# APPLIED STATISTICS in MI

## MSMI 603: Applied Statistics

### Project - Airbnb

Hypothesis Testing, Regression Models & More

MS Marketing Intelligence, University of San Francisco

1st December, 2023

# Index

# Question 1

### 1) Research Question:

Where are the cheapest and most expensive neighborhoods to rent an Airbnb in the US and Canada?

### 2) Approach Overview:

Data Observation and Criteria Identification:
In analyzing the dataset, we focused on pinpointing the average price of Airbnb listings in each neighborhood. Our criteria for identifying the cheapest and most expensive neighborhoods rely on the average price.

Methodology:
The analysis was conducted using the R programming language, leveraging the 'dplyr' package for data manipulation and summarization. The dataset was filtered to include only listings from the US and Canada. Neighborhoods were then grouped, and calculations were performed to determine the average price for each neighborhood. Also, 'ggplot2' was used for visualizing data.

Hypothesis Formulation:

- *Null Hypothesis (H0*): There is no significant difference in the average prices of Airbnb listings across neighborhoods in the US and Canada.
- *Alternative Hypothesis (H1):* The average prices of Airbnb listings vary significantly across neighborhoods in the US and Canada.

### 3) Steps for Data Analysis:

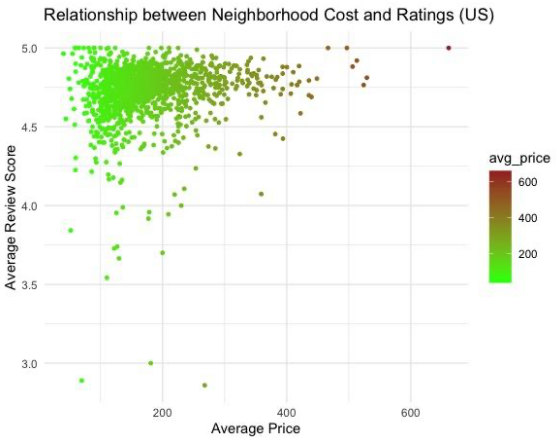
Data Exploration*:* The Airbnb listings data is imported from a CSV file into R for analysis.

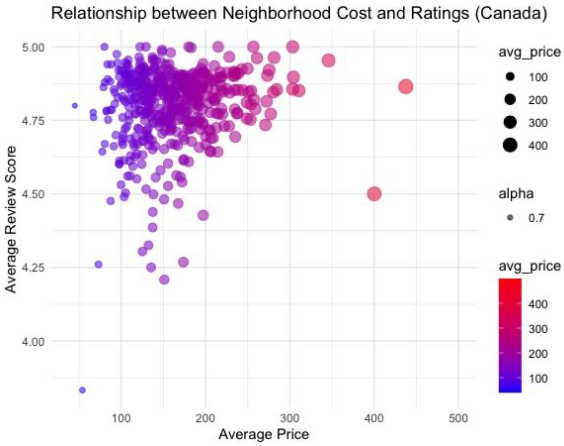
Method*:*

- Filter data as per the country for US and Canada
- Group by neighborhood and calculate average price and count of listings.
- Arrange the neighborhoods based on average price.
- Identify the cheapest and most expensive neighborhoods.

Visualizations:

- Group by neighborhood and calculate the average for US and similarly for Canada.
- Create scatter plots




**4) Result and Recommendations:**

From a statistical viewpoint, the identified neighborhoods with the lowest and highest average prices offer valuable insights for both renters and hosts.

- The cheapest neighborhoods represent budget-friendly options for travelers.
- The most expensive neighborhoods may cater to those seeking premium accommodations or hosts looking to set higher prices based on location.

The recommendations derived from this analysis can inform travelers seeking affordable options and hosts strategizing their pricing based on neighborhood characteristics.

# Question 2

1) **Research Question:**

   Is there a significant difference in the average ratings between listings hosted by superhosts and those not hosted by superhosts?

