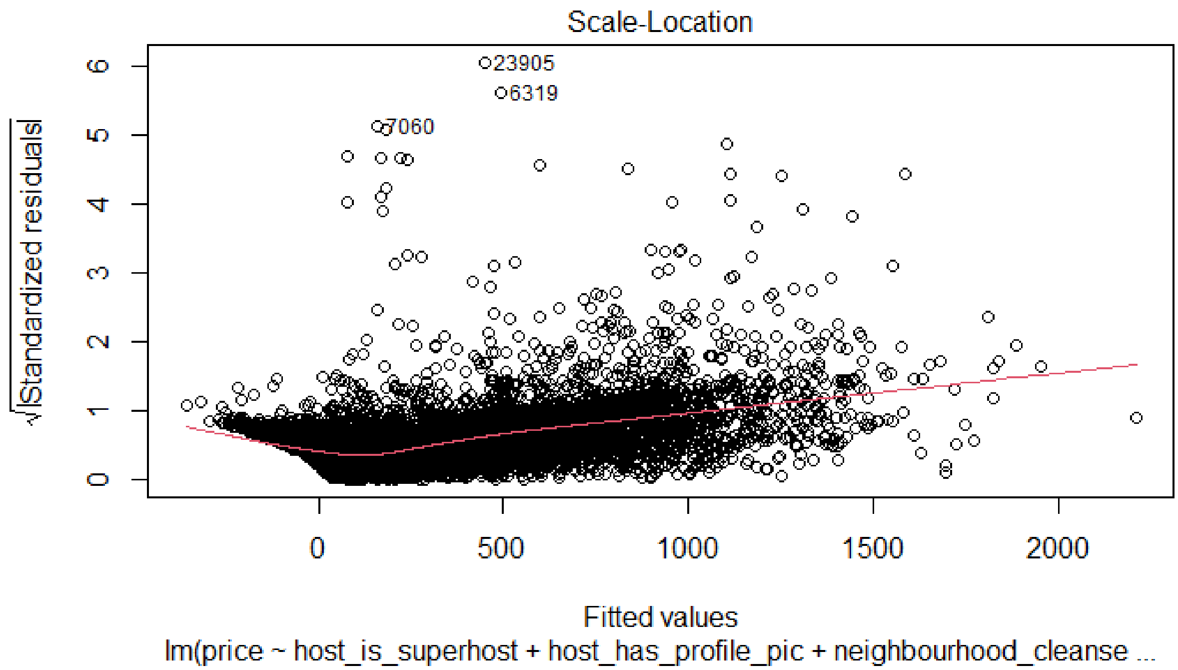
Condition Checking

Below are the conditions of the model being checked for normality and homoscedasticity.

In the first plot, checking the normality of residuals, we can see that the distribution is roughly normal, with some values trailing off on the right end, signifying that there are some outliers slightly skewing the data. We can consider the condition of normality met.



In the second plot, checking for homoscedasticity, we can see that the distribution is roughly equivalent across values of X. The residuals show homoscedasticity through an overall rectangular distribution as opposed to a wedge or patterned shape:



