Analysis of the relation between customer reviews and the price of the property

Project 2

490620014

University of Sydney | DATA1001 | April 2020

1 Executive Summary

This study explores the reviews and prices listed for three bedroom apartments . The data has been derived from Inside Airbnb; an open source data tool . No linear trend has been observed in price based on review-scores and review-scores-accuracy .This study could potentially be used by a range of buyers.

2 Full Report

2.1 Initial Data Analysis (IDA)

The source of the data is InsideAirbnb;an independent,non commercial ,open source data tool. he data has certain limitations,and is not completely reliable. 'Accuracy of the information compiled from the Airbnb site is not the responsibility of Inside Airbnb' as mentioned on their website. Reviews are submitted by people online(sometimes anonymously) leaving the data vulnerable to damage and/or falsification. Review being subjective, the data is left open to possible biases.Some reviews may be "spam" allowed by Airbnb. Individual hosts create their own listings, and the data does not goes through complete background checks.Hence,some data may be deceitful.

Reading data.

The primary objective of this report is exploring the 3 bedroom apartments.

The dimension and structure of the variables.

2.1.1 Variable classifications

Price

```
##  Factor w/ 703 levels "$0.00","$1,000.00",..: 260 388 641 174 64 284 405 136 576 470 ...
```

Review Scores Rating

```
##  int [1:1103] 87 95 90 88 100 97 99 94 95 98 ...
```

Review Scores Accuracy

```
##  int [1:1103] 9 9 9 9 10 10 10 10 10 10 ...
```

2.1.2 Changes in Variable Classification

R has classified price as factor but as price is a measurement we will change it's classification to integer.

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.
3.0 ──
```

```
## ✓ ggplot2 3.3.0     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.5
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

```
## ── Conflicts ──────────────────────────────────────── tidyverse_conflict
s() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

2.1.3 Data Wrangling

Finding IQR and quantile in order to remove the outliers from Price

Finding IQR and quantile in order to remove the outliers from Review Scores Rating

Finding IQR and quantile in order to remove the outliers from Review Scores Accuracy

Removing the outliers

Subsetting the data;removing the outliers from specific variables

2.1.4 Final dataset

2.1.4.1 The dimension of the final dataset.

```
## [1] 594 106
```

2.1.4.2 Variable classification

Classification of variables of the final dataset are the same as before.

Variable classification of specific variables:

Price

```
##  num [1:594] 250 399 185 284 400 135 496 231 199 301 ...
```

Review Scores Rating

