Assignment 3

Code

##Data Description This data was downloaded from Kaggle.com, a site that houses open source datasets. This specific dataset is titled: “New York City Airbnb Open Data,” and was originally sourced from Airbnb. From the Kaggle website (below):

##Context Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

##Content This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/downloads/new-york-city-airbnb-open-data.zip/3 (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/downloads/new-york-city-airbnb-open-data.zip/3)

##STEP 1: SET UP

##Task 1. Using filter() remove any observations that have: room_type = “Shared room” and a price greater than 500. In other words, include only non-shared rooms with prices less than or equal to 500.

Hide

```
filtered_air <- airData |>
 filter(room_type != "Shared room", price<=500)
```

Task 2. Converting a categorical variable into a numeric.

Create another variable from the factor variable, room_type, into a numeric coding scheme, where 0 = “Private room” and 1 = “Entire home/apt”. Code included below…

Hide

```
filtered_air <- filtered_air |>
  mutate(room_type_num = case_when(
    room_type == "Private room" ~ 0,
    room_type == "Entire home/apt" ~ 1))
head(filtered_air)
```

id	name	host_id	host_name	neigh
<int>	<chr>	<int>	<chr>	<chr>

id	name	host_id	host_name	neigh
<int>	<chr>	<int>	<chr>	<chr>
12539	Clean & quiet apt home by the park	2787	John	Brook
22595	Skylit Midtown Castle	2845	Jennifer	Manha
33647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manha
43831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brook
55022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manha
65099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manha

6 rows | 1-6 of 17 columns

##Answer the following questions refer to the dataframe you just filtered. (a) Looking at the different rental options, how much does a rental cost in general?

In general, any rental will cost around $133 (mean=133.21). There are a handful (9) of listings that have a price of 0, (likely listings that are no longer active), but 9 out of 46,000 would not be expected to alter the mean price.

Hide

```
favstats(filtered_air$price)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0	70	105	175	500	133.2083	88.02327	46699	0

1 row

b. Is there variation in price?

There is a lot of variation in price. Listings range from $0-500, the standard deviation of 88 tells us the variation is high and that a majority of the listings are between $45-221.

c. What do you think makes a rental cost more versus less?

I would hypothesize that room type would have the largest impact on price and then listing location (neighborhood) .

d. Create a story of a DGP that might be responsible for the variation you see in price.

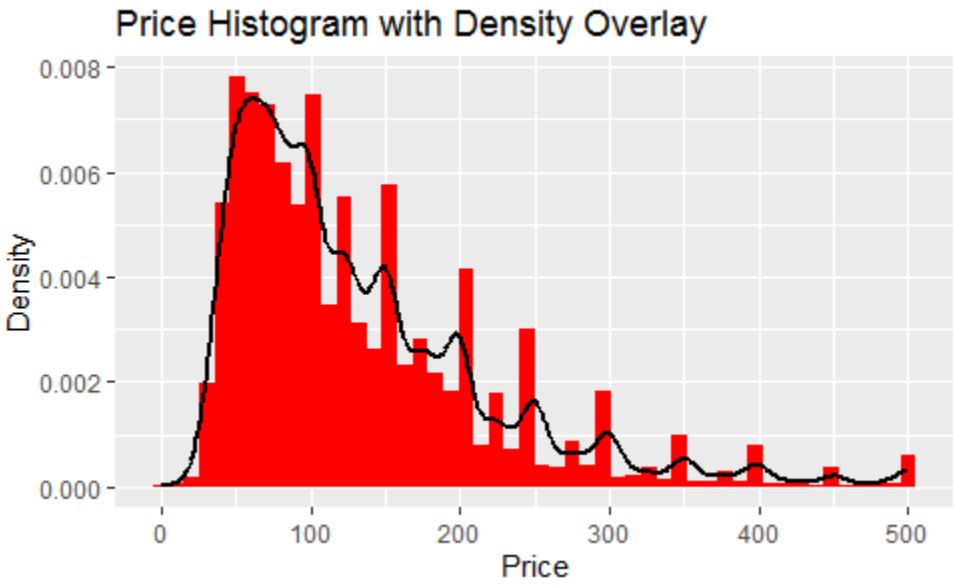
price = room type + neighborhood + other stuff

##Task 3. Exploring the Price variable

a. Plot a histogram of price.

Hide

