Introduction

We are a team of five - ***Apurva Shekhar, Karishma Visrodia, Ritu Ranjan Ravi Shankar, Sanjana Ramankandath and Vidhi Gandhi.*** We present here our data analysis, visualizations, and other interesting insights into the Airbnb data. Airbnb is an online marketplace for sharing homes and experiences, where guests who seek accommodation are matched to hosts who have spare rooms to share. Our data consists of data for six cities, New York, Washington DC, Chicago, San Francisco, Boston, and Los Angeles.

Following are the questions that we aim to answer through our analysis:

- How do prices of listings vary by the city?
- How do prices vary according to review scores in each city?
- What is the count of Airbnbs per city?
- Which room type is preferred in each city?
- Are the demand and prices of the rentals correlated?
- What factors determine the price of the Airbnb?
- What is the Airbnb density for each city?
- Are Airbnb rentals spread across each city?
- Does the popularity of a place impact the presence of rentals?
- What is the growth rate of new Airbnb hosts in the cities?
- During which month the highest number of people join Airbnb as hosts in each city?

Description of Data

The data is sourced from the Kaggle website. https://www.kaggle.com/stevezhenghp/airbnb-price-prediction.

The dataset comprises of one main table:
train - Detailed listings data showing 29 attributes for each of the listings.

Here is a detailed description of the attributes:

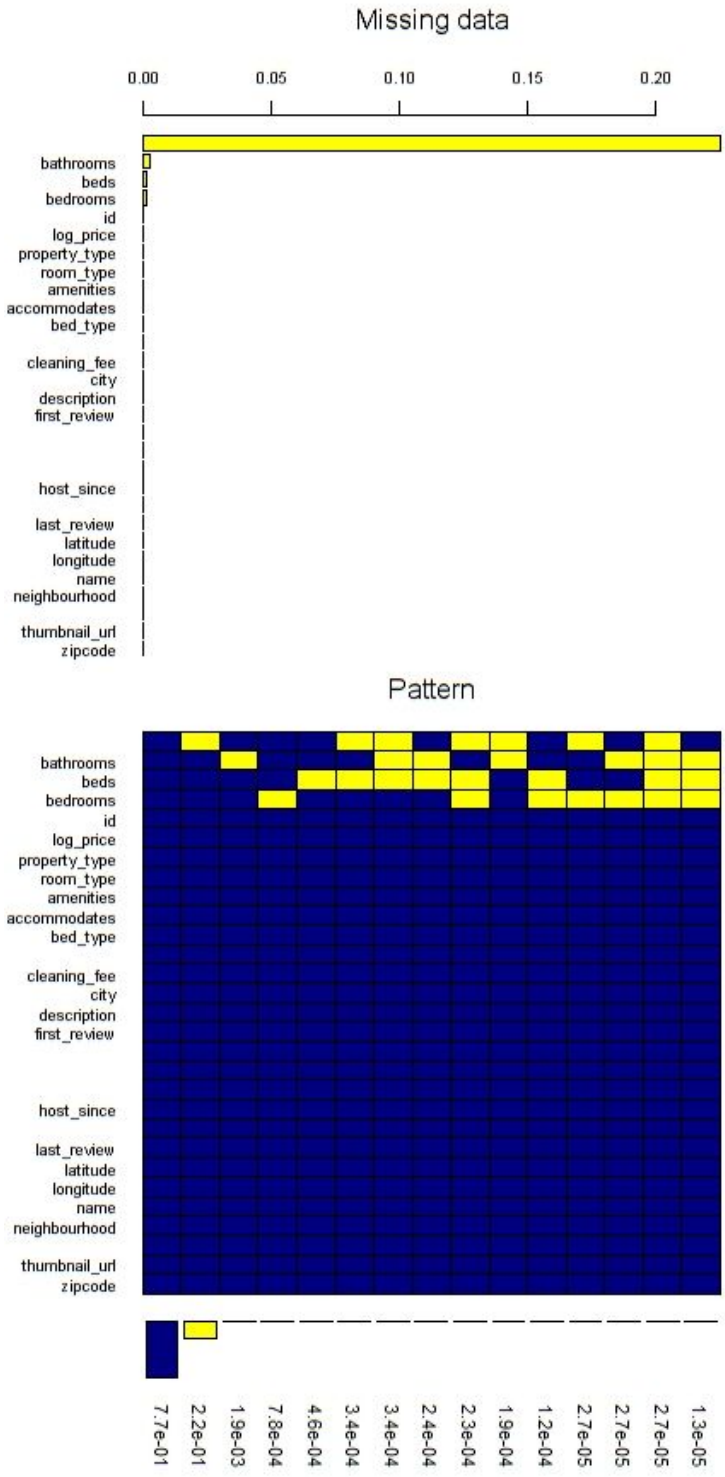
- **id:** A number that uniquely identifies the listing.
- **log_price:** The natural logarithm of the price variable.
- **property_type:**The type of property that is listed.
- **room_type:** The type of room provided in the listing (Entire home/apt, Private room, or Shared room).
- **amenities:** The types of amenities provided by the owners (TV, Wireless Internet, etc.)
- **accommodates:** The number of people that the listing can accommodate.
- **bathrooms:** The number of bathrooms in the listing.
- **bed_type:** The number of bedrooms in the listing.
- **cancellation_policy:** Whether the cancellation policy is flexible, moderate or strict.
- **cleaning_fee:** Whether the cleaning fee is applicable or not.
- **city:** The city where the property is listed.
- **description:** The description of the property on Airbnb.
- **first_review:** When was the first review posted for the particular property.
- **host_has_profile_pic:** Whether the host has a profile picture or not.

- **host_identity_verified:** Whether the host's identity has been verified or not.
- **host_response_rate:** The response rate of the hosts.
- **host_since:** The date when the host signs up to Airbnb.
- **instant_bookable:** Whether the bookings are instantly bookable or not.
- **last_review:** When was the last review posted for the particular property.
- **latitude:** The latitude of the property listing**.**
- **longitude:** The longitude of the property listing
- **name:** The name of the rental property. This is also the property title or heading for the listing on the Airbnb website.
- **neighbourhood:** The neighborhood that the listing is located in (Allston, Back Bay, Beacon Hill, Brighton Downtown, or South End.
- **number_of_reviews:** The number of people who have reviewed for a given property.
- **review_scores_rating:** The review score for a given property. This is calculated on the basis of the review provided by a customer.
- **thumbnail_url:** The Url containing the property details.
- **zip code:** The zipcode of the location where the property is located.
- **bedrooms:** Number of bedrooms in the listing.
- **beds:** Number of beds available in the listing.

A quick glance at the data shows that there are:
74,111 unique listings amongst the six cities in total. Some columns will not be used as features, such as ID and thumbnail URL, so we are left with 26 columns to process and consider as features.
There was no information about this data so we assume that since all the listings are in the US, the price (or log_price) that we are trying to predict is the general pricing per 1 night stay of the listing, in USD, not for specific dates/seasons and not including additional fees, i.e. cleaning and Airbnb service fees.

Identifying Missing Data

The below plot shows the pictorial representation of the missing values in the Airbnb dataset.



We see from the above plot that, most of the columns don't have NULL or NAs. However, we see from the below output that review_score_rating column has the highest number of missing values, followed by the bathrooms, beds and bedrooms.

```
Variables sorted by number of missings:
              Variable       Count
 review_scores_rating 0.225634521
            bathrooms 0.002698655
                 beds 0.001767619
             bedrooms 0.001227888
                   id 0.000000000
            log_price 0.000000000
        property_type 0.000000000
```

Data Preprocessing

Imputing missing values

The missing values in the dataset were imputed as the following:

- Missing values in numeric features were replaced with the mean of that column(mean imputing)
- Missing values in categorical features were replaced with the most frequently occurring value (mode imputing).
- Missing values in date columns were omitted as it couldn't be replaced with any other value.

What factors determine the price of the Airbnb?

Feature Selection

The aim is to predict the price of the Airbnb based on various independent variables(categorical variables and continuous/discrete variables) from the dataset.

We didn't use feature selection techniques to identify the relative importance of the features being considered. The features below were considered based on mere logical reasoning.

Continuous/discrete variables:

- No. of Amenities, No. of Reviews, Review Score rating, No. of Bedrooms, No. of Beds, Accommodates, log price, No of Bathrooms, length of description, length of name.

Categorical Variables:

- Room type, bed type, cancellation policy, cleaning fee, city, host identity verified, instant bookable.

Regression Model Creation – Price Prediction

- Performed a simple linear regression with log price as the Y input and the rest of the predictors as the X input.
- The function used for this purpose is lm(). This function is part of the base R package.
- The model result is shown in the screenshot below.

```
lm(formula = MyRegression_Data_2$MyData.log_price ~ ., data = MyRegression_Data_2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6949 -0.3083 -0.0292  0.2643  3.5602

Coefficients:
                                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)                              4.115e+00  4.915e-02   83.725  < 2e-16 ***
MyData.room_typePrivate room            -6.228e-01  4.229e-03 -147.252  < 2e-16 ***
MyData.room_typeShared room             -1.066e+00  1.128e-02  -94.478  < 2e-16 ***
MyData.num_aminities                     5.005e-03  2.861e-04   17.490  < 2e-16 ***
MyData.accommodates                      8.687e-02  1.616e-03   53.769  < 2e-16 ***
MyData.bathrooms                         1.458e-01  3.880e-03   37.582  < 2e-16 ***
MyData.bed_typeCouch                     1.542e-01  3.670e-02    4.203 2.64e-05 ***
MyData.bed_typeFuton                    -1.100e-02  2.800e-02   -0.393   0.6944
MyData.bed_typePull-out Sofa             4.578e-02  2.951e-02    1.551   0.1208
MyData.bed_typeReal Bed                  3.690e-02  2.209e-02    1.670   0.0948 .
MyData.cancellation_policymoderate      -4.325e-02  5.072e-03   -8.527  < 2e-16 ***
MyData.cancellation_policystrict         1.446e-03  4.695e-03    0.308   0.7581
MyData.cancellation_policysuper_strict_30 3.167e-01  4.551e-02   6.960 3.44e-12 ***
MyData.cancellation_policysuper_strict_60 7.248e-01  1.161e-01   6.243 4.32e-10 ***
MyData.cleaning_feeTRUE                 -5.783e-02  4.431e-03  -13.050  < 2e-16 ***
MyData.cityChicago                      -2.964e-01  1.134e-02  -26.131  < 2e-16 ***
MyData.cityDC                            3.380e-02  1.036e-02    3.264   0.0011 **
MyData.cityLA                           -1.411e-01  8.797e-03  -16.035  < 2e-16 ***
MyData.cityNYC                          -8.120e-03  8.625e-03   -0.941   0.3465
MyData.citySF                            3.121e-01  1.011e-02   30.868  < 2e-16 ***
MyData.num_description                  -5.084e-04  3.522e-05  -14.437  < 2e-16 ***
MyData.host_identity_verifiedf          -5.814e-02  3.504e-02   -1.659   0.0970 .
MyData.host_identity_verifiedt          -7.787e-02  3.497e-02   -2.227   0.0260 *
MyData.instant_bookablet                -6.161e-02  4.091e-03  -15.059  < 2e-16 ***
MyData.num_name                          6.404e-04  9.651e-04    0.664   0.5070
MyData.number_of_reviews                -6.511e-04  4.986e-05  -13.059  < 2e-16 ***
MyData.review_scores_rating              5.715e-03  2.589e-04   22.071  < 2e-16 ***
MyData.bedrooms                          1.428e-01  3.327e-03   42.922  < 2e-16 ***
MyData.beds                             -5.421e-02  2.578e-03  -21.026  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4779 on 74082 degrees of freedom
Multiple R-squared:  0.5565,    Adjusted R-squared:  0.5563
F-statistic:  3320 on 28 and 74082 DF,  p-value: < 2.2e-16
```

Interpreting the Model

The estimate of coefficients of the predictors shows how each predictor would affect the price.