```
##  int [1:594] 87 95 88 97 99 94 98 100 93 92 ...
```

Review Scores Accuracy

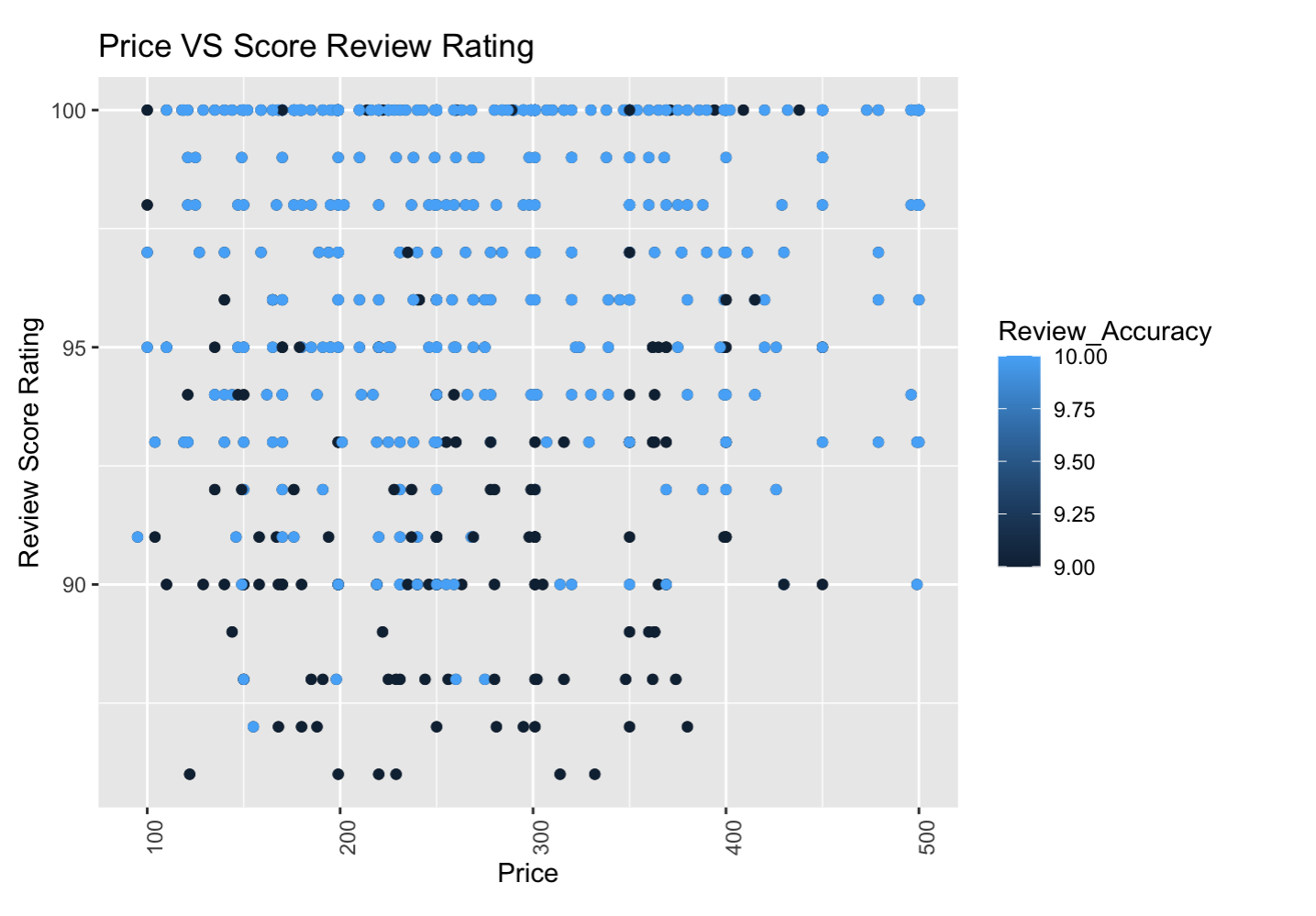
```
##  int [1:594] 9 9 9 10 10 10 10 10 9 10 ...
```

2.2 Exploring Data/ Research Question

Research Question-How are the customer reviews related to the prices of 3 bedroom apartments?

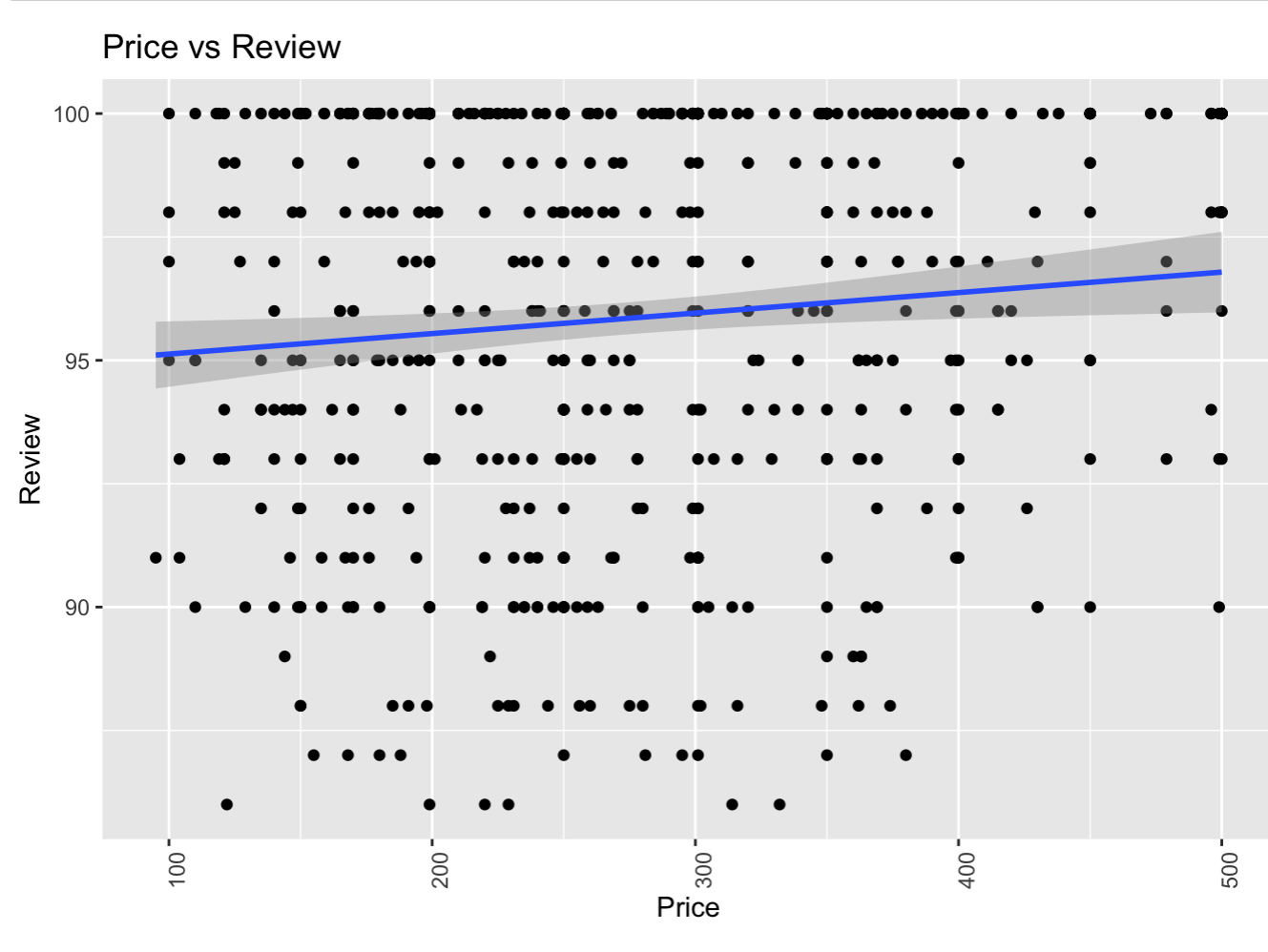
2.2.1 Graphical summaries

2.2.1.1 Scatter plot



2.2.1.2 Scatter plot and correlation coeffecient for price and review score rating

