Lab 05: Data Wrangling & Regression

Isaac Plotkin

2/18/2022

```r
airbnb <- read_csv("raw_data/listings.csv")
```

```
## Rows: 1489 Columns: 18

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): name, host_name, neighbourhood, room_type
## dbl  (11): id, host_id, latitude, longitude, price, minimum_nights, number_o...
## lgl   (2): neighbourhood_group, license
## date  (1): last_review

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1.

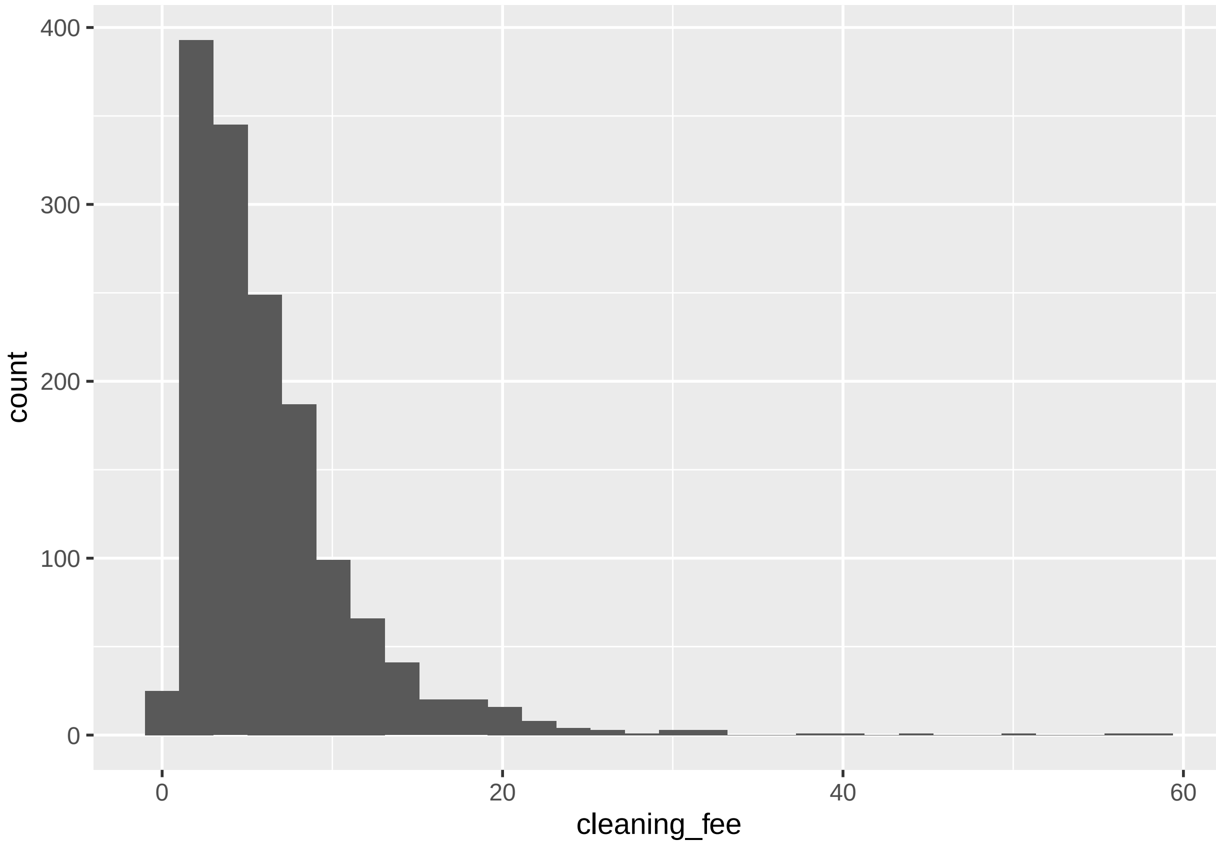
```r
airbnb <- mutate(airbnb, cleaning_fee = 0.02*price)
airbnb
```

