lab05

Dylan Scoble

2/17/2022

The Github repository for this assignment is https://github.com/dscoble/lab05

```
airbnb <- read_csv("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.
```

Exploratory Data Analysis (EDA)

Exercise 1

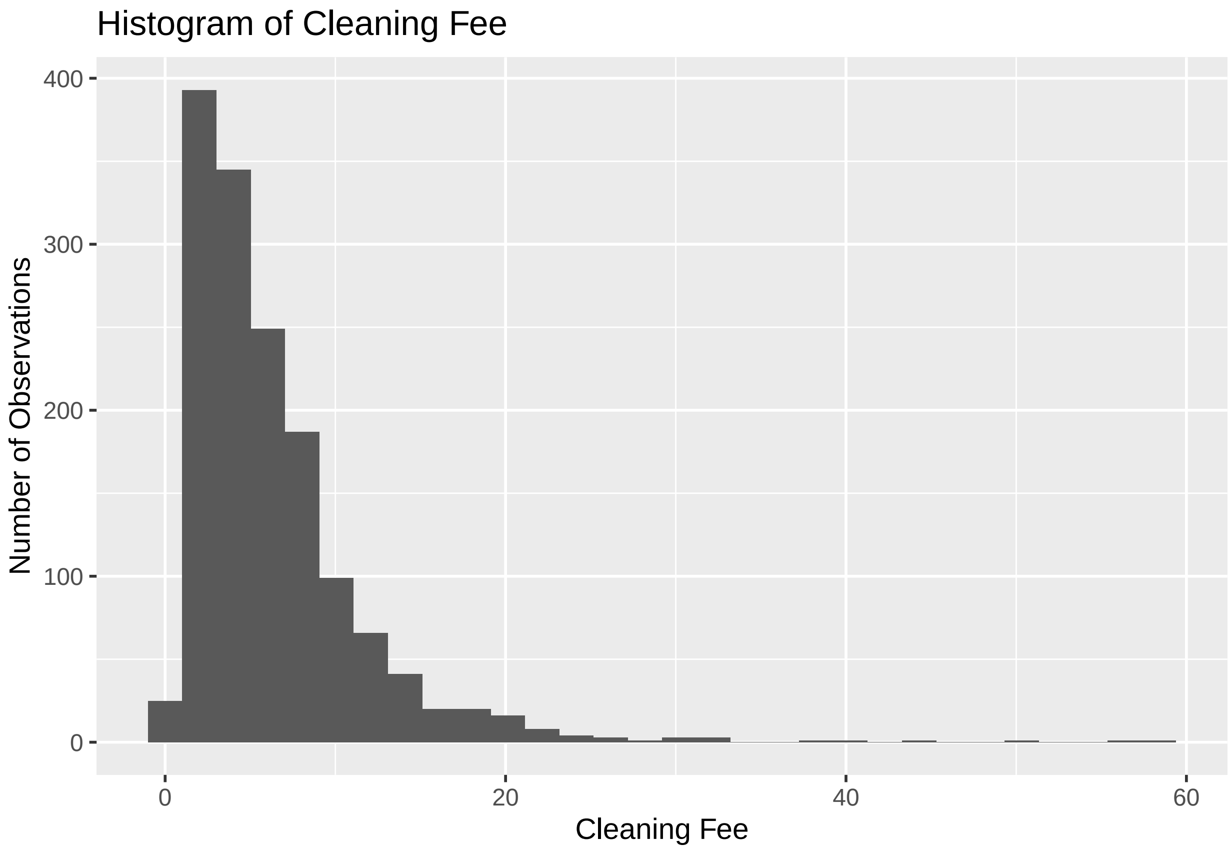
```
airbnb <- airbnb %>%
  mutate(cleaning_fee = price * 0.02)

glimpse(airbnb)
```

```
## Rows: 1,489
## Columns: 19
## $ id                             <dbl> 8357, 11879, 24548, 31721, 43785, 49520~
## $ name                           <chr> "The Mushroom Dome Retreat & LAND of Pa~
## $ host_id                        <dbl> 24281, 44764, 99532, 136376, 191477, 22~
## $ host_name                      <chr> "Kitty And Michael", "Steven", "Kerstin~
## $ neighbourhood_group            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ neighbourhood                  <chr> "Unincorporated Areas", "Unincorporated~
## $ latitude                       <dbl> 37.00939, 36.98048, 36.97191, 36.95849,~
## $ longitude                      <dbl> -121.8855, -121.8813, -121.9973, -121.9~
## $ room_type                      <chr> "Entire home/apt", "Private room", "Pri~
## $ price                          <dbl> 159, 91, 100, 196, 102, 95, 342, 155, 9~
## $ minimum_nights                 <dbl> 2, 3, 1, 4, 2, 30, 2, 10, 3, 1, 1, 1, 2~
## $ number_of_reviews              <dbl> 1623, 85, 510, 253, 495, 145, 119, 1, 4~
## $ last_review                    <date> 2021-12-28, 2021-11-23, 2021-09-29, 20~
## $ reviews_per_month              <dbl> 10.71, 0.61, 3.58, 2.28, 3.59, 1.13, 0.~
## $ calculated_host_listings_count <dbl> 2, 3, 1, 2, 1, 1, 2, 3, 1, 5, 2, 2, 1, ~
## $ availability_365               <dbl> 84, 180, 1, 350, 350, 365, 0, 44, 149, ~
## $ number_of_reviews_ltm          <dbl> 106, 30, 31, 35, 50, 0, 1, 0, 55, 84, 9~
## $ license                        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ cleaning_fee                   <dbl> 3.18, 1.82, 2.00, 3.92, 2.04, 1.90, 6.8~
```