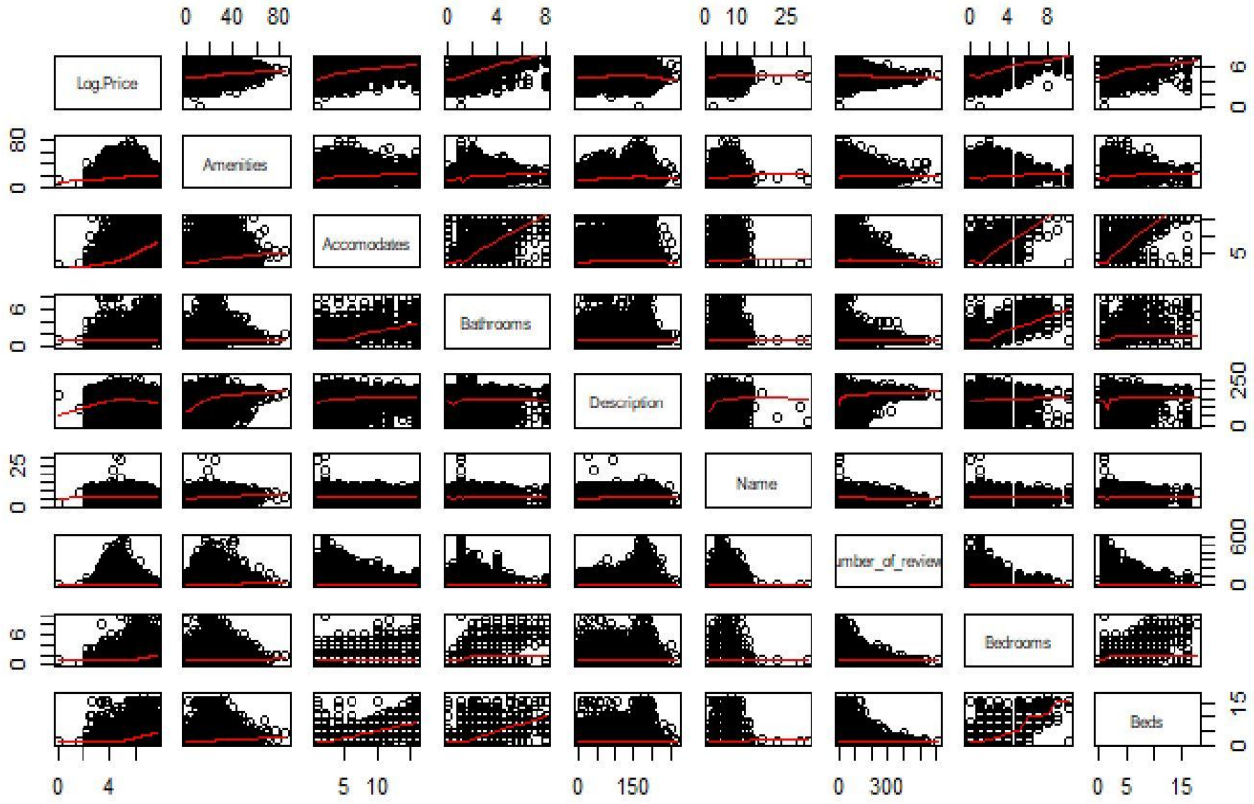
- A positive coefficient of the predictors - no. of bathrooms, no. of bedrooms, accommodates, no. of amenities, etc. shows that if these variables increase, the price would increase proportionately. Also, these variables are significant as indicated by '***'.
- Assuming, number of reviews to be the number of bookings, we see that the coefficient of number of reviews is negative. This shows an inverse relationship between price and number of reviews. As price increases, the number of bookings would go down.
- The R^2 and the adjusted R^2 is 0.56. This tells us that the actual data is close to our regression line by 56%.

One-Hot Encoding:
If we take a closer look at the regression data above, we see that R automatically does "one-hot encoding" for the categorical variables specified under the *feature selection* section.
For example: The variable room-type is a categorical variable with categories as Entire Home, Private room and Shared room. When we ran linear regression, we see that R has automatically performed one-hot

encoding : i.e. split room_type variable into room_type_private_room and room_type_shared_room and ignored Entire Home category as that category is not very significant.

9X9 Matrix plot

The below 9X9 matrix plot, shows the trend of how each variable is affected by other individual variables(didn't consider categorical variable).



In the above graph, the graphs to the left of the diagonal are the inverse of the graphs to the right of the diagonal.

From this graph, we see that with an increase in the number of bedrooms, number of beds, number of amenities available, number of people the rental accommodates, number of bathrooms, there is a direct increase in the price of the Airbnb (*refer to graphs in the first row of the matrix*).

This is consistent with the economic policy of hospitality sectors that an increase in the square feet area, and the availability of other amenities in a rental, increases its price.

Number of reviews, considering it as number of bookings, shows an inverse relationship w.r.t price. We could also see that (*from graph 5th row, 7th column)*, if the description of the rentals is detailed, there is an increase in the number of reviews (number of bookings). We could say that, customers prefer booking rentals that has detailed description of the property, the neighborhood and other details.

Are the demand and prices of the rentals correlated?

Demand and Price Analysis

We analyze the demand for Airbnb listings in six cities – NYC, SF, LA, Chicago, Boston, Washington DC. We are considering the year 2017, as that is the latest year in our dataset.

Our aim is to establish a relationship between price and demand and validate our regression model.

Since we didn't have data on the number of bookings w.r.t to price for each city, we have assumed that the number of reviews = number of bookings. We believe this assumption is a good representation because, on evaluating the first review date and the last review date we see that these two dates are approximately 3 weeks apart.

Assuming that the other predictors considered above remain unaffected we could perform the Price vs Demand Analysis for each city as below:

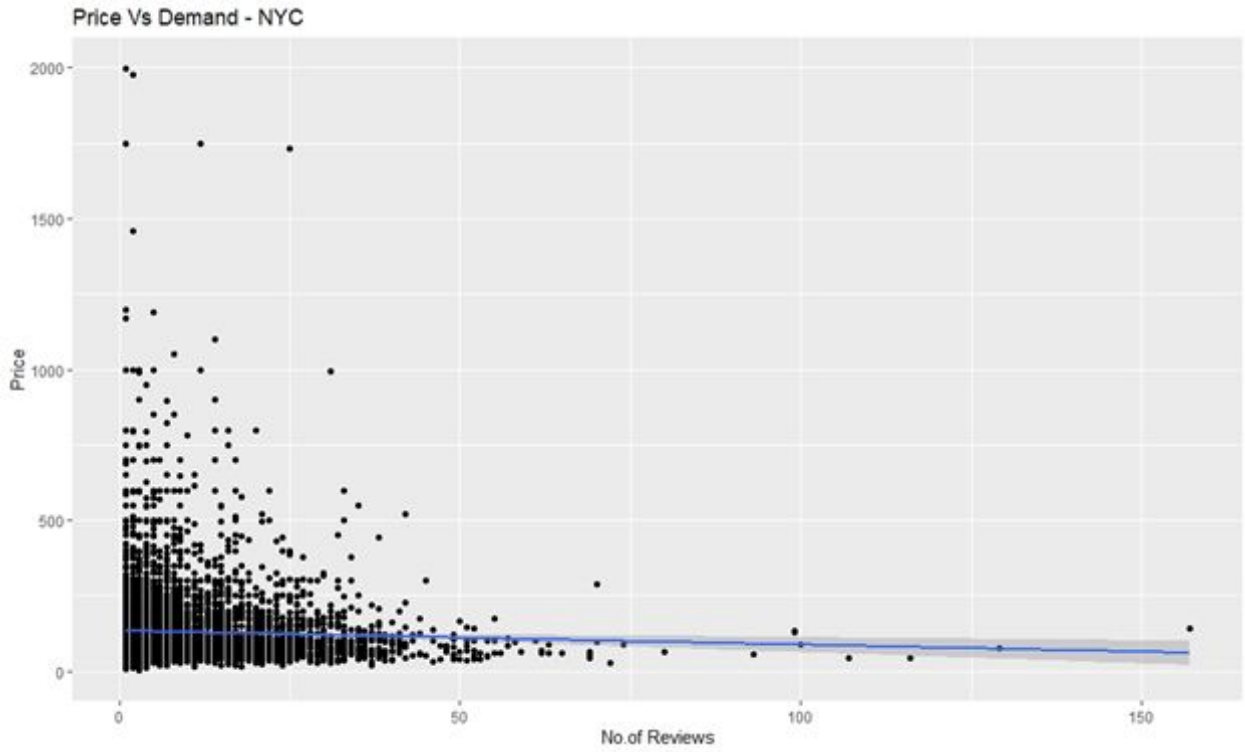
Price vs Demand Analysis – NYC

a) Subset the data from the main dataset where city = NYC , First_Review = 2017, Last_Review = 2017.

   Sample data from the subset.

Price	Number_of_reviews	First_Review	Last_Review	City
169	6	2017	2017	NYC
145	10	2017	2017	NYC
99	5	2017	2017	NYC

b) Plotting graph (ggplot2) with Price as Y axis and Number of Reviews as X axis, we get the demand line as shown by the trendline below.



