STA 160 Group 7 – Midterm Project

Rishi Bhuva,[1] Aditya Kallepalli,[2] Vishnu Rangiah,[3] and Ivan Yang

[1] *Student ID: 915218518, Email: rdbhuva@ucdavis.edu*

[2] *Student ID: 915079375, Email: arkallepalli@ucdavis.edu*

[3] *Student ID: 916562849, Email: vrrangiah@ucdavis.edu*

[4] *Student ID: 915463046, Email: igyang@ucdavis.edu*

(we will contact the first author for any question about the article)

In this report, our group performed data analysis on 2019 Airbnb listing data in New York City, USA. Apart from general exploratory data analysis, We specifically performed Natural Language Processing Methods on the listing names to gain further insight on diction in influencing different variables in our dataset. We conclude that hosts with lower-priced listings prioritize appealing towards everyday customers with borough names and positive descriptors, while hosts with more expensive listings prioritize listing property features to appeal to higher-end customers. Additionally, hosts will use different languages to target customers belonging to specific cultural groups.
(95 words)

I. INTRODUCTION

We will be analyzing Airbnb data from the "Airbnb Open New York City 2019 Data Set." The data we are currently analyzing comes from https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data.

Airbnb is an online marketplace for lodging, primarily home-stays for vacation rentals, and tourism activities. This platform is accessible through their website or mobile application.

This data set describes listing activity and other metrics in New York City, USA in 2019, such as geospatial information, average number of reviews per month, etc. Through this data set, we can get information regarding hosts, geographical availability and necessary metrics needed to make predictions and draw conclusions.

We will be performing our analysis using Python with the packages *nltk, spacy*, *word cloud* for NLP and *pandas, matplotlib, seaborn, numpy, geopandas, contextily* for EDA. In Section 2, we will be cleaning and performing exploratory analysis on the data set. Then we will explore the relationship between the names of each listing and the popularity and price of each listing using Natural Language Processing Methods in Section 4. We wanted to specifically look at word choice in listing names as they are usually the first piece of information that a customer encounters when searching for listings on the Airbnb platform. Interestingly, we observed a wide variation, ranging from language choice, image-provoking adjectives, to even the use of emojis. Examples of headings:

II. EXPLORATORY DATA ANALYSIS

A. Main Data Frame

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

FIG. 1. (Main Data Frame).

Here we find that our dataset contains 48,895 rows which represent individual listings in New York. We are also provided with 16 different columns which contain necessary factors needed for exploratory data analysis. They are: id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, and availability_365. 7 variables are of integer type, 6 variables are floats, and the others are objects.

B. Data Cleaning

Our first line of action when analyzing this data is to properly clean the dataset. Having clean data will provide us with the highest quality of information needed and therefore will provide us with the most accurate predictions and correlations. Our process for cleaning the dataset can be seen below.

1. Removing Missing and Irrelevant Information

We will first check the number of rows with missing values. We will then remove such rows depending on the missing variables. There are 4 variables with missing values. However, we've determined that all of these missing variables do not greatly affect our NLP analysis. Therefore, we've decided to keep these listings.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60	45	49	2017-10-05	0.40	1	0
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79	2	118	2017-07-21	0.99	1	0
14	6090	West Village Nest - Superhost	11975	Alina	Manhattan	West Village	40.73530	-74.00525	Entire home/apt	120	90	27	2018-10-31	0.22	1	0
20	7801	Sweet and Spacious Brooklyn Loft	21207	Chaya	Brooklyn	Williamsburg	40.71842	-73.95718	Entire home/apt	299	3	9	2011-12-28	0.07	1	0

FIG. 2. (Missing Values in Availability Feature of Data Frame).

On the other hand, while combing through our main dataframe, we discovered that there are a number of listings with 0 available days to be rented as well as 0 total reviews. We chose to extract these listings from the dataframe we will analyze, because zero availability means that the listings are not available at all. Therefore, this data would be irrelevant as we want to perform analysis on listings that have availability.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
19	7750	Huge 2 BR Upper East Cental Park	17985	Sing	Manhattan	East Harlem	40.79685	-73.94872	Entire home/apt	190	7	0	NaN	NaN	2	249
26	8700	Magnifique Suite au N de Manhattan - vue Cloîtres	26394	Claude & Sophie	Manhattan	Inwood	40.86754	-73.92639	Private room	80	4	0	NaN	NaN	1	0
36	11452	Clean and Quiet in Brooklyn	7355	Vt	Brooklyn	Bedford-Stuyvesant	40.68876	-73.94312	Private room	35	60	0	NaN	NaN	1	365
38	11943	Country space in the city	45445	Harriet	Brooklyn	Flatbush	40.63702	-73.96327	Private room	150	1	0	NaN	NaN	1	365

FIG. 3. (Missing Values in Number of Reviews Feature of Data Frame).

The data frame above consists of listings with zero number of reviews. We chose to extract these values from the data we want to analyze, because number of reviews is a factor that can tell us a lot about the popularity of the listing and is an important component to our Natural Language Processing Method which we will perform. Therefore, listings with zero number of reviews would be irrelevant data and not needed for our type of analysis.

2. Examining Outliers – Interquartile Test (IQR)

Now, we will examine the price range of all Airbnb listings by separating these listings into quartiles to get a general insight for this column alone. Along with the 25, 50, 75 % quartile calculations, we can also get information such as the mean, standard deviation and the minimum and maximum value of our price category.

	price
count	26155.0
mean	149.8851844771554
std	198.81696180412806
min	0.0
25%	70.0
50%	109.0
75%	175.0
max	9999.0

FIG. 4. (Summary Statistics of Price Feature).

