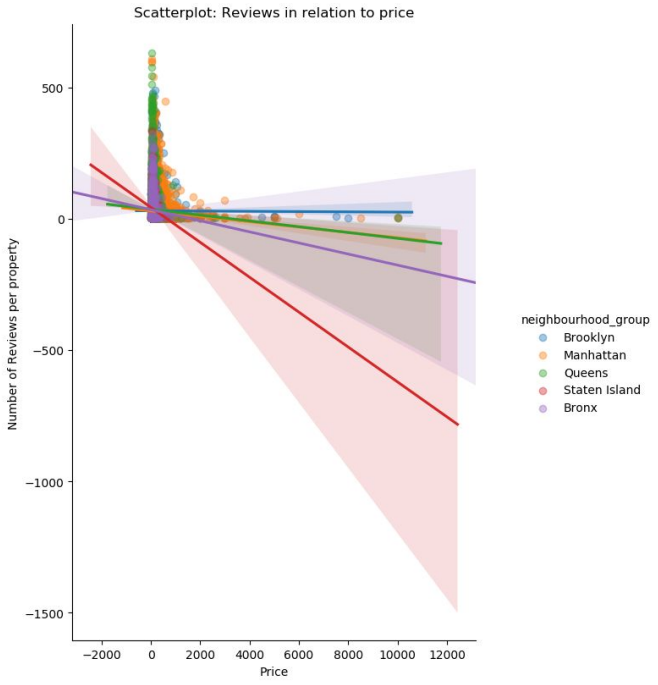
Introduction:

The dataset I chose was about Airbnb statistics in the New York City area. I got the data from user Dgomonov on Kaggle.com.[1] It is unclear whether this data is actually from Airbnb.

Data Exploration:

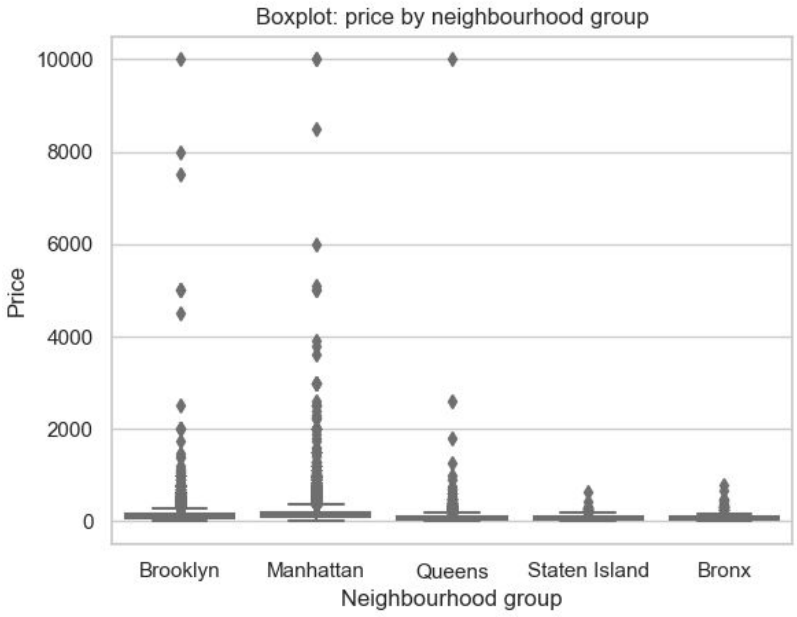
The first thing I discovered when diving into the dataset is that for the quantitative fields, there was a huge range of values. for example, the median number of available nights for an Airbnb property was 55, the mean was around 114, and the standard deviation was around 129. Very similar spreads existed with pricing for Airbnbs, the minimum nights a guest has to stay at an Airbnb, and the number of reviews. I was curious to see how all of these categories of data interacted with each other.



The graph below shows the number of reviews of a location to the price of the location. The hue is the "neighbourhood group" or borough the Airbnbs are located in. It is assumed that the number of reviews correlates to the number of bookings. The steep decline in reviews in relation to the price could indicate a couple of things: a price sensitivity in certain boroughs, or a lack of demand in certain boroughs which drives prices down. The number of boroughs which seem to have the steepest decline in the number of reviews in relation to the price are Staten Island and the Bronx. These are the furthest boroughs from tourist locations like Times Square or Downtown. The Brooklyn area seems to be level or increase in reviews depending on the price. This could be because those Airbnbs are venues for certain larger shows, and have consistently high turnover.

I used a boxplot to answer the question about pricing in certain neighborhoods as seen in the graph below. The Bronx and Staten Island have the lowest prices. It is unclear whether this is due to a shortage of guests looking to stay in the area, or a price sensitivity from guests who are looking to stay in the area.