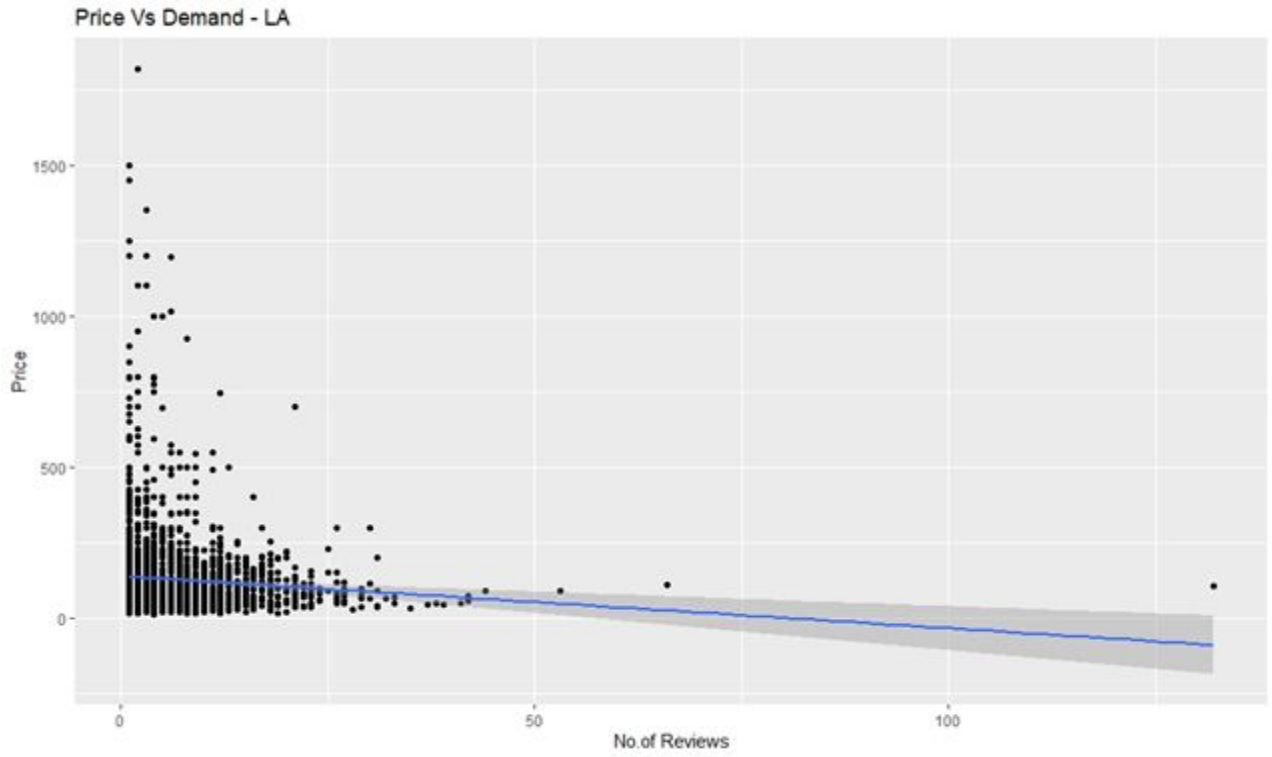
Price vs Demand Analysis – LA

a) Subset the data from the main dataset where city = LA , First_Review = 2017, Last_Review = 2017.

Sample data from the subset.

Price	Number_of_reviews	First_Review	Last_Review	City
83	15	2017	2017	LA
36	2	2017	2017	LA
142	2	2017	2017	LA

b) Plotting graph (ggplot2) with Price as Y axis and No. of Reviews as X axis, we get the demand line as shown by the trendline below.



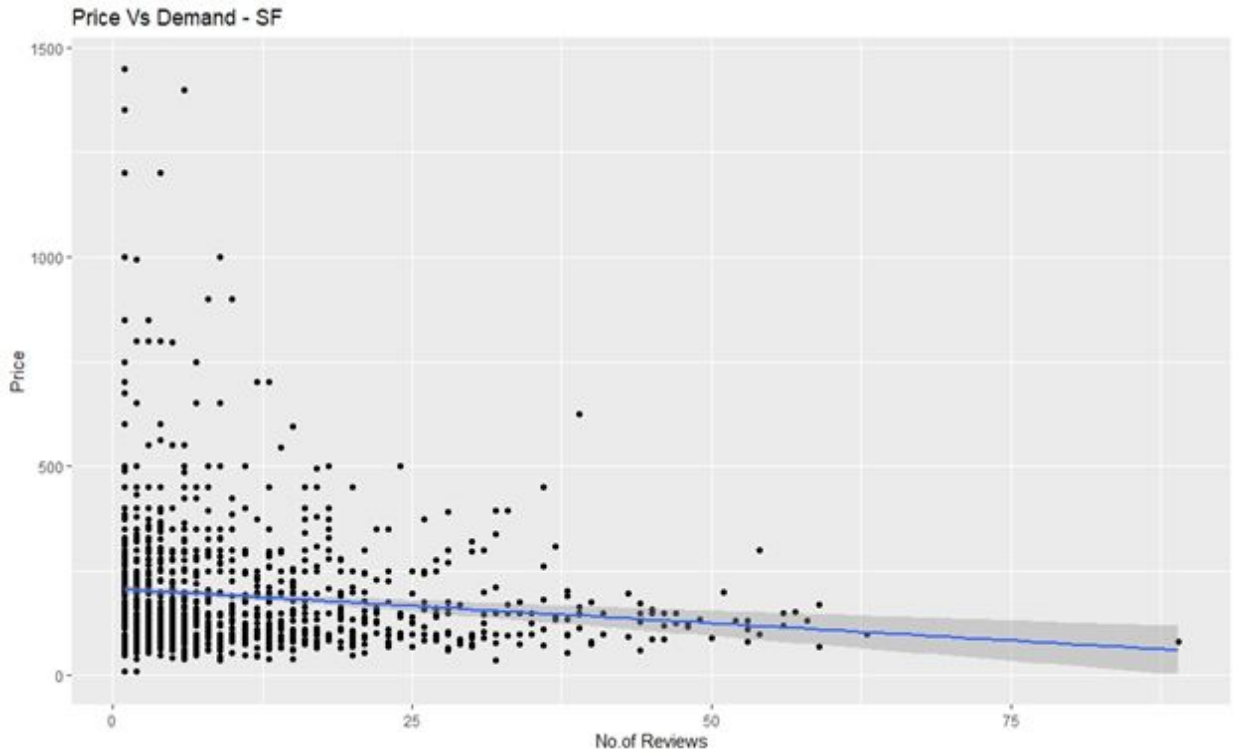
Price vs Demand Analysis – SF

a) Subset the data from the main dataset where city = SF , First_Review = 2017, Last_Review = 2017.

Sample data from the subset.

Price	Number_of_reviews	First_Review	Last_Review	City
85	3	2017	2017	SF
150	1	2017	2017	SF
270	1	2017	2017	SF

b) Plotting graph (ggplot2) with Price as Y axis and No. of Reviews as X axis, we get the demand line as shown by the trendline below.



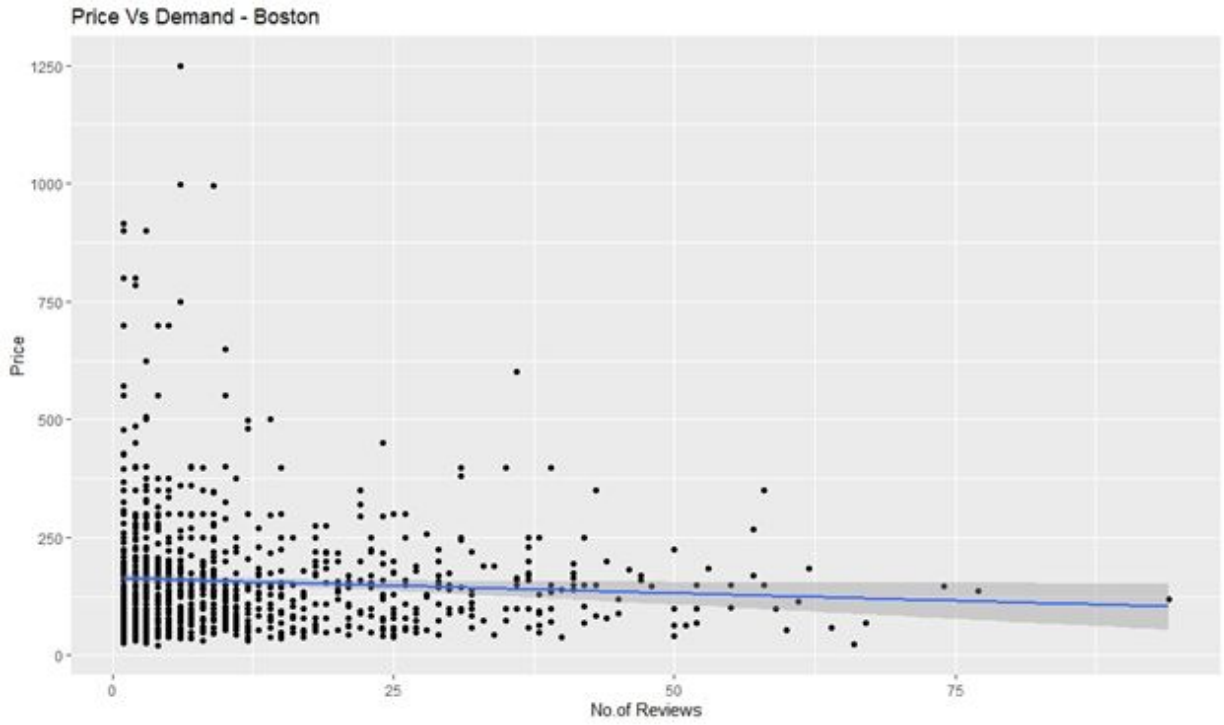
Price vs Demand Analysis – Boston

a) Subset the data from the main dataset where city = Boston , First_Review = 2017, Last_Review = 2017.

Sample data from the subset.

Price	Number_of_reviews	First_Review	Last_Review	City
125	5	2017	2017	Boston
143	2	2017	2017	Boston
55	1	2017	2017	Boston

b) Plotting graph (ggplot2) with Price as Y axis and No. of Reviews as X axis, we get the demand line as shown by the trendline below.



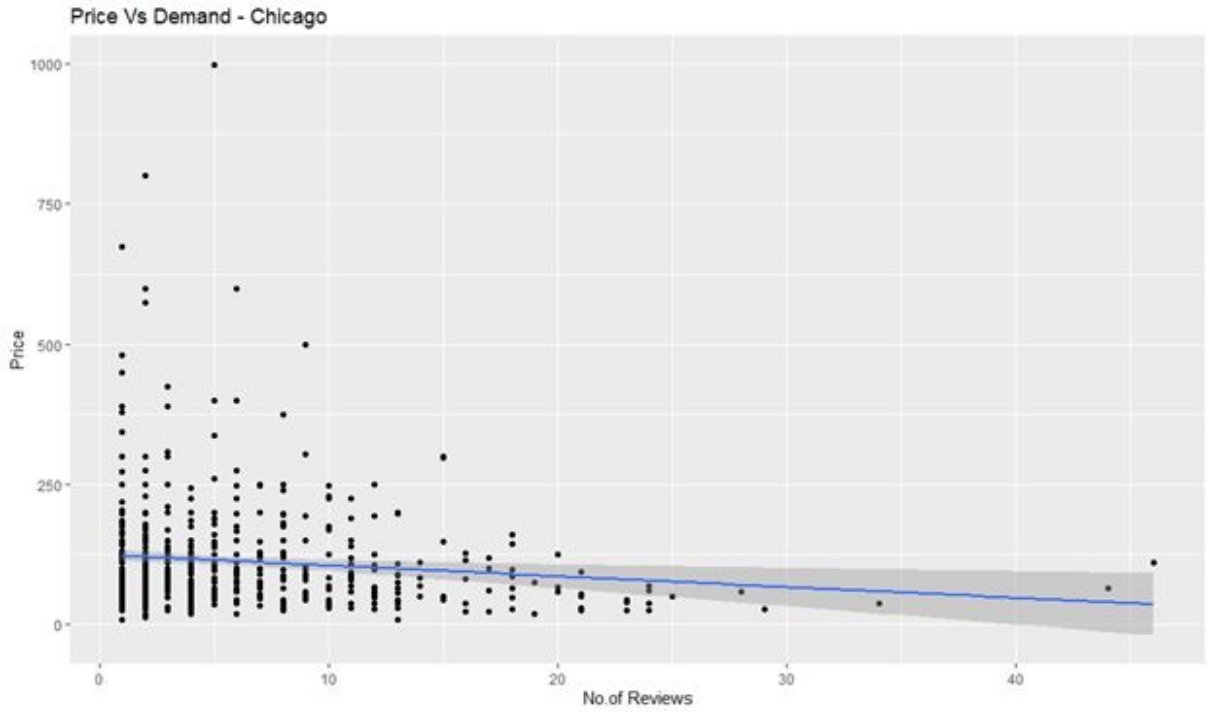
Price vs Demand Analysis – Chicago

a) Subset the data from the main dataset where city = Chicago , First_Review = 2017, Last_Review = 2017.