As observed in our quartile table above, there are surprisingly listings that are $0. Our next course of action would be to pull up these listings and do further research.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
23161	18750597	Huge Brooklyn Brownstone Living, Close to it all.	8993084	Kimberly	Brooklyn	Bedford-Stuyvesant	40.69023	-73.95428	Private room	0	4	1	2018-01-06	0.05	4	28
25433	20333471	★Hostel Style Room | Ideal Traveling Buddies★	131697576	Anisha	Bronx	East Morrisania	40.83296	-73.88668	Private room	0	2	55	2019-06-24	2.56	4	127
25634	20523843	MARTIAL LOFT 3: REDEMPTION (upstairs, 2nd room)	15787004	Martial Loft	Brooklyn	Bushwick	40.69467	-73.92433	Private room	0	2	16	2019-05-18	0.71	5	0
25753	20608117	Sunny, Quiet Room in Greenpoint	1641537	Lauren	Brooklyn	Greenpoint	40.72462	-73.94072	Private room	0	2	12	2017-10-27	0.53	2	0
25778	20624541	Modern apartment in the heart of Williamsburg	10132166	Aymeric	Brooklyn	Williamsburg	40.70838	-73.94645	Entire home/apt	0	5	3	2018-01-02	0.15	1	73
25794	20639628	Spacious comfortable master bedroom with nice ...	86327101	Adeyemi	Brooklyn	Bedford-Stuyvesant	40.68173	-73.91342	Private room	0	1	93	2019-06-15	4.28	6	176
25795	20639792	Contemporary bedroom in brownstone with nice view	86327101	Adeyemi	Brooklyn	Bedford-Stuyvesant	40.68279	-73.91170	Private room	0	1	95	2019-06-21	4.37	6	232
25796	20639914	Cozy yet spacious private brownstone bedroom	86327101	Adeyemi	Brooklyn	Bedford-Stuyvesant	40.68258	-73.91284	Private room	0	1	95	2019-06-23	4.35	6	222
26259	20933849	the best you can find	13709292	Qiuchi	Manhattan	Murray Hill	40.75091	-73.97597	Entire home/apt	0	3	0	NaN	NaN	1	0
26841	21291569	Coliving in Brooklyn! Modern design / Shared room	101970559	Sergii	Brooklyn	Bushwick	40.69211	-73.90670	Shared room	0	30	2	2019-06-22	0.11	6	333
26866	21304320	Best Coliving space ever! Shared room.	101970559	Sergii	Brooklyn	Bushwick	40.69166	-73.90928	Shared room	0	30	5	2019-05-24	0.26	6	139

FIG. 5. (Low Values in Price Feature of Data Frame).

After multiple online searches of these listings, we've come up with mixed results of these listings' robustness. Listings such as Kimberly's seem to have been removed while all three of Adeyemi's listings are still up on AirBnB and running in 2021. Additionally, there isn't any way that we can revisit AirBnB's 2019 websites to verify these listings.

One possible reason for these $0 prices may have been a deliberate act by the hosts to temporarily remove the listing from the Airbnb market when these listings were webscraped. The additional fact that among these $0 listings, 3 and 2 of them belong to the same host may further indicate that it is a host-activated anomaly.

Either way, we will remove these listings just to minimize any possibility of accruing errors based on unknown anomalies. Below are our summary statistics with entries of price = $0 excluded.

	price
count	26147.0
mean	149.93104371438406
std	198.83008554328674
min	10.0
25%	70.0
50%	109.0
75%	175.0
max	9999.0

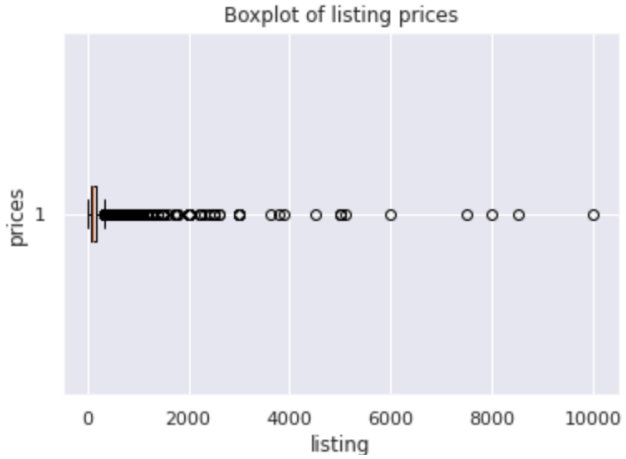


FIG. 6. (Summary Statistics of Price Feature After Cleaning).

FIG. 7. (Box plot of Price Feature).

From the boxplot regarding our modified prices above, we can see that the data is quite skewed with multiple extremely high prices heavily affecting the data. Further research also shows that some of these data points are faulty listings, such as a 2012 Superbowl private room listing.

We will now use the 1.5 Interquartile Range Test to determine outliers within price and store these listings in a different dataframe for further data analysis.

Through the 1.5 Interquartile Range Test, we can properly establish the range of outliers. This test essentially says that a data point is considered an outlier if it is more than 1.5 * IQR above the third and first quartile. Therefore low outliers are below our 25th percentile - 1.5*IQR and high outliers are above our 75th percentile + 1.5*IQR. The IQR variable is calculated using the difference between the 75th percentile and the 25th percentile. The result from this test to determine outliers is: (-87.5, 332.5). We can see that anything above the price of 332.5 dollars would be considered too expensive, and therefore would be an outlier. Since there technically cannot be any negative values for prices, we will fix the lower limit at 10 dollars which is the minimum value as seen in our quartile table. Therefore, any prices under 10 dollars would be considered low outliers. Our next course of action would be to remove these price outliers from the dataset we want to analyze.

3. *Tidied Dataset*

After removing all unnecessary information and outliers, we are left with a tidied dataset consisting of 24540 listings.