```
## # A tibble: 1,489 x 19
##       id name        host_id host_name   neighbourhood_g~ neighbourhood latitude
##    <dbl> <chr>         <dbl> <chr>       <lgl>            <chr>            <dbl>
##  1  8357 The Mushro~   24281 Kitty And ~ NA               Unincorporat~     37.0
##  2 11879 Sunny room~   44764 Steven      NA               Unincorporat~     37.0
##  3 24548 Room with ~   99532 Kerstin     NA               City of Sant~     37.0
##  4 31721 Dog Friend~  136376 Annie       NA               City of Capi~     37.0
##  5 43785 Guest bedr~  191477 Caroline    NA               City of Sant~     37.0
##  6 49520 Guest Cott~  225721 Christine   NA               Unincorporat~     37.0
##  7 54948 Modern Bea~  258675 Terry & Cl~ NA               City of Sant~     37.0
##  8 57031 Sunny in n~   44764 Steven      NA               Unincorporat~     37.0
##  9 70829 Master Bed~  360285 Maisie      NA               City of Sant~     37.0
## 10 72288 Cottage on~  366768 Quentin     NA               Unincorporat~     37.1
## # ... with 1,479 more rows, and 12 more variables: longitude <dbl>,
## #   room_type <chr>, price <dbl>, minimum_nights <dbl>,
## #   number_of_reviews <dbl>, last_review <date>, reviews_per_month <dbl>,
## #   calculated_host_listings_count <dbl>, availability_365 <dbl>,
## #   number_of_reviews_ltm <dbl>, license <lgl>, cleaning_fee <dbl>
```

2.

```r
ggplot(data = airbnb, aes(x = cleaning_fee)) +
  geom_histogram() +
  labs("Distribution of Cleaning Fee")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



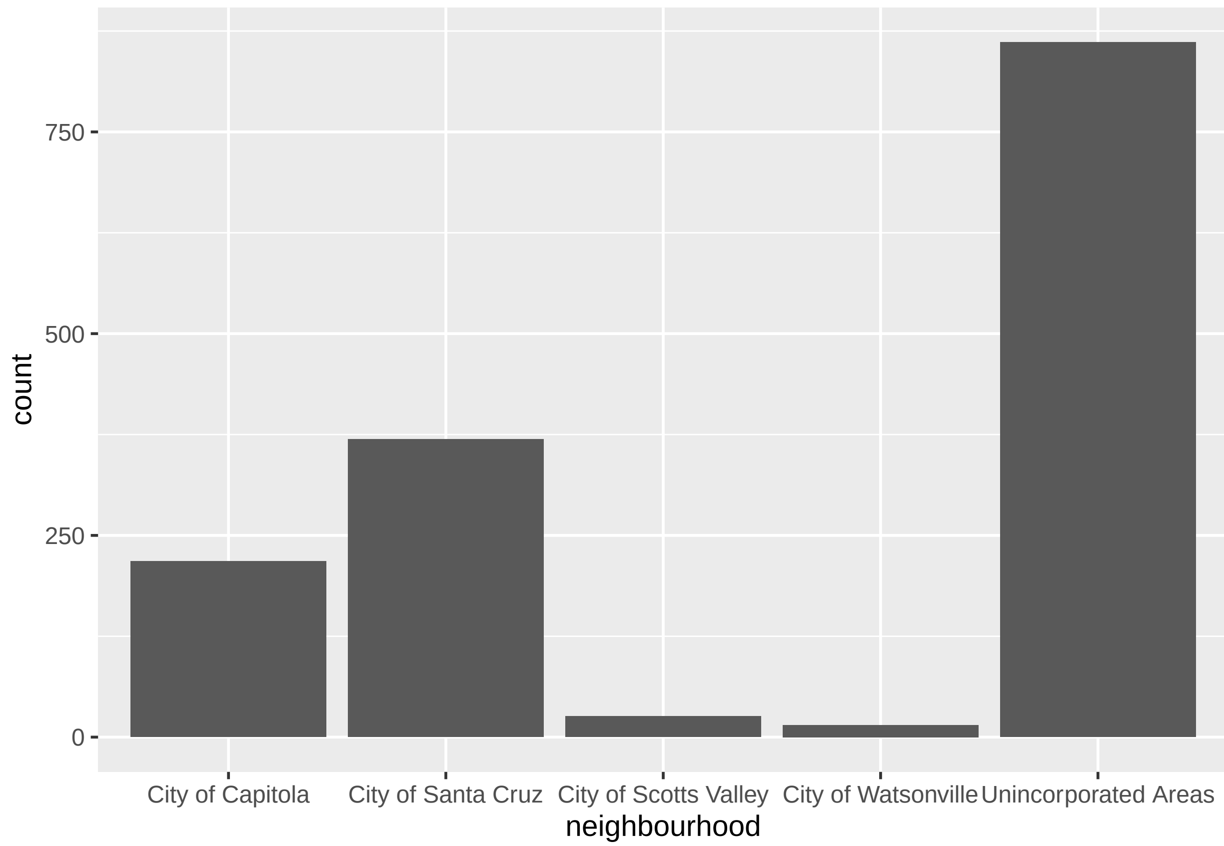

```r
airbnb %>%
  summarise(min = min(cleaning_fee),
            q1 = quantile(cleaning_fee, 0.25),
            q3 = quantile(cleaning_fee, 0.75),
            max = max(cleaning_fee),
            iqr = IQR(cleaning_fee),
            mean = mean(cleaning_fee),
            median = median(cleaning_fee),
            std_dev = sd(cleaning_fee)
            )
```

```
## # A tibble: 1 x 8
##     min    q1    q3   max   iqr  mean median std_dev
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>
## 1  0.62  2.88  8.06    59  5.18  6.38      5    5.39
```