```
## `geom_smooth()` using formula 'y ~ x'
```



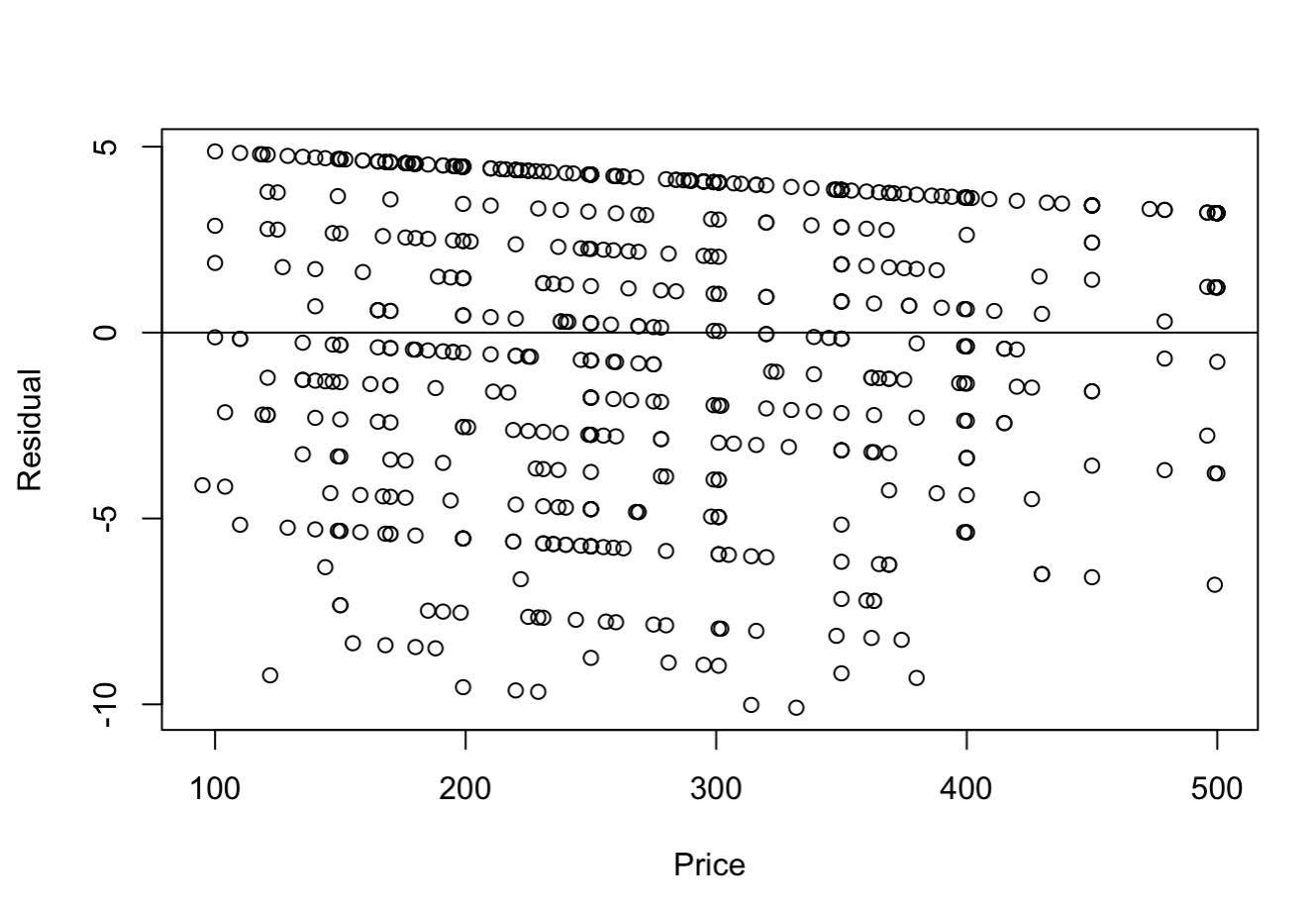