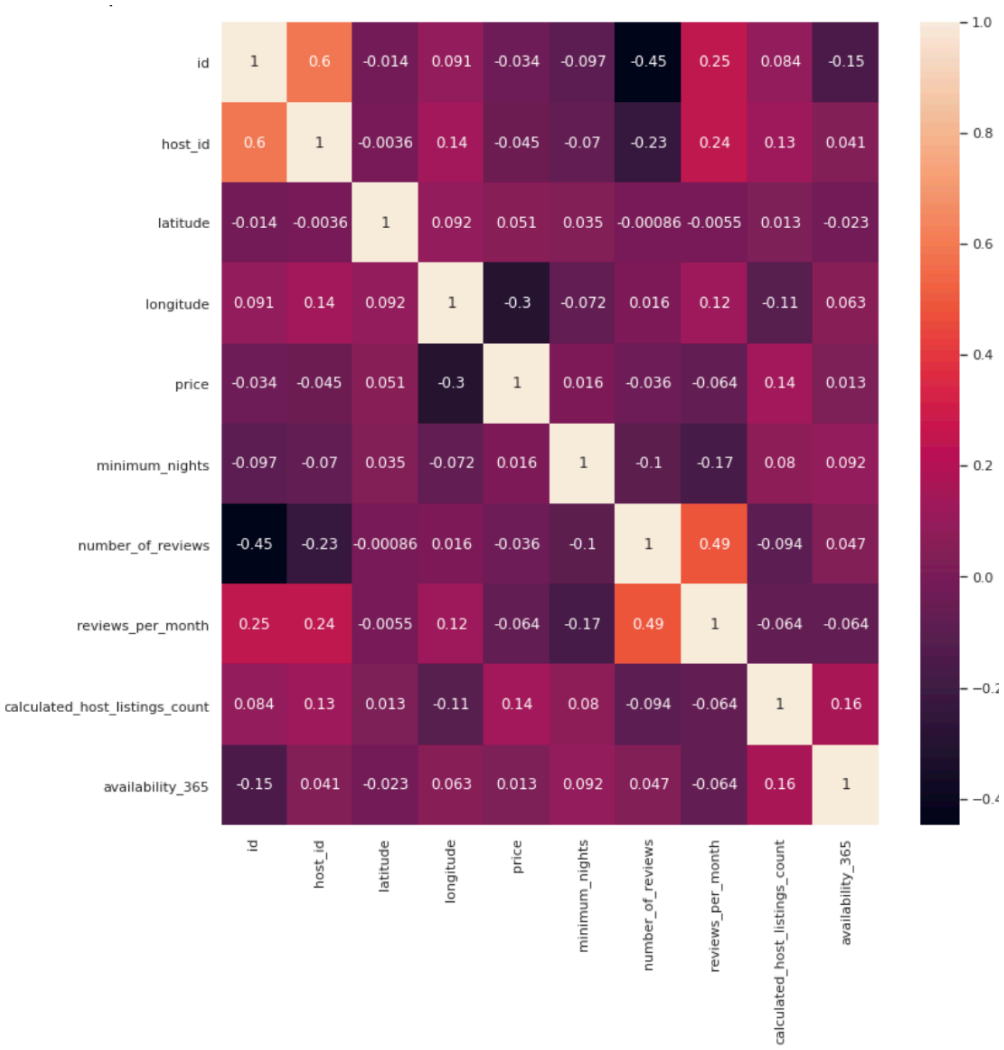


FIG. 8. (Correlation Matrix Between Variables).

Here, we can see a proper correlation matrix between the variables within our dataset. According to this plot, we can see that the more lighter the color, the higher the correlation between each variable. Vice versa, we can see that the darker the number, the lower the correlation. From this plot, we can immediately see a high correlation of 0.6 between id and host_id. Along with this, it can be seen that number of reviews and reviews per month also have a higher correlation than others at 0.49. Along

with positive correlations, we can also see negative correlations with id and number of reviews at -0.45 and with id and reviews per month at -0.45. We can also see a negative correlation between price and longitude at -0.3, although a positive correlation of 0.051 between price and latitude.

As id is a specific tag per listing, host_id is also a specific tag per host. Therefore, this positive correlation does seem to make sense as some hosts would have multiple listings under their name. Number of reviews and reviews per month also tend to have a positive correlation as each number of reviews contributes to a review per month and vice versa, therefore these two variables seem to have a direct relationship.

The lowest correlation we can see in this plot can be seen between both variables of reviews and id. The large negative correlation between them does seem accurate, as each id is specific and randomly generated and does not relate at all to reviews at all. Another negative correlation that was interesting comes from price and longitude. As longitude tends to be west and east, it can be seen that price in New York does not tend to have any sort of relationship with those specific directions. Although, price and latitude show a somewhat positive correlation, meaning that price is more correlated with listings that range from the north and south rather than the west and east.

III. GRAPHS

We will plot some graphs with our tidied dataset to gain further insight, as well as to visualize any more abnormal data points for further cleaning. We will not be subsetting these outliers from the main tidied dataset as these variables will not be our main focus for our NLP analysis. However, we will generate graphs of these outliers for further analysis.

A. Tidied Dataset

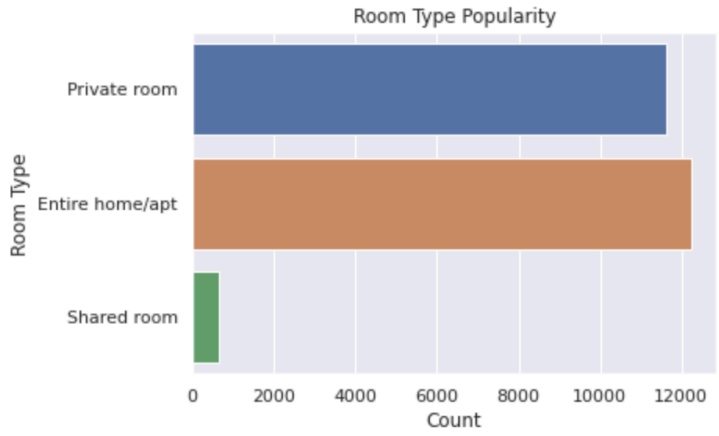


FIG. 9. (Bar Plot of Room Type Popularity).

We can see that most listings are either private rooms or the entirety of the property. This indicates that most Airbnb customers may be looking for private accommodation rather than shared spaces.

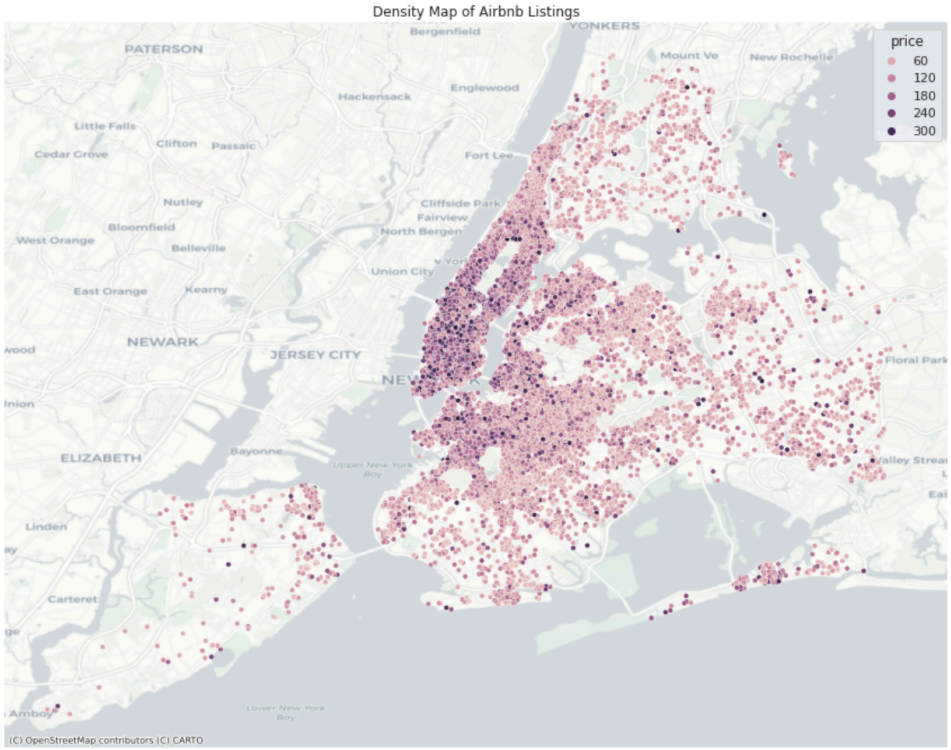


FIG. 10. (Density Map of Price of Airbnb Listings).

