# Lab 5

Tim Lanthier

2/16/2022

Github Repository: https://github.com/talanthier/lab-05

## Lab 5: Data Wrangling & Regression

The data we will be working with is Airbnb listings. The data can be obtained through the following link:

http://insideairbnb.com/get-the-data.html

For this lab we will only be using the listings data from Santa Cruz County, CA.

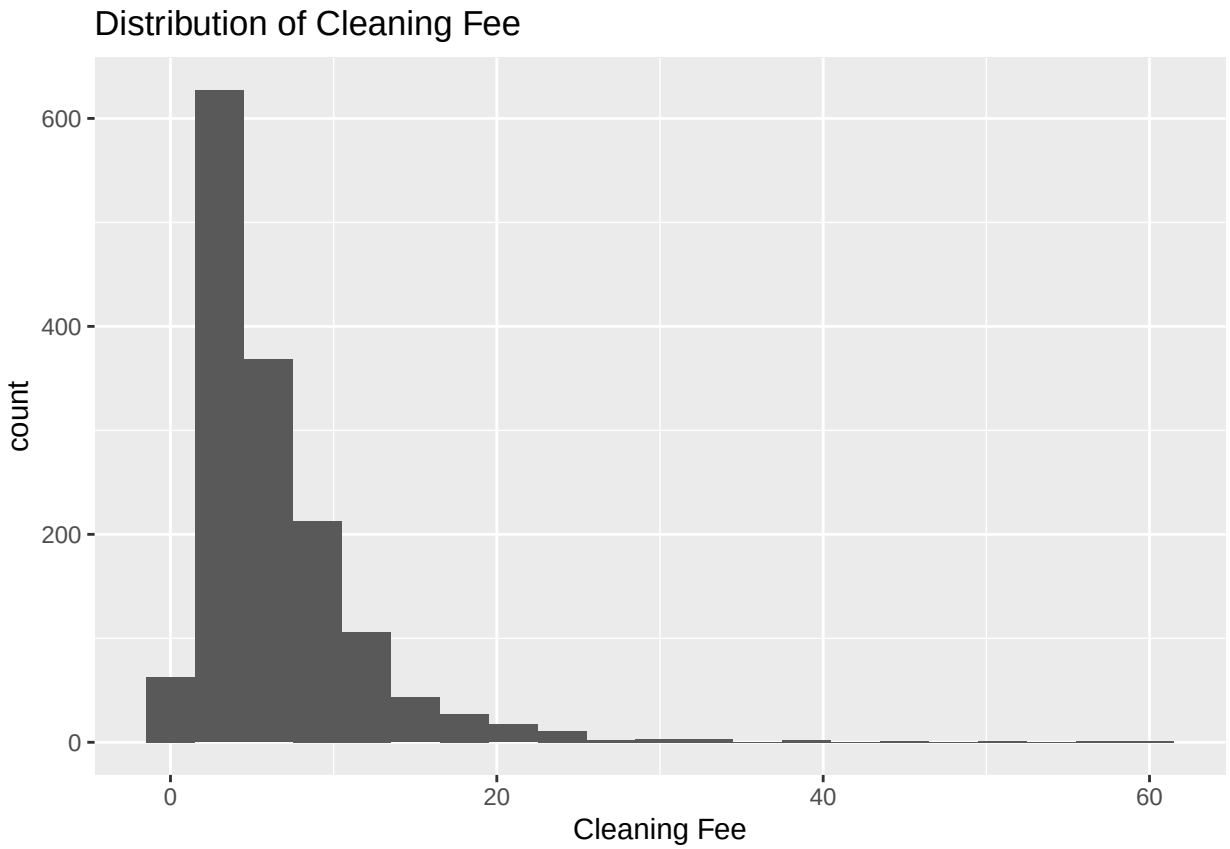
### Data Wrangling & EDA

Since some airbnb rentals have cleaning fees, we will create a new variable `cleaning_fee` which is 2% of the price per night.

```
airbnb <- read.csv('data/listings.csv') %>%
  mutate(cleaning_fee = 0.02*price)
```

The distribution for `cleaning_fee` is shown below.

```
ggplot(data = airbnb, aes(cleaning_fee)) +
  geom_histogram(binwidth = 3) +
  labs(x = 'Cleaning Fee', title = 'Distribution of Cleaning Fee')
```