2) **Approach Overview:**

Data Observation and Criteria Identification:

In examining this dataset, we focused on determining whether there is a substantial difference in the average overall ratings between listings hosted by superhosts and those not hosted by superhosts. Our criteria involved defining the independent variable as host status (superhost or not) and the dependent variable as review score ratings.

Methodology:

The analysis was conducted using the R programming language, specifically employing the 'dplyr' package. We extracted relevant data from the Airbnb listings, defined the variables, calculated the mean ratings for superhosts and non-superhosts, and then performed a T-test to assess the significance of the observed differences.

Hypothesis Formulation:

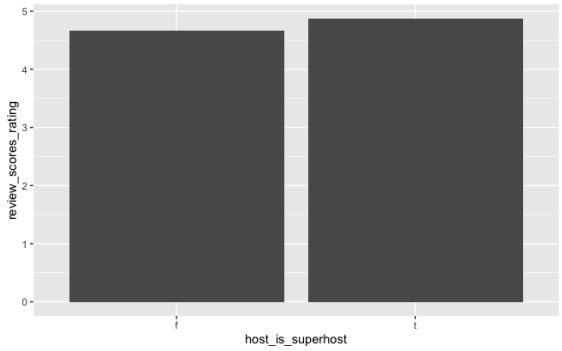
- *Null Hypothesis (H0)*: There is no significant difference in the average overall ratings between listings hosted by superhosts and those not hosted by superhosts.
- *Alternative Hypothesis (H1):* There is a significant difference in the average overall ratings between the two groups. The presence of superhosts positively influences the overall rating of Airbnb listings.

3) **Steps for Data Analysis:**

Data Exploration*:* The Airbnb listings data is imported from a CSV file into R for analysis.

Method*:*

- Define the independent variable as host status (superhost or not) and the dependent variable as overall ratings
- Create a histogram




- Calculate the mean of the overall ratings of superhost and non-superhost

- Run the T-test

P-value is very low, indicating that the difference in ratings between superhosts and non-superhosts is not due to randomness.

**4) Result and Recommendations:**

- From a statistical viewpoint, the T-test results indicate a significant difference in ratings between Airbnb listings managed by Superhosts and those not managed by superhosts. The low p-value suggests that the observed difference is unlikely due to random chance. In conclusion, understanding and leveraging the influence of Superhost status can be a strategic approach for hosts aiming to enhance their Airbnb listings' overall appeal and guest satisfaction.

- There is a significant difference in ratings between Airbnb listings managed by Superhosts and those not managed by superhosts. This difference is attributed to the Superhost status rather than random variation. In Airbnb's system, Superhosts are recognized for their exceptional hosting, often reflected in higher ratings from guests. This status is awarded based on a combination of factors such as high response rates, low cancellation rates, and high overall guest ratings. Given this, it's reasonable to conclude that an Airbnb owner seeking to improve their ratings should aim to achieve Superhost status.

## Question 3

**1) Research Question:** What price should we list a newly purchased property on Airbnb?

(Imagine I purchased the property shown in the screenshots below to rent via Airbnb. How much should I charge per night to rent this entire property? Is there anything I should try to highlight in the listing?)

**2) Approach Overview:**

Data Observation and Criteria Identification:

In examining this dataset, we focused on the properties that were exclusively situated in the City of San Francisco (Filter 1 City).

The next best criteria for figuring out was - which neighborhoods are there in our City and in which neighborhood our property was situated. (Filter 2 - Neighborhood)

Focusing on similar Airbnb listings to our *purchased house* for which we wanted to find out the Average cost per night to be charged. Further, an alternative filter was applied. We filtered out the Neighborhood data having entire homes. (Filter 3 - Entire Homes)

Therefore, we estimated Average cost per night of the bought house situated in the Lone Mountain Neighborhood, being an entire home would resonate with the average price of other entire homes in the nearby neighborhood.