From the density map above, we can see that most of the listings are concentrated around the boroughs of Brooklyn and Manhattan. We can also observe that Manhattan has a higher density of

expensive listings, and apart from standout anomalies, prices tend to be lower as listings become further from Manhattan.

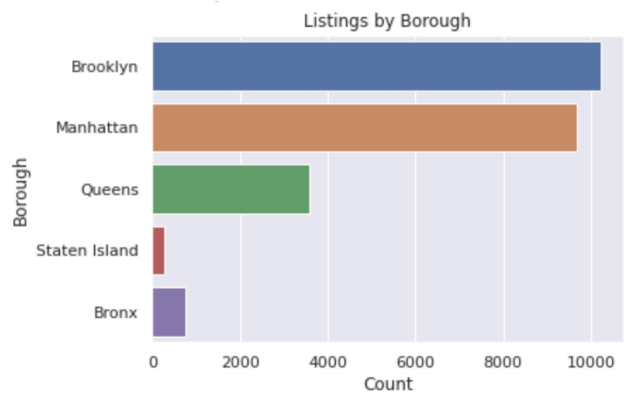


FIG. 11. (Bar Plot of Listings by Borough).

A deeper analysis shows that most listings are either in the boroughs of Brooklyn or Manhattan. This is logical given that these 2 boroughs contain the most attractions and activities in New York City. Staten Island has the lowest number of listings, which may be due to its geographical location being relatively inaccessible compared to the other boroughs.

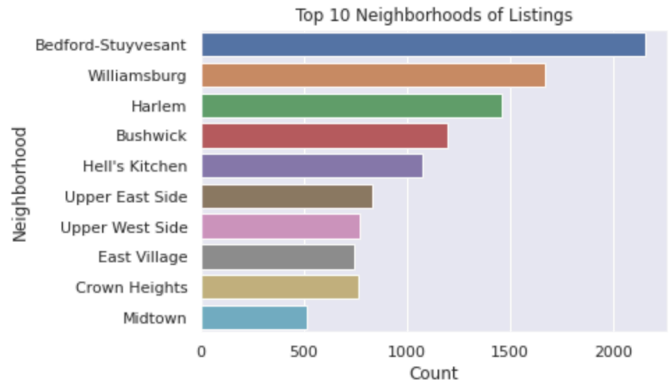
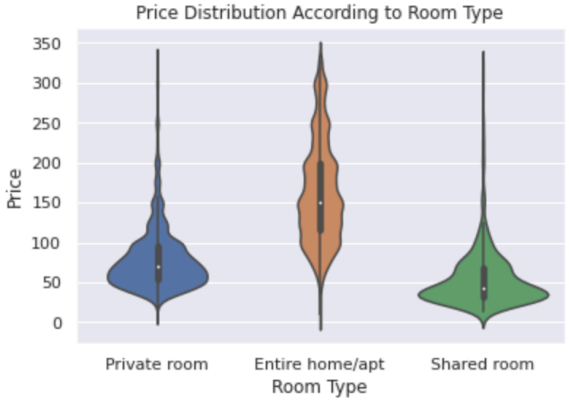


FIG. 12. (Bar Plot of Listings by Neighborhoods).

The top 4 neighborhoods of Airbnb listings are all in Brooklyn, with Bedford-Stuyvesant being the top neighborhood by quite a margin. The 5th to 8th most popular neighborhoods are in Manhattan. This may be due to affordability of Brooklyn compared to Manhattan.

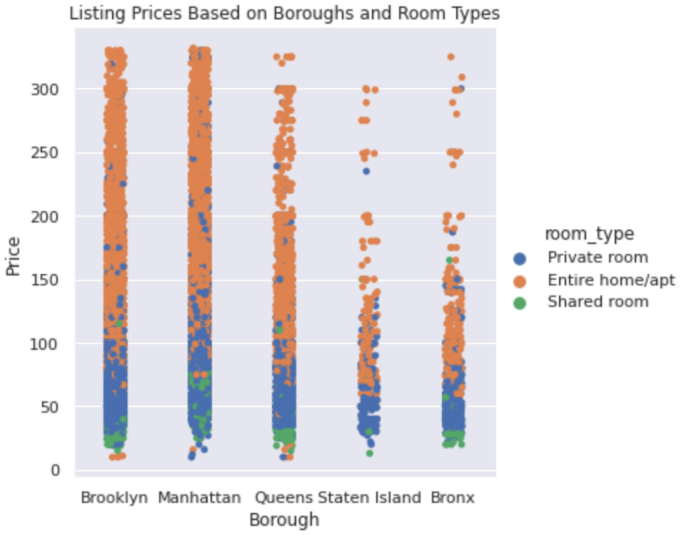


143

144 FIG. 13. (Histogram of Listing Prices). FIG. 14. (Violin Plot of Room Type).

145 From the histogram above, we can see a left-skewed unimodal distribution. A majority of the
146 listings hover around a price between 35 to 130 USD. However, there are significant spikes at 150,
147 200, 250, and 300 USD bins, as well as smaller spikes at 175, 225, 275, and 325 USD bins. We
148 believe that this may be caused by human psychology where hosts may round their prices to the
149 nearest 25 or 50 dollar for a more "aesthetically pleasing" price number.

150 The violin plot above shows the distribution of price by room type. There is much variation
151 in price within each room type. We can see that price greatly fluctuates between the type of room. It
152 can be seen that entire home/apt listings tend to be more expensive and evenly distributed across
153 the price range than both private room and shared room, although private room is slightly more
154 expensive than a shared room. A majority of the private and shared rooms tend to be around $40 to
155 $60. The logic behind the prices seem to make sense.



156

157 FIG. 15. (Listing Prices Based on Boroughs and Room Types).

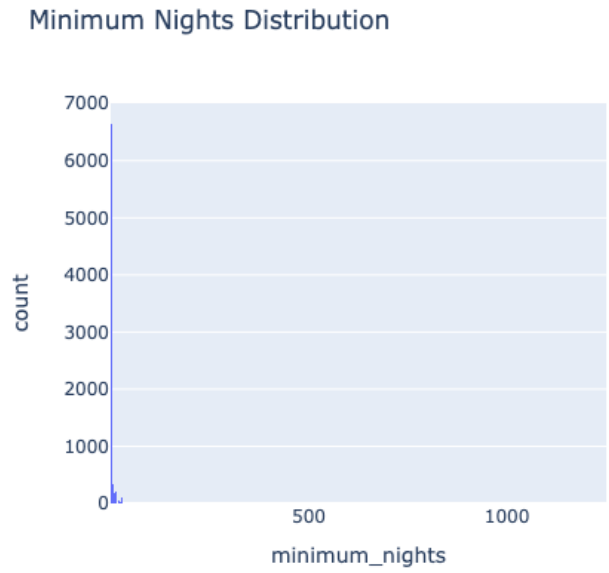
158 As also observed in our previous violin plot, there is an obvious trend that shared rooms are the

159 cheapest option in all boroughs, followed by private room and entire property. Additionally, Staten

160 Island and the Bronx have a lot less listings above $150 compared to Brooklyn and Manhattan.

161 Additionally, as seen in our density map, Manhattan has the highest density of expensive listings

162 compared to other boroughs.



