Analysis of Airbnb Dataset

CSE-544 : Project Report

Aishwarya Danoji- 111493647, Rahul Bhansali- 111401451,
Sneha Pathrose- 111447763, Gaurab Bhattacharya- 170048888,
Mohit Khemchandani - 111491667

May 2018

1 Introduction

Airbnb is a peer-to-peer online marketplace and home stay network for people to list, book or rent short-term lodging in residential properties, with the cost of such accommodation set by the property owner. The people who list bookings are the 'hosts' and the ones who rent these listings are the 'guests'. There has recently been a surge of people using Airbnb for traveling, business and home stay. Thus, it would be of great use if the hosts knew how to set a fair rental price and the factors that affected their number of bookings. Similarly, the guests could benefit with the knowledge of knowing what would be a cost-effective time to visit certain destinations and plan their trips accordingly.

Our project aims to analyze the factors that would enable Airbnb hosts to profit the most from their listings by increasing the chances of their listings getting booked. It would also help the Airbnb guests to make wise decisions regarding which listings to book and when to book them. The chief dataset used for this project was the Airbnb dataset for New York City (NYC) from 2015-17. Crime and Population datasets of NYC were also referred to for certain hypotheses. Besides this, online websites were used to get the number of subway stations and best family neighbourhoods in NYC.

Overall, we cover 5 broad questions that we wish to answer and the range of techniques covered to answer these questions include: Parametric and Non-parametric inference, Hypothesis Testing, Linear Regression, Time-series analysis and Bayesian Inference.

2 Dataset

2.1 Dataset Description

The primary dataset used is the Airbnb dataset for NYC from 2015-17 obtained from the Inside Airbnb website. This dataset is provided month-wise. In total we had 33 months' data (3 months' data was missing). Each row in the dataset represents a listing and the columns are all the features associated with that listing (e.g: bedrooms, bathrooms, is host a superhost etc). On an average, there are 30,000 rows for each month. The number of columns is 95 for all months.

We also used the NYC neighbourhood-wise population dataset from the website nyc.gov. This dataset contains the total population and persons-per-acre for each neighbourhood of NYC. We used the persons-per-acre feature to classfiy different neighbourhoods of Manhattan into highly populated or lowly populated areas which was used for certain hypotheses (like checking whether the type of area affects the number of bookings or listings).

Crime Dataset for NYC was also used from the website: addressreport.com. It contains neighbourhood -wise crime rate data of each neighbourhood of NYC and was used to classify different neighbourhoods of Manhattan into low or high crime areas based on the crime rate of that area and comparing it to national average crime rate. This information was made use in the hypothesis to check if the type of area affects the number of bookings or listings.

Besides these, manual work was done to obtain the number of subway stations in each neighbourhood of Manhattan (which was used as a feature to predict the price of a listing) through GoogleMaps. We also classified each neighbourhood of Manhattan into a family zone area or non-family zone area using rankings available from the website niche.com.

Dataset Links:
Airbnb Dataset
Crime Dataset
Population Dataset
Family neighbourhood classification

2.2 Data Preprocessing

Each topic involved separate data processing of the Airbnb dataset to extract the required columns and use those columns to obtain the required information. Some of the topics also involved using the Crime, Population, Subway or Neighbourhood information.

One key aspect of the data that was missing in the raw dataset was the number of bookings that were made for a listing for each month. To handle this problem, we made use of the column "number of reviews" of the dataset. So, firstly, we got the common listings for two consecutive months by doing a join operation on the listing IDs. Then, we subtracted the "number of reviews" of the consecutive months to obtain the count of the number of bookings made for each listing. The assumption made here is that if a guest books a listing then the guest writes a review for that listing. We feel that this assumption is reasonable enough and would not lead to an overestimation of the number of bookings made for a listing.

Another pre-processing that was done was to add the columns of whether a neighbourhood was criminal or not (using Crime dataset), whether a neighbourhood was highly or lowly populated (using Population dataset), the number of subway stations in a neighbourhood and whether a neighbourhood was a family-zone or not.

3 Topics Analysed

- Predicting listing price/ price category.
- Does type of area affect bookings and listings?
- Do events, holidays and seasons affect average price of a listing ?
- Host dependent features that affect the number of bookings.
- Predicting average number of bookings for the next month(s) according to hosts.

3.1 Predicting listing's prices/price category on Airbnb dataset

Nowadays, more and more people use Airbnb of travelling, bussiness and homestay. However, most hosts and guests remain uncertain about a fair rental price or price range. Thus, it would be useful for the guests to have a price predictor for their reference. To do this, we built a regression model which takes several features of the listing's information as input and gives the price as output.For our analysis, we decided to build the model for New York City, because we had access to a very comprehensive dataset comprising of over 300K rows(after pre-processing) distributed over 2014-2017.

We decided to include 2014-16 prices for training and testing over 2017 prices. Each row in our dataset includes many listing's features, but we narrowed down to some of the most salient features beacuse we think it is very necessary to do some exploratory analysis on the dataset, which can help us to select the best features for our model.