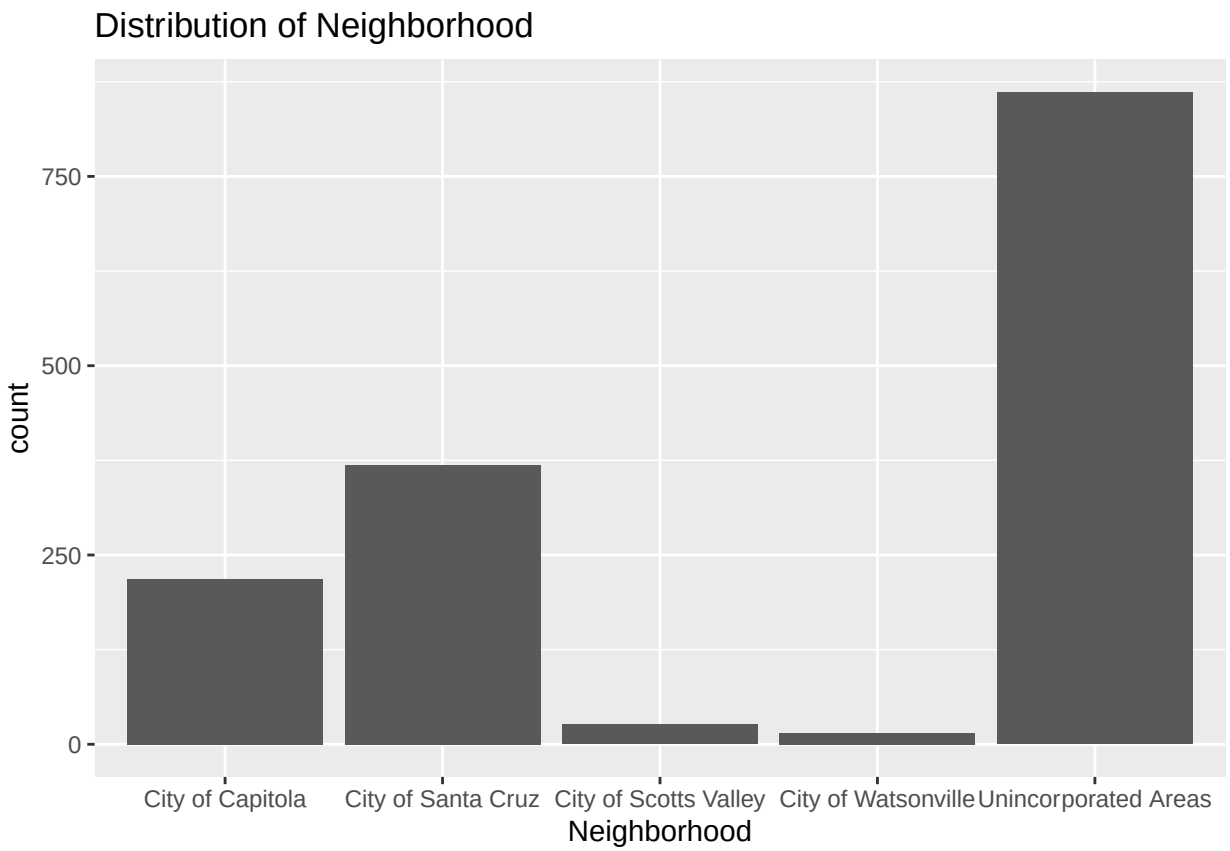

```
summarise(airbnb, mean = mean(cleaning_fee),
          sd = sd(cleaning_fee),
          min = min(cleaning_fee),
          q1 = quantile(cleaning_fee, 0.25),
          median = median(cleaning_fee),
          q3 = quantile(cleaning_fee, 0.75),
          max = max(cleaning_fee),
          IQR = q3-q1)
```

```
##       mean       sd  min   q1 median   q3 max  IQR
## 1 6.377582 5.394347 0.62 2.88      5 8.06  59 5.18
```

Looking at the above histogram, we see that `cleaning_fee` is heavily positively skewed as shown with most of the data being concentrated around 5 with a long right hand tail. This is supported by the fact that we have a median of 5 with a larger mean of 6.378. We also appear to have quite a few outliers. Looking at our quantiles, 75% of listings' cleaning fees are below $8.06, but we have a maximum cleaning fee of $59.

Now we will take a look at `neighbourhood`.

```
ggplot(data = airbnb, aes(neighbourhood)) +
  geom_bar() +
  labs(x = 'Neighborhood', title = 'Distribution of Neighborhood')
```




So we have 5 different values for `neighbourhood`: City of Capitola, City of Santa Cruz, City of Scotts Valley, City of Watsonville, and Unincorporated Areas. As shown by the above plot, most of the listings within Santa Cruz County are in Unincorporated Areas. The city of Santa Cruz and Capitola have the second and third most listings respectively. Meanwhile Scotts Valley and Watsonville have significantly fewer listings than the rest.

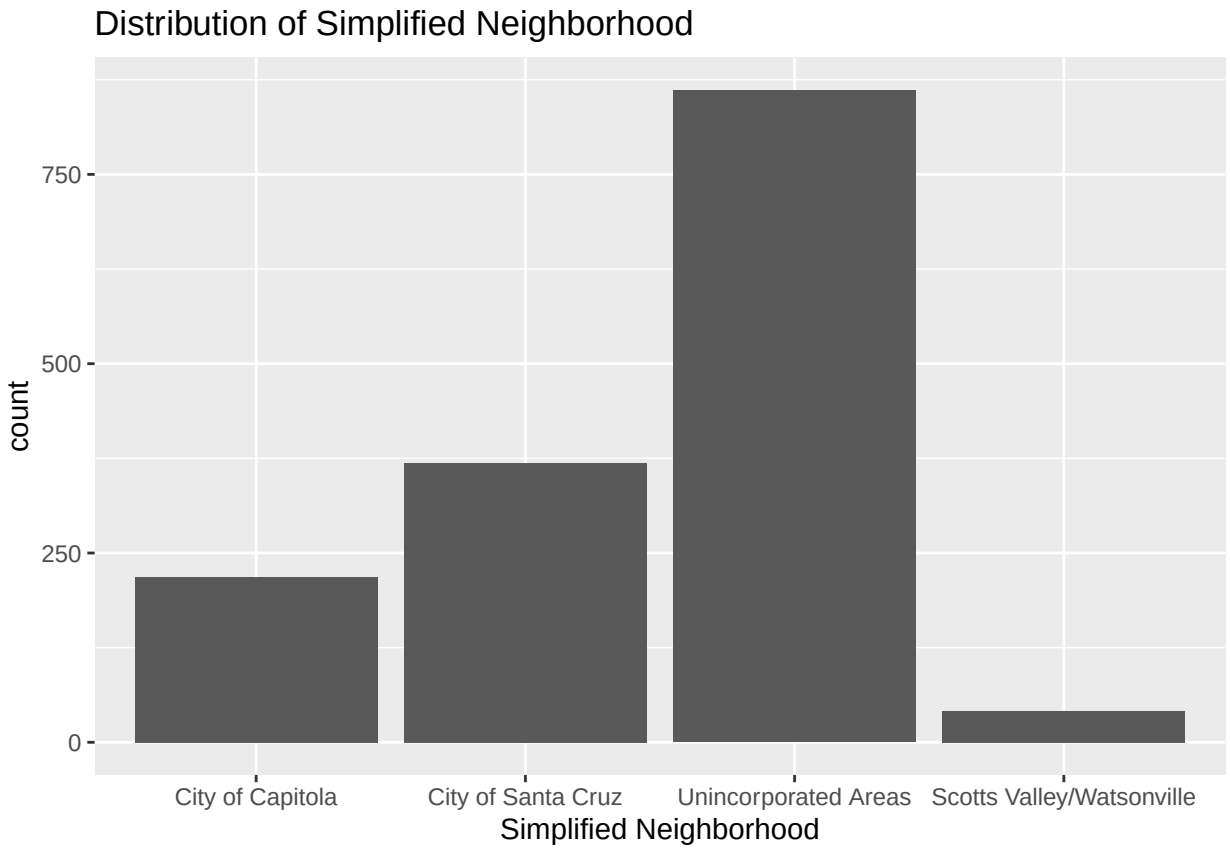
```
airbnb %>% count(neighbourhood) %>% mutate(percent = n/sum(n))
```

```
##           neighbourhood   n    percent
## 1      City of Capitola 218 0.14640698
## 2    City of Santa Cruz 369 0.24781733
## 3 City of Scotts Valley  26 0.01746138
## 4   City of Watsonville  15 0.01007388
## 5  Unincorporated Areas 861 0.57824043
```