Sample data from the subset.

Price	Number_of_reviews	First_Review	Last_Review	City
39	16	2017	2017	Chicago
65	4	2017	2017	Chicago
210	3	2017	2017	Chicago

b) Plotting graph (ggplot2) with Price as Y axis and No. of Reviews as X axis, we get the demand line as shown by the trendline below.



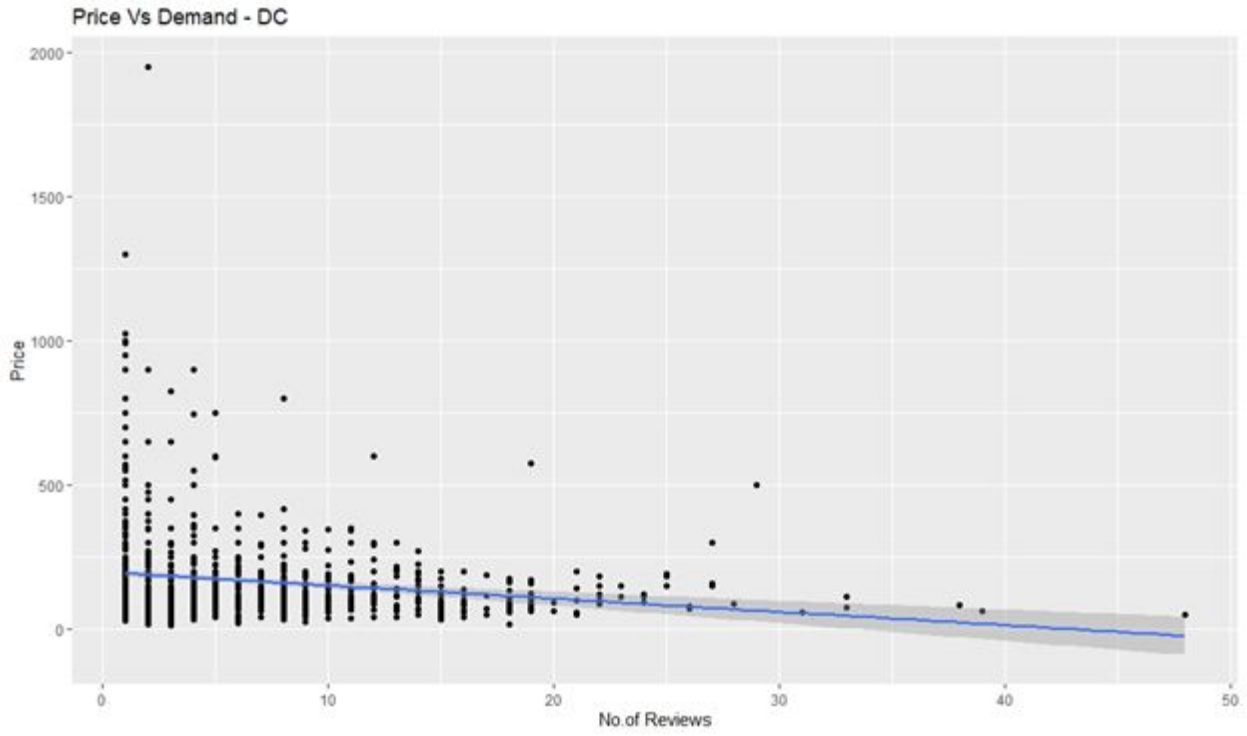
Price vs Demand Analysis – DC

a) Subset the data from the main dataset where city = DC , First_Review = 2017, Last_Review = 2017.

Sample data from the subset.

Price	Number_of_reviews	First_Review	Last_Review	City
200	13	2017	2017	DC
60	1	2017	2017	DC
130	3	2017	2017	DC

b) Plotting graph (ggplot2) with Price as Y axis and No. of Reviews as X axis, we get the demand line as shown by the trendline below.



Observation

We see from the Price vs Number of Reviews graphs above, that price and reviews are following an inverse relationship. This validates our assumption of considering the number of reviews as number of bookings. The trendline in each graph is the respective demand lines for the cities considered.

The Demand lines of NYC, LA and Boston are relatively flatter than the rest. This might indicate that a smaller change(decrease) in price might lead to an increased demand of Airbnb rentals.

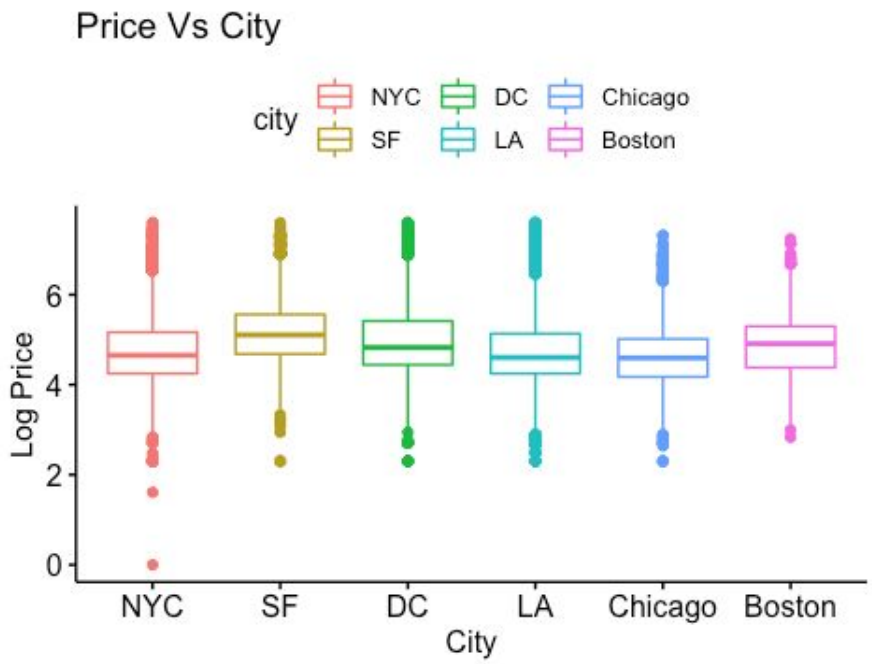
How do prices of listings vary by the city?

The log price varies from 0 to 7.6.

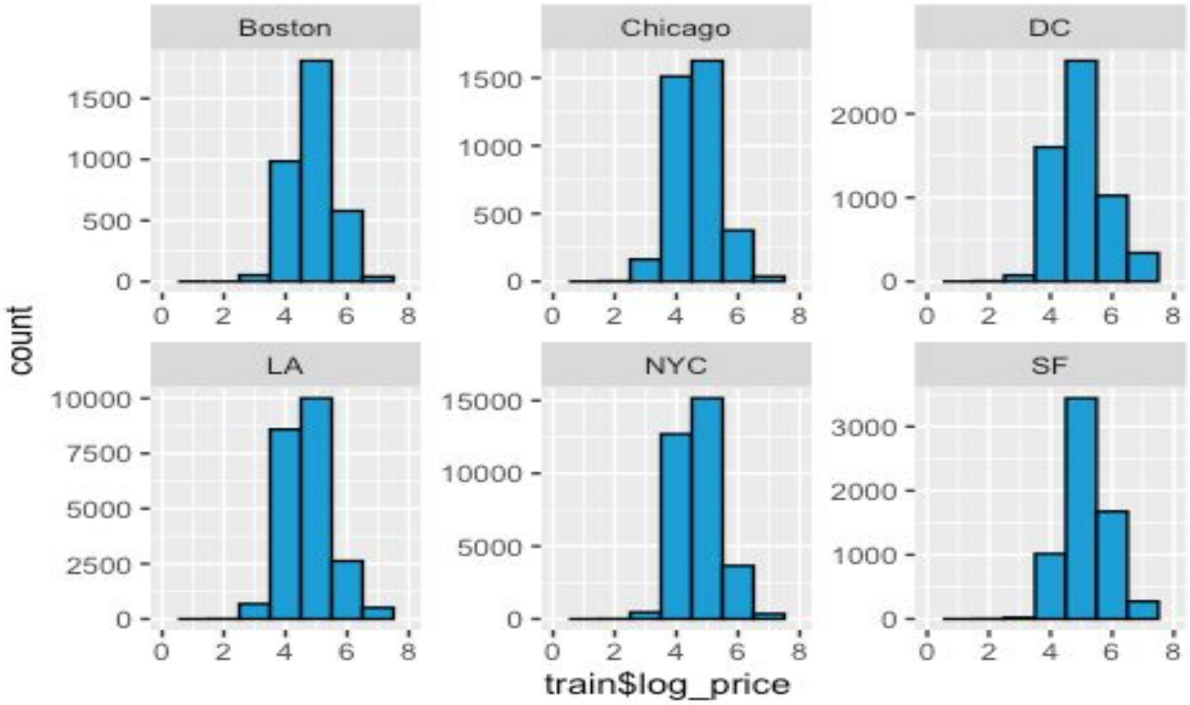
Summary of each city with respect to the log_price:

```
train$city: Boston
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.833   4.382   4.913   4.884   5.298   7.244
------------------------------------------------------------
train$city: Chicago
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.303   4.174   4.595   4.620   5.017   7.313
------------------------------------------------------------
train$city: DC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.303   4.443   4.828   4.987   5.416   7.600
------------------------------------------------------------
train$city: LA
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.303   4.248   4.605   4.720   5.136   7.600
------------------------------------------------------------
train$city: NYC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   4.248   4.654   4.719   5.165   7.600
------------------------------------------------------------
train$city: SF
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.303   4.682   5.106   5.170   5.561   7.598
```

Box plot of each city wrt log_price:



Summary of all the six cities and log_price:



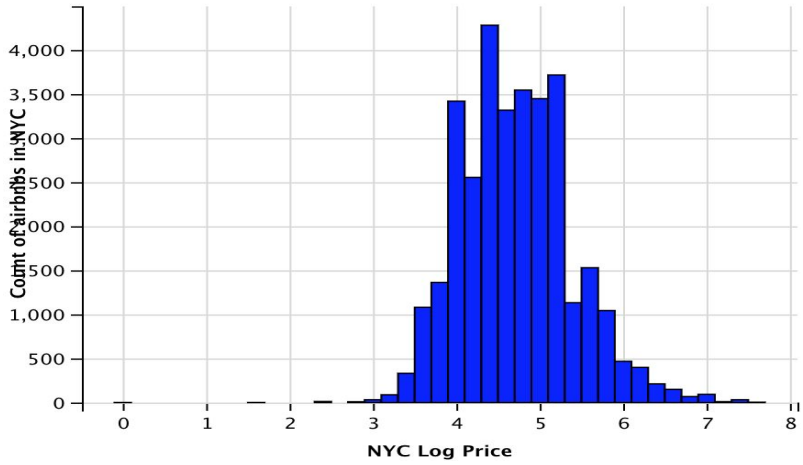
As we see from the summary and the plots above, the average cost of the rental is least in Chicago (4.6). Boston and Chicago have the most variation in terms of the price of rental properties. SF has the highest average cost of rental(5.17). NYC has cheapest rental option (from 1.6).