```
## [1] 0.1004594
```

2.2.1.3 Linear model summary and Residual plot

Linear model summary:

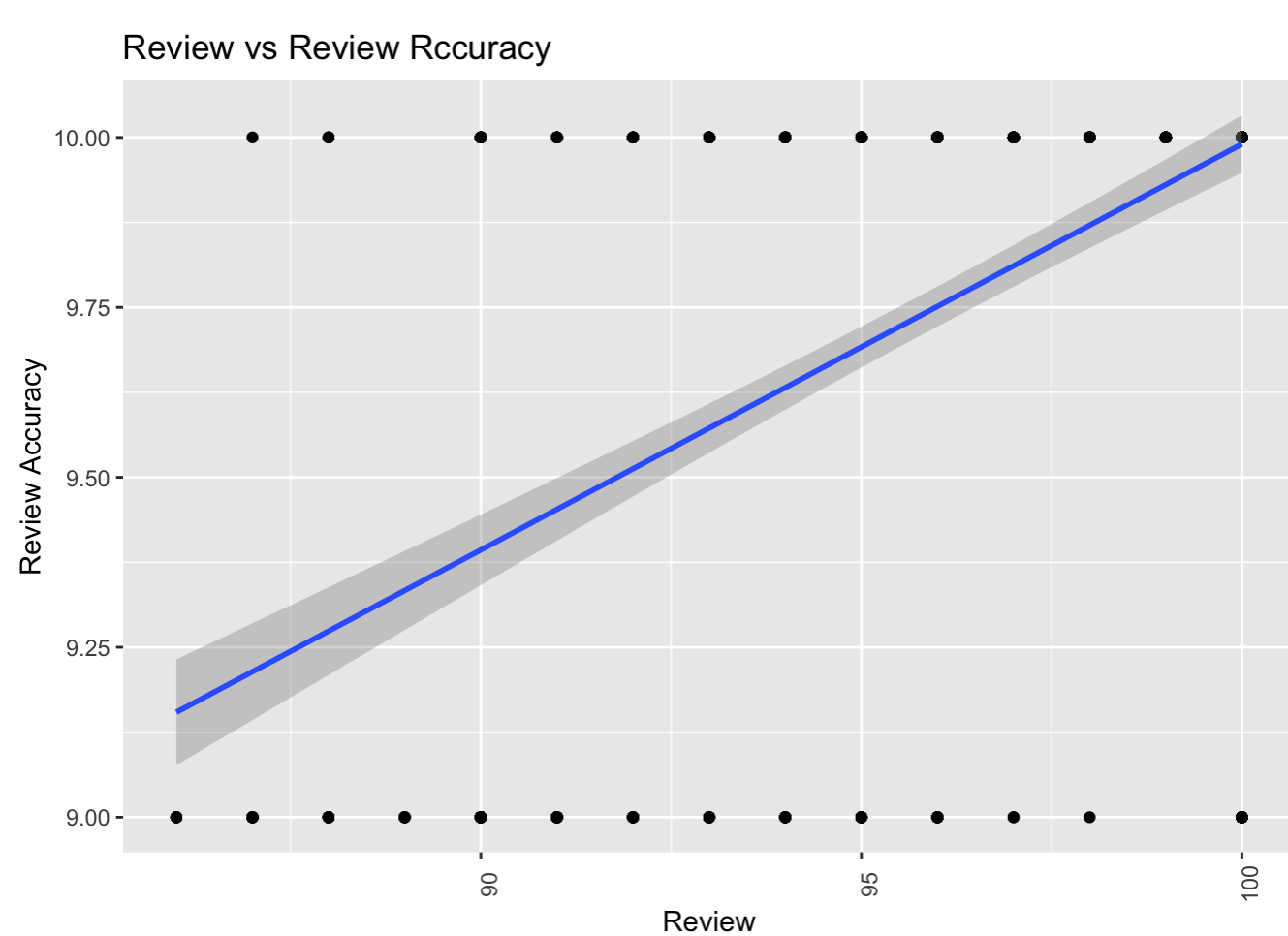
```
##
## Call:
## lm(formula = Review_Score ~ Price_of_Property)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0899  -2.9799   0.6033   3.8354   4.8731
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       94.711879   0.491779 192.590   <2e-16 ***
## Price_of_Property  0.004151   0.001690   2.457   0.0143 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.017 on 592 degrees of freedom
## Multiple R-squared:  0.01009,    Adjusted R-squared:  0.00842
## F-statistic: 6.035 on 1 and 592 DF,  p-value: 0.01431
```

Residual Plot:



2.2.1.4 Scatter plot and correlation coeffecient for review score rating and review score accuracy