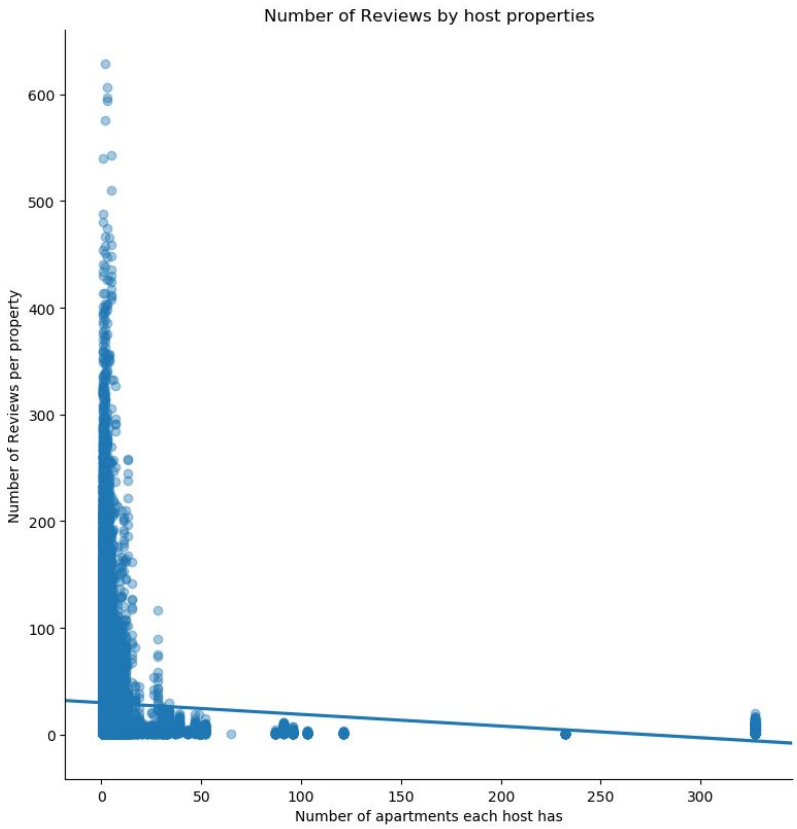
[1] https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data



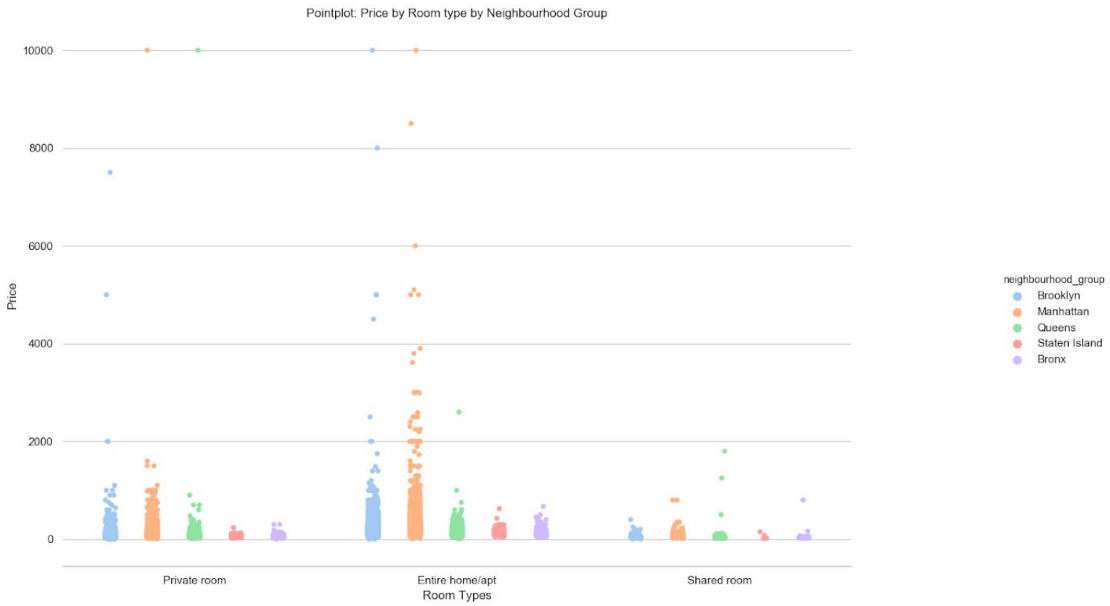
I was curious about the number of reviews as it correlates to the number of properties a host had. The assumption was that hosts with fewer properties would be more reliant on guests to provide them with an advantage on Airbnb by providing them with good reviews. This assumption bears out in the graph below.