163

164 FIG. 16. (Histogram of Minimum Nights).

Most listings in 2019 NYC have a minimum night requirement between 1 to 7 nights, with an average minimum night requirement of 6.64 and 2 nights being the highest. There is an apparent spike of listings with a minimum night requirement of 30 nights.

There seems to be multiple outliers, with one as extreme as 1250 minimum nights. This wide range of minimum night requirements allude to the existence of not just long term vacation rental properties but also potentially long term residence.

We will determine these outliers again using the 1.5 IQR test. We obtain a lower limit of -156.5 and 161.5 nights. Since there technically be any negative values for minimum nights, we will fix the lower limit at 0. We end up with 47 outlying listings with extreme values of minimum nights.

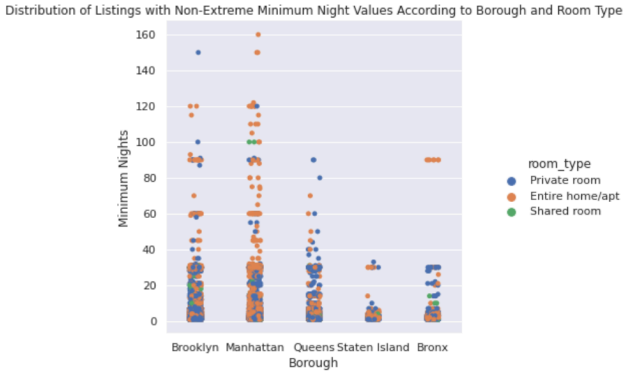


FIG. 17. (Minimum Nights based on Borough and Room Types).

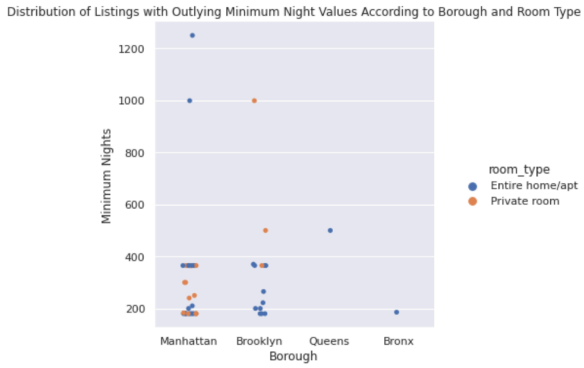


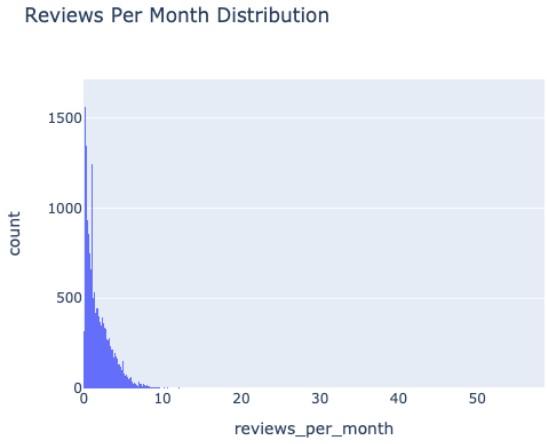
FIG. 18. (Outlying Minimum Nights based on Borough and Room Type.)

From the graph, it seems like Staten Island and the Bronx are not popular boroughs for long term rentals longer than 30 days. Further investigation of the few orange anomalies in the Bronx reveal that they are all from the same host, Sasha (Host ID: 2988712), who seems to have multiple 90-day listings around Claremont Village and Mount Hope in the Bronx.

Most longer-term rentals in Brooklyn, Manhattan, and Queens are of entire properties. Interesting anomalies include the 90-day minimum night requirement listings for a shared room in Manhattan. Further investigation shows that these listings (Host IDs: 21628183 and 23184420) are both from

LaGuardia Houses Public Housing Development looking for long term housemates. We believe that these listings are a result of hosts trying to bypass New York's legal barrier of subletting public housing for additional income, adding onto the high accessibility of Airbnb for anyone to put out a listing.

As expected, most extreme values of minimum night requirement are from listings in Manhattan and Brooklyn, with Queens and the Bronx each only having 1 listing and Staten Island not having any. Both listings in Queens and the Bronx seem to be long term rental leasings, just like most of the other leasings in this plot.



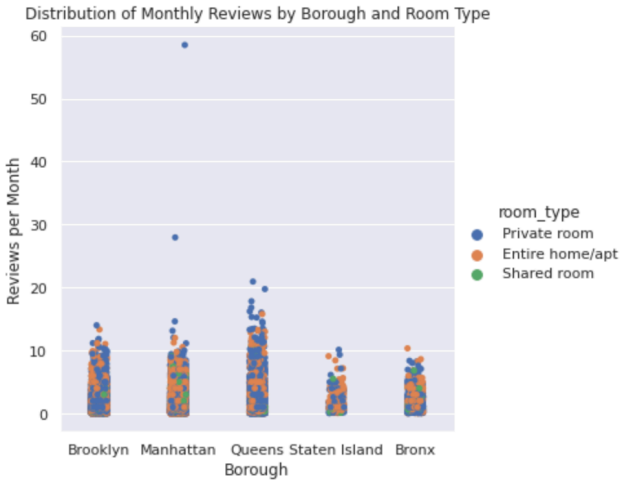


FIG. 19. (Histogram of Price Feature).

FIG. 20. (Reviews per Month based on Borough and Room Type).

The number of reviews are heavily skewed to the left with listings getting an average of 1.83 reviews per month. There are obvious outliers, with a listing getting as many as 58.5 average reviews per month.

We can see that most boroughs have a similar pattern of reviews per month, with 2 anomalies in Manhattan. Further analysis reveals that both listings (ID: 32678719 and 32678720) are by the same host Row NYC (Host ID: 244361589) which is a hotel in New York City. They may have more reviews as they may be more attractive to customers as an established hotel chain.

