CIS 450 Project Proposal - Group 24

1. **Team Members**
   a. Aakash Jajoo - aakashj1@sas.upenn.edu - GitHub: aakashjajoo1
   b. Karan Jaisingh - karanj@seas.upenn.edu - GitHub: kjaisingh
   c. Yathushan Nadanapathan - yathu@seas.upenn.edu - GitHub: yathu-n
   d. Kush Pandey - kpandey@seas.upenn.edu - GitHub: kushpandey1811

2. **Application Description**
   We are planning on using an airbnb dataset along with dataset on some cities to give recommendations about locations/airbnbs a traveller can stay in depending on their preferences. Additionally, we are going to analyze the correlations between airbnbs (price, rating, reviews, etc) and the number of noise complaints in NYC.

3. **Dataset Description**

Dataset #1: AirBNB Data - https://www.kaggle.com/kritikseth/us-airbnb-open-data

- This dataset contains information about Airbnb listings in the US. The attributes in the dataset include the name of the host (and the ID of the host in their database), location of the listing (neighborhood and latitude/longitude), type of property and the cost per night. Based on some exploratory analysis, there are 30 unique cities in the dataset and about 226,000 listings. About 45,756 of the listings are in New York City and these are the ones we will focus on, as explained below.

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.575600e+04	4.575600e+04	45756.000000	45756.000000	45756.000000	45756.000000	45756.000000	35123.000000	45756.000000	45756.000000
mean	2.309535e+07	8.437272e+07	40.729633	-73.950747	149.557413	8.767703	22.560189	0.883917	5.964813	120.740733
std	1.382572e+07	9.920451e+07	0.055025	0.047776	315.367989	23.938680	46.491671	1.270743	22.240225	142.653576
min	2.595000e+03	2.438000e+03	40.508680	-74.239860	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	1.061546e+07	9.250917e+06	40.690030	-73.983150	65.000000	2.000000	1.000000	0.120000	1.000000	0.000000
50%	2.224709e+07	3.726838e+07	40.724550	-73.955385	100.000000	3.000000	4.000000	0.360000	1.000000	52.000000
75%	3.593987e+07	1.359344e+08	40.763160	-73.933990	160.000000	6.000000	21.000000	1.160000	2.000000	251.000000
max	4.533040e+07	3.665571e+08	40.912140	-73.712990	10000.000000	1250.000000	746.000000	44.060000	238.000000	365.000000

- This is a summary chart of all the listings in the dataset. The data still needs to be cleaned, which is why we see listings with a price of 0 and other noisy data. The number of reviews vary between 0 and 966, while the price varies between 0 (corrupted data) and 24,999 dollars per night. The other statistics can be seen in the table.

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	2.260300e+05	2.260300e+05	226030.000000	226030.000000	226030.000000	2.260300e+05	226030.000000	177428.00000	226030.000000	226030.000000
mean	2.547176e+07	9.352385e+07	35.662829	-103.220662	219.716529	4.525490e+02	34.506530	1.43145	16.698562	159.314856
std	1.317814e+07	9.827422e+07	6.849855	26.222091	570.353609	2.103376e+05	63.602914	1.68321	51.068966	140.179628
min	1.090000e+02	2.300000e+01	18.920990	-159.714900	0.000000	1.000000e+00	0.000000	0.01000	1.000000	0.000000
25%	1.515890e+07	1.399275e+07	32.761783	-118.598115	75.000000	1.000000e+00	1.000000	0.23000	1.000000	0.000000
50%	2.590916e+07	5.138266e+07	37.261125	-97.817200	121.000000	2.000000e+00	8.000000	0.81000	2.000000	140.000000
75%	3.772624e+07	1.497179e+08	40.724038	-76.919322	201.000000	7.000000e+00	39.000000	2.06000	6.000000	311.000000
max	4.556085e+07	3.679176e+08	47.734620	-70.995950	24999.000000	1.000000e+08	966.000000	44.06000	593.000000	365.000000

- This is a summary chart of the listings in New York City. The data still needs to be cleaned, which is why we see listings with a price of 0 and other noisy data. The number of reviews vary

between 0 and 746, while the price varies between 0 (corrupted data) and 10,000 dollars per night.

Dataset #2: New York Party Data - https://www.kaggle.com/somesnm/partynyc?select=bar_locations.csv

- A dataset on the number of noise complaints in the 5 boroughs of NYC between 2015 and 2017. Additionally, we have a dataset on 2000 bars in NYC; specifically, we have the bar's latitude and longitude and the number of complaints they've had.
- This dataset contains 225,000 noise complaints (rows). Attributes included in the dataset: city, borough, latitude, longitude, date that the complaint was open, date that the complaint was closed, and location type of the complaint. The average number of complaints that a bar had between 2015 and 2017 is 37.
- Statistics on the party dataset:

	Incident Zip	Latitude	Longitude	num_calls
count	2440.000000	2440.000000	2440.000000	2440.000000
mean	10631.856967	40.733985	-73.952497	37.025820
std	591.657847	0.066349	0.060292	59.641884
min	10001.000000	40.511255	-74.251277	10.000000
25%	10019.000000	40.702064	-73.987822	14.000000
50%	10463.000000	40.728351	-73.957928	21.000000
75%	11217.000000	40.765226	-73.925422	40.000000
max	11694.000000	40.910201	-73.709219	1513.000000

4. **Queries List**

Query 1: Average prices and ratings of AirBNBs given a certain geographical location (based on longitude and latitude) or area.

Query 2: Average prices and ratings of AirBNBs given a set of filtering queries, such as room type, host rating and other metrics.

Query 3: Average minimum number of nights a stay needs to be booked for in a particular neighbourhood.

Query 4: Average party activity in a certain geographical location in New York City (based on longitude and latitude) or suburb.

Query 5: Recommended AirBNBs based on the amount of partying that a given individual desires (currently have the data for NYC and plan to add more cities if we can find the relevant data).