As shown in the output above, Unincorporated areas, the City of Santa Cruz, and the City of Capitola account for 57.8%, 24.8%, and 14.6% of listings in Santa Cruz County respectively. Meanwhile the cities of Scotts Valley and Watsonville combined account for less than 3% of Santa Cruz County Airbnb listings. Seeing as Scotts Valley and Watsonville account for so few listings, we will lump them into a single category as the new variable `neigh_simp`.

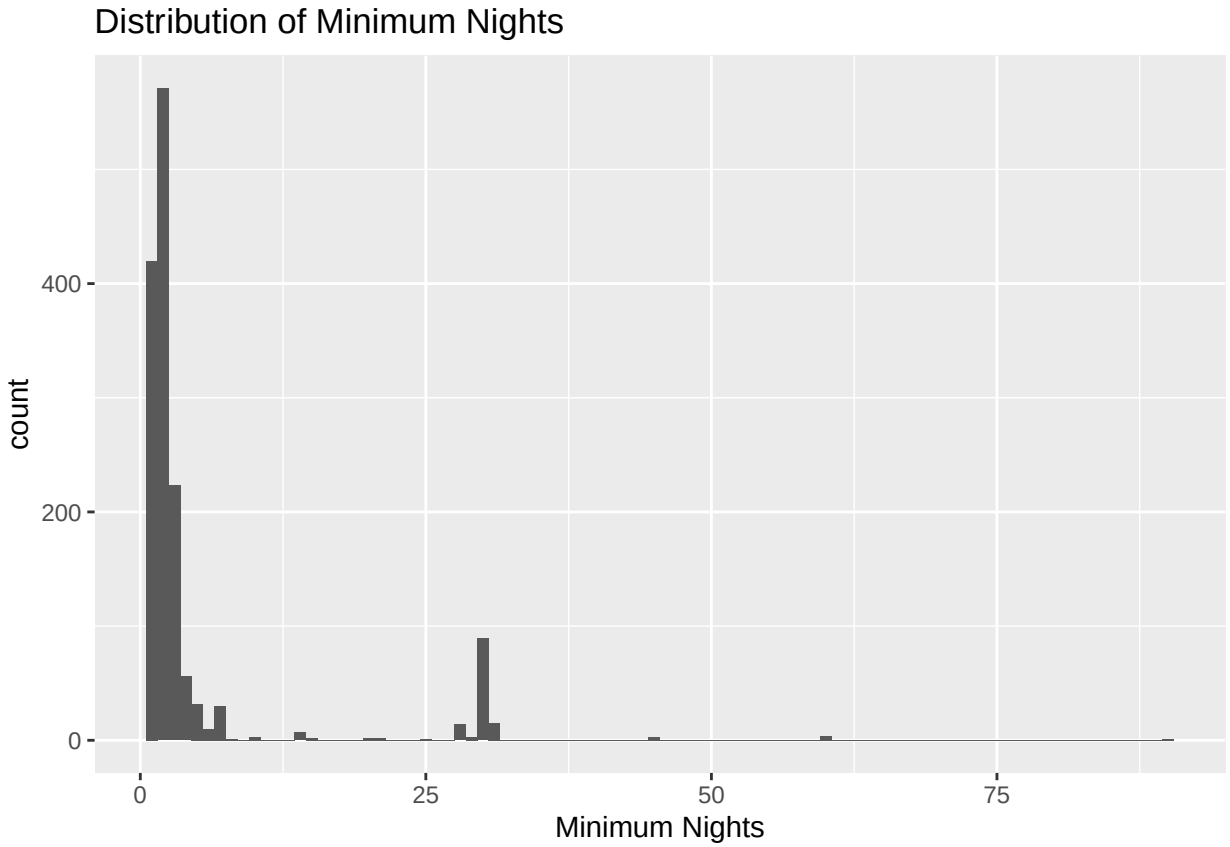
```
airbnb <- airbnb %>%
  mutate(neigh_simp = fct_lump_n(neighbourhood,3,other_level = 'Scotts Valley/Watsonville'))

ggplot(data = airbnb, aes(neigh_simp)) +
  geom_bar() +
  labs(x = 'Simplified Neighborhood', title = 'Distribution of Simplified Neighborhood')
```




Now we will take a look at `minimum_nights`.

```
ggplot(data = airbnb, aes(minimum_nights)) +
  geom_histogram(binwidth = 1) +
  labs(x = 'Minimum Nights', title = 'Distribution of Minimum Nights')
```