Additionally, they may have multiple rooms available under the same listing, which would explain why they have listings getting 58.5 reviews in a span of 30 or 31 days.

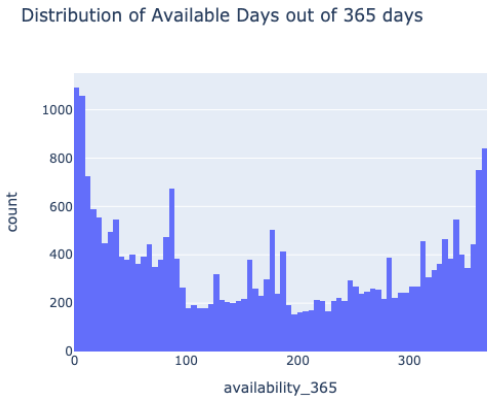


FIG. 21. (Histogram of Available Days).

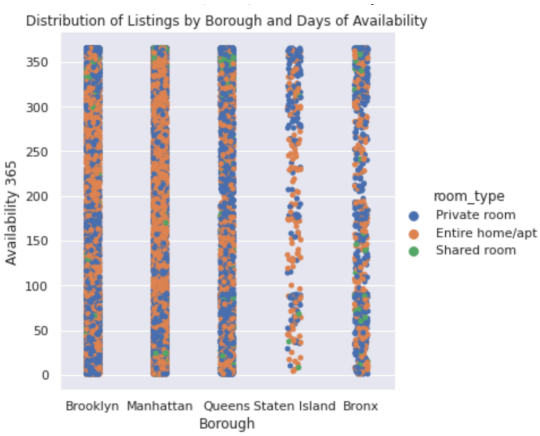


FIG. 22. (Availability based on Borough and Room Types).

We can see a bimodal distribution with spikes at both ends. This tells us that on one extreme, most NYC hosts were not leasing out their place during 2019, or maybe for a duration not more than a week. On the other extreme, there were hosts that had listings available for almost the entire year of 2019.

All boroughs had availability ranging from 0 to 365, with not much of a differentiation between room types. It is interesting to note that a majority of Staten Island listings were available for more days of the year compared to other boroughs.

B. Price Outliers

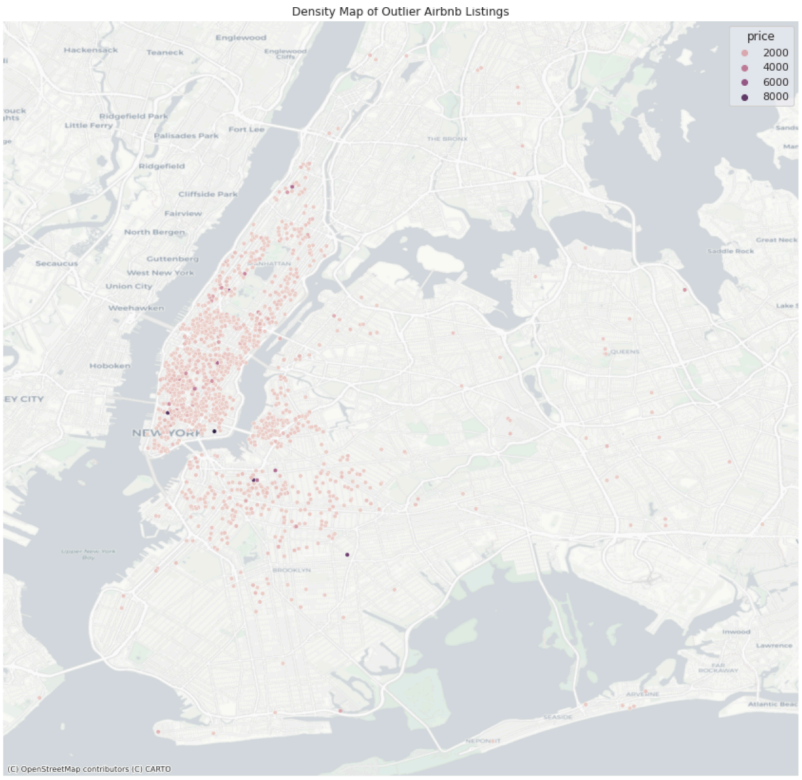


FIG. 23. (Density Map of Price Outliers of Listings).

Most listings in our outlying price range falls around $2000. There does not seem to be an apparent deviation of geographical correlation between expensive Airbnb listing prices compared to our tidied dataset, with a majority of the listings also concentrated in the Manhattan area. There is also only one Staten Island listing under this price range, which supports our above findings that the average Staten Island listing prices are much lower compared to the other 4 boroughs.

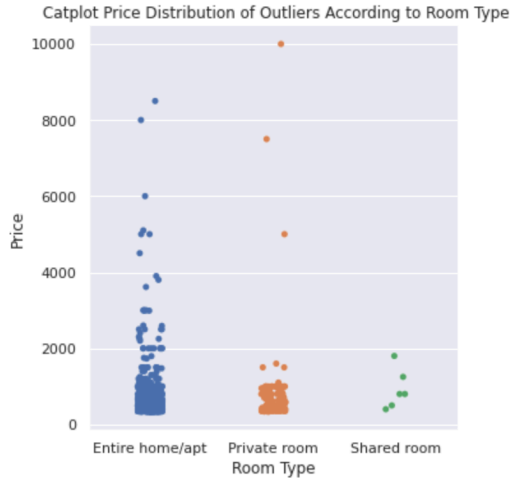


FIG. 24. (Catplot of Price Features based on Room Type).

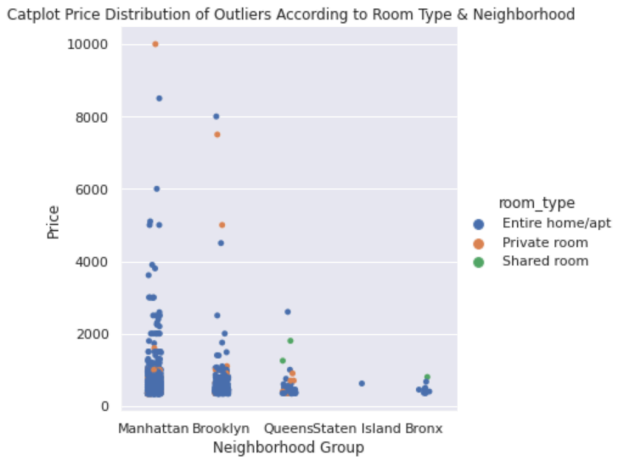


FIG. 25. (Catplot of Outlying Price Features based on Room Type).