Exercise 2

```
ggplot(data = airbnb, aes(x = cleaning_fee)) +
  geom_histogram() +
  labs(x = "Cleaning Fee",
       y = "Number of Observations",
       title = "Histogram of Cleaning Fee")
```

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



```
airbnb %>%
  summarise(mean = mean(cleaning_fee),
            median = median(cleaning_fee),
            std_dev = sd(cleaning_fee),
            iqr = IQR(cleaning_fee))
```

```
## # A tibble: 1 x 4
##    mean median std_dev   iqr
##   <dbl>  <dbl>   <dbl> <dbl>
## 1  6.38      5    5.39  5.18
```

The distribution of cleaning fee seems to be skewed to the right. There are multiple cases of outliers, which skews the mean value to be significantly larger than the median value.

Exercise 3

```
count(airbnb, neighbourhood)
```

```
## # A tibble: 5 x 2
##   neighbourhood              n
##   <chr>                  <int>
## 1 City of Capitola         218
## 2 City of Santa Cruz       369
## 3 City of Scotts Valley     26
## 4 City of Watsonville       15
## 5 Unincorporated Areas     861
```

There are five different categories of neighbourhood in the dataset, as shown by the table above. The three most common neighbourhoods represented in this dataset are "Unincorporated Areas", "City of Santa Cruz", and "City of Capitola". These three neighbourhoods make up 97.25% of the total number of observations.

Exercise 4

```
airbnb <- airbnb %>%
  mutate(neigh_simp = fct_lump_n(neighbourhood, n=3))
count(airbnb, neigh_simp)
```

```
## # A tibble: 4 x 2
##   neigh_simp               n
##   <fct>                <int>
## 1 City of Capitola       218
## 2 City of Santa Cruz     369
## 3 Unincorporated Areas   861
## 4 Other                   41
```

Exercise 5

```
count(airbnb, minimum_nights)
```

```
## # A tibble: 21 x 2
##    minimum_nights     n
##             <dbl> <int>
##  1              1   420
##  2              2   571
##  3              3   223
##  4              4    56
##  5              5    32
##  6              6    10
##  7              7    30
##  8              8     1
##  9             10     3
## 10             14     7
## # ... with 11 more rows
```

The four most common values for the minimum_nights variable are 1, 2, 3, and 30. The seemingly unusual value of 30 minimum nights can be easily explained by the fact that these rentals are designated for long-term residents only, not tourists.

```
airbnb <- airbnb %>%
  filter(minimum_nights <= 3)
count(airbnb,minimum_nights)
```

```
## # A tibble: 3 x 2
##   minimum_nights     n
##            <dbl> <int>
## 1              1   420
## 2              2   571
## 3              3   223
```

Regression

Exercise 6

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

Exercise 7

```
model <- lm(price_3_nights ~ neigh_simp + number_of_reviews + reviews_per_month, data = airbnb)

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_simpUnincorporated Areas	-312.632	65.758	-4.754	0.000	-441.652	-183.613
neigh_simpOther	-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

Exercise 8

The coefficient of number_of_reviews in the model above shows that as the number of reviews is increased, the price of a three night stay decreases at a rate of $0.437 per review.

The confidence interval of number_of_reviews shows that in a repeated sampling of the data, the sample coefficient of number of reviews was between -0.834 and -0.040 95% of the time. In other words, we are 95% confident that the number of reviews affects the price of a three night stay at a rate between -$0.834 and -$0.04

Exercise 9

The coefficient of neigh_simpCity of Santa Cruz in the model above shows that compared to listings in the City of Capitola, the price of a three night stay in the City of Santa Cruz is $208 cheaper on average.

The confidence interval of neigh_simpCity of Santa Cruz shows that in a repeated sampling of the data, the sample coefficient of listings in the City of Santa Cruz was between -356.966 and -59.036 95% of the time. In other words, we are 95% confident that the price of a three night stay in the City of Santa Cruz is between $356.966 and $59.036 cheaper.

Exercise 10

The intercept of this model represents data for the city of Capitola. For example, the coefficient of the intercept represents the mean price of three nights in Capitola. The rest of the model's coefficients and confidence intervals are all dependent on the intercept, which makes the intercept meaningful.

Exercise 11

For this exercise, we are considering a rental in Scotts Valley. Scotts Valley is not represented in our original model because we used the neigh_simp variable instead of the neighbourhood variable. A new model is created below.