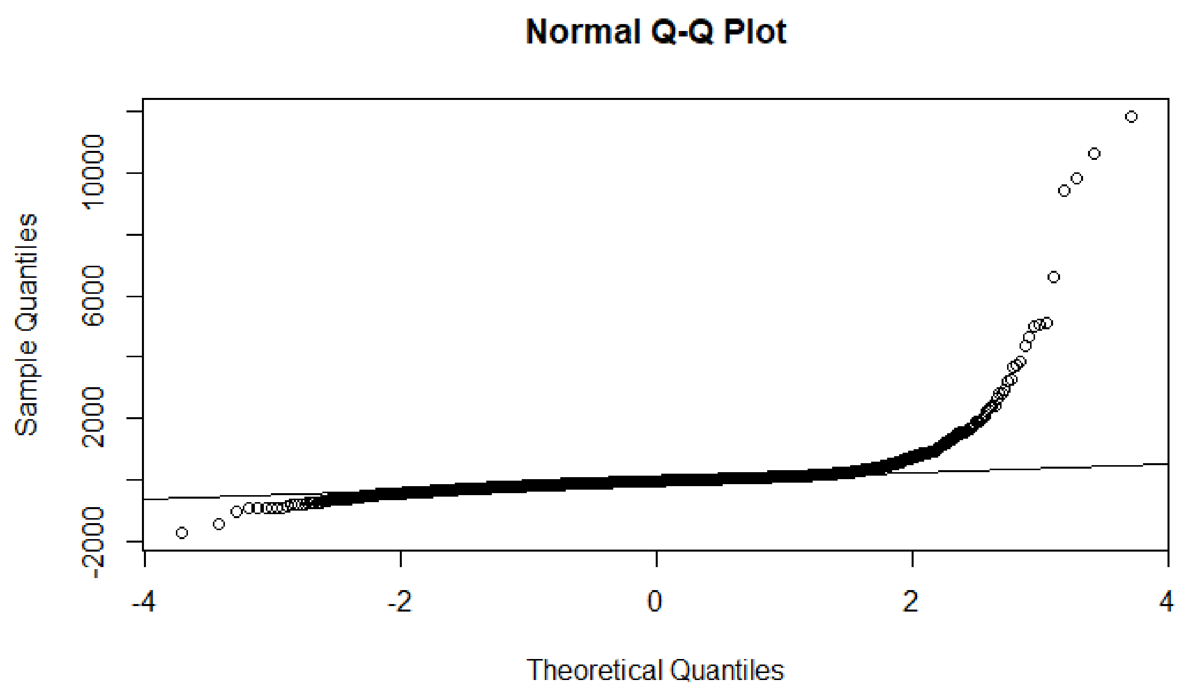
Regression Model 3 - Lasso

```{r}
coefficients(lasso_model)
```

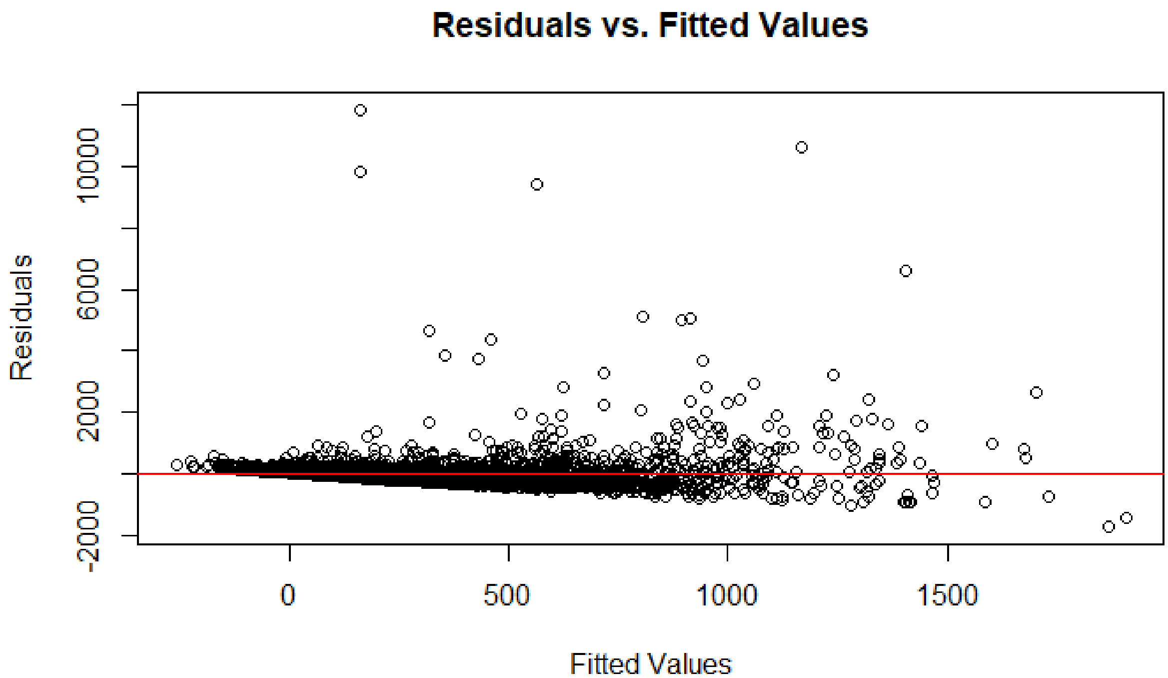
```
54 x 1 sparse Matrix of class "dgCMatrix"
                                                 s0
(Intercept)                             -127.021391972
(Intercept)                                .
host_is_superhostt                        58.241753578
host_has_profile_pict                    -39.036295080
host_identity_verifiedt                    3.945914084
neighbourhood_cleansedAuburn              -1.534836671
neighbourhood_cleansedBankstown          -71.960238701
neighbourhood_cleansedBlacktown         -149.968582472
neighbourhood_cleansedBurwood            -46.418793140
neighbourhood_cleansedCamden            -205.194691609
neighbourhood_cleansedCampbelltown      -143.075277524
neighbourhood_cleansedCanada Bay           .
neighbourhood_cleansedCanterbury         -60.286506312
neighbourhood_cleansedCity Of Kogarah    -52.159745882
neighbourhood_cleansedFairfield         -152.059616743
neighbourhood_cleansedHolroyd           -147.046363681
neighbourhood_cleansedHornsby            -44.495198190
neighbourhood_cleansedHunters Hill       -49.959911578
neighbourhood_cleansedHurstville         -68.282036353
neighbourhood_cleansedKu-Ring-Gai        -38.655698864
neighbourhood_cleansedLane Cove            .
neighbourhood_cleansedLeichhardt          43.180673599
neighbourhood_cleansedLiverpool         -199.744819488
neighbourhood_cleansedManly              109.869733221
neighbourhood_cleansedMarrickville         1.327886889