As seen from the plot above, most of the outlying expensive listings are either for entire properties or private rooms. Interestingly, the most expensive listing is a 10000USD private room in Queens. This listing is now defunct, but we can infer that it might be a long term rental situation as the listing also indicates a minimum night requirement of 100 nights.

From this plot, it can be seen that the majority of outliers for price, seem to be under the room type of an entire home/apt. The home/apt room type is also the majority for each neighborhood group. We can also see that Manhattan tends to have the most number of price outliers, while Brooklyn has the second highest. The logic that entire home/apt are prevalent seem to make sense, as an expensive listing would be an entire home rather than a private or shared room. We can also see that Staten Island tends to have the least amount of outliers that are too expensive and that majority of the room type of outliers in Staten Island are entire homes/apartments.

IV. NATURAL LANGUAGE PROCESSING (NLP)

We will be exploring the name feature of the Airbnb dataset in depth using NLP techniques. With the use of the popular NLP library Spacy we are able to clean the listing names and extract

relevant information from them such as the most frequently used words in listings as well as non-English symbols used in the listings.

In our analysis of the names of the listings, we provide visualizations to show the relationship of the name listings across price subgroups and popularity subgroups which we equated to be the number of reviews per month.

Additionally, we explore the unique listing names which contain non-English symbols and plot their locations to see if there is any correlation between the language used in the listing and the listing location. This correlation may be present in listing names that contain Chinese characters and are potentially located in Chinatown.

FIG. 26. (Word Cloud of Tidydf Data Set).

An initial analysis of our raw dataframe's listing names shows that the words with the most frequencies are neighborhood and borough names, type of room, and general location indicators. It would make sense that Brooklyn and Manhattan are some of the most frequent words because they are the 2 most popular boroughs as seen in our EDA. We can also infer that hosts use listing names to capitalize on different features of the property, from location indicators such as "heart" and "near", number of bedrooms, and different landmarks and neighborhood names. Additionally, it is

interesting to note the many different variations of phrases containing "cozy", such as "cozy room", "cozy bedroom", etc.

A. Data Cleaning – Listing Names

Apart from ensuring that we are only comparing active listings in 2019, we also observed many spelling errors and large input variations which would greatly affect the quality of our NLP analysis. We will thus clean and create subgroups of the listing names based on their language (English and non-English) and the presence of non-alphabetical characters, particularly emojis.

We begin by converting our name column into a list of strings. Then using spacy's NLP package we iterate through our list element of listing names and tokenize each name. We are able to identify which words in each name contain punctuation or are part of spacy's stop words and we discard those from each name. We also recognized that iterating through the list of names reduced our computation time significantly than iterating on the name column directly. Another aspect of cleaning which we kept in mind was not disregarding special characters since these characters will later be used to identify non-english listings as well as listings that contain emojis. We did choose to omit the words "apt", "apartment", "bedroom" from each of the names by adding them to spacy's stop words since we believe that these names did not provide any relevant meaning and in our initial finds were present across many subgroups of both popularity and price.

Here, we used NLP to get us the frequency of words within our prices category. As seen from the plots below, we established price groups and got the word frequencies for each group. We essentially used the summary statistics to create each group. Therefore, the low price group consists of prices in the 25th percentile, medium in the 50th percentile, and high in the 75th percentile. The

"Extra Price" group consists of outliers in our price entries. These outliers include prices that are above the 75th percentile and prices below the 25th percentile.

B. Data Analysis

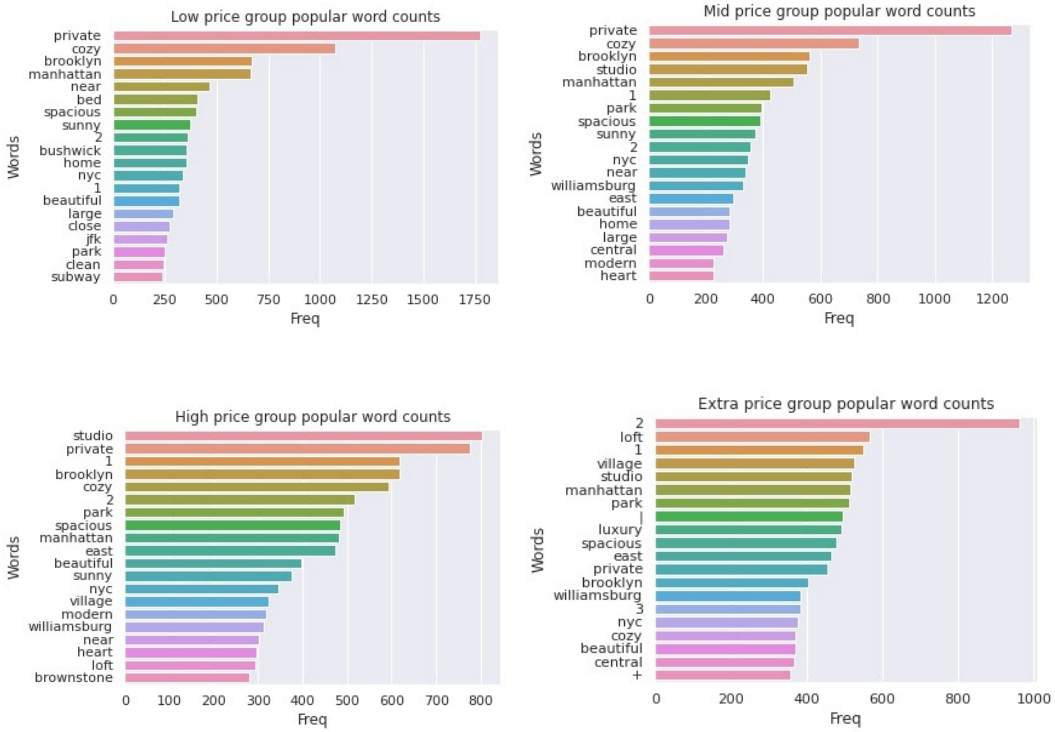


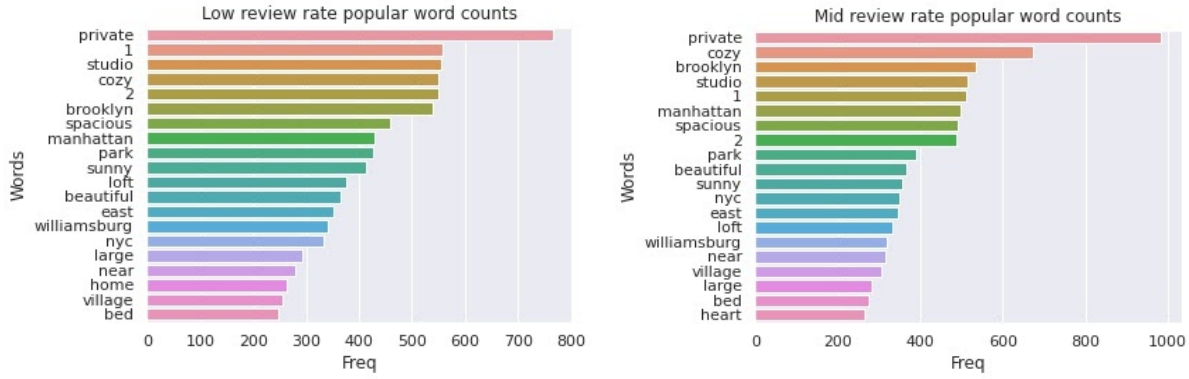
FIG. 27. (Count plots of Word Frequency on Type of Price).