The graph and summary statistics show that cleaning_fee is a right skewed distribution. The mean > median and there is a longer tail on the right side of the distribution.

3.

```
ggplot(data = airbnb, aes(x = neighbourhood)) +
  geom_bar() +
  labs("Distribution of Neighbourhood")
```




```
common_hoods <- sum(airbnb$neighbourhood == 'City of Capitola' | airbnb$neighbourhood == 'City of Santa
total_hoods <- nrow(airbnb)

# % of top 3 neighborhoods
common_hoods/total_hoods
```

```
## [1] 0.9724647
```

There are 5 categories of neighborhood in the dataset. The 3 most common neighborhoods are Capitola, Santa Cruz and Unincorporated Areas. They make up 97.24% of the total neighborhoods.

4.

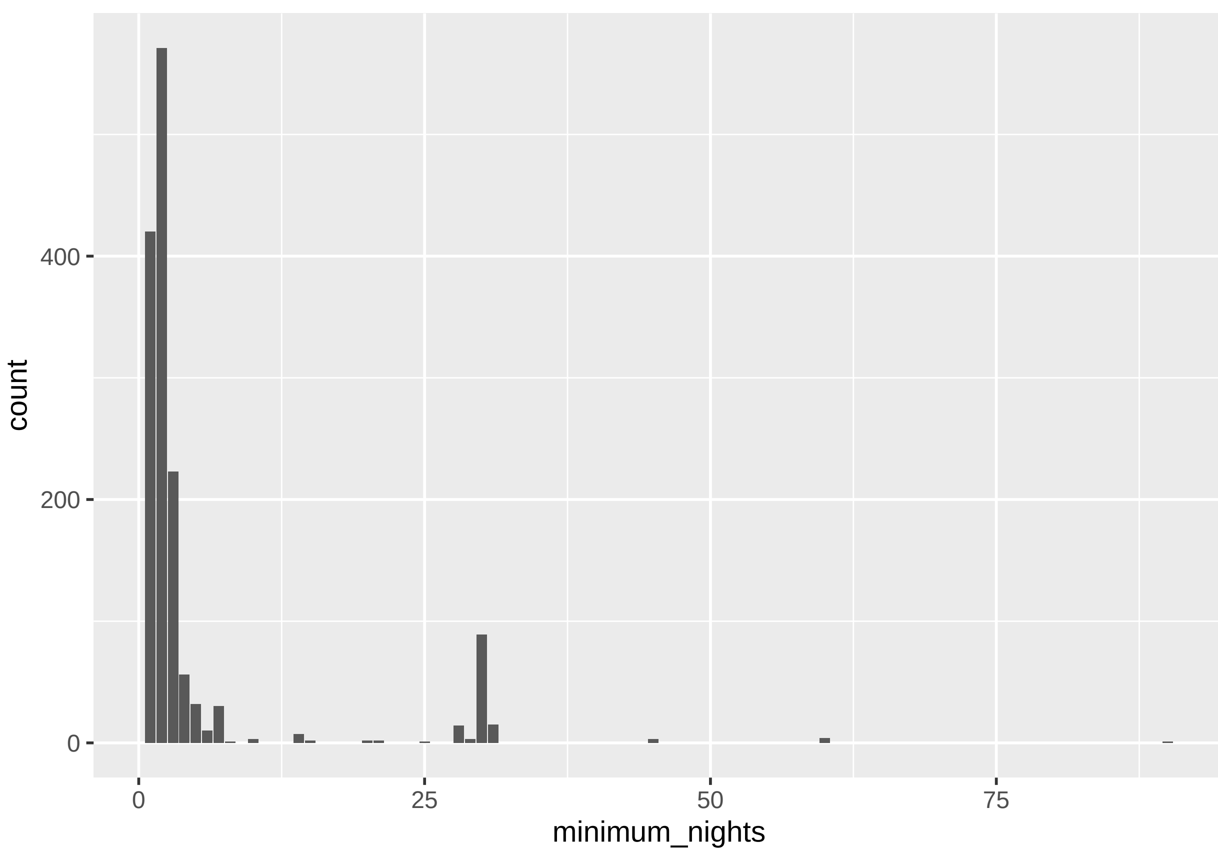
```
airbnb <- mutate(airbnb, neigh_simp = fct_recode(neighbourhood, "Other" = "City of Scotts Valley", "Othe

airbnb
```

```
## # A tibble: 1,489 x 20
##       id name        host_id host_name   neighbourhood_g~ neighbourhood latitude
##    <dbl> <chr>         <dbl> <chr>       <lgl>            <chr>            <dbl>
##  1  8357 The Mushro~   24281 Kitty And ~ NA               Unincorporat~     37.0
##  2 11879 Sunny room~   44764 Steven      NA               Unincorporat~     37.0
##  3 