neighbourhood_cleansedMosman             324.798538153
neighbourhood_cleansedNorth Sydney        64.447660379
neighbourhood_cleansedParramatta         -78.858592627
neighbourhood_cleansedPenrith           -143.903146329
neighbourhood_cleansedPittwater          358.443414905
neighbourhood_cleansedRandwick            26.045580764
neighbourhood_cleansedRockdale           -20.016258421
neighbourhood_cleansedRyde               -39.402296910
neighbourhood_cleansedStrathfield        -33.207312663
neighbourhood_cleansedSutherland Shire    12.730351968
neighbourhood_cleansedSydney              73.343467561
neighbourhood_cleansedThe Hills Shire    -10.643178648
neighbourhood_cleansedWarringah           47.912468712
neighbourhood_cleansedWaverley           108.169479154
neighbourhood_cleansedWilloughby           6.269895343
neighbourhood_cleansedWoollahra          274.448614465
room_typeHotel room                       85.557020093
room_typePrivate room                      .
room_typeShared room                     -27.875868728
accommodates                              81.918133686
beds                                       3.697924966
minimum_nights_avg_ntm                     0.726289905
maximum_nights_avg_ntm                     .
availability_365                           0.440874457
number_of_reviews                         -0.565451734
instant_bookablet                        -43.197225693
hostfor                                   -0.005740466
host_about_word_count                      0.534443116
bathrooms_count                            7.169395762
```