What is the count of airbnbs per city?

Boston	Chicago	DC	LA	NYC	SF
3468	3719	5688	22453	32349	6434

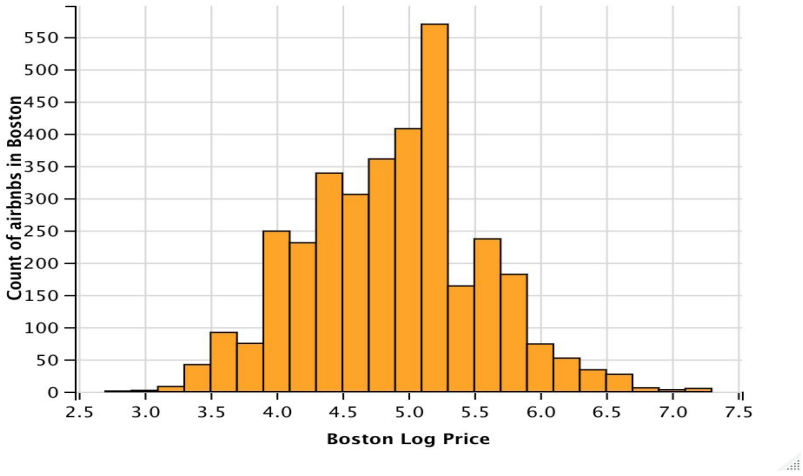
For the data of airbnbs per city, we selected log_price and city attributes.
Summary of NYC data along with the plot of the count of Airbnbs in NYC wrt log_price:

```
     city              log_price
 Length:32349       Min.   :0.000
 Class :character   1st Qu.:4.248
 Mode  :character   Median :4.654
                    Mean   :4.719
                    3rd Qu.:5.165
                    Max.   :7.600
```



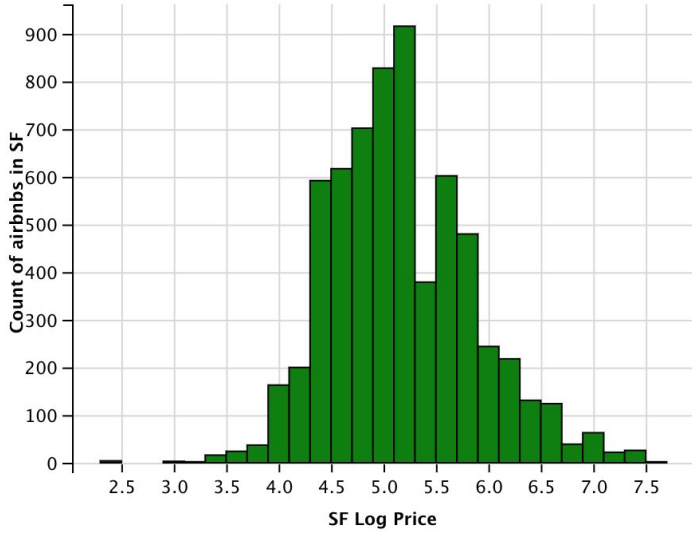
Summary of **Boston** data along with the plot of the count of Airbnbs in Boston wrt log_price:

```
     city              log_price
 Length:3468        Min.   :2.833
 Class :character   1st Qu.:4.382
 Mode  :character   Median :4.913
                    Mean   :4.884
                    3rd Qu.:5.298
                    Max.   :7.244
```



Summary of **SF** data along with the plot of the count of Airbnbs in Boston wrt log_price:

```
     city              log_price
 Length:6434        Min.   :2.303
 Class :character   1st Qu.:4.682
 Mode  :character   Median :5.106
                    Mean   :5.170
                    3rd Qu.:5.561
                    Max.   :7.598
```

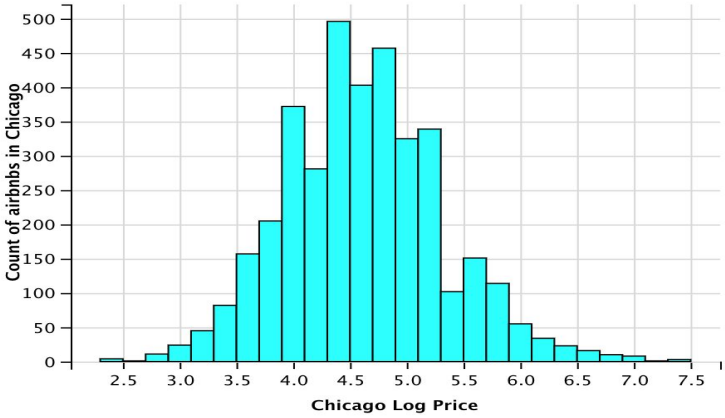


Summary of **DC** data along with the plot of the count of Airbnbs in Boston wrt log_price:

```
    city              log_price
Length:5688        Min.   :2.303
Class :character   1st Qu.:4.443
Mode  :character   Median :4.828
                   Mean   :4.987
                   3rd Qu.:5.416
                   Max.   :7.600
```

Summary of **Chicago** data along with the plot of the count of Airbnbs in Boston wrt log_price:

```
    city              log_price
Length:3719        Min.   :2.303
Class :character   1st Qu.:4.174
Mode  :character   Median :4.595
                   Mean   :4.620
                   3rd Qu.:5.017
                   Max.   :7.313
```

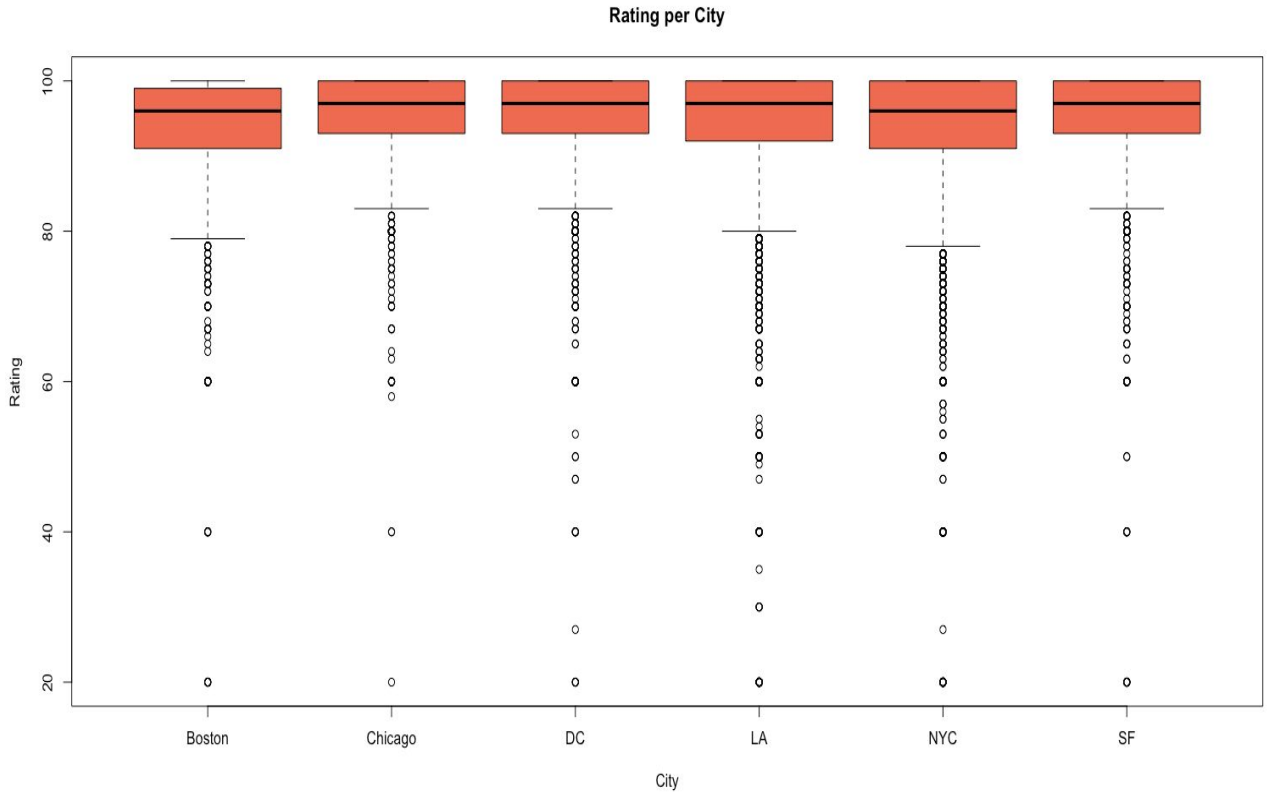


As we see from the above data, NYC has the most number of Airbnb rentals (32349). Also, as seen from all the individual city vs price plots, NYC has the least variation. Most of the rentals for NYC are priced between 3 and 6.5. The listing costs (log_price) are largely in line with the cities.

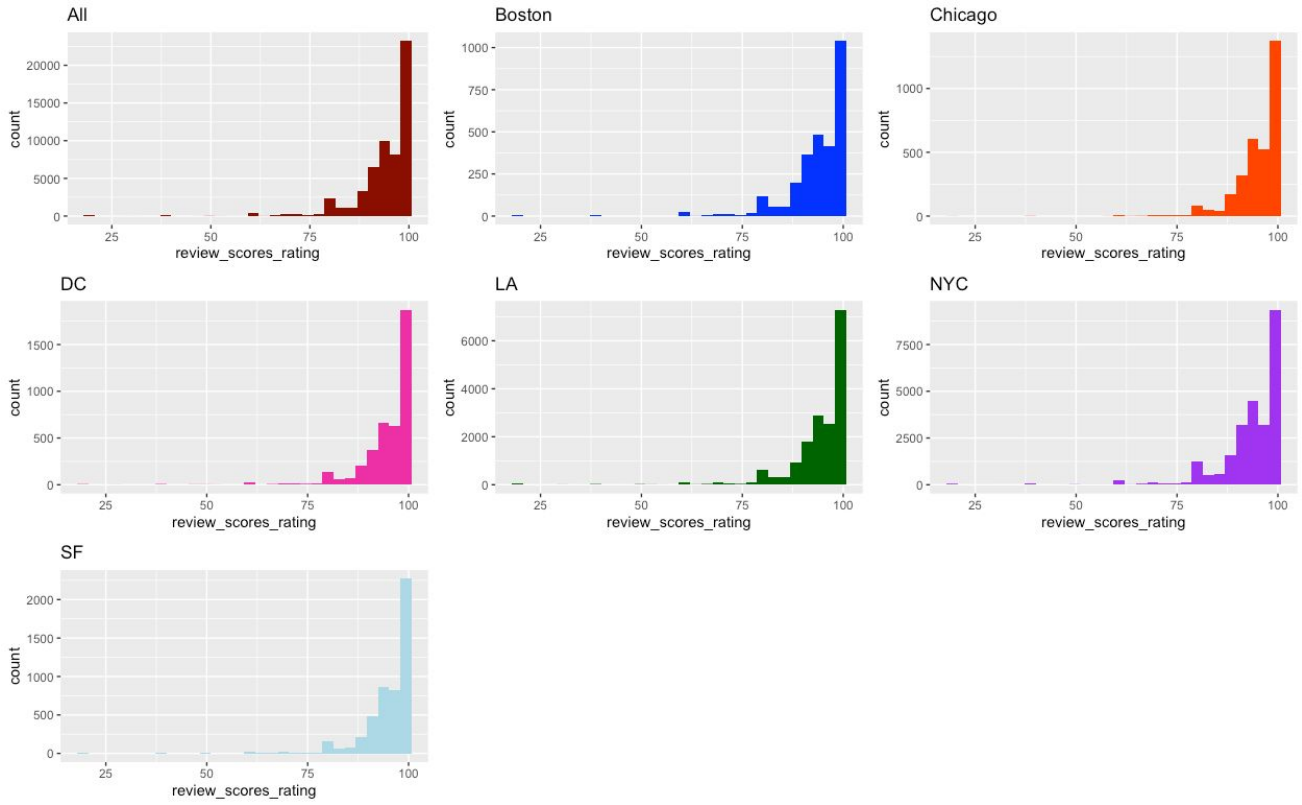
How do review scores vary in each city?

