COGS9: Introduction to Data Science
Final Project
Due date: Wednesday 2019 December 11 23:59:59
Grading: 10% of overall course grade. 40 points total.
Completed as a group. One submission per group on Gradescope.

Group Member Information:

Please read COGS 9 team policies to best understand how to approach group work and to understand what the expectations are of you in COGS 9.

First Name	Last Name	PID
Xue	**Wang**	**A15908778**
Cain	**Chen**	**A15894106**
Xueer	**Zeng**	**A16167190**
Xinyue	**Chen**	**A15452339**
Muhan	**He**	**A15480974**

Question

What is the relationship between factors such as location, room type, and price, relative to the number of bookings per month for Airbnb in New York City?

Hypothesis

We believe that the higher the popularity of the location, the more expensive the price since we assume that in those regions with high popularity such as Manhattan would have a large demand for rooms. In general, we think that the larger the room, the more expensive the price because larger rooms usually occupy more people. In conclusion, generally the higher the price, the lower the number of bookings.

Background Information

https://www.digitaltrends.com/home/what-is-airbnb/

https://www.businessinsider.com/how-airbnb-was-founded-a-visual-history-2016-2#even-during-y-combinator-they-still-got-rejected-famously-by-investors-fred-wilson-of-union-square-ventures-admitted-in-2011-that-he-had-failed-to-look-past-the-air-bed-and-breakfast-name-and-see-the-business-14

Airbnb was founded by Joe Gebbia, Brian Chesky, and Nathan Blecharczyk in 2008 and its purpose was to offer people with someone's home to stay instead of a hotel. Before Airbnb became a $31 billion company, it was a simple idea of two young men who wanted to earn some money on rent. In 2007, Joe

Gebbia and Brian Chesky couldn't afford their San Francisco rent. As a result of sold-out hotels and their idea of selling their own place as a bed and breakfast to make money, " they saw a potential market for the idea and developed a website called airbednbreakfast.com" (Rawes and Coomes). Their first guests, two men and one woman, each paid $80 to stay on the air mattress. After receiving exciting feedback from their first guests, Gebbia and Chesky got together with their old roommate, Nathan Blecharczyk, to build it into a company. Airbnb was not successful from the beginning and it experienced at least three separate launches because the investors were not convinced by their plan.With their ambitious project plan, the cheap hotel shortage, and people's demands, the founders continued with their plan and more and more hosts in New York write reviews and photograph their houses. In 2009, Air Bed & Breakfast became Airbnb and marked the company's turning point.

However, problems began to rise after Airbnb became popular. Some areas, such as New York, "threatened to ban Airbnb and short-term rentals in 2014 and fine every host" (Aydin). Specifically, some cities employed laws that prevent hosts from renting their place without their presence for a certain number of days. Despite all the troubles, Airbnb continued expanding and improving their business.

Currently, Airbnb has over 6 million listings in roughly 190 countries. The range of information that can be seen for a listing include "the size of the space and its amenities, check-in and pricing information, a detailed description of the space, house rules, safety features, and availability" along with "reviews from other guests and information about the hosts" (Rawes and Coomes). When the customer is satisfied with the listing, he/she can request the listing, enter their basic information, and wait for approval from the host. The price of a listing is determined by factors such "location, quality of the listing, and the amenities" (Rawes and Coomes).

Data

In [2]:
```
import numpy as np
from datascience import *
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
from client.api.notebook import Notebook
```

In [40]:
```
table = Table().read_table("AB_NYC_2019.csv")
table = table.drop("host_name").drop("latitude", "longitude", "last_review","availability_365")
neighbourhood_group = table.column("neighbourhood_group")
neighbourhood = table.column("neighbourhood")
table = table.with_columns("neighbourhood_group", neighbourhood,  "neighbourhood", neighbourhood_group).relabeled("neig
table
```

Out[40]:

name	host_id	neighbourhood	neighbourhood_group	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	host_listings_count
Clean & quiet apt home by the park	2787	Kensington	Brooklyn	Private room	149	1	9	0.21	6
Skylit Midtown Castle	2845	Midtown	Manhattan	Entire home/apt	225	1	45	0.38	2
THE VILLAGE OF HARLEM....NEW YORK !	4632	Harlem	Manhattan	Private room	150	3	0	nan	1
Cozy Entire Floor of Brownstone	4869	Clinton Hill	Brooklyn	Entire home/apt	89	1	270	4.64	1
Entire Apt: Spacious Studio/Loft by central park	7192	East Harlem	Manhattan	Entire home/apt	80	10	9	0.1	1
Large Cozy 1 BR Apartment In Midtown East	7322	Murray Hill	Manhattan	Entire home/apt	200	3	74	0.59	1
BlissArtsSpace!	7356	Bedford-Stuyvesant	Brooklyn	Private room	60	45	49	0.4	1
Large Furnished Room Near B'way	8967	Hell's Kitchen	Manhattan	Private room	79	2	430	3.47	1
Cozy Clean Guest Room - Family Apt	7490	Upper West Side	Manhattan	Private room	79	2	118	0.99	1
Cozy Clean Guest Room - Family Apt	7490	Upper West Side	Manhattan	Private room	79	2	118	0.99	1
Cute & Cozy Lower East Side 1 bdrm	7549	Chinatown	Manhattan	Entire home/apt	150	1	160	1.33	4

... (48885 rows omitted)

In [ ]:

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

According to our project question: “What is the relationship between factors such as location, room type, and price, relative to the number of bookings per month for Airbnb in New York City?”,the ideal dataset should include the detailed locations(neighborhood, neighborhood_group), the room types, the price, and the stats about the number of bookings ; also, the data should be easy enough to visualize in order to show the relationships between the number of bookings and the factors which would also provide clear explanations to the readers.