```
ggplot(filtered_air, aes(x = price)) +
    geom_histogram(aes(y = ..density..), bins = 50, fill = "red") +
    geom_density(color = "black", size = 1) +
    labs(title = "Price Histogram with Density Overlay",
    x = "Price", y = "Density")
```



b. Discuss the shape, center, spread, and weirdness.

The histogram peaks at $50 (mode), is skewed heavily right and shows a pattern of more listings with prices in multiples of 50 (100, 150, 200, etc.)

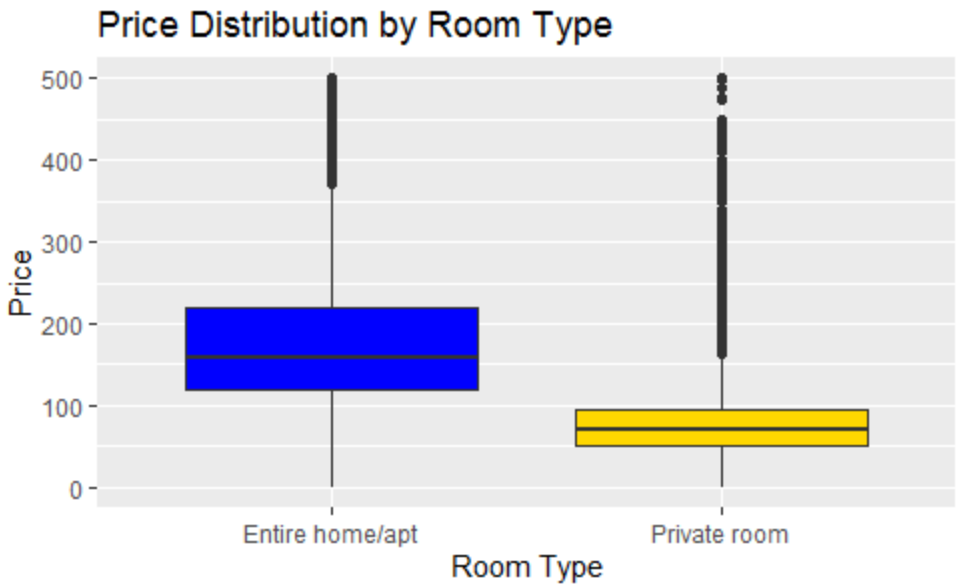
c. Provide a five-number summary of price. (from favstats above)

min Q1 Q2 Q3 max

0 70 105 175 500

Hide

```
ggplot(filtered_air, aes(x = factor(room_type), y = price)) +
  geom_boxplot(fill = c("blue", "gold")) +
  labs(title = "Price Distribution by Room Type",
       x = "Room Type",
       y = "Price")
```



d. Create a visualization to see price (outcome) by the different room_type categories.

e. What do you observe is the relationship between price and room type, based on the graphic?

Overall listings for entire homes/apartment are more expensive than private rooms. It makes sense that private home/apt listings would be more than a private room in an occupied listing. The chart shows the median price for an entire home/apt is around $160, while the private room is around $75. The IQR for a home/apt ranges from $115-220, while a room ranges from $50-95. The outliers are plentiful on the high end for both room types.

##Conclusion

The relationship between price and room type indicates that entire homes/apartments are more expensive and show greater price variability compared to private rooms. This difference likely reflects extra space, privacy, and amenities offered in an entire home/apartment.

##Task 4: Fitting Models

a. Fit an empty model using the linear model function: lm(y ~ NULL, data=dataframe).
- Do this using price as your outcome variable.
- Save is to an object called priceEmpty.
- Review the results of your model to answer (b - e)

Hide

```
priceEmpty <- lm(price ~ NULL, data = filtered_air)
summary(priceEmpty)
```

```
Call:
lm(formula = price ~ NULL, data = filtered_air)

Residuals:
    Min      1Q  Median      3Q     Max
-133.21  -63.21  -28.21   41.79  366.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 133.2083     0.4073     327   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.02 on 46698 degrees of freedom
```

b. Based on your model output, what is the value of the intercept?

Intercept = $133.21

c. What does the intercept represent?

This is the mean value across all observations of price in the filtered data.

Hide

```
anova(priceEmpty)
```

```
Analysis of Variance Table

Response: price
             Df    Sum Sq Mean Sq F value Pr(>F)
Residuals 46698 361820610  7748.1
```

d. Use anova() to find the value of sums of squares from your priceEmpty model.

sums of squares = 361820610

e. What does the sums of squares represent, in your own words?

The total variability of all the price data from our mean (squared to remove the negative values)

f. Define residuals, sums of squares, and standard deviations in your own words.

#Residuals: The differences between observed values and the values predicted by a model. They represent the deviation of the observed data points from the model’s predicted value of the mean.

#Sum of Squares: A measure of the total variability from a dataset. The sum of the squared differences between each observation and the mean of the observations.

#Standard Deviations: A measure of the amount of variation in a dataset. It is the average distance of each observation from the mean; the average of the deviations.

##Task 6: Z-Scores (a) Using z-scores, compare the price, $275, between the room_types, “Private room”=0 and “Entire home/apt” = 1.

Hide