One trend I saw developing was this hard “L” shape. A few properties, property owners, and units with the highest reviews dragged the graphs to far extremes in almost every category.

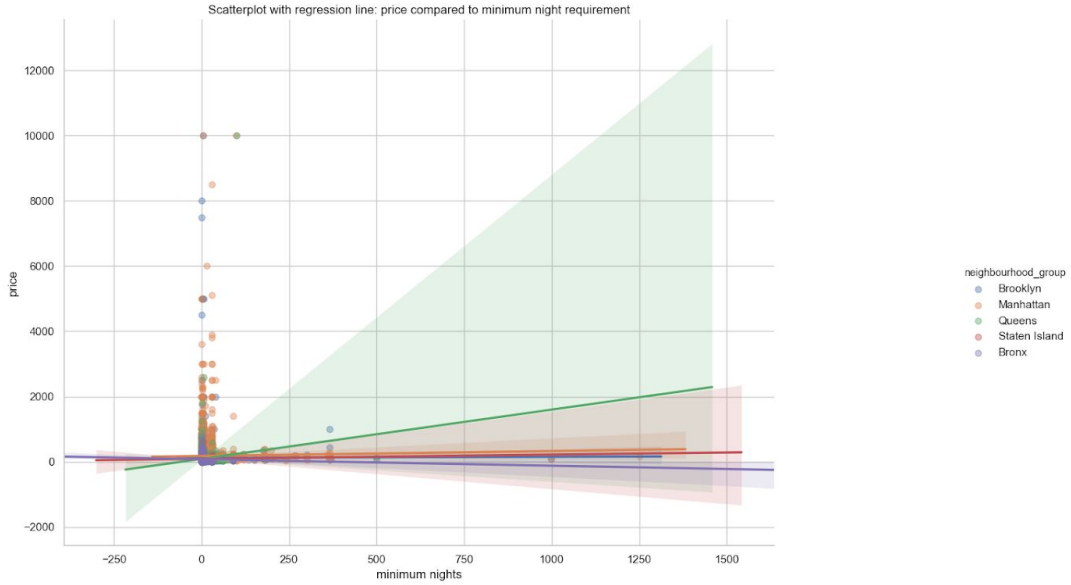
lown of



how Airbnb rooms change depending on what borough they are located in shows two interesting data points: Manhattan has the most even distribution of prices for each Airbnb type than any other, and whole apartments seem to be the room type with the greatest price variation. This isn’t to say they had a perfectly equal distribution of prices across all room types.



Curious about how neighborhoods affect pricing and minimum nights required, I created the following graph.



In this graph it becomes clear that not all long term stays are more expensive than short term stays, except in Queens.

What makes Staten Island such a price sensitive environment, Manhattan so price diverse per room type, and Queens so expensive per increase in minimum nights required?

Research Proposal

The Problem: Do guests book Airbnbs for different reasons? What are those reasons, and how do we better match guests with appropriate venues in order to maximize bookings city wide?

Potential Solution: Adding specific uses to descriptions for properties will help users better find spaces that are right for them.

Null Hypothesis: Adding specific uses to descriptions will not affect bookings or will lower the number of bookings.

Method for testing: Divide hosts into two groups. Group A will be the control. Group B will be asked to tag their properties with one of the following: vacation, short term stay, long term stay, venue, and business. Divide users up into two groups. Group A will be the control. Group B will be asked to select tags that describe the reason for them booking an Airbnb. The tags are as follows: vacation, short term stay, long term stay, venue, and business. Group B for guests will only be able to see Group B host listings.

We will compute the amount of guests who book on Airbnb spaces in Group A and Group B using an A/B test. The experiment will last for six months, From January 1st to July 1st. This date range has a variety of holidays, as well as down periods.

If guest bookings increases by 10% during those six months, we will implement the tagging system website wide. If the tagging system shows less than 10% growth but more than no growth, the experiment will continue for another six months. If there is no growth or a decrease in bookings, there will be a complete reversion to the old way of booking.

We will also be observing different uses of Airbnbs across boroughs. This way we can see if certain boroughs are performing better with a certain type of space. We will measure the booking rate as it relates to the number of properties listed per borough per property type. We will be using a one-tailed t-test, checking the amount of more bookings, and will consider the hypothesis correct if the p-value is less than .05.We will calculate the t-value using the following code:

```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
from scipy import stats

Experiment_data = pd.read_csv(PATH)
#calculating the t-value
#0 stands for no tag, any value above zero stands for a tag
stats.ttest_ind(airbnb_data[airbnb_data.tags == 0].revenue,
```

airbnb_data[airbnb_data.tags != 0].revenue)

If this is the case, we will inform property owners in the lowest performing borough to change their space tags to reflect the space tags of the better performing boroughs.