24548 Room with ~   99532 Kerstin     NA               City of Sant~     37.0
##  4 31721 Dog Friend~  136376 Annie       NA               City of Capi~     37.0
##  5 43785 Guest bedr~  191477 Caroline    NA               City of Sant~     37.0
##  6 49520 Guest Cott~  225721 Christine   NA               Unincorporat~     37.0
##  7 54948 Modern Bea~  258675 Terry & Cl~ NA               City of Sant~     37.0
##  8 57031 Sunny in n~   44764 Steven      NA               Unincorporat~     37.0
##  9 70829 Master Bed~  360285 Maisie      NA               City of Sant~     37.0
## 10 72288 Cottage on~  366768 Quentin     NA               Unincorporat~     37.1
## # ... with 1,479 more rows, and 13 more variables: longitude <dbl>,
## #   room_type <chr>, price <dbl>, minimum_nights <dbl>,
## #   number_of_reviews <dbl>, last_review <date>, reviews_per_month <dbl>,
## #   calculated_host_listings_count <dbl>, availability_365 <dbl>,
## #   number_of_reviews_ltm <dbl>, license <lgl>, cleaning_fee <dbl>,
## #   neigh_simp <fct>
```

5.

```
ggplot(data = airbnb, aes(x = minimum_nights)) +
  geom_bar() +
  labs("Distribution of Neighbourhood")
```



```r
min_nights_table <- table(airbnb$minimum_nights)
min_nights_table
```