Methodology:

The analysis was performed using the R programming language, utilizing packages such as 'dplyr' and 'ggplot2'. The data was refined to align with specified criteria, and computations were made to ascertain the nightly rental charge for the entire property.

Hypothesis Formulation:

- *Null Hypothesis (H0):* The average nightly price charged for entire home listings on Airbnb is not significantly influenced by the number of reviews the property has received.
- *Alternative Hypothesis (H1)*: Number of Reviews impact the average nightly price charged for the entire home's listings on Airbnb.

**3) Steps for Data Analysis:**

Data Exploration*:* The Airbnb listings data is imported from a CSV file into R for analysis.

Method 1*: Applying the First Set of Filters and Running the Regression between Price and No. of reviews.*

The first method filters the dataset based on primary criteria:

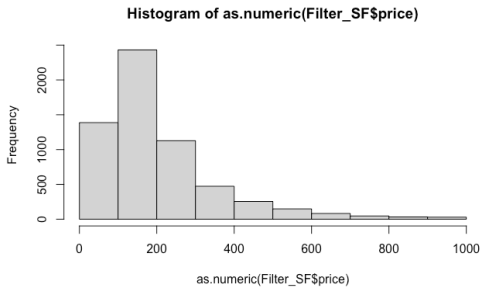
- Selection of San Francisco city out of the whole dataset

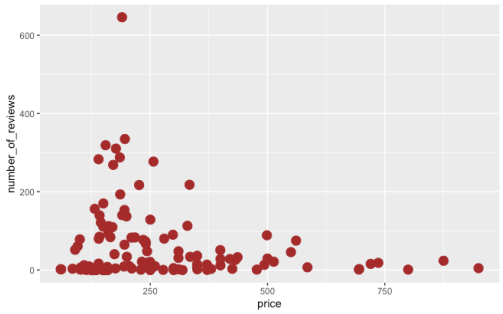
● Followed by the unique Neighborhoods within the SF City (“Golden Gate Park", "Inner Richmond")

● Filtering out the listings as Entire Homes from the filtered neighborhoods and Plotted the Histograms for each of the following steps above and also Plotted a scatter plot for _the No. of reviews and Price per night and Ran the regression

● Also Categorizing no. of reviews into different categories in order to find out the price for Properties without reviews (assuming new properties have no reviews)

Visualizations:




2. Method 2: *Applying the First Set of Filters and Running the Regression btw Price and accommodates.*

   Using the same filtered data, we run the regression between the Price and accommodates.

**4) Result and recommendations:**

1) When we ran the regression with no. of reviews to determine the price, we found out the P value was high approx. 14% which indicated that the No. of Reviews coefficient was not statistically significant and was not a significant predictor for price. Thus our Alternate Hypothesis was rejected and Null Hypothesis stands true.
2) The next step we ran a regression analysis with accommodates in order to predict price. The interpretations were as follows -
   P value was 1.68e-13 ***, hence we can say that the relationship was statistically significant. And Accommodates was a significant predictor for Price
   Price of the property when there is no accommodation is $63.32. With every single accommodation the price increases by $46. 82..
   As per the property description in Lone Mountain (Screenshot details) , the property has 2 beds and 2 baths and assuming there would be 4 accommodates, the price of this listing should be:

   Price = Intercept + (Accommodates estimates * no. of accommodates)
   Price = 63.32 + (46.82 *4) = 250.60$

Recommendations

- Focusing on accommodation capacity for pricing strategy, since the number of accommodations is a significant predictor of price, we should focus on this aspect when setting prices for Airbnb properties.
- Properties with higher accommodation capacities can be priced higher. Given that the number of reviews was not a significant predictor of price, you should not heavily factor in the number of reviews when setting prices. However, it's still important to encourage guests to leave reviews, as they can influence other aspects of the rental experience, like perceived trustworthiness and attractiveness to potential guests.
- While this analysis focused on the number of accommodations and reviews, other factors might also significantly impact price. Future analyses could explore other variables like location, amenities, or seasonality.