Condition Checking

In the first plot, checking the normality of residuals, we can see that the distribution is roughly normal, with some values trailing off on the right end, signifying that there are some outliers slightly skewing the data. We can consider the condition of normality met.



In the second plot, checking for homoscedasticity, we can see that the distribution is roughly equivalent across values of X. The residuals show homoscedasticity through an overall rectangular distribution as opposed to a wedge or patterned shape:



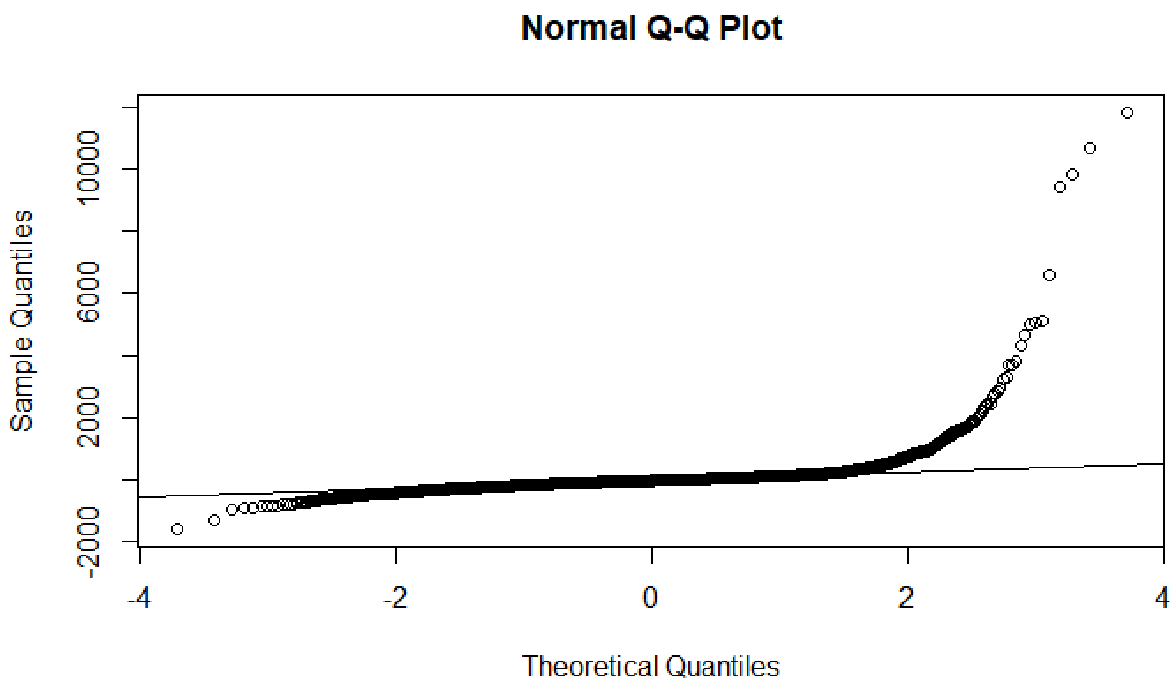
Regression Model 4 - Ridge Regression

```{r}
coefficients(ridge_model)
```