We used review scores rating to find out which city has the highest number of highly rated listings. The box-plot shows that Chicago, DC and SF are the top three highly rated cities. These cities have the biggest fraction of listings that are very highly rated.

Even though NYC has many listings which are on the costlier side, the median review score is NYC is low compared to other cities.

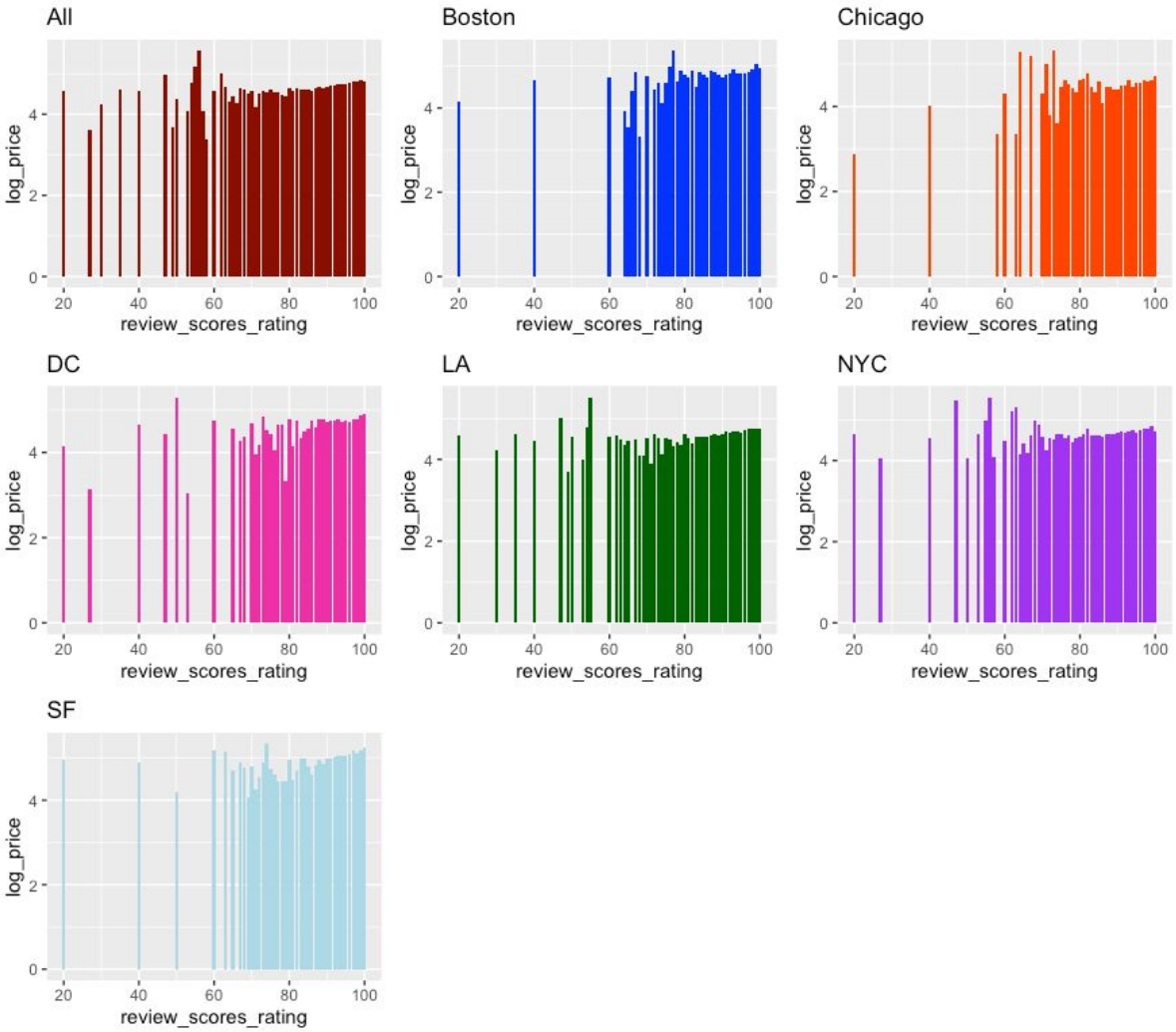


We plotted the review scores rating of all listings in each city as well as across all cities. We found that a large fraction of listings have high rating across all cities. This could either mean that customers usually log into the system to give good ratings or that hosts with low ratings quickly go out of the market.



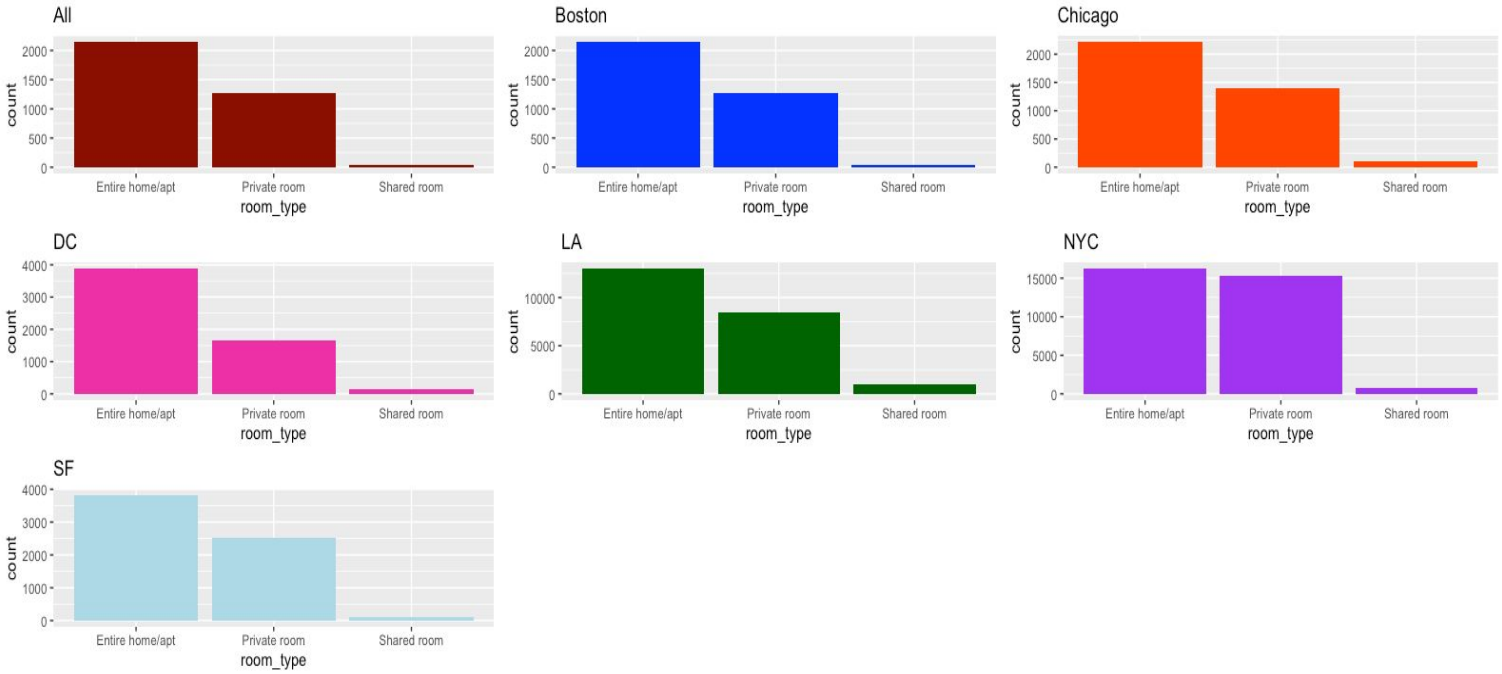
How do prices vary according to review scores in each city?

We plotted review rating in each city against log price to find out how prices vary according to rating in each city. We observed that log price decreases with review scores at the top of the spectrum. The lower half of review spectrum is quite unpredictable. Sometimes the price of very poorly rated properties is more than highly rated properties as well.



Room Preference In Each City:

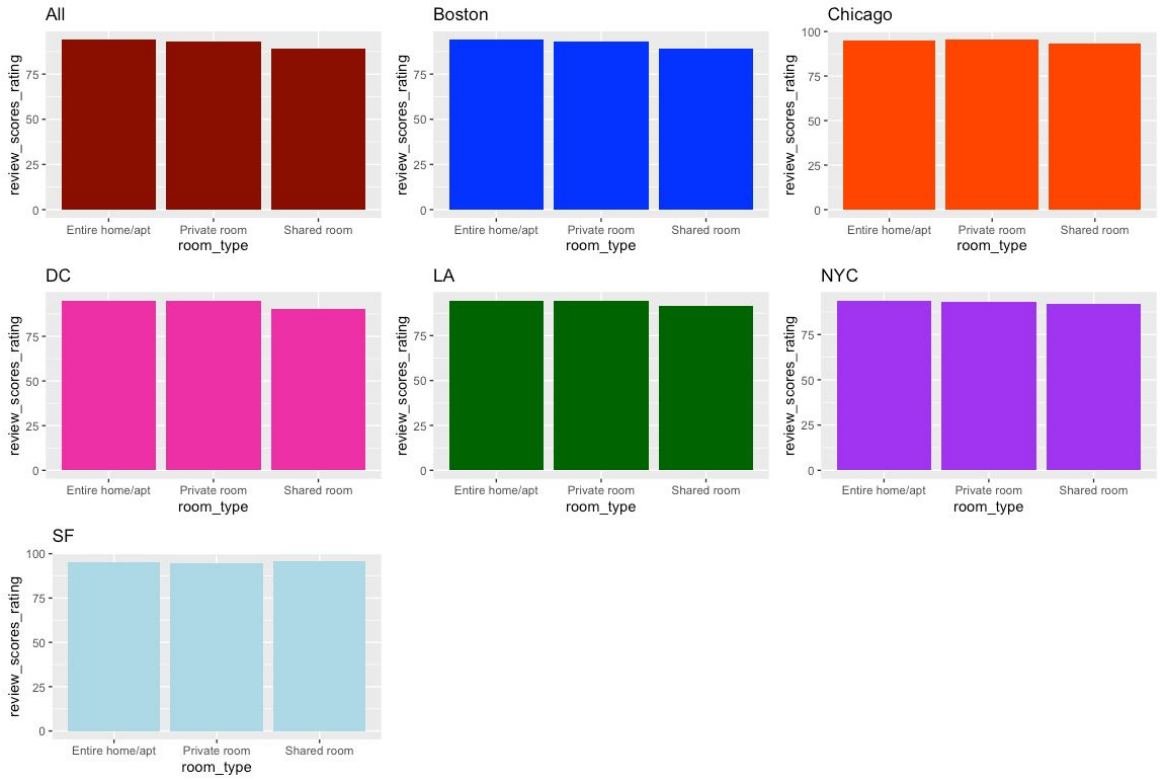
To find out the room preference across all cities, we plotted the number of listed room of each type for every city.