# Question 4:

**1.Research Question:**

Can you find a better property in San Francisco to purchase as a short-term rental?

**2.Approach:**

Data Observation and Criteria Identification:

The criterion for short-term rental is decided through:

1.Minimum nights availability: We are assuming that the properties in San Francisco where the minimum nights to rent is >=30, are short term rentals. (We had initially decided that we would keep the criteria of minimum nights >=30 and Maximum nights <=200, however this was leading to a very small sample size)
2. listing which is not a hotel room since we are looking at rental occupancy (tenants)

Methodology:

- The research was done on R and packages like 'dplyr' and 'ggplot2' were employed. The two data sets used here are Airbnb listings and Airbnb calendar.
- The data was filtered to find relevant columns. The Airbnb listings and Calendar data were clubbed basis unique ids where minimum nights>=30 and specific neighborhoods- "Pacific Heights", Daly City", "Noe Valley", "Inner Richmond", "Downtown/Civic Center" were used for analysis.
- The question is about "better" properties to purchase as short-term rentals. This has been defined through "review_scores_location".

Hypothesis Formulation:

- *Null Hypothesis (H0):* The presence of an Airbnb listing in a specific neighborhood has no significant effect on the review score for location. Essentially, knowing the neighborhoods does not contribute to a difference in the review score.
- Alternative *Hypothesis (H1)*: The review score location is different for different neighborhoods.

**3.Steps for Data Analysis:**

Data Exploration

- Airbnb Calendar data set was filtered based on minimum nights' availability >=30
- This data was clubbed with Airbnb Listings based on distinct ids.
- After this filtering for listings where the entire home, shared rooms and private rooms are available.
- Running Analysis: We ran the regression analysis where Review_score_location is predicted by neighborhoods to decide which neighborhood is better.

**4. Results & Recommendations**

The review_score_location is different in different neighborhoods. The P value for all the neighborhoods is low which shows that the relationship is impactful. If we look at the results to determine "better neighborhoods" then:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | **4.556** | **0.08883** | **51.284** | **< 2e-16** |
| **neighborhoodDowntown/Civic Center** | **-0.519** | **0.1088** | **-4.77** | **3.22E-06** |
| **neighborhoodInner Richmond** | **0.272** | **0.13538** | **2.01** | **0.046** |
| **neighborhoodNoe Valley** | **0.332** | **0.1183** | **2.803** | **0.005** |
| **neighborhoodPacific Heights** | **0.356** | **0.130** | **2.741** | **0.007** |

Recommendations

- Neighborhoods like Noe Valley, Pacific Heights, and Inner Richmond, which have positive estimates, should be highlighted in your marketing and promotional materials.
- Downtown/Civic Center has a negative estimate, indicating lower review scores. Investigate the specific reasons behind this, such as safety concerns, noise levels, or lack of amenities. Develop strategies to mitigate these issues, such as providing detailed local guides, improving property security, offering soundproof rooms for noise cancelation.
- Adjusting pricing strategy based on neighborhood scores. Higher-scoring neighborhoods might justify higher prices due to their higher guest satisfaction, while lower-scoring areas might need more competitive pricing to attract bookings.

# Question 5:

**1.Research Question:**

Do people have a positive perception for the Airbnb listings that are cheaper?

**2.Approach:**

Data Observation and Criteria Identification:

We have defined positive perception through “sentiment”. Specific neighborhoods- "Palo Alto","Los Altos","Los Gatos”, “Mountain View”, “Redwood City" were used for analysis.

Methodology:

The research was done on R and packages like ‘dplyr’ and ‘ggplot2’ were employed.
The data was filtered to find relevant columns like – listings id, price, sentiment, neighborhood, and city. We made visualizations to find out the relationship between the 2 variables and then ran regression to check the impact.