```
## `geom_smooth()` using formula 'y ~ x'
```



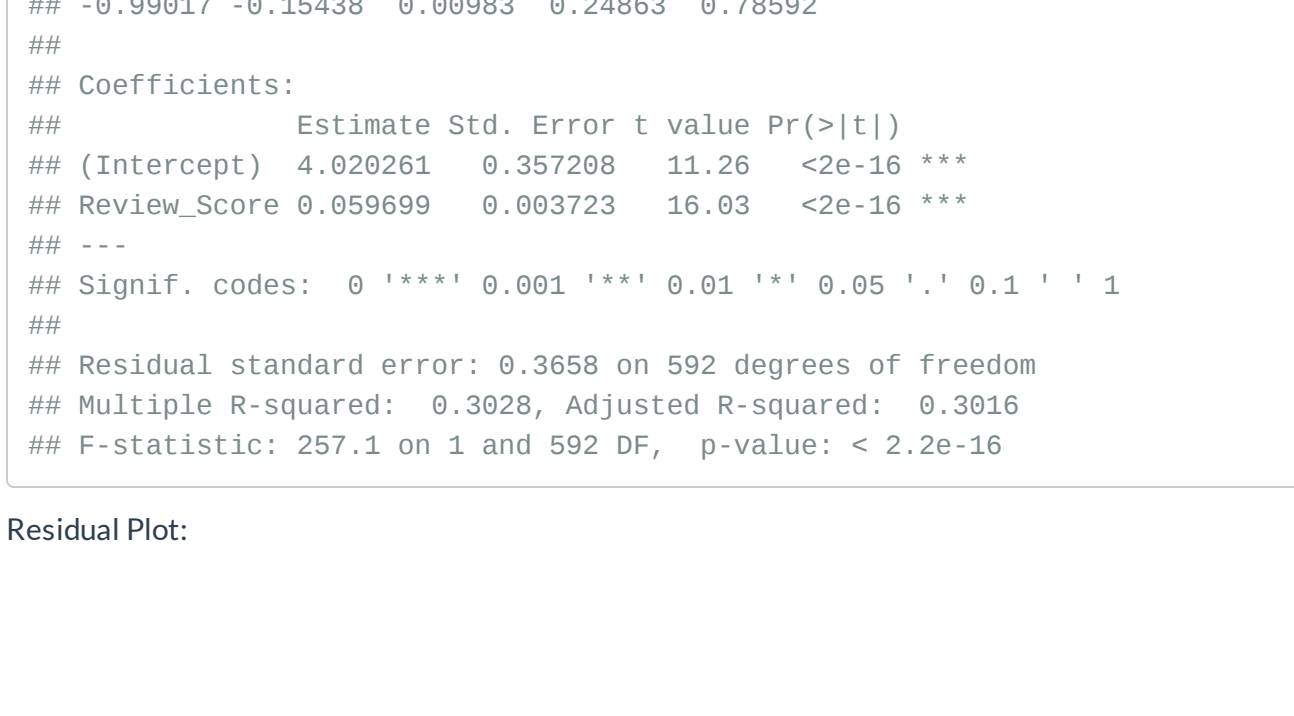