3.1.1 Hypothesis 1

Our model for predicting and recommending listing prices was a multi-linear regression model, in which the estimated price was calculated as the dot product of learned weights and features of Airbnb listings. We decided to use below features, for our predictive modelling:

- Neighbourhood
- Bourough
- Room-Type
- Number of Beds
- Number of Bathrooms
- Number of Bedrooms
- Number of Reviews
- Number of Amenities
- Property-Type
- Accommodates
- Number of Metro Stations
- Distance from Center Point of Neighbourhood

We removed the feature "Accommodates", as it was highly positively correlated to "No of beds", we observed this by building correlation matrix between features. We transformed certain features like "Distance from Center" into categorical feature (far,near). We used label encoding to transform each categorical column into numerical column. In the end, we did feature scaling using min-max normalization for each column.

As we can see from Figure 1, testing results are not that great, there are very large residuals, and also the SSE and MAPE scores of the linear model are huge. Adjusted R-square score for our linear model was 0.42. We can also see that the residual plot (Figure 2) is non-symmetric, resulting in low homoscedasticity.

So we can conclude that there may not be a linear relationship between listing features and prices. We know that for most of the cases, a linear model can always be outperformed by other low bias and high variance models.

3.1.2 Hypothesis 2

As seen, our linear model was not good enough for predicting prices, we decided to rank our most

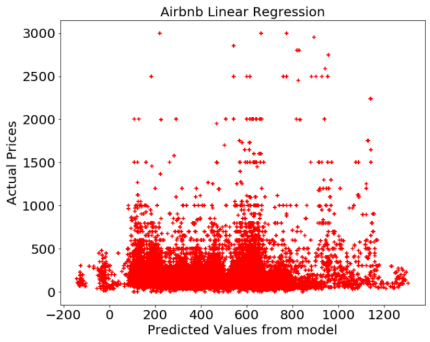

Figure 1: Linear Regression Testing.

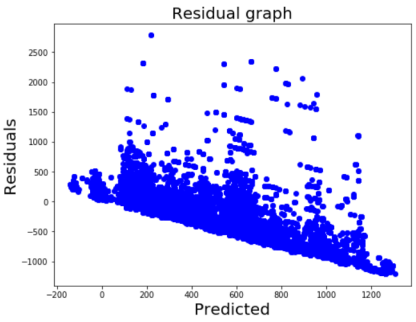

Figure 2: Residual Plot.

important features, so that hosts gets to know what are the important features affecting price.We performed feature ranking on features listed in **3.1.1**. We did this using the linear model build in above section, we dropped feature one by one from our model and testing the model recording the SSE and MAPE scores. We did feature ranking based on those score i.e. a feature was ranked higher, which resulted in high score by dropping it. Below are the top 6 features resulting from our testing.

1. Property-Type
2. Number of Bathrooms
3. Number of Bedrooms
4. Distance from center of Neighbourhood
5. Number of Amenities
6. Room-Type

3.1.3 Hypothesis 3

As already seen, linear regression is not a good model for predicting prices, so we decided to categorize price ranges into low ($\leq$200), medium($>$200 $\leq$700) and high($>$700) based on features as listed in **3.1.1**. Such an analysis can help hosts decide on the price ranges for their respective listing. We decided to use Bayesian inference for classification of prices under the assumption that the listing's features are conditionally independent given the class of price. We know that using Bayesian

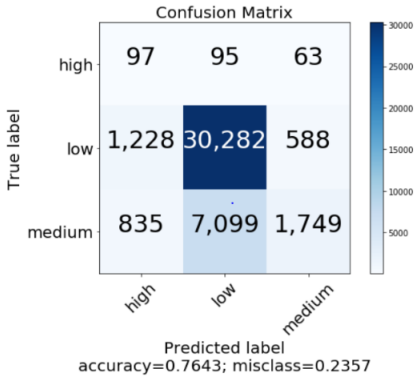

Figure 3: Bayesian Test Results

inference we can built a generative probabilistic model for classification of prices.

We further pre-processed our dataset used in linear model, to make sure that for each feature, values are same in both training and testing set so that our probabilities don't push down to zero for an unseen feature value. However we can always overcome this problem by adding smoothing factor in our probabilistic model. Further, we narrowed down our analyses for prices distributed over 2016-2017, out of which 70% of data was utilized for training and 30% for testing.

Predicting classes of price for testing dataset using Bayesian inference requires data likelihood and prior probabilities to be calculated, this was done using training dataset. Figure 3 shows the results of our probabilistic model.

We were able to achieve around 77% accuracy for testing dataset. As can be seen from the confusion matrix, for the low class price range, we were correctly able to predict 30k rows out of 33k rows. Hence we can conclude that Bayesian Inference is a reasonable choice for predicting listing's price range.

3.2 Does type of area affect bookings and listings?

It is intuitive to think that different types of areas (e.g.: Criminal v/s Non-Criminal) should not follow the same type of distribution for the number of bookings as well as the number of listings in that type of area. Thus we wanted to check if this intuition is true based on the given dataset. For this topic, the neighbourhoods in Manhattan were roughly categorized into the following types :

1) Criminal vs Non-criminal
2) Family oriented vs Non-family oriented
3) High population vs Low population

Figures 4-7 show the trend in number of bookings and listings from the dataset, over the year

2017

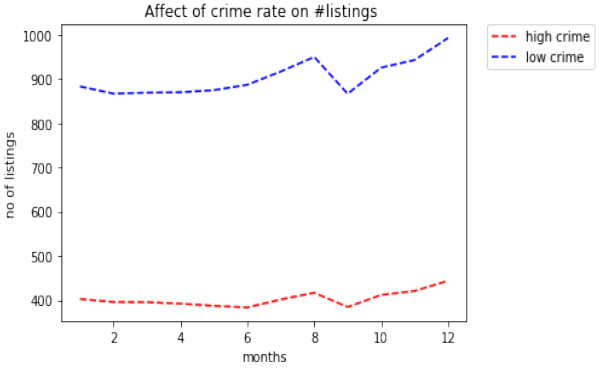

Figure 4: Trend for number of listings in high and low crime areas over year 2017.