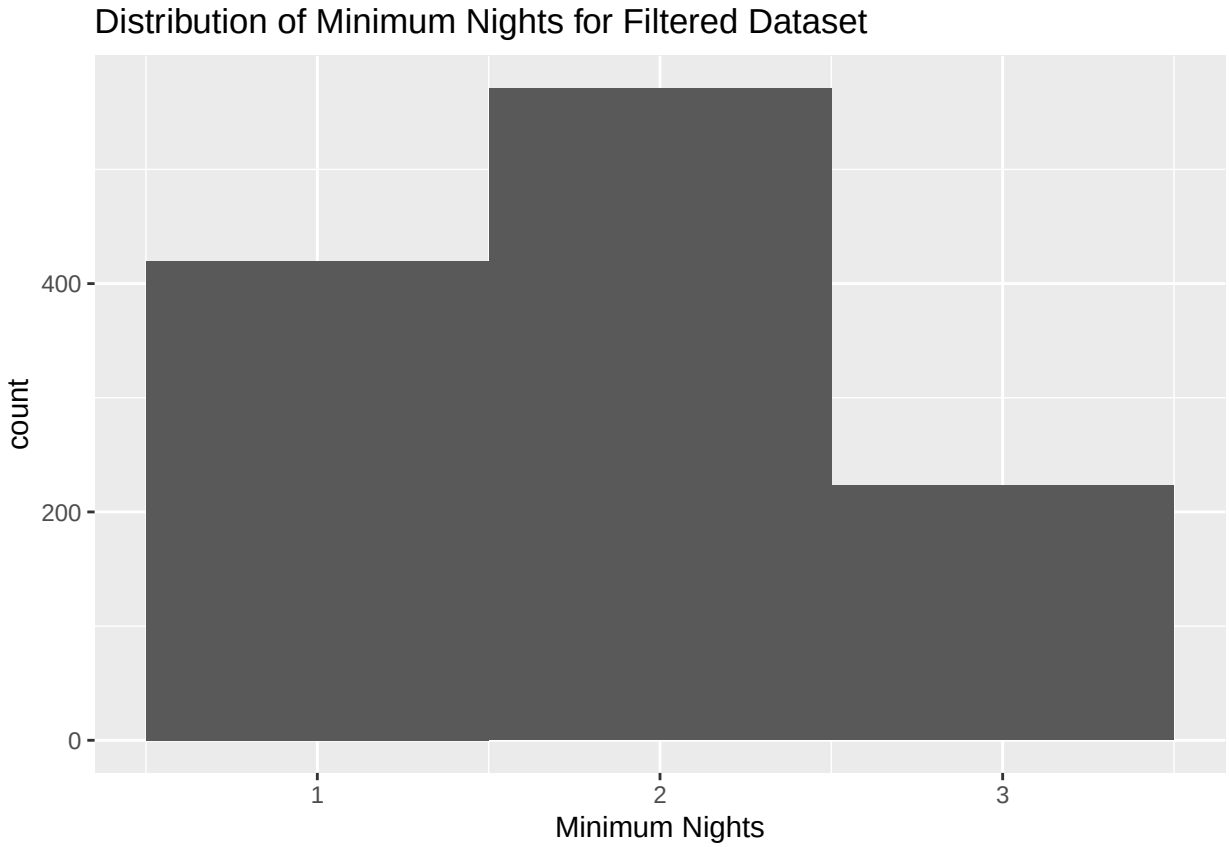



```
sort(summary(as.factor(airbnb$minimum_nights)),decreasing = TRUE)
```

```
##   2   1   3  30   4   5   7  31  28   6  14  60  10  29  45  15  20  21   8  25
## 571 420 223  89  56  32  30  15  14  10   7   4   3   3   3   2   2   2   1   1
##  90
##   1
```

The distribution for `minimum_nights` is shown above. Once again, the distribution for minimum number of nights is positively skewed with most of the data concentrated near 0, but we do have a spike around the 30 mark. Looking at the counts for each category, the 4 most common values for minimum nights is 2,1,3, and 30. The one that stands out of course is 30 which is the unusual spike we noticed earlier. While listings with 1,2, or 3 minimum nights are likely used for tourists, those with a value of 30 minimum nights is likely intended for those who are looking for more long term housing in the Santa Cruz area. Since we want to focus on listings intended for travel purposes, for the remainder of the lab we will filter out observations with a minimum nights of over 3.

```
airbnb <- airbnb %>% filter(minimum_nights <= 3)
ggplot(airbnb, aes(minimum_nights)) +
  geom_histogram(binwidth = 1) +
  labs(x = 'Minimum Nights', title = 'Distribution of Minimum Nights for Filtered Dataset')
```




We will use this dataset for the remainder of the lab.

**Regression**

We will start by defining a new variable `price_3_nights` which is the price to stay at each location for 3 nights.

```
airbnb <- airbnb %>%
  mutate(price_3_nights = 3*price+cleaning_fee)
```