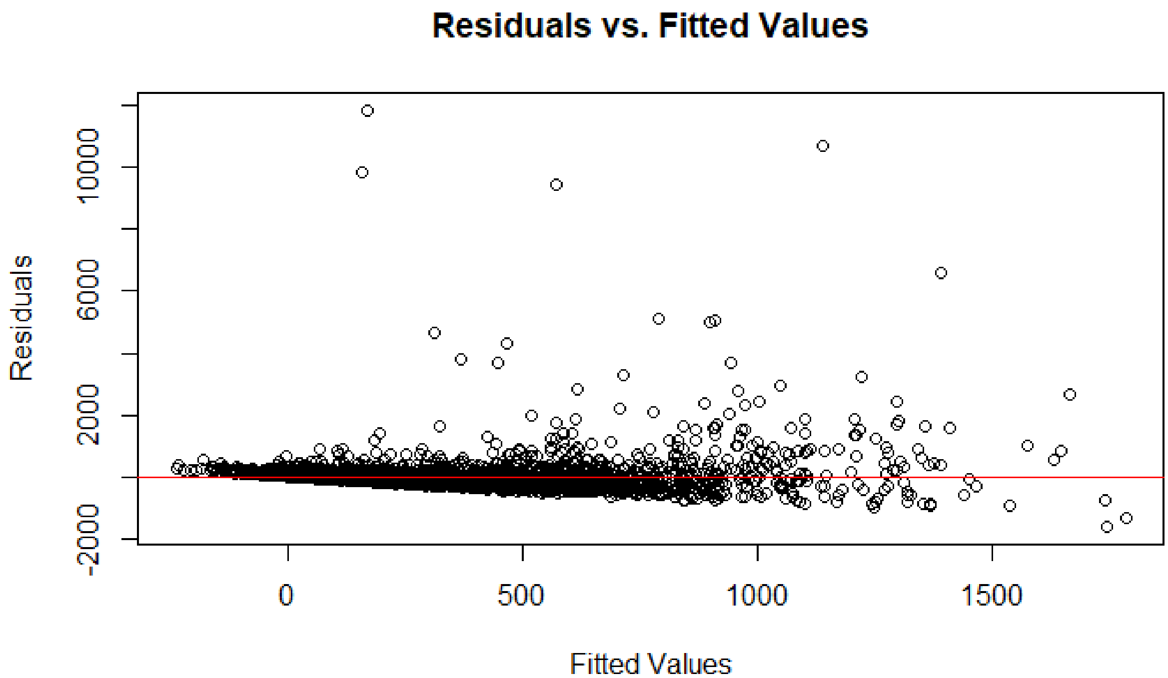
```
54 x 1 sparse Matrix of class "dgCMatrix"
                                                         s0
(Intercept)                               -79.70022691025679
(Intercept)                                 .
host_is_superhostt                         57.49951554155518
host_has_profile_pict                     -41.19494280759766
host_identity_verifiedt                     7.54885348805301
neighbourhood_cleansedAuburn              -34.76859982875165
neighbourhood_cleansedBankstown          -104.42460252178894
neighbourhood_cleansedBlacktown          -174.23026175311193
neighbourhood_cleansedBurwood             -80.56422547487989
neighbourhood_cleansedCamden             -229.82123548533949
neighbourhood_cleansedCampbelltown       -177.79362556056290
neighbourhood_cleansedCanada Bay          -33.09066860220167
neighbourhood_cleansedCanterbury          -91.57981412797251
neighbourhood_cleansedCity Of Kogarah     -83.07933644826187
neighbourhood_cleansedFairfield          -180.40363604072542
neighbourhood_cleansedHolroyd            -175.56279649244476
neighbourhood_cleansedHornsby             -77.10715460325241
neighbourhood_cleansedHunters Hill        -97.41754310213042
neighbourhood_cleansedHurstville         -102.29901268558017
neighbourhood_cleansedKu-Ring-Gai         -70.28252060210940
neighbourhood_cleansedLane Cove           -22.71397468469628
neighbourhood_cleansedLeichhardt           18.78270805125885
neighbourhood_cleansedLiverpool          -222.75137300487432
neighbourhood_cleansedManly                80.96338675685212
neighbourhood_cleansedMarrickville        -23.87398711703386
neighbourhood_cleansedMosman              292.28795800777829
neighbourhood_cleansedNorth Sydney         36.54231788653213
neighbourhood_cleansedParramatta         -106.78632780656716
neighbourhood_cleansedPenrith            -171.00736866373072
neighbourhood_cleansedPittwater           324.31184305837257
neighbourhood_cleansedRandwick             -0.98775921269761
neighbourhood_cleansedRockdale            -52.82515290064804
neighbourhood_cleansedRyde                -70.52369758897653
neighbourhood_cleansedStrathfield         -66.92386682016777
neighbourhood_cleansedSutherland Shire     -9.58221753430572
neighbourhood_cleansedSydney               44.13371589855461
neighbourhood_cleansedThe Hills Shire     -40.93509670560472
neighbourhood_cleansedWarringah            22.15669828275574
neighbourhood_cleansedWaverley             78.87173538925727
neighbourhood_cleansedWilloughby          -16.61264158024802
neighbourhood_cleansedWoollahra           240.84549466229089
room_typeHotel room                        86.10770917856608
room_typePrivate room                      -9.56997972249741
room_typeShared room                      -50.90520730454664
accommodates                               69.88444588046455
beds                                       18.25825577565410
minimum_nights_avg_ntm                      0.69143365006888
maximum_nights_avg_ntm                      0.00000001818548
availability_365                            0.42889882188484
number_of_reviews                          -0.56987852151415
instant_bookablet                         -43.55888143990754
hostfor                                    -0.00634464589715
host_about_word_count                       0.53298194774367
bathrooms_count                             7.30841786044161
```