```
air_bnb.stats <- favstats(price ~ room_type, data = filtered_air)
air_bnb.stats
```

room_type <chr>	min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
Entire home/apt	0	119	158	220	500	179.72822	88.53081	24516	0
Private room	0	50	70	95	500	81.79592	51.12975	22183	0

2 rows

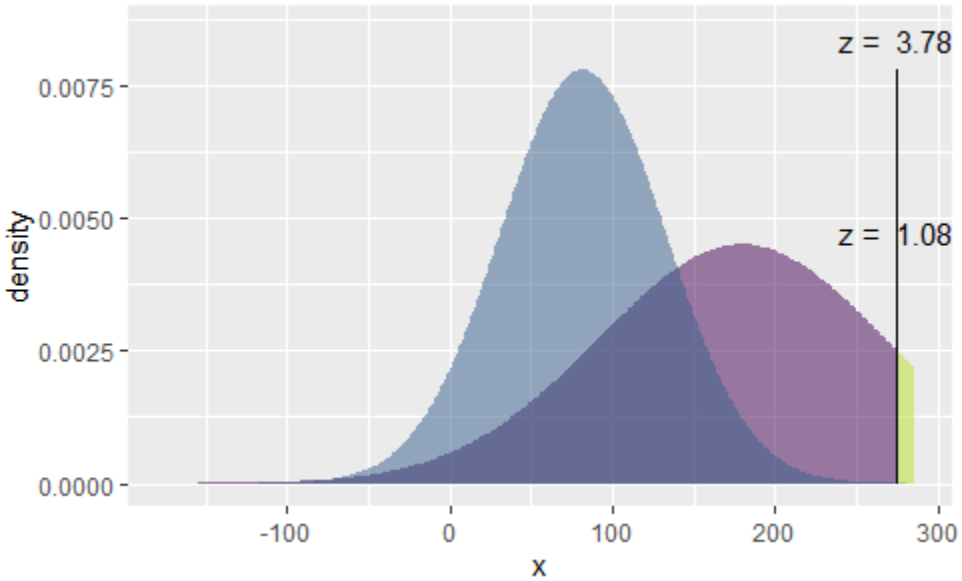
Hide

```
xpnorm(275, mean = air_bnb.stats$mean, sd = air_bnb.stats$sd, alpha=.5)
```

```
If X ~ N(179.7, 88.53), then
If X ~ N( 81.8, 51.13), then

    P(X <= 275) = P(Z <= 1.076) = 0.8591    P(X <= 275) = P(Z <= 3.779) = 0.9999
    P(X >  275) = P(Z >  1.076) = 1.409e-01 P(X >  275) = P(Z >  3.779) = 7.882e-05
```

```
[1] 0.8590683 0.9999212
```



b. What z-scores did you get for the two room_types?

Entire z-score = 1.08 private z-score = 3.78

c. Interpret the graph you just produced.

The graph shows both distribution curves of the price based on room type: blue for “Private room” and purple for “Entire home/apt”. The line and at 275 show us a visual depiction of the 14% (1-.8591) of listings below the “entire home/apt” curve that are more expensive than 275 (.01% for “private room”). The z-scores tell us that 275 is 3.78 sd above the “private room” 81.80 mean (peak of the blue curve) and 1.08 sd above the “entire home/apt” mean of 179.73. The high z-score for “private room” tells us it is significantly above the mean price for that room type, while it is close to 1 standard deviation above the price for an “entire home/apt.” By comparing the z-scores, one can conclude that the price of $275 is much more unusual for a private room than for an entire home/apartment. This can help understand how common or uncommon a specific price point is relative to different categories of listings.

Let’s calculate these manually!

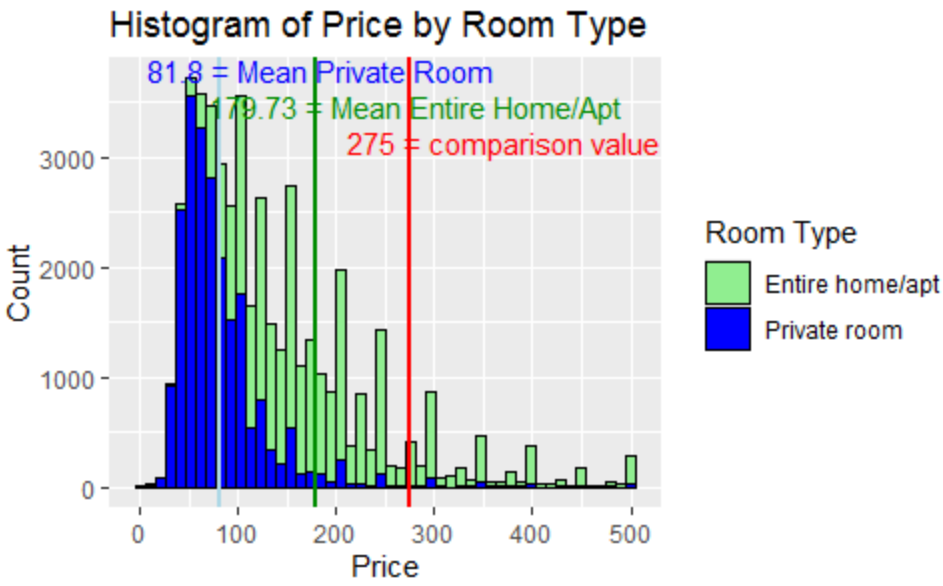
Hide