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

The 4 most common values for minimum_nights are 1, 2, 3, and 30 nights. 30 minimum nights stands out. The most likely intended purpose of 30 minimum nights is to require people to rent the house for at least a month so the landlords do not have to find new renters every week.

```r
airbnb_travel <- airbnb %>%
  filter(minimum_nights<=3)

airbnb_travel
```

```
## # A tibble: 1,214 x 20
##        id name       host_id host_name  neighbourhood_gr~ neighbourhood latitude
##     <dbl> <chr>        <dbl> <chr>      <lgl>             <chr>            <dbl>
##  1   8357 The Mushr~   24281 Kitty And~ NA                Unincorporat~     37.0
##  2  11879 Sunny roo~   44764 Steven     NA                Unincorporat~     37.0
##  3  24548 Room with~   99532 Kerstin    NA                City of Sant~     37.0
##  4  43785 Guest bed~  191477 Caroline   NA                City of Sant~     37.0
##  5  54948 Modern Be~  258675 Terry & C~ NA                City of Sant~     37.0
##  6  70829 Master Be~  360285 Maisie     NA                City of Sant~     37.0
##  7  72288 Cottage o~  366768 Quentin    NA                Unincorporat~     37.1
##  8 126012 Santa Cru~  625642 Mary Jane  NA                Unincorporat~     37.0
##  9 153903 Redwood T~  625642 Mary Jane  NA                Unincorporat~     37.0
## 10 183564 Apple Orc~  880252 Jay & Sib~ NA                Unincorporat~     37.0
## # ... with 1,204 more rows, and 13 more variables: longitude <dbl>,
## #   room_type <chr>, price <dbl>, minimum_nights <dbl>,
## #   number_of_reviews <dbl>, last_review <date>, reviews_per_month <dbl>,
## #   calculated_host_listings_count <dbl>, availability_365 <dbl>,
## #   number_of_reviews_ltm <dbl>, license <lgl>, cleaning_fee <dbl>,
## #   neigh_simp <fct>
```

6.

```r
airbnb_travel <- mutate(airbnb_travel, price_3_nights = 3*price + cleaning_fee)

airbnb_travel
```

```
## # A tibble: 1,214 x 21
##        id name       host_id host_name  neighbourhood_gr~ neighbourhood latitude
##     <dbl> <chr>        <dbl> <chr>      <lgl>             <chr>            <dbl>
##  1   8357 The Mushr~   24281 Kitty And~ NA                Unincorporat~     37.0
##  2  11879 Sunny roo~   44764 Steven     NA                Unincorporat~     37.0
##  3  24548 Room with~   99532 Kerstin    NA                City of Sant~     37.0
##  4  43785 Guest bed~  191477 Caroline   NA                City of Sant~     37.0
```

```
## 5  54948 Modern Be~  258675 Terry & C~ NA              City of Sant~     37.0
## 6  70829 Master Be~  360285 Maisie     NA              City of Sant~     37.0
## 7  72288 Cottage o~  366768 Quentin    NA              Unincorporat~     37.1
## 8 126012 Santa Cru~  625642 Mary Jane  NA              Unincorporat~     37.0
## 9 153903 Redwood T~  625642 Mary Jane  NA              Unincorporat~     37.0
## 10 183564 Apple Orc~  880252 Jay & Sib~ NA              Unincorporat~     37.0
## # ... with 1,204 more rows, and 14 more variables: longitude <dbl>,
## #   room_type <chr>, price <dbl>, minimum_nights <dbl>,
## #   number_of_reviews <dbl>, last_review <date>, reviews_per_month <dbl>,
## #   calculated_host_listings_count <dbl>, availability_365 <dbl>,
## #   number_of_reviews_ltm <dbl>, license <lgl>, cleaning_fee <dbl>,
## #   neigh_simp <fct>, price_3_nights <dbl>
```

7.

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

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_simpOther	-671.550	159.777	-4.203	0.000	-985.040	-358.059
neigh_simpUnincorporated Areas	-312.632	65.758	-4.754	0.000	-441.652	-183.613
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

8. The coefficient of number of reviews shows that there is a $0.44 decrease in price_3_nights for every new review. The 95% confidence interval shows that there is a 95% chance that the coefficient for number of reviews will be between -0.834 and -0.040 if we repeated the sampling.

9. The coefficient of neigh_simpCity of Santa Cruz shows that there is a $208 decrease in price_3_nights if the airbnb is located in Santa Cruz The 95% confidence interval shows that there is a 95% chance that the coefficient for neigh_simpCity of Santa Cruz will be between -356.966 and -59.036 if we repeated the sampling.