Condition Checking

In the first plot, checking the normality of residuals, we can see that the distribution is roughly normal, with some values trailing off on the right end, signifying that there are some outliers slightly skewing the data. We can consider the condition of normality met.



In the second plot, checking for homoscedasticity, we can see that the distribution is roughly equivalent across values of X. The residuals show homoscedasticity through an overall rectangular distribution as opposed to a wedge or patterned shape:



Regression Model 5 - Elastic Net

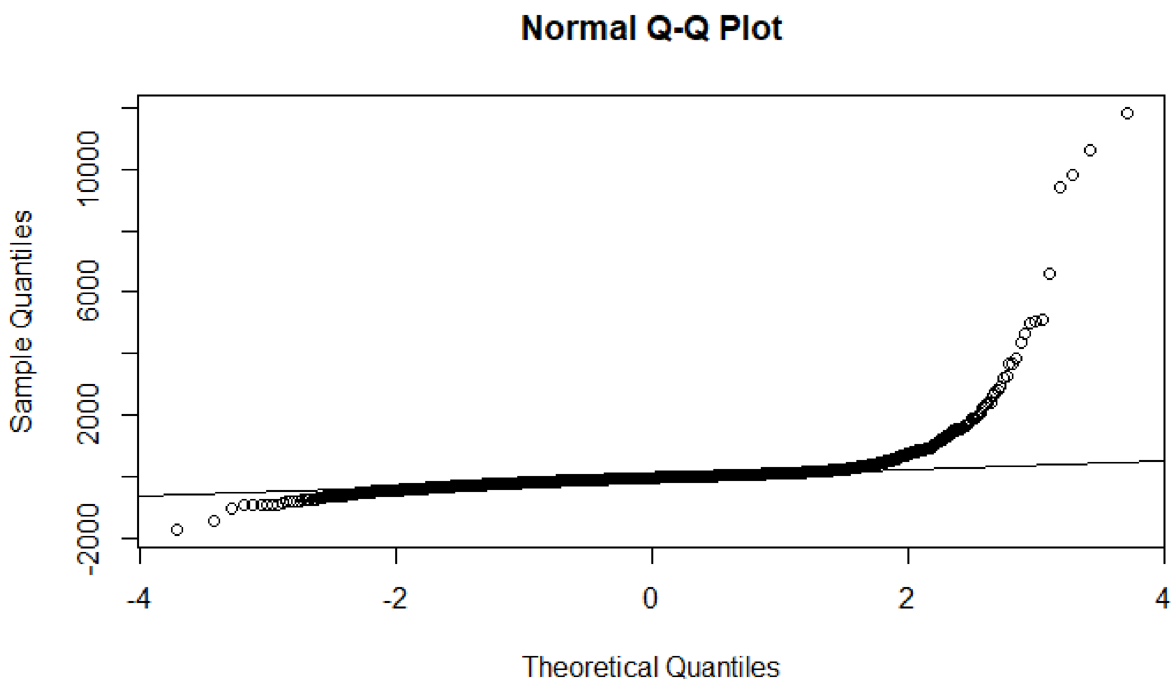
```{r}
coefficients(en_model)
```

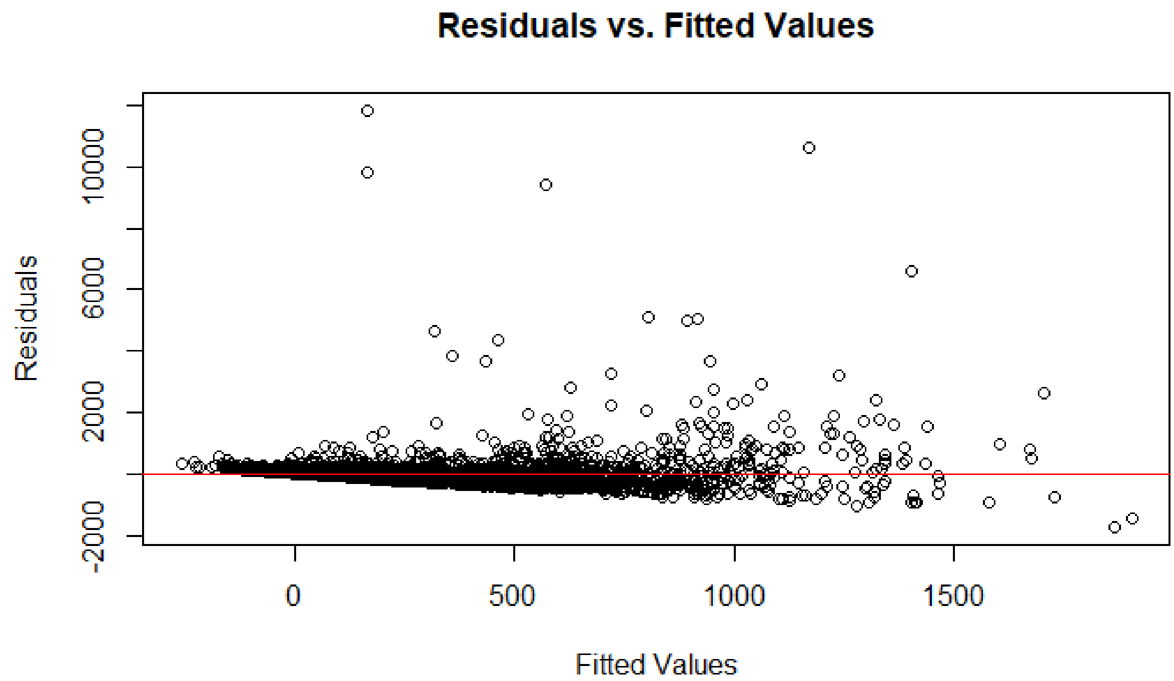
```
54 x 1 sparse Matrix of class "dgCMatrix"
                                                       s0
(Intercept)                            -129.9513690436901356
(Intercept)                               .
host_is_superhostt                       59.7051256748312156
host_has_profile_pict                   -41.2981981281808359
host_identity_verifiedt                   5.0305434090319991
neighbourhood_cleansedAuburn              .
neighbourhood_cleansedBankstown         -70.9290713294349473
neighbourhood_cleansedBlacktown        -148.4266197974990007
neighbourhood_cleansedBurwood           -44.2651757150756424
neighbourhood_cleansedCamden           -206.9488678349176780
neighbourhood_cleansedCampbelltown     -144.4305277128029275
neighbourhood_cleansedCanada Bay          .
neighbourhood_cleansedCanterbury        -58.3014839251223265
neighbourhood_cleansedCity Of Kogarah   -51.3628961728836586
neighbourhood_cleansedFairfield        -151.7419360642617221
neighbourhood_cleansedHolroyd          -148.3859248453629505
neighbourhood_cleansedHornsby           -42.0001465976951920
neighbourhood_cleansedHunters Hill      -53.7955021536839766
neighbourhood_cleansedHurstville        -67.3823531244625684
neighbourhood_cleansedKu-Ring-Gai       -36.3116061671961745
neighbourhood_cleansedLane Cove           7.4940458463433286
neighbourhood_cleansedLeichhardt         51.3790339834299843
neighbourhood_cleansedLiverpool        -198.9200645224177038
neighbourhood_cleansedManly             117.2473041336882602
neighbourhood_cleansedMarrickville        9.3501250574829804
neighbourhood_cleansedMosman            333.2370149722328847
neighbourhood_cleansedNorth Sydney       71.9532041065427421
neighbourhood_cleansedParramatta        -76.4603550689048888
neighbourhood_cleansedPenrith          -143.0719357766099051
neighbourhood_cleansedPittwater         365.2428820082163270
neighbourhood_cleansedRandwick           33.0591510856891588
neighbourhood_cleansedRockdale          -17.1541014494996169