Now we will create a linear regression model for price_3_nights.

```
model1 <- lm(price_3_nights ~ neigh_simp + number_of_reviews + reviews_per_month, data = airbnb)
tidy(model1, conf.int = TRUE) %>%
  kable(digits = 3, format = 'markdown')
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 1475.380 | 65.136 | 22.651 | 0.000 | 1347.580 | 1603.181 |
| neigh_simpCity of Santa Cruz | -208.001 | 75.923 | -2.740 | 0.006 | -356.966 | -59.036 |
| neigh_simpUnincorporated Areas | -312.632 | 65.758 | -4.754 | 0.000 | -441.652 | -183.613 |
| neigh_simpScotts Valley/Watsonville | -671.550 | 159.777 | -4.203 | 0.000 | -985.040 | -358.059 |
| number_of_reviews | -0.437 | 0.202 | -2.158 | 0.031 | -0.834 | -0.040 |
| reviews_per_month | -85.171 | 12.564 | -6.779 | 0.000 | -109.821 | -60.520 |

For `number_of_reviews`, we have a coefficient of -0.437 and a confidence interval of (-0.832, -0.040). This means that holding all else constant, if the number of reviews for a listing were to increase by 1, we would

expect the price of the listing for 3 nights to decrease by $0.437. As for the confidence interval, this means we are 95% confident that the true value for the coefficient of `number_of_reviews` lies within the interval (-0.832, -0.040).

For `neigh_simpCity of Santa Cruz`, we have a coefficient of -208.001 and a confidence interval of(-356.966, -59.036). This means that keeping all other characteristics of the listing the same, we would expect that the listing that is in the City of Santa Cruz would have a 3 day price of $208.001 less than the same listing which is located in the City of Capitola. According to the confidence interval, we are 95% confident that this coefficient representing the difference in 3 day price for the same listings between the City of Santa Cruz and Capitola woud lie between -$356.966 and -$59.036.

Interpreting the intercept, we would say that with a listing which is in the City of Capitola which has 0 reviews and 0 reviews per month, we would expect that the listing would have a 3 night price of $1475.38. Seeing that it is possible that a listing has these characteristics, this interpretation makes sense in the context of this problem. It is essentially stating what we should expect to pay for a listing if we have no review information.

Now suppose we find an Airbnb which has 5.14 reviews per month, 10 total reviews, and is in the Scotts Valley.

```
new_obs <- data.frame(neigh_simp = 'Scotts Valley/Watsonville', number_of_reviews = 10, reviews_per_mon
predict(model1, new_obs, interval = 'prediction')
```

```
##        fit       lwr      upr
## 1 361.6874 -1121.781 1845.156
```

Using our model, we would predict that we would pay $361.68 for a 3 night stay at this listing. We also have a confidence interval of (-1121.781, 1845.156). We would say that we are 95% confident that the price of an Airbnb listing in Scotts Valley with 5.14 reviews per month and 10 total reviews is between -$1121.78 and $1845.16. Seeing as 3 day prices cannot be negative, the interpretation of this confidence interval doesn't make sense.

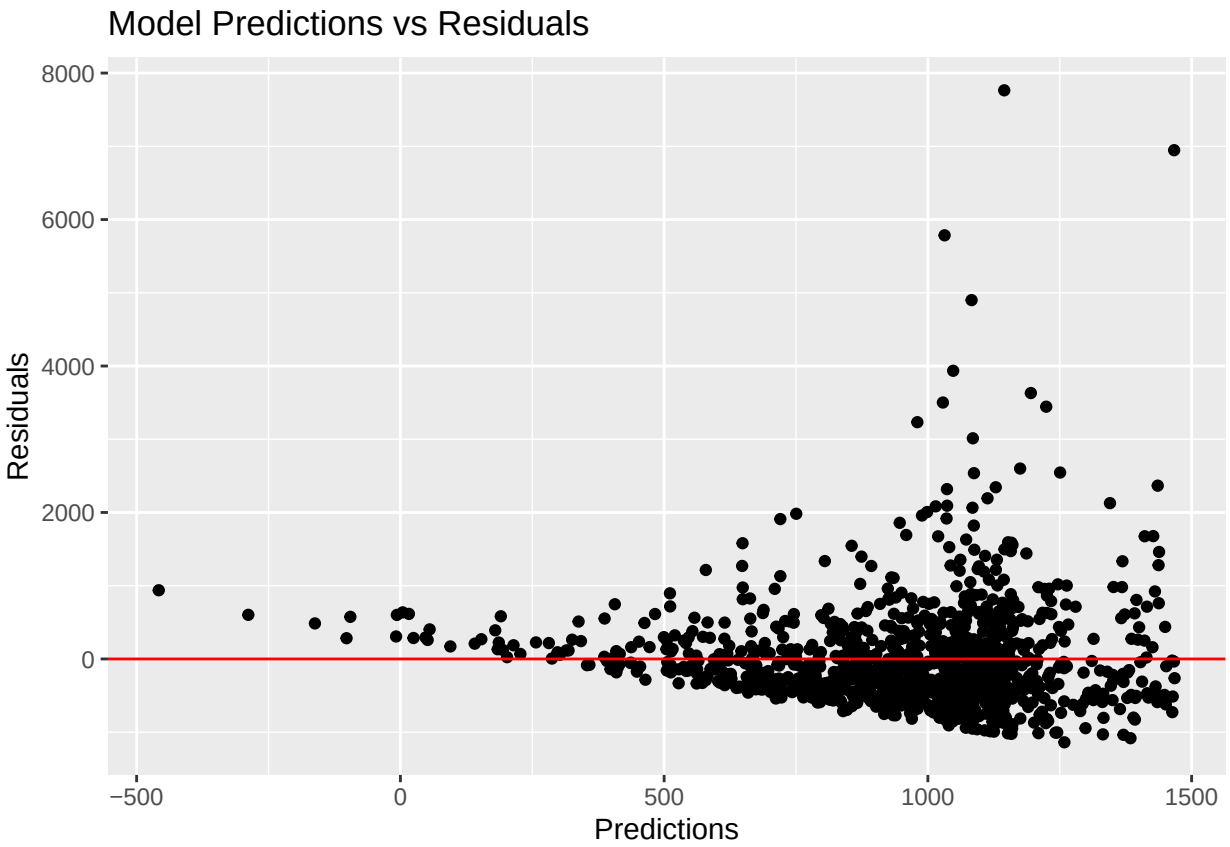
### Checking Assumptions

Now we will check the assumptions for our model. We will start by looking at the linearity and constant variance assumptions.

