IE6600 CAPSTONE REPORT

Analysis of NYC Airbnb Data

Team

Yash Nikhare
Suhan Li
Omkar Narkar

INTRODUCTION

Airbnb is a vacation rental company that operates an online marketplace and primarily deals with short-term, temporary rentals of privately owned properties. Since all of our capstone members have first-hand experience with Airbnb's services, we were interested in analyzing their data. For the purpose of this project, we found Airbnb's rentals in New York City and its 5 boroughs. The dataset was sourced from kaggle.com and describes listings in NYC in 2019. This is a cleaned version of the raw data that was released by Airbnb, and therefore can be vouched for its accuracy and relevance. We begin our project by first understanding the nuances of the data and how users respond to various listings. We then look into price predictability given a variety of numerical and categorical variables using regression and machine learning methods.

DATA

There are **48895 rows** in the dataset, each of which represents a unique listing. There are **16 columns**, each of which represents a unique feature, which has been described in the table below:

Column Name	Description	Type
id	Airbnb's unique identifier for the listing	int64
name	Name of the listing	string
host_id	Airbnb's unique identifier for the host/user	int64
host_name	Name of the host. Usually just the first name(s).	string
neighborhood_group	The neighborhood group geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	string
neighborhood	The neighborhood is geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	string
latitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	float64
longitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	float64
room_type	[Entire home/apt|Private room|Shared room|Hotel] All homes are grouped into the following three room types: Entire place/ Private room/ Shared room/ Entire place	string
price	daily price in local currency	int64
minimum_nights	minimum number of night stay for the listing (calendar rules may be different)	int64
number_of_reviews	The number of reviews the listing has	int64
last_review	The date of the last/newest review	object
reviews_per_month	The number of reviews the listing has over the lifetime of the listing	float64
calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.	int64
availability_365	The availability of the listing 365 days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.	int64

METHODOLOGY

We sourced a robust dataset from Kaggle, meticulously cleaning it to ensure reliability by eliminating inconsistencies and missing values. Our analysis focused on key variables like neighborhood groupings, room types, and host listing counts, chosen for their impact on pricing and engagement. To validate hypotheses effectively, we employed varied graph types:

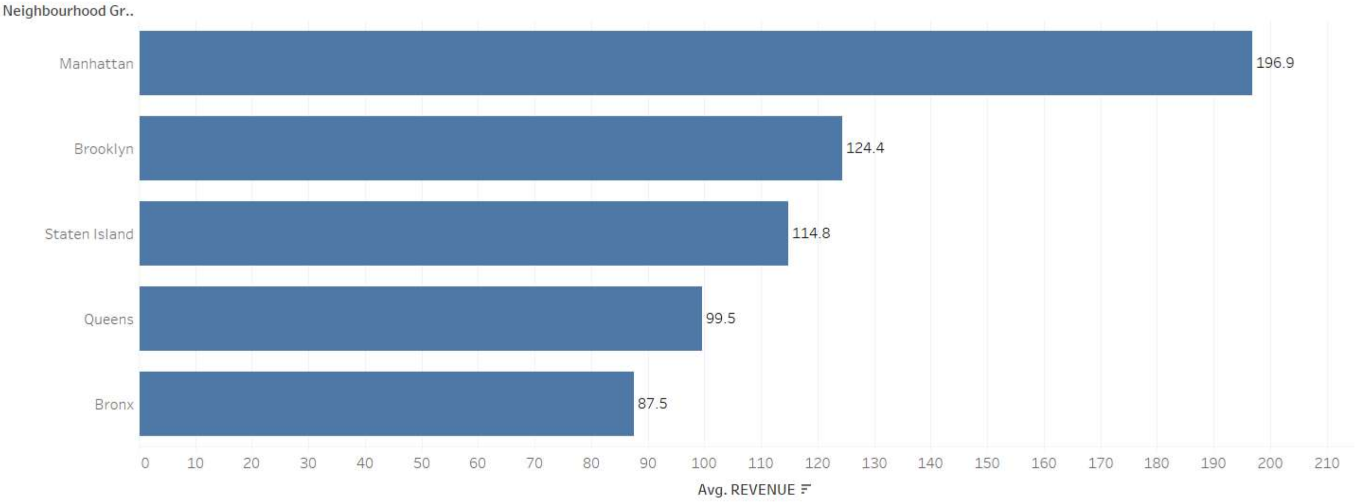
- Heat Maps: Revealing spatial pricing trends across neighborhoods through average price distribution visualization.
- Bar Charts: Comparing revenue across room types and neighborhoods to unveil performance and preference distinctions in different market segments.
- Time Series Analysis: Using line graphs to track revenue over time, aiding in identifying seasonal trends and growth paths.
- Descriptive Statistics: Displaying key metrics like total revenue, reviews, and occupancy rates for a quick market health snapshot.