We see that in all cities most number of room types is "Entire home". The ratio of number of "Entire home" listings to "Private room" listings is also similar across all cities except New York. New York has an exceptionally large number of "Private Room" listings. This can be attributed to the lack of space in NYC. Also, LA has the most number of "Shared room" options, closely followed by New York.

Highly Rated Room Types In Each City:

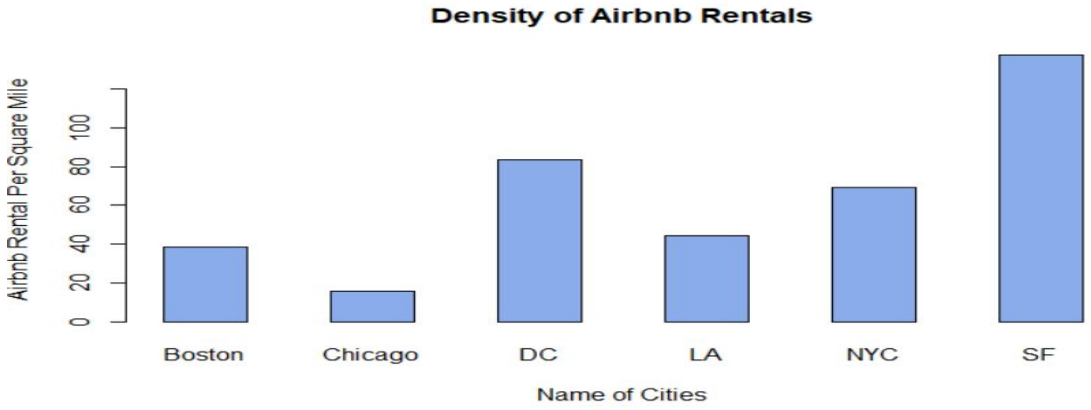
We compared the room type in each city as well as across all cities to find out the most highly rated room type. We observed that “Shared rooms” are highly rated in SF, even though the most pricey listings are in NYC. This might be due to more budget travellers who come to SF to explore the area.



Airbnb rentals per square mile for each city

The total number of Airbnb rentals is the highest for the city of New York(**32349)** while Boston(**3468)** and Chicago(**3719)** have the lowest number of rentals. However, if we calculate the number of rentals per square mile for each city, San Francisco seems to be leading the chart followed by DC and New York. Chicago seems to be in the lower half for both metrics which seem to denote that the city is not as widely adopted by Airbnb hosts compared with the rest of the cities.

```
Airbnb rental per square mile for each city
     Boston   Chicago        DC        LA       NYC       SF
A  38.69240  15.89316  83.23090  44.63817  69.12179 137.27331
```

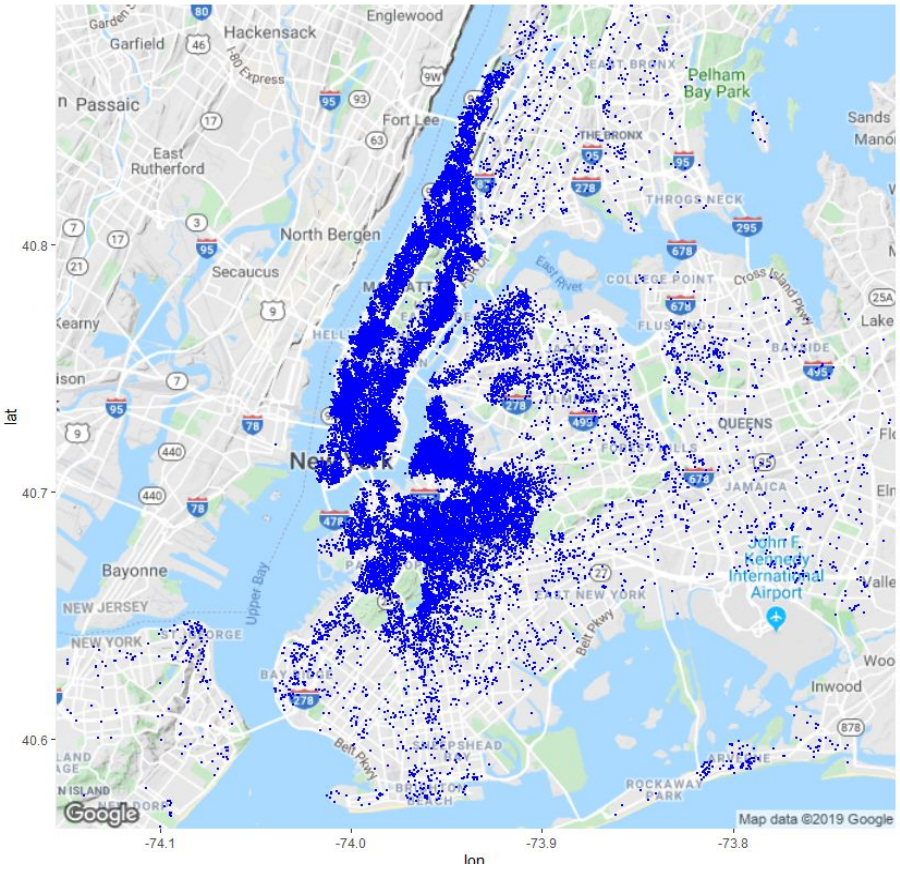


Density of Airbnb rentals based on zip code

NEW YORK

If we look at the density of the Airbnb rentals in the city of New York, we can observe that the rentals are concentrated near Manhattan and Brooklyn. In Manhattan, this could be attributed to the presence of tourist spots like the Statue of Liberty, Central Park, Times Square, and New York downtown. Brooklyn also has a fairly high number of rentals due to tourist spots like Brooklyn Bridge Park, Brooklyn Botanic Garden, etc.

Surprisingly Airbnb rentals near JFK Airport are sparse. This gives an indication that tourists tend to prefer rentals in the heart of the city.

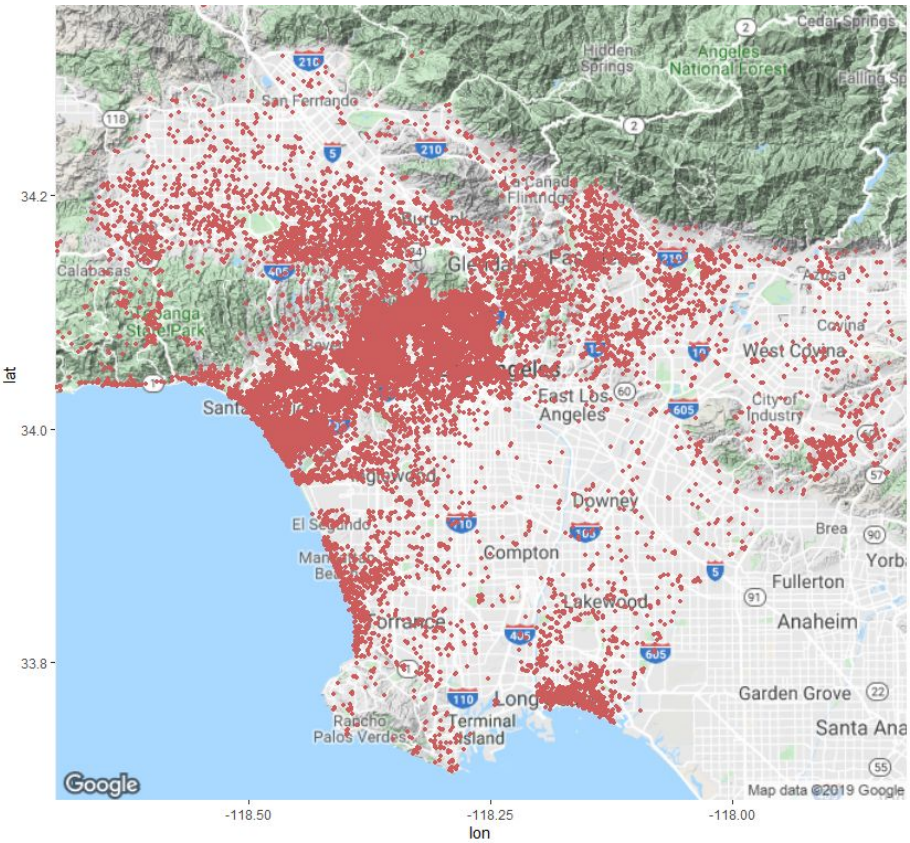


LOS ANGELES

Los Angeles is a Southern California city and the center of the nation's film and television industry. There are plenty of Airbnb rentals near the areas of Beverley Area, Hollywood, Santa Monica and which are some of the most popular attraction in the USA. Los Angeles Airport surrounding is also having more Airbnb's.

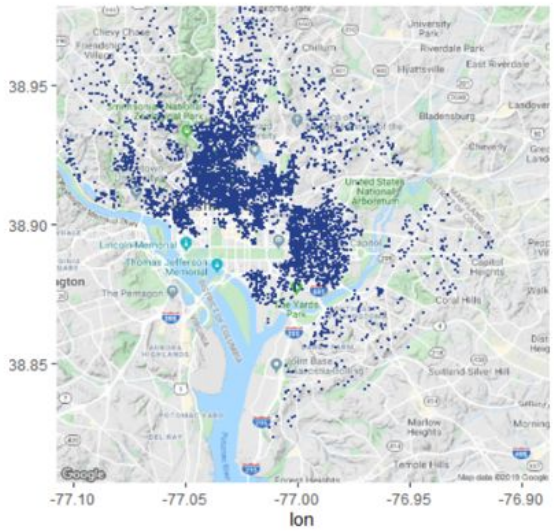
One observed interesting fact from this plotted map is the abundance of Airbnb rentals along the coastal areas, specifically near the popular beaches like Santa Monica, Long Beach.

Compared to JFK, LAX airport has far more rentals around it. This could be due to the fact that the airport is closer to Santa Monica which is a convenience for tourists.

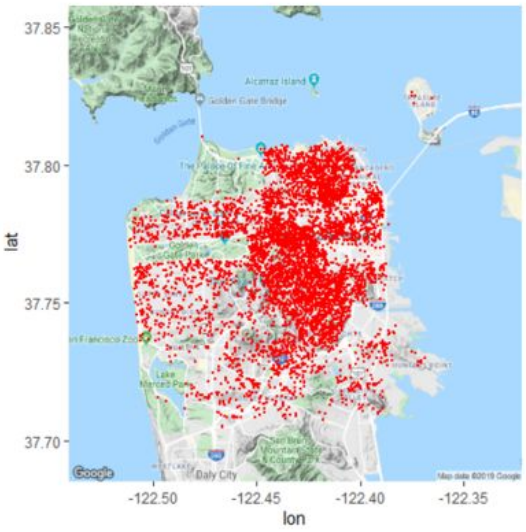


WASHINGTON DC & SAN FRANCISCO

While DC and SF have fewer number of rentals (5688 and 6434) they still have higher density compared to NYC. For instance SF has the highest Airbnb rental per square mile. In DC airbnb rentals are packed near the Capitol and White House.