```
airbnb_model1 <- augment(model1)
glimpse(airbnb_model1)
```

```
## Rows: 1,143
## Columns: 11
## $ .rownames         <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "~
## $ price_3_nights    <dbl> 480.18, 274.82, 302.00, 308.04, 1032.84, 283.88, 634~
## $ neigh_simp        <fct> Unincorporated Areas, Unincorporated Areas, City of ~
## $ number_of_reviews <int> 1623, 85, 510, 495, 119, 446, 637, 542, 820, 340, 48~
## $ reviews_per_month <dbl> 10.71, 0.61, 3.58, 3.59, 0.88, 3.36, 4.84, 4.24, 6.4~
## $ .fitted           <dbl> -458.0866, 1073.6800, 739.7850, 745.4828, 1140.4696,~
## $ .resid            <dbl> 938.26661, -798.85997, -437.78503, -437.44284, -107.~
## $ .hat              <dbl> 0.123372355, 0.002346508, 0.013920203, 0.013110303, ~
## $ .sigma            <dbl> 739.9700, 740.1868, 740.4516, 740.4519, 740.5602, 74~
## $ .cooksd           <dbl> 4.298749e-02, 4.576209e-04, 8.345364e-04, 7.834660e-~
## $ .std.resid        <dbl> 1.35377137, -1.08045689, -0.59556823, -0.59485847, -~
```

```
ggplot(data = airbnb_model1, aes(x = .fitted, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, color = 'red') +
  labs(x = 'Predictions', y = 'Residuals', title = 'Model Predictions vs Residuals')
```



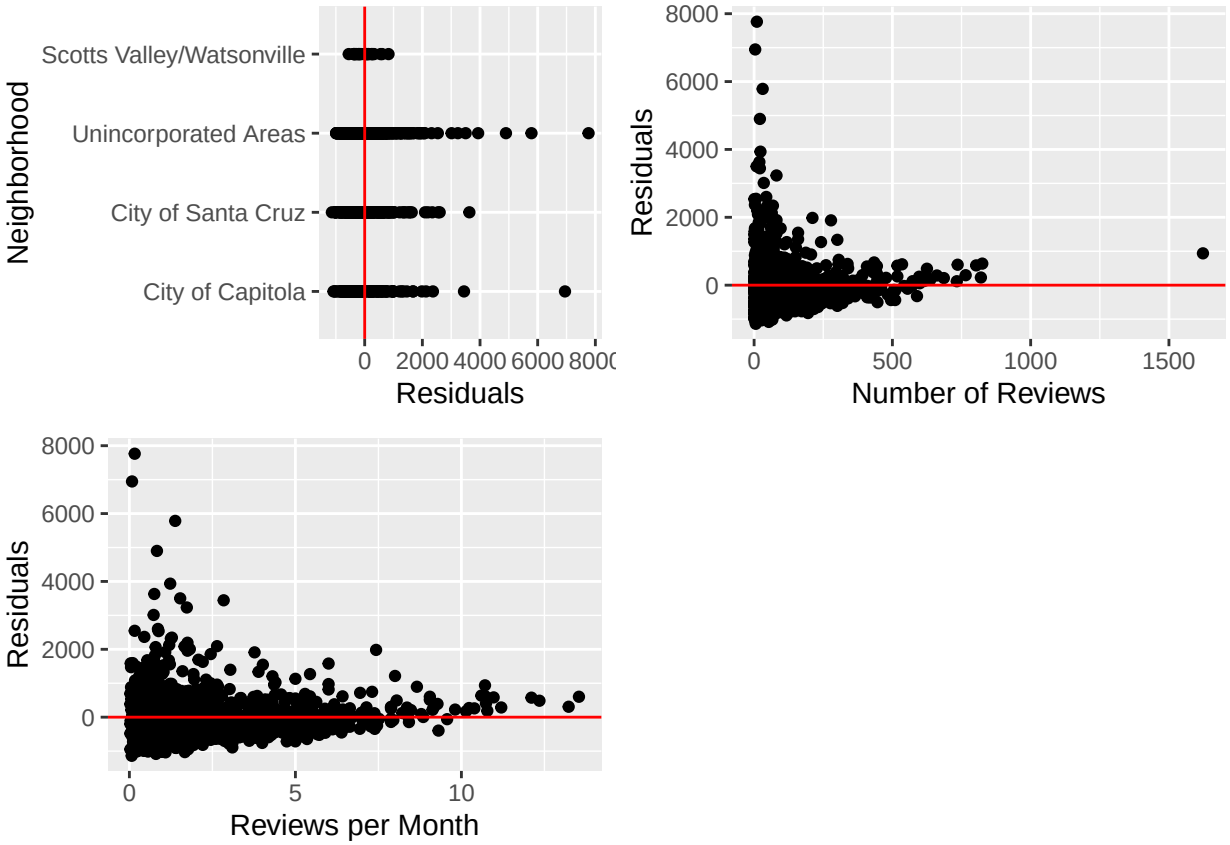

Looking at the plot of our predictions against the residuals, it looks like there is a pattern. With very low predictions, we tend to have positive residuals. Meanwhile with larger predictions most of the residuals lie below 0.