```r
model2 <- lm(price_3_nights ~ neighbourhood + number_of_reviews + reviews_per_month, data = airbnb)

tidy(model2, conf.int = TRUE) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1474.992	65.192	22.625	0.000	1347.082	1602.902
neighbourhoodCity of Santa Cruz	-208.012	75.955	-2.739	0.006	-357.041	-58.984
neighbourhoodCity of Scotts Valley	-696.147	200.983	-3.464	0.001	-1090.486	-301.808

term	estimate	std.error	statistic	p.value	conf.low	conf.high
neighbourhoodCity of Watsonville	-634.943	241.718	-2.627	0.009	-1109.206	-160.680
neighbourhoodUnincorporated Areas	-312.628	65.785	-4.752	0.000	-441.702	-183.554
number_of_reviews	-0.436	0.202	-2.156	0.031	-0.834	-0.039
reviews_per_month	-85.032	12.588	-6.755	0.000	-109.730	-60.335

This model allows us to calculate that a rental in Scotts Valley that has 10 reviews and 5.14 reviews per month will have a three-night price of:

estimate = intercept_estimate + scotts_valley_estimate + (10 * number_of_reviews_estimate) + (5.14 * reviews_per_month_estimate) estimate = 1474.992 - 696.147 - 4.36 - 437.064 = $337.42

and a confidence interval between:

conf_int_low = intercept_ci_low + scotts_valley_ci_low + (10 * number_of_reviews_ci_low) + (5.14 * reviews_per_month_ci_low) conf_int_low = 1347.082 - 1090.486 - 8.34 - 564.012 = -$315.76

and

conf_int_high = intercept_ci_high + scotts_valley_ci_high + (10 * number_of_reviews_ci_high) + (5.14 * reviews_per_month_ci_high) conf_int_high = 1602.902 - 301.808 - 0.39 - 310.122 = $990.58

Exercise 12

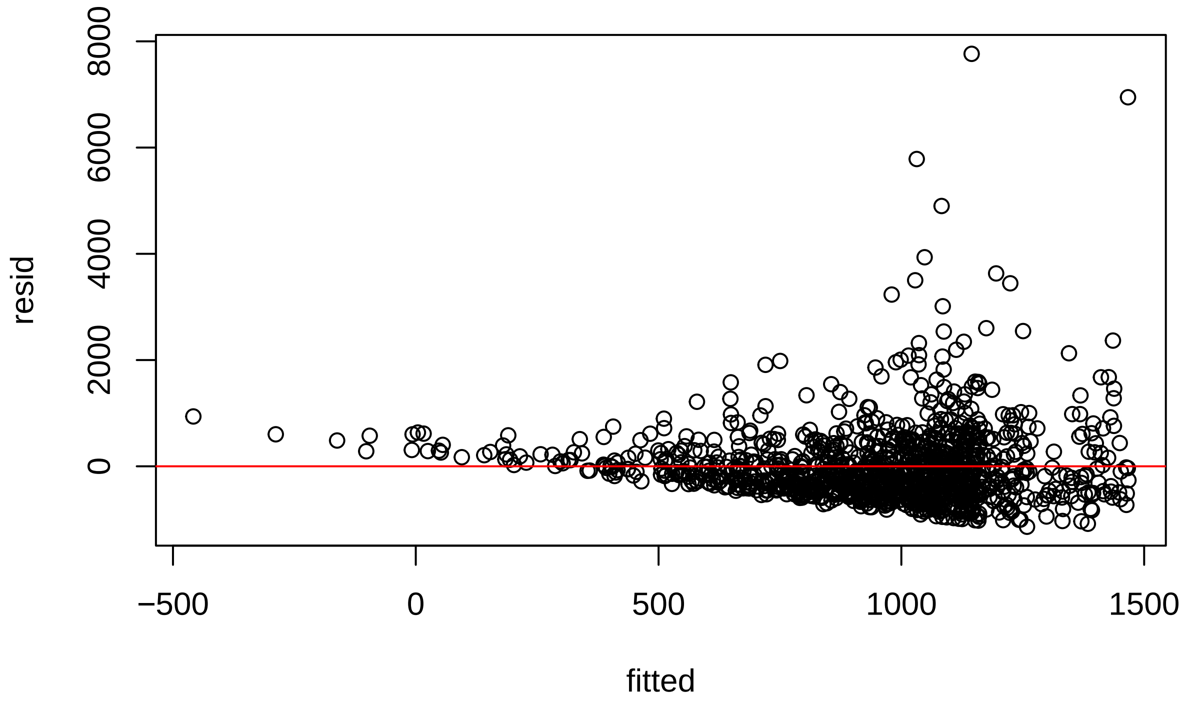
The assumptions we check will require the use of the model's fitted values and residulas. The following code retrieves this data.

```
resid <- model$residuals
fitted <- model$fitted.values
```