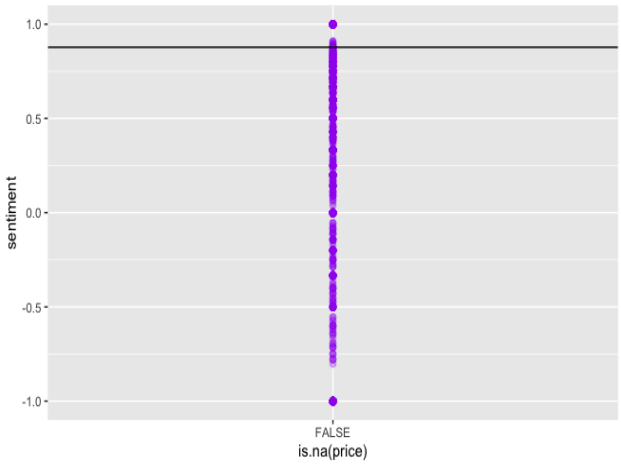
Hypothesis Formulation:

- Null Hypothesis (H0): There is no significant impact of the price on the sentiments.
- Alternative Hypothesis (H1): High Prices lead to low sentiment.

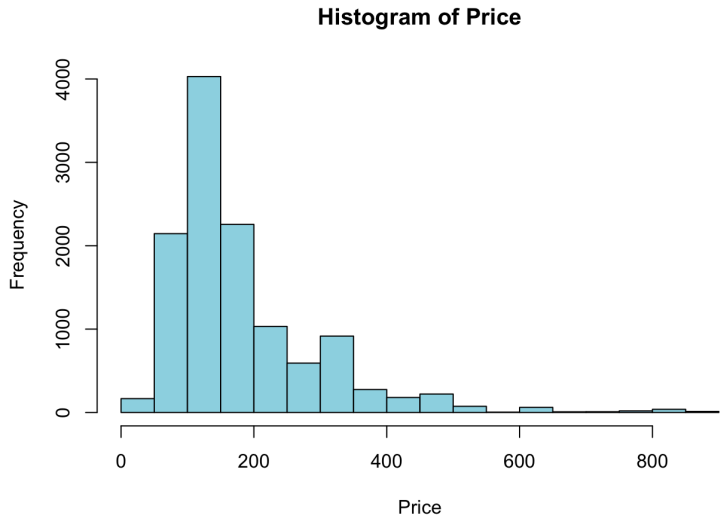
**3.Steps for Data Analysis:**

Data Exploration

- After identifying relevant columns, we filtered the data for specific neighborhoods.
- Plotted Null Hypothesis**:** the mean sentiment is 0.878 (Price=0)

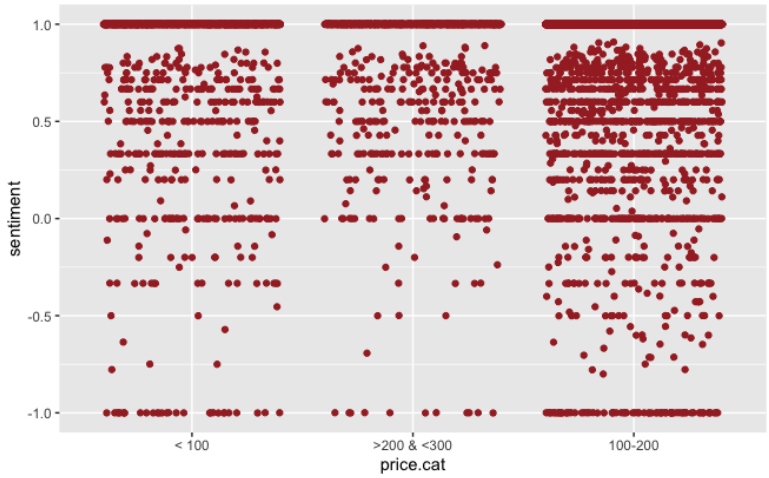



- Plotted Alternate Hypothesis




- Running Analysis

We ran the regression analysis and results are discussed below. However, we wanted to check if sentiments vary at different Price categories, so we created three categories of Price ('<$100', '$100-$200', '$200-$300')




There are several houses in the category '$100-$200' at a sentiment ranging between 0.5 & 1. We ran regression analysis to get a clearer analysis.