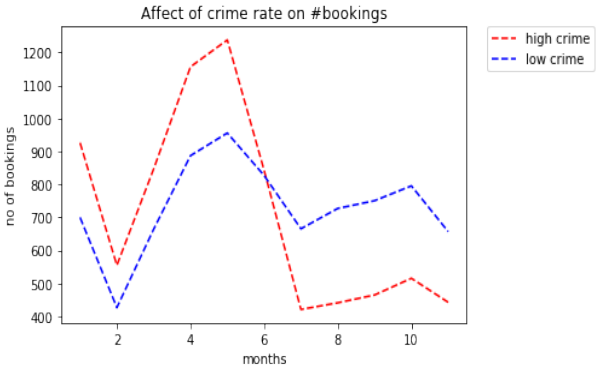

Figure 5: Trend for number of bookings in high and low crime areas over year 2017.

We analyzed how the type of area affects the number of bookings , number of listings and the type of apartment that got booked or were listed in the area.

Methods used: Non-parametric inference was used to get the empirical CDF. **KS Test** was used to compare the distribution of the Hypothesis. And **Permutation Test** was used to confirm further whether the results are the same as obtained from KS test or not.
The critical value considered for KS-test for n=12 (number of months in a year) and alpha=0.05 (99% accuracy) is 0.3754,
Alpha for permutation test : 0.05
Null Hypothesis : H_0: The distributions of average number of bookings/listings is same for the two types of areas
Alternative Hypothesis: H_1: The distributions of average number of bookings/listings is different for the two types of areas

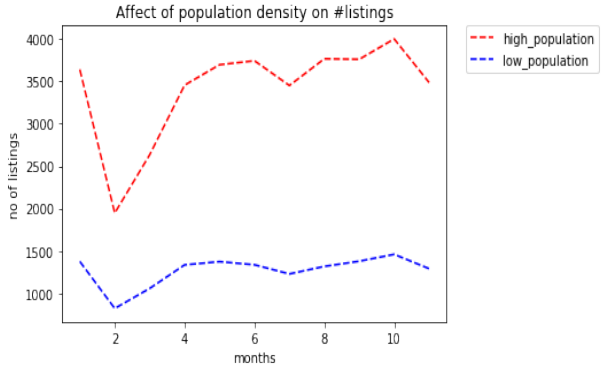

Figure 6: Trend for number of listings in highly and sparsely populated areas over year 2017.

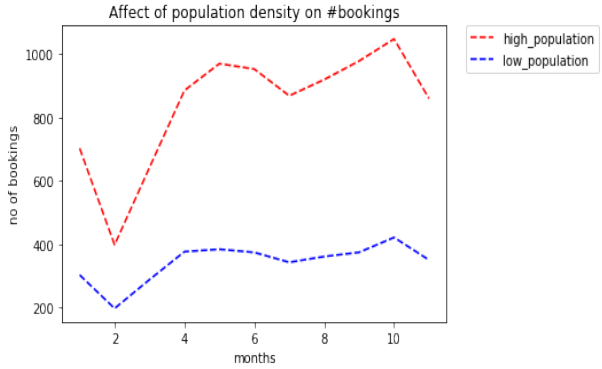

Figure 7: Trend for number of bookings in highly and sparsely populated area over year 2017.

3.2.1 Hypothesis 1: Crime rate in the area

We separated the 32 neighbourhood of Manhattan into high and low crime areas based on the data obtained from the site addressreport.com. 14 neighbourhoods in Manhattan fell under low crime criteria and the rest under high crime.Though the number of areas with high crime are more we saw that as predicted the number of listings and number of bookings done there were less as compared to low crime areas. These results were true even from permutation and ks-test . We compared the distribution of number of bookings over the 12 months in 2017 for low and high crime areas using ks and permutation test and found that the distribution are in fact different. Which means the crime rate in the area effects the number of listings and bookings done in that area. Thus if a host wants to buy a property and list them on Airbnb , it is better if the property is present in low crime area as the chances of it getting booked is higher as seen from the graphs obtained above.

Parameters affected	Perm-test
No.bookings	statistic:0.859, H_0 Accepted
Avg.No.listings	statistic:3e-06, H_0 Rejected
No.listings-entire-apartment	statistic:4e-06, H_0 Rejected
No.listings-private-room	statistic:1e-06, H_0 Rejected
No.listings-shared-room	statistic:1e-06, H_0 Rejected

Table 1: Permutation test results- How crime rate affects bookings and listings in Manhattan for year 2017.