```
## [1] 0.5502394
```

2.2.1.5 Linear model summary and Residual plot

Linear model summary:

```
##
## Call:
## lm(formula = Review_Accuracy ~ Review_Score)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99017 -0.15438  0.00983  0.24863  0.78592
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.020261   0.357208   11.26   <2e-16 ***
## Review_Score 0.059699   0.003723   16.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3658 on 592 degrees of freedom
## Multiple R-squared:  0.3028, Adjusted R-squared:  0.3016
## F-statistic: 257.1 on 1 and 592 DF,  p-value: < 2.2e-16
```

Residual Plot:



2.2.1.6 Multiple linear regression model

Summary of multiple linear regression model:

```
##
## Call:
## lm(formula = Price_of_Property ~ Review_Accuracy + Review_Score)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -188.74  -79.90  -14.47   69.62  241.68
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       59.075    104.688   0.564   0.5728
## Review_Accuracy   -4.449     10.932  -0.407   0.6842
## Review_Score       2.697      1.186   2.274   0.0233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.29 on 591 degrees of freedom
## Multiple R-squared:  0.01037,    Adjusted R-squared:  0.00702
## F-statistic: 3.096 on 2 and 591 DF,  p-value: 0.04595
```

2.2.2 Analysis of data

The plot for the price and the review score rating is scattered.They have a weak correlation coeffecient which suggests that there is no linear trend for the price of a 3 bedroom apartment and its review score rating.It can be due to the fact that review score rating is a subjective measure and can differ based on a person's personal perception of things.The ratings given by a person are also open to several biases.The correlation between review scores rating and review scores accuracy has a positive correlation.The residual plot for the model suggests that there is no linear trend in the data.Multiple linear regression model for the price,review score rating and review scores accuracy has low adjusted R-squared as compared to the model for price and review scores rating.

2.2.3 Summary

The price, review score rating and review score accuracy are not linearly associated.

2.3 Domain knowledge and research

Airbnb is a privately owned accommodation rental website. It is online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any real estate listings, nor does it host events; it acts as broker, receiving commissions from each booking. The dataset used for this project comes from Insideairbnb.com, an anti-Airbnb lobby group that scrapes Airbnb listings, reviews and calendar data from multiple cities. This report explores the relation between price and review of 3 bedroom apartments.Consumer reviews may reflect not only perceived quality but also the difference between quality and price (perceived value).

2.4 References

Agarwal & Peshin,2018, Exploratory Data Analysis and Visualization of Airbnb Dataset,Columbia.edu, http://www.columbia.edu/~sg3637/airbnb_final_analysis.html

Chawla,2019,Data Analysis on the AirBnb Dataset,Medium, ,https://medium.com/ml2vec/data-analysis-on-the-airbnb-dataset-e0be9254eeb9

Chesky & Gebbia & Blecharkzyc ,2008, Airbnb Official Website, https://www.airbnb.com.au/

Cox, 2014, Inside Airbnb,http://insideairbnb.com/

ggplot visualisations, http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

Gupta,2019, Airbnb Rental Listings Dataset Mining,Medium,https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec

Li and Lorin M.,Hitt MIS Quarterly,Vol. 34, No. 4 (December 2010),Price Effects in Online Product Reviews: An Analytical Model and Empirical Analysis, pp. 809-831

2019,Confounding, https://catalogofbias.org/biases/confounding/