## 4. Results & Recommendations

Results _As per the Regression analysis run initially, Price does not have an impact on the sentiment (P-value= 0.217) and the impact that we can see due to Price maybe random and just out of chance.

After categorizing Prices, the sentiment for Price category <$100 was 0.876 at a very low P-value whereas P-value for price category '$100-$200' was very high at 0.64, hence the impact of this category on Price is just

due to chance. For the price category '>$200 & <$300' the P-value is 0.489 where we can say that this particular category has an impact on sentiment, and this may not be just because of chance. (though it is very close to 0.005 threshold)

Recommendations

- In higher price segments, focusing on exclusive amenities or experiences that justify the higher price, while in lower segments, emphasize value-for-money and essential comforts.Focus on Lower Price Segment for Positive Sentiment: Given that the sentiment for the price category <$100 is positive (0.876) with a very low P-value, it suggests that guests are particularly satisfied with properties in this price range.

- Consider offering more options in this price segment or adjusting prices of some properties to fall within this range to boost overall guest satisfaction.

- To understand customer perception of Airbnb, sentiment might be influenced by other factors like location, neighborhoods. Additionally, competitive pricing can be used by the Airbnb listings to improve the overall sentiment. Focusing on overall service quality, property maintenance, and guest engagement are equally important for maintaining high satisfaction levels.

# *Appendix*

**Question 1:**

```
#Filter data for the US
Airbnb_US <- Airbnb_Listings %>%
  filter(country == "United States of America")

#Group by neighborhood and calculate average price and count of listings
us_neighborhood_prices <- Airbnb_US %>%
  group_by(neighborhood) %>%
  summarise(avg_price = mean(price, na.rm = TRUE),
            count_listings = n())

#Identify cheapest neighborhoods
cheapest_us_neighborhoods <- us_neighborhood_prices %>%
  arrange(avg_price) %>%
  head(10)


#Identify most expensive neighborhoods
most_expensive_us_neighborhoods <- us_neighborhood_prices %>%
  arrange(desc(avg_price)) %>%
  head(10)


#Filter data for Canada
Airbnb_Canada <- Airbnb_Listings %>%
  filter(country == "Canada")

#Group by neighborhood and calculate average price and count of listings
canada_neighborhood_prices <- Airbnb_Canada %>%
  group_by(neighborhood) %>%
  summarise(avg_price = mean(price, na.rm = TRUE),
            count_listings = n())

#Identify cheapest neighborhoods
cheapest_canada_neighborhoods <- canada_neighborhood_prices %>%
  arrange(avg_price) %>%
  head(10)


#Identify most expensive neighborhoods
most_expensive_canada_neighborhoods <- canada_neighborhood_prices %>%
  arrange(desc(avg_price)) %>%
  head(10)


#Visualizations for US-
```

```
#Group by neighborhood and calculate the average for US
us_neighborhood_data <- Airbnb_US %>%
  group_by(neighborhood) %>%
  summarise(avg_price = mean(price, na.rm = TRUE),
            avg_review_score = mean(review_scores_rating, na.rm = TRUE))


#Create a scatter plot
ggplot(us_neighborhood_data, aes(x = avg_price, y = avg_review_score, color = avg_price, size =
avg_price)) +
  geom_point(size = 1) +
  scale_color_gradient(low = "green", high = "brown") +
  labs(title = "Relationship between Neighborhood Cost and Ratings (US)",
       x = "Average Price",
       y = "Average Review Score") +
  theme_minimal()

#Visualizations for Canada

# Group by neighborhood and calculate the average for Canada
canada_neighborhood_data <- Airbnb_Canada %>%
  group_by(neighborhood) %>%
  summarise(avg_price = mean(price, na.rm = TRUE),
            avg_review_score = mean(review_scores_rating, na.rm = TRUE))

#Create a scatter plot

ggplot(canada_neighborhood_data, aes(x = avg_price, y = avg_review_score, color = avg_price, size =
avg_price, alpha = 0.7)) +
  geom_point() +
  scale_color_gradient(low = "blue", high = "red") +
  scale_size_continuous(range = c(1, 5)) +
  labs(title = "Relationship between Neighborhood Cost and Ratings (Canada)",
       x = "Average Price",
       y = "Average Review Score") +
  theme_minimal()
```