So far, we have found 48895 observations of different Airbnbs in New York City. For each of these observations, we have columns: host_id, neighbourhood, neighbourhood_group, room_type, price, minimum_nights, number_of_reviews, reviews_per_month, host_listings_count. However, our data is still not ideal enough because we are short of the direct indicator of the number of bookings; instead we only have the number of reviews, which is not a comprehensive substitute. We also need some indicators of the

popularity of each neighbourhood group. Moreover, due to the lack of certain variables, the visualization is not completed yet.

Ethical Considerations

Fairness

The dataset included almost all different registered Airbnb owners whose housing estate(s) is(are) in New York. Considering this may not be all of them, we limit our research result generalization in only Airbnb housing estates in New York. At the same time, if we ever find other excluded data samples, we will combine those with the current dataset and re-do the data analysis process within the new dataset. What's more, random individual samples have the same possibilities to be drawn from the dataset while doing the analysis. With the usage of Jupyter Datahub, we defined all the classifications with firm references of number or results calculated from existing data points to avoid biases that appeared in artificial analysis.

Transparency

We downloaded our dataset online from kaggle.com in the form of .csv. According to the description on the website, this public dataset we used for the project is a part from Airbnb, and the original source can be found on this website. The information can be assured valid and accurate since the application requires authenticity when the people sign up to be Airbnb owners. This open dataset has a C00 license, making it a part of the public domain, allowing us to freely use, reuse, redistribute, and perform the work all without asking for permission. It does not contain information about specific individuals or any kind of information related to individual privacy. All the data included are ensured to be legal and can be viewed by anyone if wanted.

Accountability

The dataset includes data samples from all districts in New York(locations), all room types and specific price changes resulting from special events(e.g. national festivals, political elections, extreme weather changes) that we can find now. When doing the data analysis, we will classify them and make sure we consider these possible confounding variables. We were alert of other possible confounding variables and checked our completed process from time to time. We will go back to our analysis if we find other factors that need to be considered or have impacts on the data we found.

Analysis Proposal

As mentioned above, we need more data on direct indicators of booking of each airbnb room. We could use API to get the data from the airbnb website. As shown in the data section, we used Jupyterhub to clean our data such as deleting irrelevant information, making the columns in appropriate format, and so on. Since we have information about the location of each airbnb room, we plan on drawing a choropleth map showing the numbers of bookings. Based on our data, we are going to convert our independent variables such as

location, room type, and price to a weighted score and find the correlation between the weighted score and the number of ratings. We plan on building the training dataset with our data and find a different dataset via web scraping and make it the test data. With machine learning, we will be able to use a regression to predict how location, room type, and price will affect the number of bookings. Once that is done, we will have a clear idea on the relationship between location, room type, price and the number of bookings.

Discussion

Due to the fact that we are substituting the number of reviews for the number of bookings, it is possible that the results of our proposed analysis is completely different from the actual relationship between location, room type, price, and the number of bookings. Since some of the Airbnb are missing the factor number of reviews per month, represented as “nan” in the data, we have to remove these Airbnbs from our dataset, which can cause potential bias or skew in the dataset. In addition, we have to make sure that we do not have the bias to see the result that we expect and what has happened historically. One of the limitations of using a choropleth is that choropleth is used for displaying big pictures and it is not efficient at showing subtle differences. Therefore, if the number of bookings for different regions differ slightly, it is hard for viewers to actually see the difference. To address this, we can make a barplot in displaying the differences. While dealing with data sets, some data samples could be excluded or repeated in the old data set. To ensure fairness, we defined specific classifications for each existing data point with associated value via Jupyter Datahub, and then re-do the data analysis. To avoid possible risk of leaking individual privacy, we downloaded dataset online from kaggle.com, which is a free-usable public website without any individual information. To deal with some possible confounding variables, we checked the completed process and do the analysis again to ensure the data we used is accountable.

Group Participation

Xue Wang (PID: A15908778), Xueer Zeng (PID: A16167190), Xinyue Chen (PID: A15452339), Muhan He(PID: A15480974) and Cain Chen (PID: A15894106) worked together to formulate the question and the hypothesis. Xue Wang and Cain Chen worked on the data section and the analysis proposal section. Xinyue Chen worked on the background section and answered 3 of the 4 questions in the discussion section. Muhan He worked on the background section and answered question 4 in the discussion section.