10. The intercept is the base value for an airbnb located in Capitola with no reviews. This seems like a meaningful interpretation.

11.

```
# visit_SC <- data.frame(neigh_simp = "Other", number_of_reviews = 10, reviews_per_month = 5.14)
predict(model, data.frame(neigh_simp = "Other", number_of_reviews = 10, reviews_per_month = 5.14), inte
```

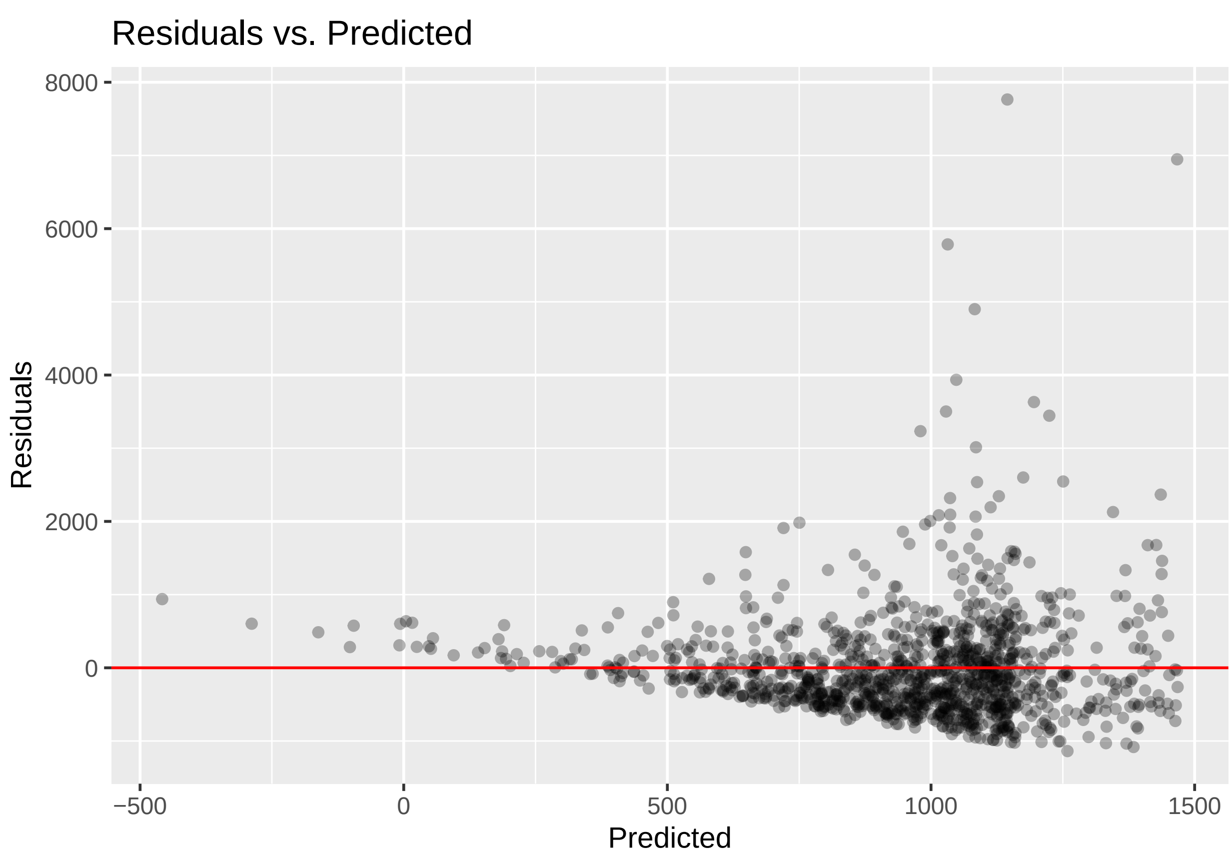
```
##        fit      lwr      upr
## 1 361.6874 59.63618 663.7387
```

12. Linearity

```r
airbnb_aug <- augment(model)
glimpse(airbnb_aug)
```

```
## 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  <dbl> 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, -~
```

```r
ggplot(data = airbnb_aug, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.3) +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Predicted", y = "Residuals",
       title = "Residuals vs. Predicted")
```