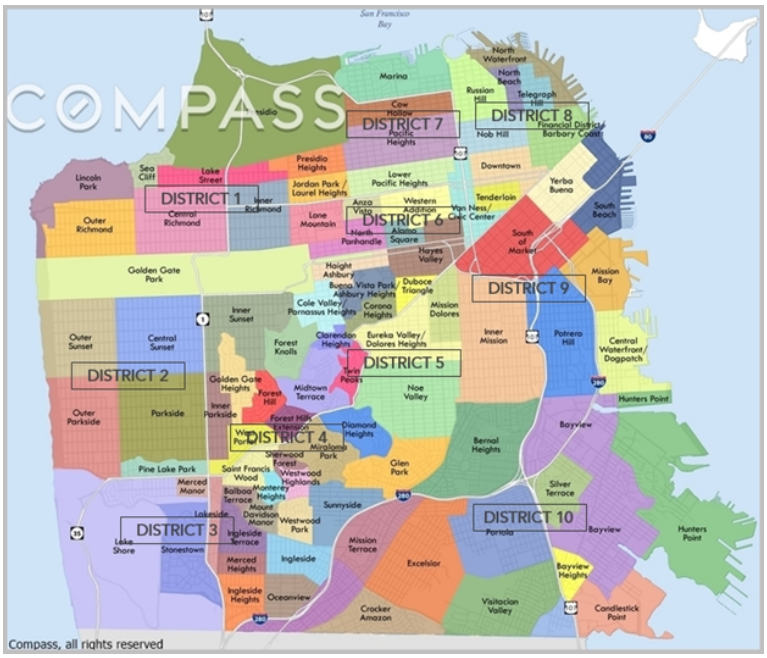
**Question 2:**

```
#Filtering the data
Filter_data1<-Raw_data_Hyp1[Raw_data_Hyp1$city %in%
c("SanDiego","SanFrancisco","SanMateo","SantaCruz"),]

# Create a histogram

ABH1 |>
  group_by(host_is_superhost) |>
  summarise(review_scores_rating = mean(review_scores_rating, na.rm = F)) |>
  ggplot(
```

```
  aes(x = host_is_superhost,
      y = review_scores_rating)) +
geom_bar(stat = "identity")
```

**Question 3**

```
# Installing the packages
#List of packages
pkgs <- c('dplyr', 'ggplot2','reshape', 'car')

# Check for packages that are not installed
new.pkgs <- pkgs[!(pkgs %in% installed.packages()[,"Package"])]

# Install the ones we need
if(length(new.pkgs)) install.packages(new.pkgs)

#Load them all in
lapply(pkgs, library, character.only = TRUE)

# Remove lists
rm(pkgs, new.pkgs)

#Reading the data into R and naming it as Air1
Air1 = read.csv("Airbnb_Listings.csv")


#Filter 1 - Filtered the data by City
Filter_SF<-Air1[Air1$city %in% c("SanFrancisco"),]

#Plot Histogram
hist(as.numeric(Filter_SF$price))
```




```
#Filter 2 - Filtered the data by Neighbourhood
Filter_SF2<-Filter_SF[Filter_SF$neighborhood %in% c("Golden Gate Park", "Inner Richmond"),]

#Plot Histogram
hist(as.numeric(Filter_SF2$price))

#Filter 2 - Filtered the data by entire homes
Filter_SF3<-Filter_SF2[Filter_SF2$entire_home %in% c("1"),]
hist(as.numeric(Filter_F3$price))


#Plotting scatter plot
ggplot(Filter_SF3,
       aes(x = price,
           y = number_of_reviews)) +
  geom_point(position = "jitter",color="brown", size=4) +
  geom_smooth(formula = 'y~x',
              method = NA,
              se = F,
              color = 'orange')
#Running regression with No. of reviews
summary(lm(data = Filter_SF3 ,
           price ~ number_of_reviews))

# Categorizing the reviews
Filter_SF3$number_of_reviews.cat <- ifelse(Filter_SF3$number_of_reviews < 10, "< 10",
                                           ifelse(Filter_SF3$number_of_reviews >= 10 & Filter_SF3$number_of_reviews <=
50, "10-50",
```

```
                                    ifelse(Filter_SF3$number_of_reviews > 50 & Filter_SF3$number_of_reviews
<= 100, ">50 & <=100", NA_character_)))


#Running regression with accommodates
summary(lm(data = Filter_SF3 ,
        price ~ accommodates))
```