The linearity assumption is satisfied because there is a linear relationship between the all of the predictor variables and the response variable.

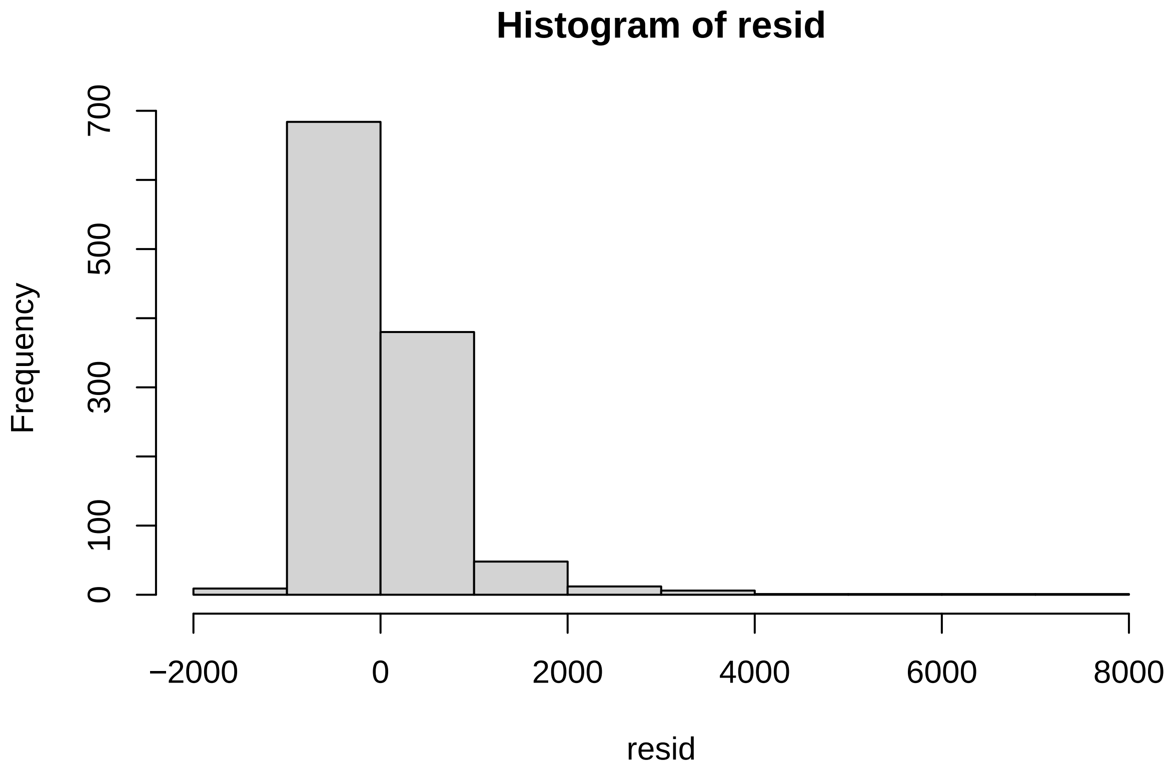
The plot below allows us to reject our assumption of constant variance. The relationship between th model's residulas and its fitted values obviously fans to the right. Because there is no "cloud pattern", the constant variance assumption is not satisfied.

```
plot(fitted, resid)
abline(h=0, col="red")
```



The plot below helps us verify that the normality assumption is satisfied. Despite a few outliers, the histogram of the model's residual values follows a relatively normal distribution, showing that the normality assumption is satisfied.

```
hist(resid)
```


Histogram of resid
Frequency
0
100
300
500
700
−2000
0
2000
4000
6000
8000
resid