```
p1 <- ggplot(data = airbnb_model1, aes(x = neigh_simp, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, color = 'red') +
  labs(x = 'Neighborhood', y ='Residuals')+
  coord_flip()

p2 <- ggplot(data = airbnb_model1, aes(x = number_of_reviews, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, color = 'red') +
  labs(x = 'Number of Reviews', y = 'Residuals')

p3 <- ggplot(data = airbnb_model1, aes(x = reviews_per_month, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0, color = 'red') +
  labs(x = 'Reviews per Month', y = 'Residuals')

plot_grid(p1,p2,p3)
```

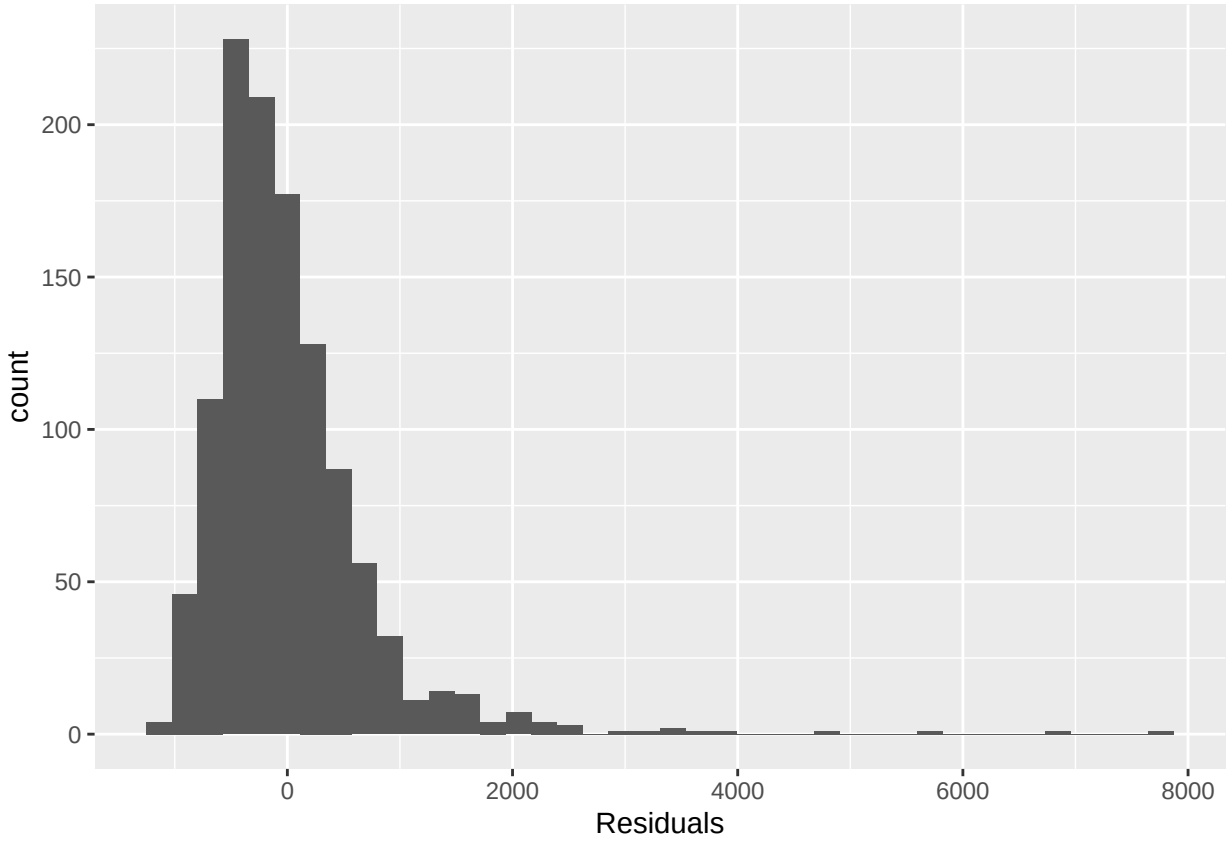



Once again, we can see a pattern between reviews per month as well as number of reviews and the residuals. Since there is a clear pattern between the residuals and some of the predictor variables as well as the predictions, we would claim that the linearity assumptions is not satisfied and a linear regression model is not optimal for modeling the relationship between these variables.

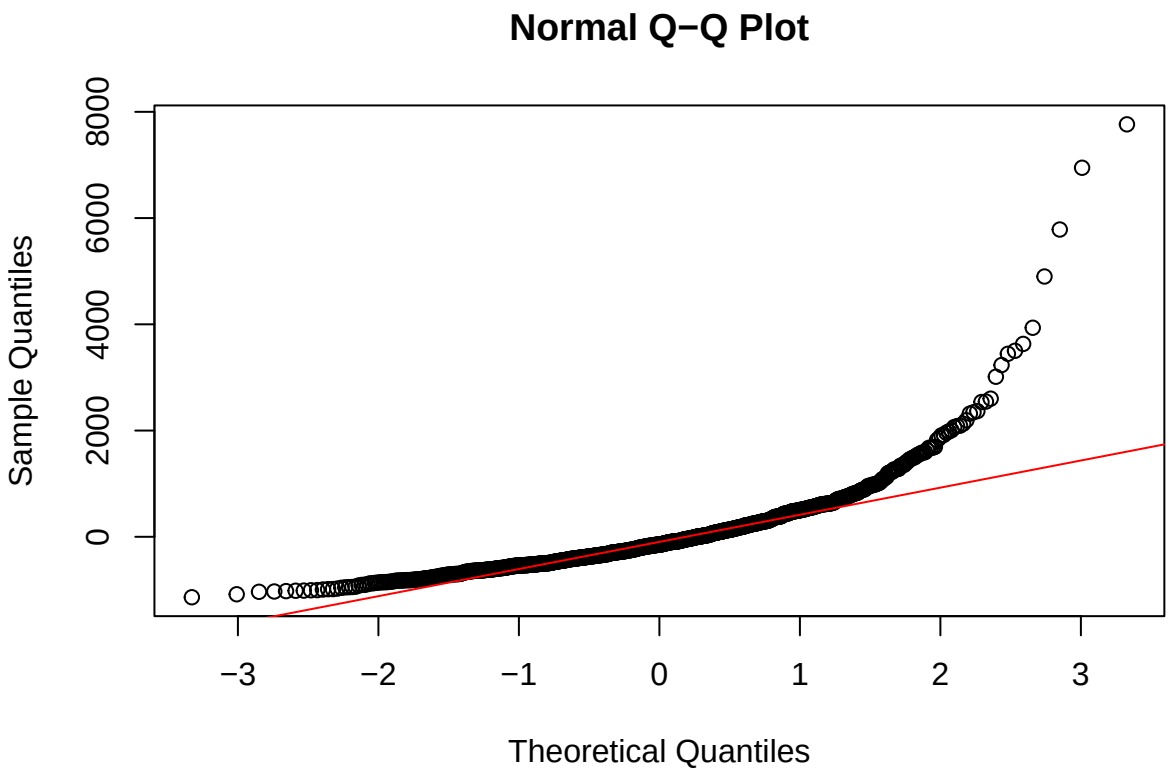
Looking at the plot of our model's residuals against its predictions, we see that with lower predicted values, we have a very tight spread for the residuals, but for larger predictions we have a much larger spread. Seeing as the spread around the horizontal line at 0 is not constant for different values for our prediction, the constant variance assumption has been violated.

Now we will evaluate whether the normality assumption has been satisfied.