**Question 4**

```
#Read the data in R, listings & calendar
Airbnb_sf_calendar<-read.csv("Airbnb_SFcalendar .csv")
Airbnb_listings<-read.csv("Airbnb_listings.csv")

#Select short-term rentals, minimum_nights>=30 days (short-term rental availability))
sf30nights_unique_ids <- Airbnb_sf_calendar %>%
  filter(minimum_nights >= 30) %>%
  distinct(listing_id)

#Join the calander listing ids with Airbnb listing ids
airbnb_joined_data <- inner_join(sf60nights_unique_ids, Airbnb_Listings, by = "listing_id")
str(airbnb_joined_data)
count(airbnb_joined_data)

#Filter to remove hotels
airbnb_joined_data_final<-airbnb_joined_data |>
  filter(entire_home == 1 | private_room == 1 | shared_room == 1)

# Filter for few neighborhoods
airbnb_joined_reduced<-airbnb_joined_data_final[airbnb_joined_data_final$neighborhood %in% c("Pacific
Heights","Daly City", "Noe Valley","Inner Richmond","Downtown/Civic Center"),]

#Run regression
summary(lm(data = airbnb_joined_reduced,
          review_scores_location~neighborhood))
```

**Question 5**

```
#Read the data in R (Airbnb_Reviews)
Airbnb_Reviews<-read.csv("Airbnb_Reviews.csv")

# Identify relevant columns
Raw_data_Hyp2<-Airbnb_Reviews[,c( "city","sentiment","price","neighborhood")]

#Filtering the data
Filter_data2<-Raw_data_Hyp2[Raw_data_Hyp2$neighborhood %in% c("Palo Alto","Los Altos",
"Los Gatos","Mountain View","Redwood City"),]

#Mean for "Sentiment" when null is true (Price=0)
```

```
mean(ABH2$sentiment,na.rm = TRUE)
 #Plotting the null hypothesis which states there's no impact of Price on the Sentiments
ggplot(ABH2,
       aes(x = is.na(price),
           y = sentiment)) +
  geom_point(color="purple", alpha=0.3, size=2) +
  geom_hline(yintercept =
             mean(ABH2$sentiment,na.rm = TRUE))

 #Plotting Alternate hypothesis
hist(ABH2$price, col = "lightblue", main = "Histogram of Price", xlab = "Price")

 #Running Regression analysis
summary(lm(data = ABH2,
           sentiment ~ price))

 #Categorizing by Price
ABH2$price.cat <- ifelse(
  ABH2$price < 100, "< 100",
  ifelse(ABH2$price >= 100 & ABH2$price <= 200, "100-200",
         ifelse(ABH2$price > 200 & ABH 2$price<=300, ">200 & <300", NA_character_))


#Run the Analysis after categorizing
summary(lm(data = ABH2, sentiment ~ price.cat))
```