Parameters affected	KS test
No.bookings	Pvalue:0.1473, H_0 Rejected
Avg.No.listings	Pvalue:2.3e-06, H_0 Rejected
No.listings-entire-apartment	Pvalue: 2.3e-06,H_0 Rejected
No.listings-private-room	Pvalue: 2.3e-06, H_0 Rejected
No.listings-shared-room	Pvalue: 2.3e-06, H_0 Rejected

Table 2: KS-test results- How crime rate affects bookings and listings in Manhattan for year 2017.

3.2.2 Hypothesis 2: Population Density of the area

Based on the population data obtained from nyc.gov we classified 32 neighbourhoods in Manhattan as highly or sparsely populated. We wanted to check if the population density of the area affects the number of bookings or number of listings done in the area. From the results of permutation test and KS test the Null hypotheses got rejected.This means that population density does affect the number of listings and bookings in the ares as you can see in the graph above and table-2 below.

Parameters affected	Permutation test
NumberOfBookings	statistic:2.1e-05, H_0 Rejected
Avg.Number of listings	statistic:3e-06 H_0 Rejected

Table 3:Permutation test results- How population density affects bookings and listings in Manhattan for year 2017.

Parameters affected	KS test
Number of bookings	P value: 5.9012e-05 H_0 Rejected
Avg. Number of listings	P value: 6.60115e-06 H_0 Rejected

Table 4:KS test results- How population density affects bookings and listings in Manhattan for year 2017.

We wanted to further analyze the data for "unusual suspects": (in Manhattan) i.e. try to find examples of areas which have:

1) High Pop-density + High Crime: But have more number of average bookings.
We applied an Upper Tail Wald's Test. Areas for which the Null Hypothesis is rejected are the "unusual suspects".
Null Hypothesis:H_0: The avg. no. of bookings is the same for the (high pop. + high crime area) and Manhattan.
Alternative Hypothesis:H_1 : The avg. no. of bookings in the (high pop. + high crime area) is more than the avg. no. of bookings in Manhattan.
We found that the **Gramercy** neighbourhood of Manhattan was the "unusual suspect" and it is a High Population and High Crime Area **yet** number of average bookings are more. A probable reason could be its high population as well as its location in Lower Manhattan which is a touristy area and thus the bookings are expected to be more regardless of anything.

2) High Pop-density + Low Crime: But have less number of average bookings.
Similar to the previous method. Here: Null Hypothesis: H_0: The avg. no. of bookings is the same for the (high pop. + low crime area) and Manhattan.
Alternative Hypothesis: H_1 : The avg. no. of bookings in the (high pop. + low crime area) is less than the avg. no. of bookings in Manhattan.
We found no such neighbourhood in Manhattan.

3.2.3 Hypothesis 3: Family oriented or not

We classified the areas into family oriented and non-family oriented using data from niche.com. We wanted to see how the number of bookings and listings get affected by family orientation of the area. This information can be used by the host to decide what type of listing he or she should make in a particular area. E.g.: if the type of area is a "family one" (nearby Central Park) then probably entire homes or apartments should be listed instead of private rooms or shared rooms

as it increases the chances of getting a listing rented. From both Permutation and KS-test we found that family orientation of area does affect the type of listings that get booked . As seen from the table the null hypothesis got rejected for all the cases.

Parameters affected	Permutation test
No.of bookings	statistic:0.0067 H_0 Rejected
Avg.No.of listings	statistic:3.7e-05 H_0 Rejected
No.of listings for entire apartment	statistic:0.0 , H_0 Rejected
No.of listings for private room	statistic:0.0 , H_0 Rejected
No.of listings for shared room	statistic:6.2e-05 H_0 Rejected

Table 5: Permutation test results- How family/non-family oriented areas affects bookings and listings in Manhattan for year 2017.

Parameters affected	KS test
Number of bookings	P value:0.02801, H_0 Rejected
Avg. Number of listings	P value:7.16110, H_0 Rejected
No.of listings for entire apartment	P value:8.1193e-07, H_0 Rejected
No.of listings for private room	P value:8.129e-07, H_0 Rejected
No.of listings for shared room	p value:7.1616e-06, H_0 Rejected

Table 6: KS test results- How family/non-family oriented areas affects bookings and listings in Manhattan for year 2017.

3.3 Do events, holidays and seasons affect average price of a listing?

We examined if holidays, seasons and popular events(recurring and one-time) in New York City affect the average price of a listing. We also considered each event separately to analyze if individual events affects the average price. It is useful to know if events, holidays and seasons affect average price of a listing so that host know when to spike prices to make profit. Also, guests can plan their visit accordingly.

Methods used:
Parametric Inference:
We want to check if prices are normally distributed in order to determine if we can apply paired t-test or not. On plotting the CDF and PDF of prices in the event and non-event duration months, we see a distribution equivalent to the normal distribution.