```
#The Manual way!

#Step 1: run fav stats
favstats_man <- favstats(price ~ room_type, data = filtered_air)

#Step 2: save the favstats output for the mean and sd of each room type.
# Private Room mean
private_m <- favstats_man[favstats_man$room_type == "Private room", "mean"]
# Private Room standard deviation
private_sd <- favstats_man[favstats_man$room_type == "Private room", "sd"]
#Whole House mean
entire_m <- favstats_man[favstats_man$room_type == "Entire home/apt", "mean"]
#Whole House standard deviation
entire_sd <- favstats_man[favstats_man$room_type == "Entire home/apt", "sd"]

#Step 3: Plot a histogram, fill by room type and create vertical lines to indicate the values
we wish to compare: 275, mean of room1, mean of room 2
ggplot(filtered_air, aes(x = price, fill = room_type)) +
  geom_histogram(bins = 50, color = "black") +
  geom_vline(aes(xintercept = 275), color = "red", size = 1) +
  geom_vline(aes(xintercept = private_m), color = "lightblue", size = 1) +
  geom_vline(aes(xintercept = entire_m), color = "green4", size = 1) +
  labs(title = "Histogram of Price by Room Type", x = "Price", y = "Count", fill = "Room Typ
e") +
  scale_fill_manual(values = c("Private room" = "blue", "Entire home/apt" = "lightgreen")) +
  annotate("text", x = 275, y = Inf, label = "275 = comparison value", color = "red", vjust =
4, hjust = .2) +
  annotate("text", x = private_m, y = Inf, label = paste(round(private_m, 2),"= Mean Private
Room"), color = "blue", vjust = 1, hjust = .21) +
  annotate("text", x = entire_m, y = Inf, label = paste(round(entire_m, 2),"= Mean Entire Hom
e/Apt"), color = "green4", vjust = 2.5, hjust = .26)
```



Hide

```
#Step 4: Calculate z-scores for Private Room types (Room0) and Whole House types (Room1)
price <- 275
z_private <- (price - private_m) / private_sd
z_entire <- (price - entire_m) / entire_sd

#Room0 = Private Room
z_private
```

```
[1] 3.778701
```

Hide

```
#Room1 = Whole House
z_entire
```

```
[1] 1.076143
```

d. Interpret what they mean, in relation to each distribution. Discuss this interpretation of z-scores for the price, $275, in relation to your DGP in question 2.

The high z-score (3.78) indicates $275 is much higher than the average price for private rooms, suggesting it is an outlier and not a typical price for this category. The moderate z-score (1.07) suggests $275 is a lot closer to the average price for an entire home/apartment, making it a more common price point for this category. Room type and some other stuff does explain the variation in price in the data.

e. How might the room type explain some of the variation in price?

Room type explains some of the variation in price as an entire home/apartment will normally command a higher price than a private room due to increased space, privacy, and included amenities. This will result in higher average prices and greater price variability within the category compared to private room.

f. Would knowing the room type help you make a better prediction about price? Why or why not?

Yes, generally, knowing the room type would make it easier to guess the price because room type significantly influences pricing. Entire homes/apartments typically have higher prices ($180) than private rooms ($82) due to the additional space and amenities. By considering room type, predictions will be more accurate, reflecting the inherent value differences in price listings.

##END ASSIGNMENT