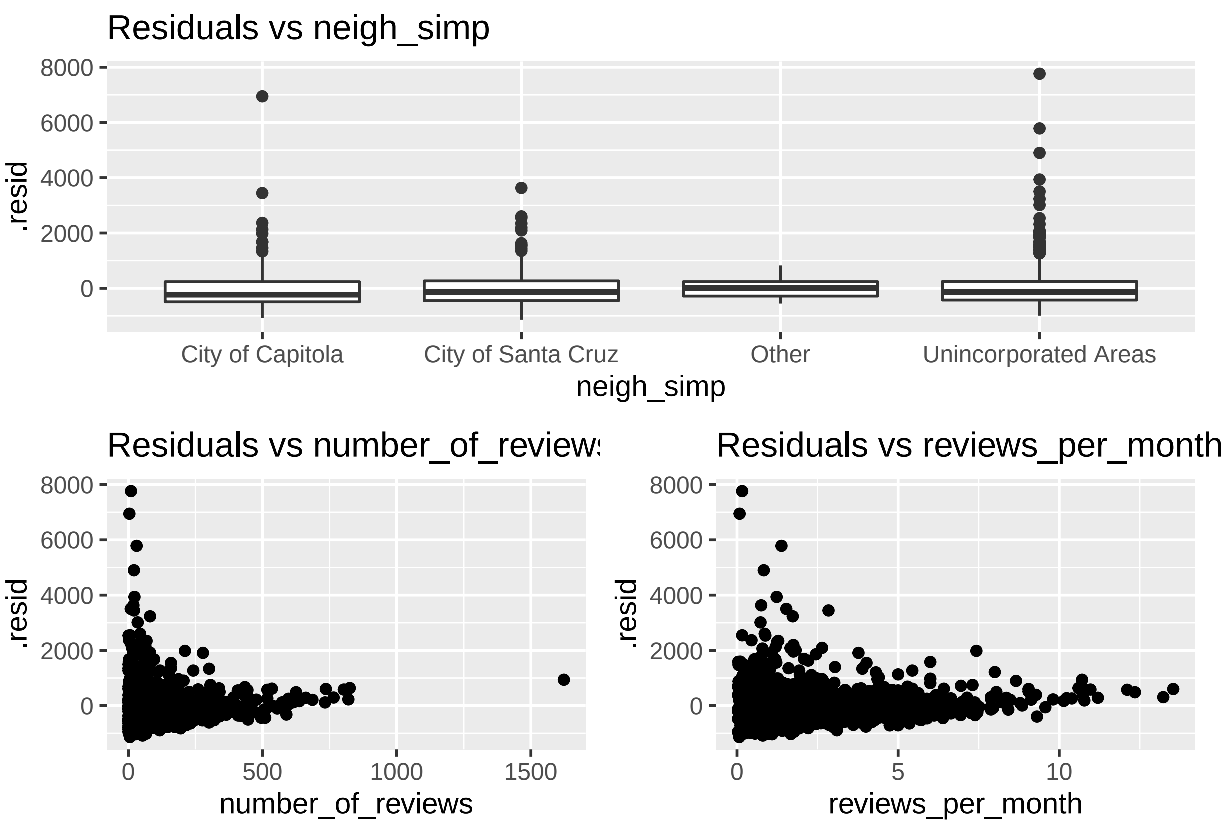


```
p1 <- ggplot(data = airbnb_aug, aes(x = neigh_simp, y = .resid)) +
  geom_boxplot() +
  labs(title = "Residuals vs neigh_simp")

p2 <- ggplot(data = airbnb_aug, aes(x = number_of_reviews, y = .resid)) +
  geom_point() +
  labs(title = "Residuals vs number_of_reviews")

p3 <- ggplot(data = airbnb_aug, aes(x = reviews_per_month, y = .resid)) +
  geom_point() +
  labs(title = "Residuals vs reviews_per_month")

p1/(p2+p3)
```



The model does not pass the linearity assumption therefore I would not be confident on interpreting the results of my model.