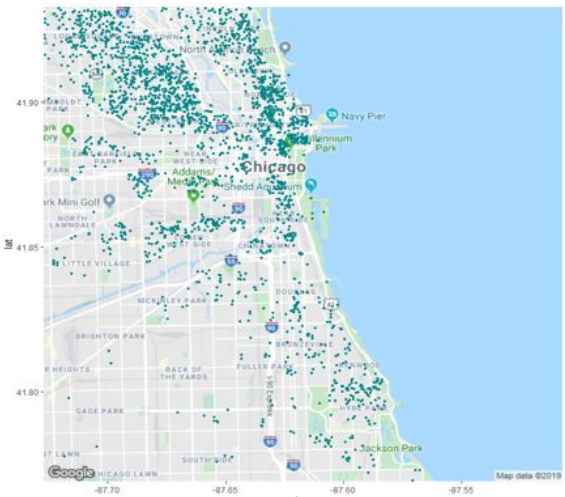
Washington DC



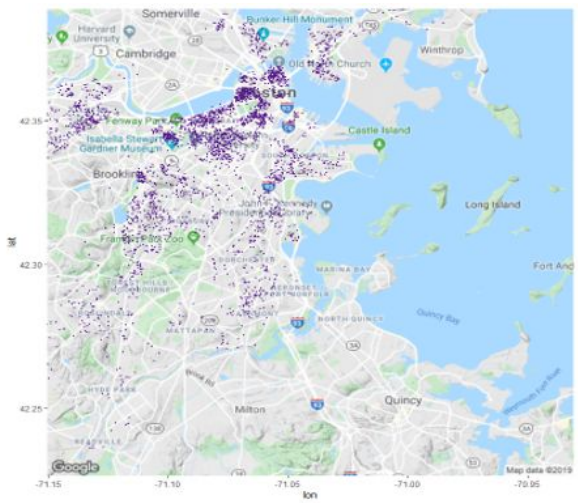
San Fransisco

CHICAGO & BOSTON

Both Boston and Chicago have fewer number of Airbnb rentals but as observed above Chicago seems to have lower number of Airbnb rentals per square mile. In both the cities Airbnb are more concentrated in the downtown area.

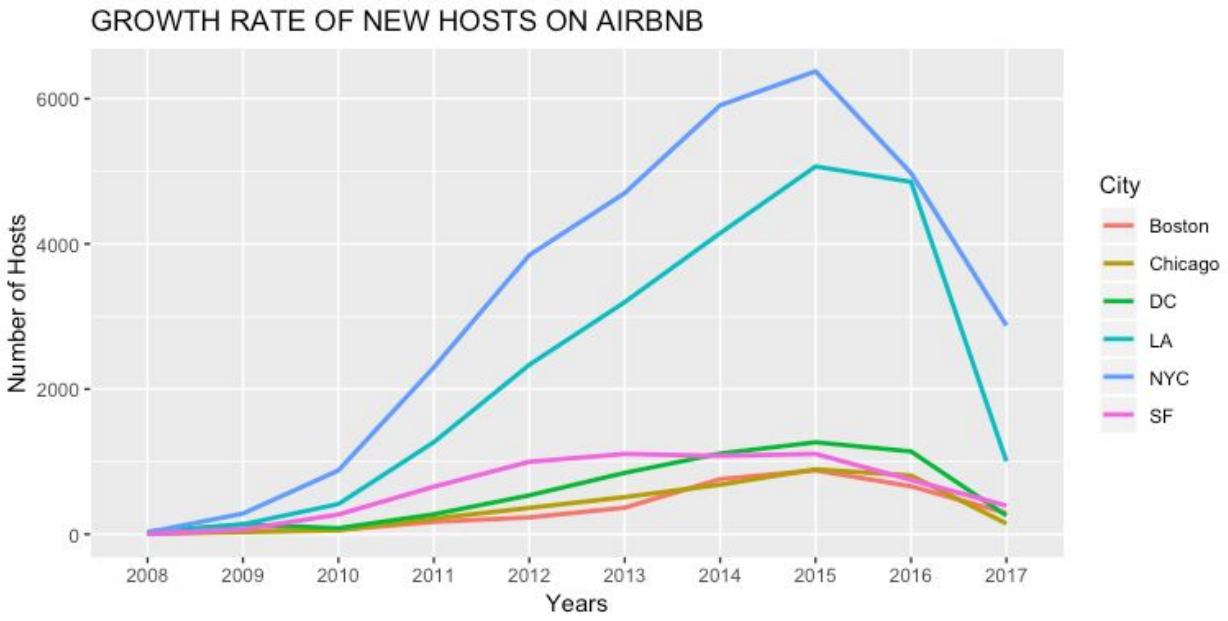


Chicago



Boston

What is the growth rate of new hosts in the cities ?



The above graph depicts the number of new hosts signing up with Airbnb from various cities every year from 2008 to 2017. It can be observed that there had been a large increase in the number of hosts signing up for NYC and LA, whereas the figures for new hosts from the other cities: Washington DC, SF, Chicago & Boston had only shown a slight growth.

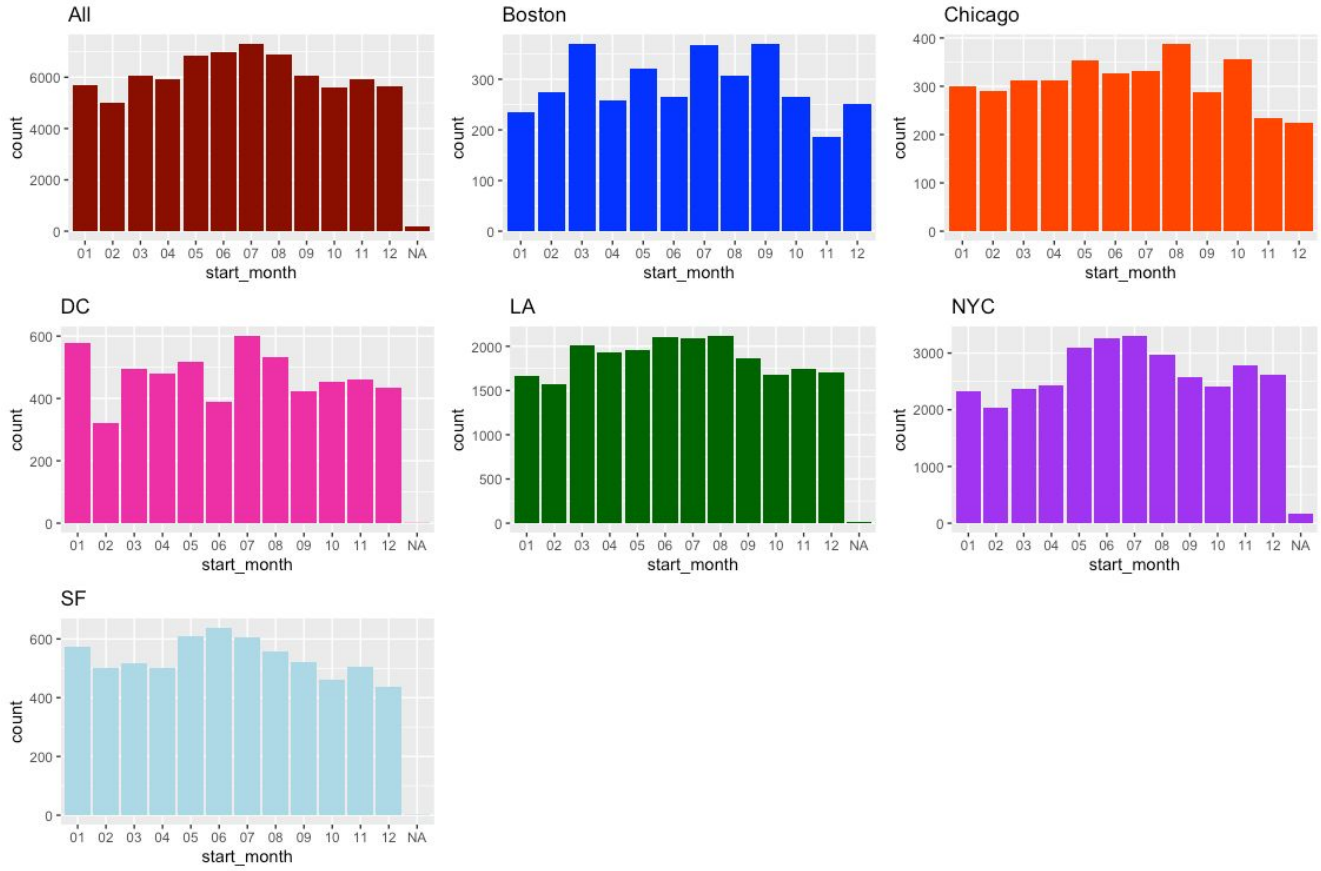
The total number of new hosts for NYC had risen considerably till 2015, after which it started showing a decline. In 2015 there were approximately 6300 new hosts in NYC. The number of new hosts for NYC climbed to approximately 850, and then doubled in only one year later in 2011. Since 2011, there has been a steady increase till 2015, after which the number of new hosts signing up declined rapidly.

The number of new hosts for LA followed a similar trend as NYC, showing a steady increase in the numbers from 2011 to 2015, after which the numbers dropped down considerably. The number of new hosts signing up was at its peak during the year 2015.

For the remaining cities, the number of new hosts signing up increased slowly from 2011 to 2015. After 2015, these cities too started to show a similar pattern as to NYC and LA as new hosts signing up on Airbnb started to decline.

During which month highest number of people join airbnb as hosts in each city?

We plotted the bar graph to find out during which month highest number of people join airbnb as hosts in each city. We counted the number of people who joined airbnb every month across all cities.



We see that across all cities, the number of people who join as hosts is relatively high in summer. This trend seems to be in line with the fact that a large number of foreigner tourists visit the USA during summer. As NYC, LA and SF are major tourist attractions in the US, the number of people who join as hosts every summer is relatively higher in these cities than in other US cities.

Conclusion

Through this exploratory data analysis and visualization project, we gained several interesting insights into the Airbnb rental market.

Besides gaining interesting insights into the Airbnb rental market , we acquired several technical and soft skills along the way. Dealing with multiple data formats helped us strengthen our skills in data manipulation and cleaning. We learned how to work on different R frameworks and libraries to create interactive visualization. Working in a team environment gave us an opportunity to collaborate with our classmates, understand different viewpoints and learn from each other.

We want to expand our analysis to multiple cities and compare patterns and trends amongst more cities including more attributes for analysis. From the insights we have derived, we would also like to build predictive models using different features from the dataset. Lastly, we hope to implement the visualizations and techniques used in this project to many other fields and datasets.

References

1) https://github.com/dkahle/ggmap