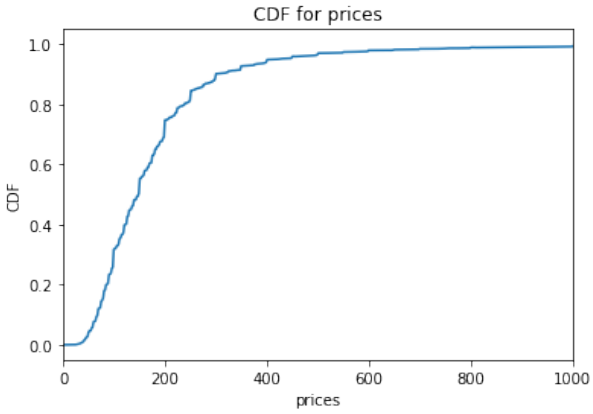

Figure 8: CDF of prices

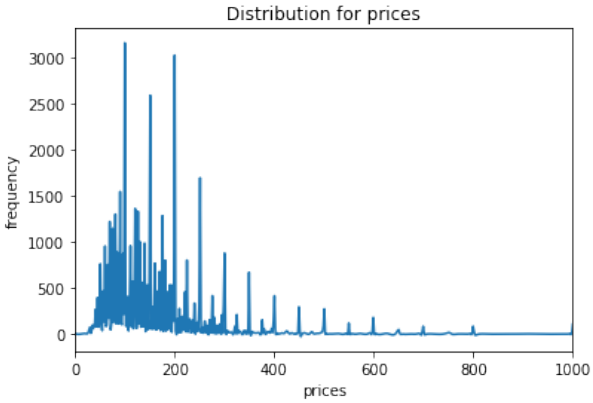

Figure 9: PDF of prices

Hence, we use **Maximum Likelihood estimator** to find the parameters for the normal price distribution. The parameters $\boldsymbol{\mu_{MLE}}$and $\boldsymbol{\sigma_{MLE}}$were estimated as 152.45 and 76.4 respectively.

K-S Test:
We check if price distribution is same as the normal distribution with the parameters obtained from parametric inference. The null and alternate hypothesis are given by:
H_0: The price distribution is same as normal distribution with parameters μ_{MLE} and σ_{MLE}
H_1: The price distribution is not same as normal distribution with parameters μ_{MLE} and σ_{MLE}
If null hypothesis is accepted, then we can apply paired t-test to check if average price is same during event and non-event months. Null Hypothesis was accepted.

2-population Wald's Test:
We did not apply the paired-t test to check if average prices are the same during event and non-event months as the number of samples is large. Paired-t test requires sample size to be small. Instead, we used the 2-population Wald's test throughout. This test is applicable here as there are n number of listings and the price of each listing is independent of another, so there are n i.i.d random variables. And since the estimator is the sample mean, through CLT, the estimator becomes asymptotically normal.

3.3.1 Hypothesis 1: Are average prices different during events/holidays?

We analyze if the average price of a listing in event duration months is same/different as the average price in non-event duration months. We consider only those listings that are common in the event and non-event duration months over 3 years.The null and alternate hypothesis are given by:
H_0: The average price of a listing in event duration months is same as average price of a listing in non-event duration months
H_1: The average price of a listing in event duration months is not same as average price of a listing in non-event duration months.

Event/Holiday	Event Timespan
Comic Con	Every October
Xmas and New Year's Eve	Every December
Papal Visit	September 2015
New York City Marathon	Every November
Thanksgiving	Every November

Table 1: Events Considered

Result	$\boldsymbol{H_0}$**rejected**
Wald's statistic	2.04
p-value	0.04
Confidence-Interval	(0.0678, 3.388)

Table 2: Results

Conclusion: The null hypothesis is rejected which implies that at 5% level of significance the data provides sufficient evidence that average price of a listing is different in event duration months and non-event duration months over 3 years.

Event /Holiday	**Result**	**Wald-stat**	**p-value**	**CI**
Comic Con October Manhattan	H_0 accepted	1.89	0.059	(-0.101, 5.51)
Xmas New Year Eve December All Boroughs	H_0 accepted	1.38	0.167	(-0.51, 2.95)
Papal Visit September 2015 Manhattan	H_0 accepted	1.92	0.055	(-0.12, 11.02)
NYC Marathon, Thanksgiving November All boroughs	$\boldsymbol{H_0}$**rejected**	2.297	0.022	(0.297, 3.753)

Table 3: Events Considered

3.3.2 Hypothesis 2: Individual events/ holidays average price effect

We examine if the average price of a listing in any individual event/holiday is same as the average price in other non-event duration months. We consider only those listings that are common in the particular event and other non-event duration months over 3 years. In case an event occurs in a particular neighbourhood, we only consider the listings in those neighbourhoods. We also compute p-value for each event to understand the extent by which H_0 is rejected. This hypothesis is useful to determine if any particular event/holiday affect the average price when compared to others.The null and alternate hypothesis are given by:
H_0: The average price of a listing in the particular event duration month is same as average price of a listing in non-event duration months
H_1: The average price of a listing in the particular event duration month is not same as average price of a listing in non-event duration months
Conclusion: The null hypothesis was accepted for all considered events and holidays except for the month of November which includes NYC-Marathon and Thanksgiving. This implies the average price in November is different from other months average price, while Christmas, New Year's eve, Comic Con have same average price as other non-event duration months.