HYPOTHESIS TESTING and KEY INSIGHT

We were mostly interested in analyzing the relationship between the price of a listing with factors such as the neighborhood where the listing is located, average monthly ratings for the listing, the minimum number of nights for the listing, and the number of listings by the host. This would help us better understand some of the dynamics of the dataset and set the stage for further exploratory analysis.

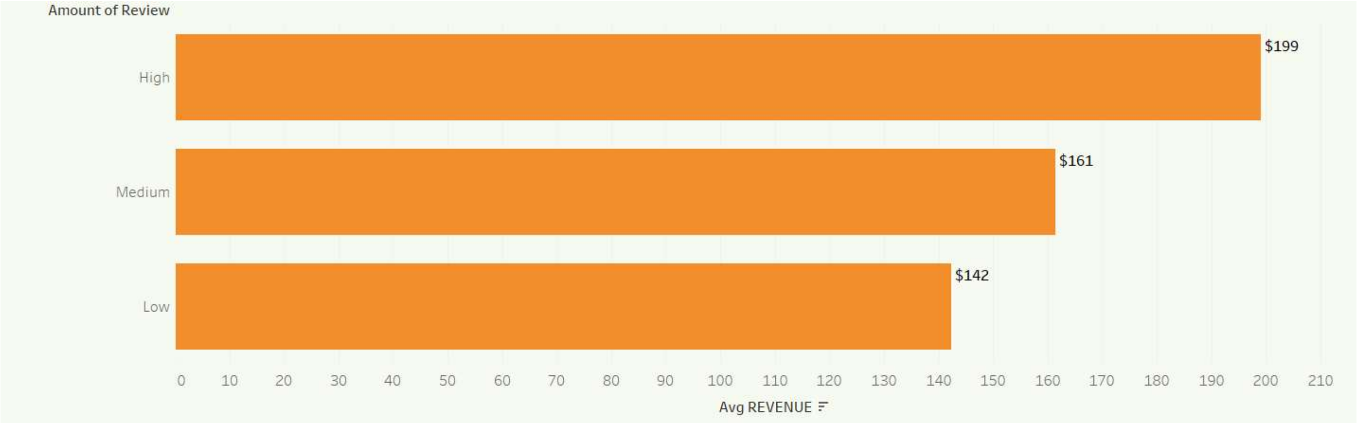
Hypothesis Test 1: Does the price of the listings vary by neighborhood groups (i.e boroughs)?

First, we divided the listings borough-wise and removed listings where the price was not mentioned or was zero. Our **null hypothesis** is that the price of listings does not vary across the boroughs. We felt the collective dataset size and comparable number of listings in Manhattan and Brooklyn was sufficiently large to mitigate this. We reject the null hypothesis. Therefore, there **is evidence that suggests the price of listings varies** by the boroughs they are located in. This result is not surprising as we know some neighborhoods in New York tend to be more expensive given general desirability, location, better maintenance, and facilities/amenities.



Hypothesis Test 2: Do listings with more reviews have higher prices than listings with less reviews?

For this test we performed a median split and split the listings into two groups: listings with reviews more than and less than the media. Our **null hypothesis is that listings with more reviews do not have higher prices** than listings with fewer reviews. Our alternative hypothesis is that listings with more ratings have higher prices. We reject the null hypothesis. Thus, the data suggests that listings with more reviews have higher prices than listings with less reviews. While the number of reviews per listing was available, the nature of these reviews (positive/negative/neutral) was not. Therefore, we are unable to make any claims regarding the sentiment of Airbnb reviews and price - perhaps some listings exceeded expectations and garnered (positive) reviews while average to below average listings received no attention. This is simply a conjecture, and a solid conclusion cannot be drawn without further information or actual reviews for the listings.



Hypothesis Test 3: Do listings having number of minimum nights on weekend or weekday listings have different prices?

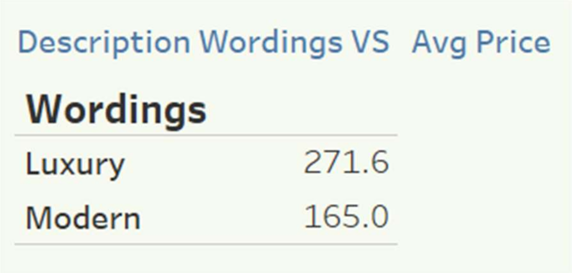
Again, we split to our data in order to create two groups: listings on weekday and weekend. Our **null hypothesis is that listings on weekday and weekend do not have different prices.** By looking into the dashboard hence leading us to **reject the null hypothesis**. Thus, we can conclude that listings on weekday and weekend have a difference in prices. Higher demand on weekends can lead to increased prices for listings with shorter minimum nights.

Hypothesis Test 4: Do hosts with more listings charge higher prices than hosts with less listings?