neighbourhood_cleansedRyde              -36.5254596988682678
neighbourhood_cleansedStrathfield       -32.1578015385325173
neighbourhood_cleansedSutherland Shire   20.6783334547648892
neighbourhood_cleansedSydney             79.8052264604984032
neighbourhood_cleansedThe Hills Shire    -8.9806050628710850
neighbourhood_cleansedWarringah          55.2732296476222231
neighbourhood_cleansedWaverley          115.1664486848738278
neighbourhood_cleansedWilloughby         14.4695689813976323
neighbourhood_cleansedWoollahra         282.0453304677982942
room_typeHotel room                      90.9146818171272031
room_typePrivate room                     0.3733528441502093
room_typeShared room                    -30.8257429173374149
accommodates                             81.6763504378296830
beds                                      4.1726302601454535
minimum_nights_avg_ntm                    0.7391271837880700
maximum_nights_avg_ntm                    0.0000000002608229
availability_365                          0.4445685908711189
number_of_reviews                        -0.5740860657890728
instant_bookablet                       -43.9354270344919584
hostfor                                  -0.0068136742385417
host_about_word_count                     0.5381319563930707
bathrooms_count                           7.2032966515484436
```

Condition Checking

In the first plot, checking the normality of residuals, we can see that the distribution is roughly normal, with some values trailing off on the right end, signifying that there are some outliers slightly skewing the data. We can consider the condition of normality met.

In the second plot, checking for homoscedasticity, we can see that the distribution is roughly equivalent across values of X. The residuals show homoscedasticity through an overall rectangular distribution as opposed to a wedge or patterned shape:





Comparison of Regression Results:

	Linear(Full Model)	Linear(Stepwise Selection)	LASSO	Ridge	Elastic Net
MSE	375316.2	375306.5	375478.4	375717.7	375502.7
RMSE	612.6388	612.6227	612.7629	612.9581	612.7828
MAE	180.2903	180.1962	178.4866	177.1752	178.3627

Based on the Mean Squared Error (MSE), Root Mean Squared Error (RSME), and Mean Absolute Error (MAE), the best overall fit model is the Stepwise Selection based on it having the lowest overall MSE and RMSE. We prioritized the MSE and RMSE since our model is reasonably resistant to outliers. It is worth noting, though, that LASSO, Ridge, and Elastic Net regression all had materially lower values for MAE, so, based on one's priorities, they may choose Ridge as the best-fit model since it will be more resistant to outliers as it has the lowest MAE.