3.3.3 Hypothesis 3: Seasonal average price effect

We examine if the average price of a listing is same in two different seasons or not. We consider only those listings that are common in both the seasons under consideration. This hypothesis

is useful to determine if seasons affect the average price of a listing.The null and alternate hypothesis are given by:
H_0: The average price of a listing in the a season(say,Summer) is same as average price of a listing in another season(say, Winter)
H_1: The average price of a listing in the a season is not same as average price of a listing in another season

Season	**Result**	**Wald's statistic**	**p-value**	**CI**
Summer (June, July, Aug) Winter (Dec, Jan, Feb)	H_0 accepted	0.276	0.782	(-1.46, 1.93)
Fall Sep, Oct, Nov Spring Mar, April, May	$\boldsymbol{H_0}$ **rejected**	-2.839	0.0045	(-5.01,-0.93)

Table 4: Seasons Considered

Conclusion: The null hypothesis was accepted for Summer vs Winter and rejected for Fall vs Spring. This implies that at 5% level of significance and p-value of 0.0045, the average price in Fall is different from the average price of a listing in Spring. However, it is the same in Summer and Winter.

3.4 Host dependent features that affect the number of bookings

In this topic, we want to look at the "non-obvious" features (under the control of the host) that can maximize the host's potential to get the listing booked. The "obvious" features include: price of the listing, location of the listing, reviews and ratings of the listing. But what are some of the unclearer and hidden features that a host can concentrate on to help the listing get booked. We analyze this. The proposed features are binary-valued and are the following:

- Host response rate: '1' for the host having a good response rate to queries and '0' for not.
- Host is superhost: '1' for being a superhost (i.e. highly rated(4.8+), responsiveness(¿0.9), active(has at least 10 stays per year) and never cancels) and '0' for not being a superhost.
- Host identity verified: Whether the host's account is verified ('1') on Airbnb or not ('0')
- Cancellation policy: Strict ('0') or flexible ('1') cancellation policy of the listing
- Instant bookable: Is the listing instantly bookable ('1') or not ('0')

Methods and Techniques Used:

Since the dataset is given month-wise, for each month (of 2017), we get **only the listings that got booked** for that month along with the columns of the above proposed features. We then collate the data of all months together and this is the final dataset on which the hypothesis testing was applied.

Each hypothesis is for a single host-related feature and checks whether that feature affects the average number of bookings or not. This is done via **Wald's Test**. The Wald's test can be applied here since, for each hypothesis, there are n (no. of observations) i.i.d Bernoulli random variables (feature) and the random variable taking the value '1' is considered a "success" and '0' a "failure". And since the estimator is the sample mean (no. of 1's / n), through CLT, the estimator becomes asymptotically normal. The alpha value used was 0.05.

Once Wald's Test is applied, we wanted to confirm that a "good" value of each host feature does not decrease (negatively-impact) the average number of bookings. For example: A good response rate should not decrease the avg.number of bookings, as that is more harmful. It does not matter how positively the avg number of bookings are affected by a good response rate. For us, the thing that matters is that it should not decrease. So, we further apply a **lower tail Wald's Test** where:
H_0 : Avg no. of bookings for a high response rate = Avg. no. of bookings for all kinds of responses.
H_1 : Avg no. of bookings for a high response rate < Avg. no. of bookings for all kinds of responses.

3.4.1 Hypothesis 1: Does host response rate affect the average number of bookings

$\mathbf{H_0}$: No affect.
$\mathbf{H_1}$: There is an affect.
Wald's Test Result: Null Hypothesis Rejected.
Wald's Statistic (absolute value): 307.281
Wald's Test CI: [0.808, 0.812]
p-Value: 0.0
Lower Tailed Wald's Test:
$\mathbf{H_0}$: Avg no. of bookings for a high response rate = Avg. no. of bookings for all kinds of responses.
$\mathbf{H_1}$: Avg no. of bookings for a high response rate < Avg. no. of bookings for all kinds of responses.
Lower-Tail Wald's Test Result: Null Hypothesis Accepted.
Wald's Statistic: 12.930
Overall Result: A good host response rate positively affects the number of bookings.

3.4.2 Hypothesis 2: Does host is superhost affect the average number of bookings

H_0 : No affect.
H_0 : There is an affect.
Wald's Test Result: Null Hypothesis Rejected.
Wald's Statistic (absolute value): 267.201
Wald's Test CI: [0.214, 0.218]
p-Value: 0.0
Lower Tailed Wald's Test:
H_0 : Avg no. of bookings for host is a superhost = Avg. no. of bookings for whether host is a superhost or not.
H_1 : Avg no. of bookings for a host is a superhost < Avg. no. of bookings for whether host is a superhost or not.
Lower-Tail Wald's Test Result: Null Hypothesis Accepted.
Wald's Statistic: 30.375
Overall Result: A host being a superhost positively affects the number of bookings.

3.4.3 Hypothesis 3: Does host identity verified affect the average number of bookings

H_0 : No affect.
H_1 : There is an affect.
Wald's Test Result: Null Hypothesis Rejected.
Wald's Statistic (absolute value): 177.339
Wald's Test CI: [0.705, 0.710]
p-Value: 0.0
Lower Tailed Wald's Test:
H_0 : Avg no. of bookings for a host identity verified host = Avg. no. of bookings for all types of host.
H_1 : Avg no. of bookings for a host identity verified host < Avg. no. of bookings for all types of host.
Lower-Tail Wald's Test Result: Null Hypothesis Rejected.
Wald's Statistic: -3.917
Overall Result: A host identity verified negatively impacts the number of bookings (Strange).