First, we performed a median split of the data, dividing into two groups: those with low host listing counts and those with high host listing counts. Then we removed the duplicate entries in our data so that only unique host ids are considered. **Our null hypothesis is that listings by hosts with more listings are not priced higher than listings by hosts with less listings**. From dashboard insights leading us to **fail reject the null hypothesis**. Thus, listings by hosts with more listings are priced lesser than listings by hosts with less listings. **Quality Over Quantity**: Hosts with fewer listings might prioritize quality over quantity, maintaining premium properties and offering enhanced amenities or unique experiences, justifying higher prices.

Hypothesis Test 5: Are listings that contain the word 'luxury' priced higher than listings that contain the word 'modern'?

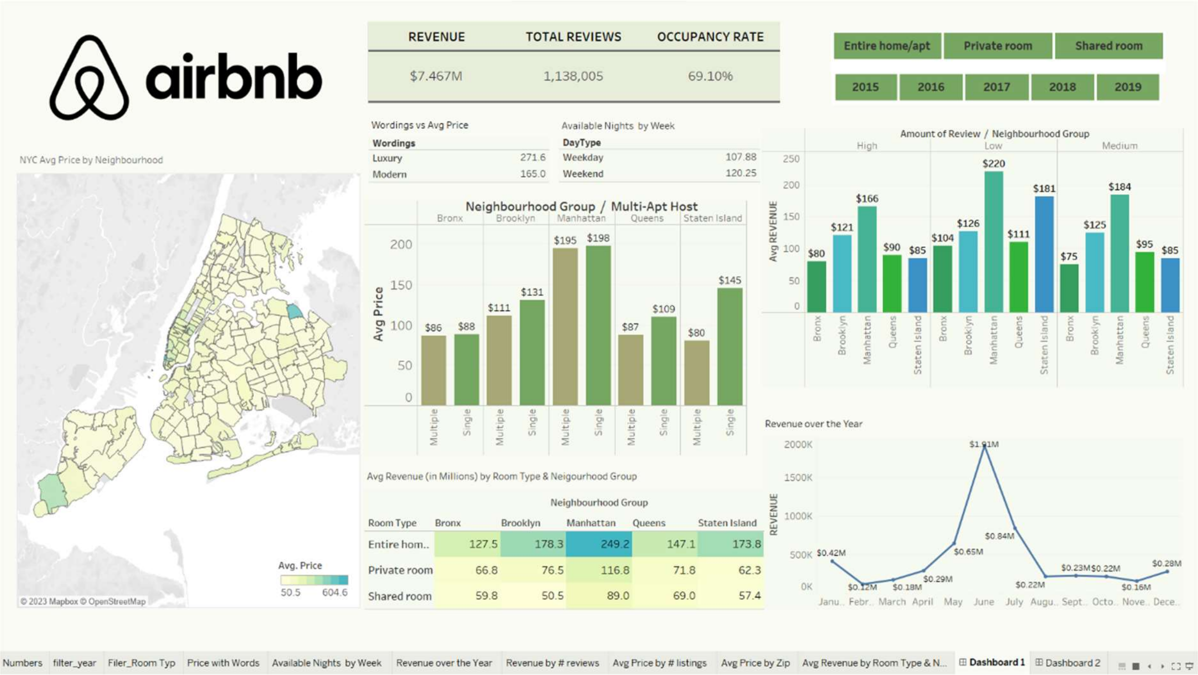
While we were perusing the data, we realized that a lot of listings mentioned the words 'luxury' and 'modern'. So, we were curious as to whether there is a significant difference in prices between listings that contain these two words. Furthermore, from a significance testing standpoint, the number of listings with each of these words was in the context of 2,000, making our two testing groups even in magnitude. **Our null hypothesis is that listings that contain the word 'luxury' are not priced higher than listings that contain the word 'modern'**. While, looking into the dashboard hence **we reject the null hypothesis**. Thus, this suggests listings that contain the word 'luxury' are priced higher than listings that contain the word 'modern'. In fact, the mean price for the 'luxury' group was in the context of $271 a night versus $165 for the 'modern' group.



Description Wordings VS Avg Price

Wordings	
Luxury	271.6
Modern	165.0

FINAL VISUALIZATION



REVENUE	TOTAL REVIEWS	OCCUPANCY RATE
$7.467M	1,138,005	69.10%

Wordings vs Avg Price

Wordings	
Luxury	271.6
Modern	165.0

Available Nights by Week

DayType	
Weekday	107.88
Weekend	120.25

Avg Revenue (in Millions) by Room Type & Neigourhood Group

	Neighbourhood Group				
Room Type	Bronx	Brooklyn	Manhattan	Queens	Staten Island
Entire hom..	127.5	178.3	249.2	147.1	173.8
Private room	66.8	76.5	116.8	71.8	62.3
Shared room	59.8	50.5	89.0	69.0	57.4