The Stepwise selection model is

Price = *90.03(neighbourhood_cleansedSydney) +*
127.53(neighbourhood_cleansedWaverley) +
295.35(neighbourhood_cleansedWoollahra) +
348.49(neighbourhood_cleansedMosman) +
378.45(neighbourhood_cleansedPittwater) +
130.95(neighbourhood_cleansedManly) + 84.21(accommodates) +
0.75(minimum_nights_avg_ntm) + 0.46(availability_365) +
0.55(host_about_word_count) + 6(host_is_superhost) + 7.24(bathoom_count) -
143.19(neighbourhood_cleansedBlacktown) -
206.91(neighbourhood_cleansedCamden) -
148.11(neighbourhood_cleansedFairfield) -
193.12(neighbourhood_cleansedLiverpool) -
138.22(neighbourhood_cleansedPenrith) - 43.91(instant_bookable) -
0.57(number_of_reviews) - 0.008(hostfor) - 174.53

Note: Many of the above variables are mutually exclusive (a listing can't be in two separate neighbourhoods simultaneously, for example). These are Boolean True (1) or False (0) data, so if a listing falls into one neighbourhood, it will only consider that variable, and all other neighbourhoods are set to 0.

Model Application

To test our model, we compared the predicted value with a few actual values from our Airbnb listings. Below are the results:

Actual listing price: $470

Output of prediction:

```
       fit       lwr      upr
2 480.0073 -524.7322 1484.747
```

Actual price: $110

Output of prediction:

```
      fit       lwr      upr
3 -133.0698 -1140.068 873.9281
```

Actual price: $130

Output of prediction:

```
      fit       lwr      upr
4 133.8099 -870.5461 1138.166
```

The model is a reasonably good predictor of price, given the statistically significant variables with some variance. This aligns with the calculated adjusted R-squared value of 0.2619. Our model can predict approximately 26.19% of the price variance for Airbnb listings in Sydney, Australia.

Conclusion

Guiding Question 1 - *Do different neighbourhoods and room types affect price?*

From our analysis of mean prices between the neighbourhoods in Sydney, we can conclude from these data that most neighbourhoods have significantly different mean prices than other neighbourhoods. However, the neighbourhood of Woollahra does not have a statistically different mean price than any other neighbourhood. When choosing an Airbnb in Sydney, we must pay attention to which neighbourhood we are staying in since the majority have significantly different mean prices.

When looking at the mean price by room type, based on these data, we can conclude that there is not a significant difference in mean price between a shared room or a private room, and there is also not a significant difference in mean price between a hotel room and an entire home/apartment. Since the mean prices are not significantly different when deciding between one of these combinations of rooms, we should be able to pick whichever we prefer without worrying about the price. All other combinations of room types have statistically significant differences in mean price.

Guiding Question 2 - *Can pricing be modelled as a linear regression of multiple factors such as the number of beds, bathrooms, neighbourhood, etc.?*

It is possible from our analysis of modelling Airbnb price as a factor of other variables. Given multiple factors in our dataset and various models, we can achieve the best model for predicting price using stepwise selection. The model that we get for this is

Price = *90.03(neighbourhood_cleansedSydney) + 127.53(neighbourhood_cleansedWaverley) + 295.35(neighbourhood_cleansedWoollahra) + 348.49(neighbourhood_cleansedMosman) + 378.45(neighbourhood_cleansedPittwater) + 130.95(neighbourhood_cleansedManly) + 84.21(accommodates) + 0.75(minimum_nights_avg_ntm) + 0.46(availability_365) + 0.55(host_about_word_count) + 6(host_is_superhost) + 7.24(bathoom_count) - 143.19(neighbourhood_cleansedBlacktown) - 206.91(neighbourhood_cleansedCamden) - 148.11(neighbourhood_cleansedFairfield) - 193.12(neighbourhood_cleansedLiverpool) - 138.22(neighbourhood_cleansedPenrith) - 43.91(instant_bookable) - 0.57(number_of_reviews) - 0.008(hostfor) - 174.53*

Using this model, we can achieve the best results from our dataset to predict the price of the Airbnbs.

References

Get the data. (n.d.). Retrieved 3 October 2023, from http://insideairbnb.com/get-the-data/

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of Data Manipulation_. R package version 1.1.1, <https://CRAN.R-project.org/package=dplyr>.

Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL https://www.jstatsoft.org/v33/i01/.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Hamner B, Frasco M (2018). _Metrics: Evaluation Metrics for Machine Learning_. R package version 0.1.4, <https://CRAN.R-project.org/package=Metrics>.