3.4.4 Hypothesis 4: Does host cancellation policy affect the average number of bookings

H_0 : No affect.
H_1 : There is an affect.
Wald's Test Result: Null Hypothesis Rejected.
Wald's Statistic (absolute value): 62.988
Wald's Test CI: [0.577, 0.582]
p-Value: 0.0
Lower Tailed Wald's Test:
H_0 : Avg no. of bookings for a soft cancellation policy = Avg. no. of bookings for all types of cancellation policy.
H_1 : Avg no. of bookings for a soft cancellation policy < Avg. no. of bookings for all types of cancellation policy.
Lower-Tail Wald's Test Result: Null Hypothesis Rejected.
Wald's Statistic: -10.583
Overall Result: A soft cancellation policy positively affects the number of bookings.

3.4.5 Hypothesis 5: Does host's listing instantly bookable policy affect the average number of bookings

H_0 : No affect.
H_1 : There is an affect.
Wald's Test Result: Null Hypothesis Rejected.
Wald's Statistic (absolute value): 144.579
Wald's Test CI: [0.323, 0.327]
p-Value: 0.0
Lower Tailed Wald's Test:
H_0 : Avg no. of bookings for a listing instantly bookable = Avg. no. of bookings for all types of bookable listings.
H_1 : Avg no. of bookings for a listing instantly bookable < Avg. no. of bookings for all types of bookable listings.
Lower-Tail Wald's Test Result: Null Hypothesis Accepted.
Wald's Statistic: 48.844
Overall Result: An instantly bookable listing positively impacts the number of bookings.

3.5 Predicting average number of bookings for the next month(s) according to hosts

Predicting the average number of bookings to be made for the next month(s) could be useful for a host in knowing whether or not a listing would get booked as well as how many listings would get booked (if a host has more than one listing). Hosts can make use of this prediction information and make better decisions about their listings. For e.g.: Hosts can change their listing price or other amenities (like Wi-Fi, free internet, etc) to increase his/her chances of getting the listing booked. The prediction information can be used by Airbnb to advertise those listings which have a higher chances of getting booked.We have used the below hypotheses for this activity:

3.5.1 Hypothesis 1

Check how numberOfReviews for each listingId of every given hostId for each consecutive months for every year affects the of bookings to be made in the Airbnb listings

Methods used: We predict the average bookings for the next months(s) using the pre-processed data for the last 22 months.
The below methods were used:

1. Autoregression - AR
2. Exponentially Weighted Moving Average - EWMA
3. Last Observed
4. Simple Moving Average

As per our prediction Autoregression is the best model among all the other methods used.

3.5.2 Autoregression - AR

Autoregression is used to predict the average bookings for the next 11 months using the data for last 22 months (data separated by 70% training and 30% testing). We have used the columns number of reviews, hostId and listingId for this activity.
Parameters used:
1)AR values set to 11 (AR = 11)
2)Actual average number of bookings for last 33 months(separated 70% data(22 months) for training and 30%(11 months) for testing)
3)Input data for last 33 months(separated 70% data(22 months) for training and 30%(11 months) for testing)

Result of Autoregression:

Average Error	Accuracy
3.873435076740053	96.32897013559031
11.49592951933543	89.16908306918971
6.467117162214385	70.13362290023208
5.913299990973125	87.94894327362589
5.909905703738332	76.3375352068645

Average Error and Accuracy for Autoregression - AR

host_id	number_of_ listings	number_of_ actual_ bookings_ 042017	number_o f_predicte d_booking s_042017	number_of _actual_ bookings_ 052017	number_of_ predicted _bookings_ 052017	number_of_ actual_ bookings_ 062017	number_of_ predicted_ bookings_ 062017
210746	3	0	0.13	0	0.037	1	0.09326374
839679	3	0.67	1.29	1.33	0.439	0.67	1.0960449
4297106	3	3.33	1.91	2.67	4.324	6.33	5.9549675
7136700	3	1.33	3.27	1.33	4.338	2.33	2.8687973
25237492	10	0.1	0.38	0.1	0.034	0.2	0.3569594