```
ggplot(data = airbnb_model1, aes(.resid)) +
  geom_histogram(bins = 40) +
  labs(x='Residuals')
```



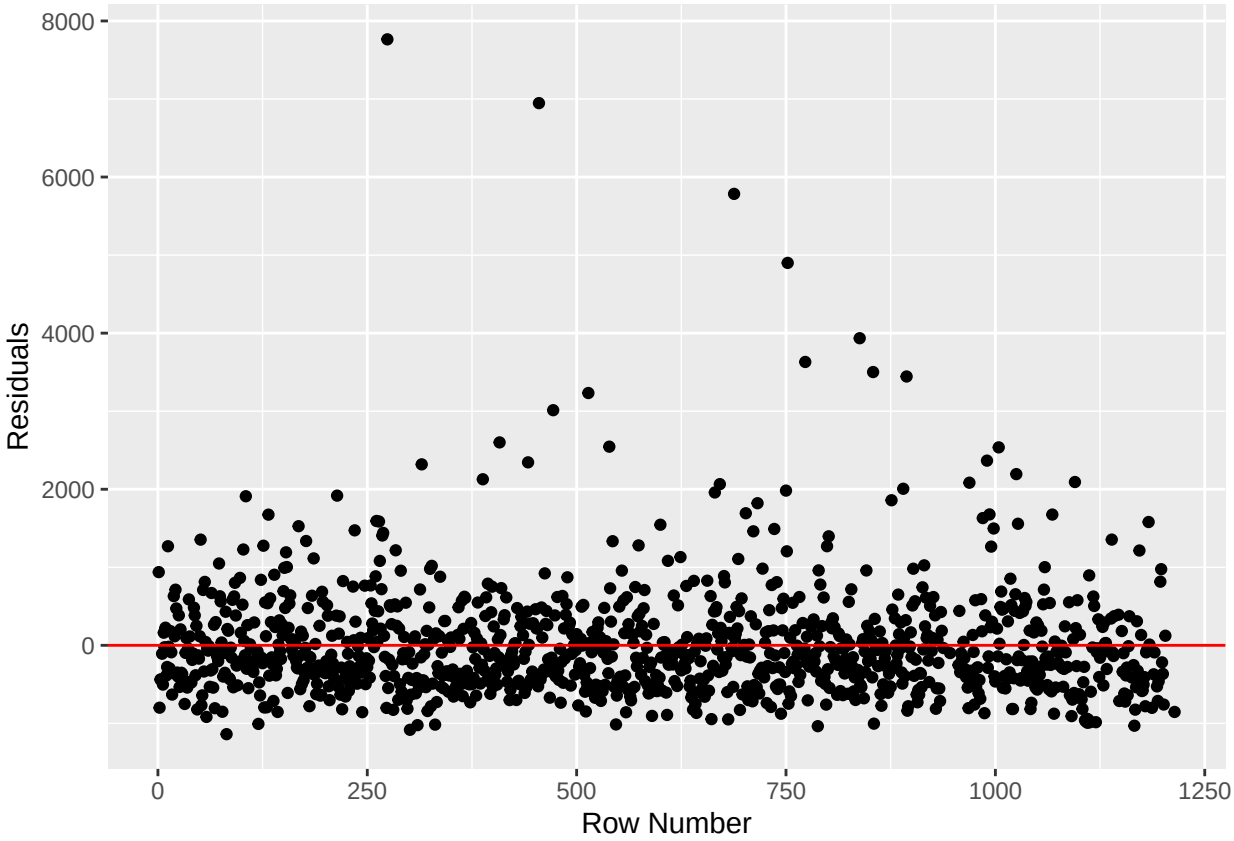

```
qqnorm(airbnb_model1$.resid)
qqline(airbnb_model1$.resid, col = 'red')
```




Looking at the histogram for the residuals, we see that the distribution of the residuals is positively skewed shown by the long right tail. Also, the mode of the distribution appears to be below 0. Also, the normal QQ-plot does not follow a straight diagonal line. While within the theoretical quantiles of about -1 to 1, the residuals seem to be follow the normal distribution quite closely, outside this range the qqplot deviates quite a bit from the diagonal line shown. So in the tails our data does not follow the normal distribution at all. Hence the normality assumption is not satisfied.

Next we will check the independence assumption.

```
ggplot(airbnb_model1, aes(x = as.integer(.rownames), y=.resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = 'red') +
  labs(x = 'Row Number', y = 'Residuals')
```




Above we have a plot of the residuals in the order which the data was put into the dataframe. It looks like we have a random scatter around 0, so assuming that the data was put into the dataset in the same order in which it was collected, we don't have any evidence that observations are dependent. So we would claim that the independence assumption has been satisfied since we have no evidence to say otherwise.

Since not all assumptions are satisfied, we cannot conduct inference on the results of our model.