Figure 10: AR - prediction

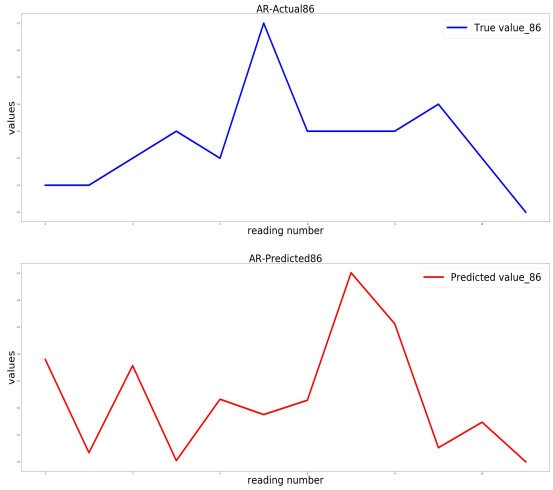

Figure 11: True vs Predicted

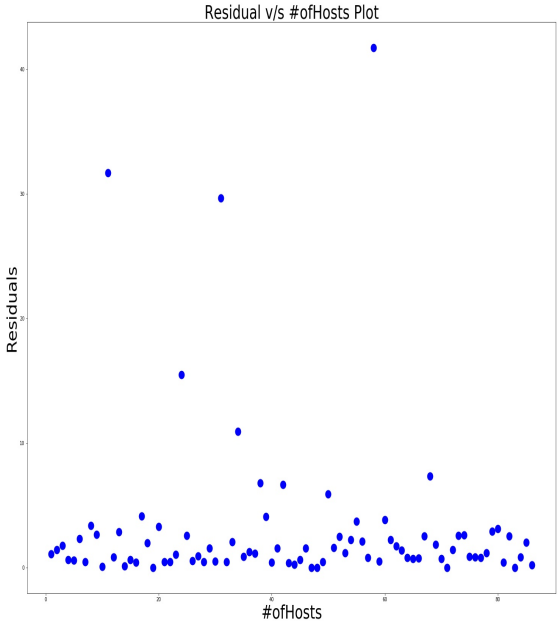

Figure 12: Residual Plot

3.5.3 Exponentially Weighted Moving Average - EWMA

Exponentially Weighted Moving Average used for making average number of booking prediction using the alpha value and preprocessed data. Like above for this activity we have used 70% data(22 months) data as training and 30%(11 months) data as testing.

Parameters used:

1)Alpha = 0.8

2)Actual average number of bookings for last 33 months(separated as 70%(22 months) for training and 30%(11 months) for testing)

3)Input data for last 33 months(separated as 70%(22 months) for training and 30%(11 months) for testing)

Result of EWMA:

Average Error	Accuracy
8.507903003030354	86.4915296255824
8.031499185495907	88.81608078295645
5.03022021924456	95.72703461898001
5.486079492916662	94.30803286684288
3.8369496998687587	90.20702936981296

Average Error and Accuracy for EWMA

3.5.4 Last Observed and Simple Moving Average

The other prediction method includes Last Observed which use the input value of last 12 months data and predict for next month and Simple Moving Average method which predicts the next months average based on average prediction for last 12 months.

Parameters used:

1)Actual average number of bookings for last 22 months(11 months for training and 11 months for testing)

2)Input data for last 22 months(11 months for training and 11 months for testing)

Result of Last Observed:

Average Error	Accuracy
7.986111111111112	71.08433505138719
8.555555543965763	78.75061363524189
3.9236111682322288	75.0267927286326
3.0069444378217067	52.68987409676177
8.798611016737091	92.87245403915357

Average Error and Accuracy for Last Observed

Result of Simple Moving Average:

Average Error	Accuracy
6.706367350286907	83.70278369949524
5.3443863226307755	82.22030408852685
4.988452212678062	55.89435602489271
4.854114549027549	68.35815607543161
2.9867238418923487	95.18625615518303

Average Error and Accuracy for Simple Moving Average

Conclusion: We were accurately able to predict the the number of bookings for each host over the next 11 months. Best model was obtained for Autoregression.

4 Prior Work on Airbnb Data

The InsideAirbnb dataset has been mainly used for visualization of different listing types across cities. However, there has been some relevant prior work performed on the InsideAirbnb dataset.

We found a paper (Predicting listing price on Airbnb dataset) that was aimed at predicting prices of Airbnb listings in London based on linear regression, similar to our attempt of predicting listing prices in New York City. However, the difference in both the approaches is that the paper used Text-Analysis features (bag-of-words models etc) besides categorical and numerical features. However, we did not use any Text-Analysis features. Also, R^2 was used as the evaluation metric in the paper but we used SSE and MAPE.

Another relevant analysis that we found was a Github project (Project Link) on the Airbnb dataset for Boston to analyze seasonal patterns of prices, as well as if holidays affected average prices of listings on a github. This is similar to a hypothesis in our third topic, however, the methods used on github include analysis through histograms and line graphs. However, the methods we used are based on Hypothesis testing like 2 population Wald's Test.

The third prior work that we found is a paper (Explore the Spatial Relationship between Airbnb Rental and Crime) that tries to analyze whether there exists a spatial relationship between crime and Airbnb listings for Florida. For e.g.: is there a positive relationship between Airbnb listing density and property crime and if there is a negative relationship between Airbnb listing density and violent crime. This is somewhat similar to a hypothesis of our second topic where we try to analyze whehther or not a criminal or non-criminal area affects the number of bookings and listings for the area. However, the difference lies in the methods used. The paper uses methods such as Geographically-weighted Regression and

Spatial autocorrelation. However, we applied Hypothesis testing methods to arrive at our conclusion.

5 FutureWork

A future prospect in predicting price would be to analyze the geographical features at a finer level and understanding how content of reviews affect the rental price. While analyzing correlations between crime and Airbnb, it would be interesting to check if tourism intensity level in an area has an effect on crime's impact on Airbnb. Another interesting aspect would be to check if the type of listing alters the relationship with crimes. For the fourth topic that we analyzed related to host features and bookings, it would also be useful to perform an Upper-Tailed Wald's Test to further confirm the relationship between certain host-related features and how they impact the bookings. One major limitation in our data is that we do not have listing information on a daily basis, so it is not possible to analyze short-term rental features.

A broader category that could be explored is examining the impact of Airbnb on Hotel industry like how Airbnb occupancy rate affects the hotel occupancy rate, if Airbnb prices affect the hotel average prices.

GitHub link of Project Code: Airbnb Data Analysis project