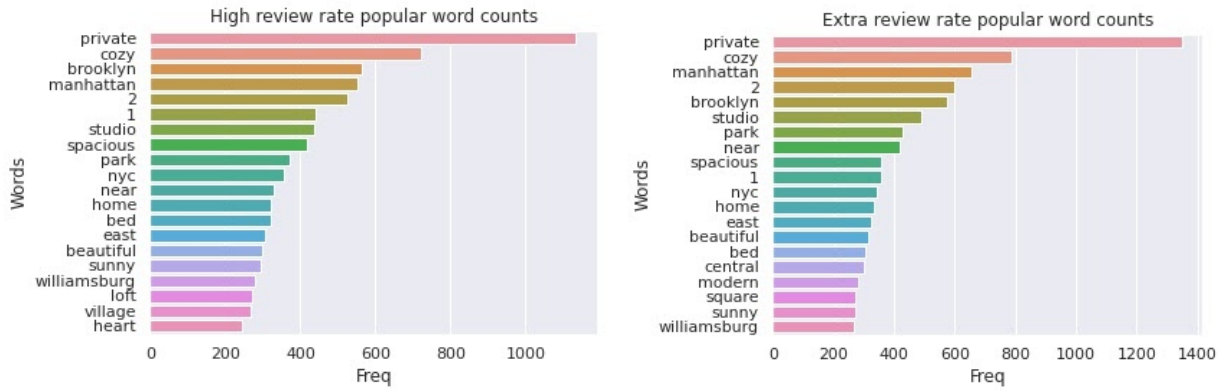
From the above 4 bar charts, we can see that the term "private" is the most common, and "cozy" is also the second most common term across the lower two price ranges. Referring back to the word cloud, this would make sense as we concluded that the most frequent term across all listings was "private room." The prevalence of this term indicates just how important privacy is to everyday Airbnb customers in New York City. This is understandable as given New York City's high population density, people would especially want a homey place to unplug from the hustle and bustle.

In comparison, many of the common terms across the higher two price brackets are of property features such as "1" and "2", most likely referring to the number of rooms or "bed"[rooms], "studio", etc. Additionally, there are terms that one would associate with more expensive properties, such as "luxury", "loft", and "brownstone". However, it seems like all Airbnb renters want properties that are "spacious", "beautiful", and "sunny".

There are some common geographical terms across all price ranges, such as "nyc", "manhattan", "brooklyn", "park", etc. Some of these terms' frequencies are clearly correlated with price, such as "park", most likely referring to proximity with Central Park as "central" is also a term that appears in the mid and extra price range. We can also see a correlation between lower-priced listings and convenience in transportation accessibility, as only the lower price bracket has the terms "jfk" and "subway". There are also specific neighborhood names that correlate to price, such as "bushwick" with lower prices and "williamsburg" with higher prices.

In general, we can also observe that hosts with lower-priced listings prioritize appealing towards everyday customers with borough names and positive descriptors, while hosts with more

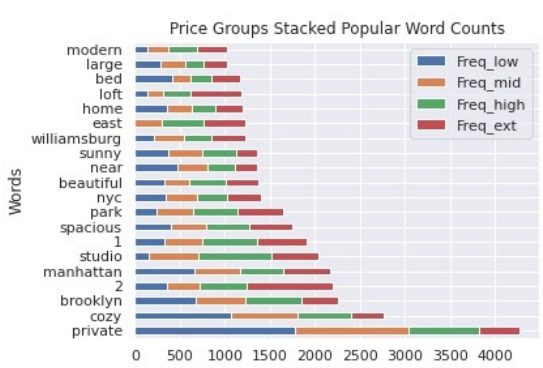




305

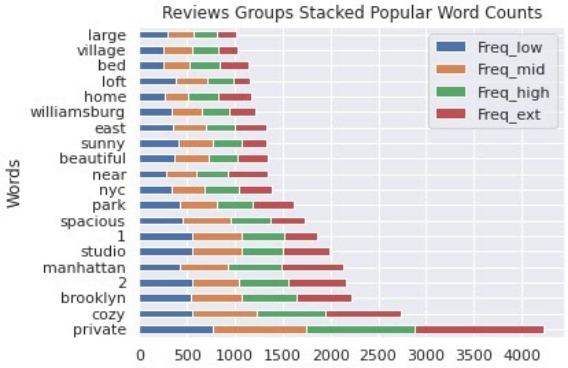
306 FIG. 28. (Count plots of Word Frequency on types of Review Rates).

307 From the above 4 bar charts, we can see that the first two most universal common terms
308 "private" and "cozy" seem to have no effect on determining whether a renter would leave a review.
309 Unique terms such as "modern" and "square" in listing names suggest that listings with high review
310 rates tend to be more modern and/or mention Times Square.



311

312 FIG. 29a. (Stacked plot of Word Frequency on Price).



313 FIG. 29b. (Stacked plot of Word Frequency on Reviews).

314 These stacked plots above show the frequencies of the most popular words among the 4
315 price and review groups. The stacked bar plots allow us to see differences between each group.
316 Looking at the review groups stacked bar plots we can notice that the frequency of each word of
317 each group looks very close to each other. While the price groups have noticeably different

frequencies among the same words. This is likely because there is an even distribution of review scores between each of the different price categories. Looking deeper into the stacked plot for word counts based on price groups, we notice that the lower-priced listings have much higher usage of the words "private" and "cozy" compared to the other groups.

C. Finding Special Characters in Data Set

When looking through the list of names, we discovered certain listing names with emojis. To further explore this we used simple python techniques to identify all the listing names which contained characters that could not be encoded as ASCII characters. Apart from 230 names containing emojis, we found names with accented letters such as é in French listings. We also found 160 non-English listing names in total. Among this subset was a big category of listings which included East Asian characters, specifically Chinese. We reasoned that this was due to New York having a prevalent East Asian community and having one of the most recognized Chinatowns in the world.

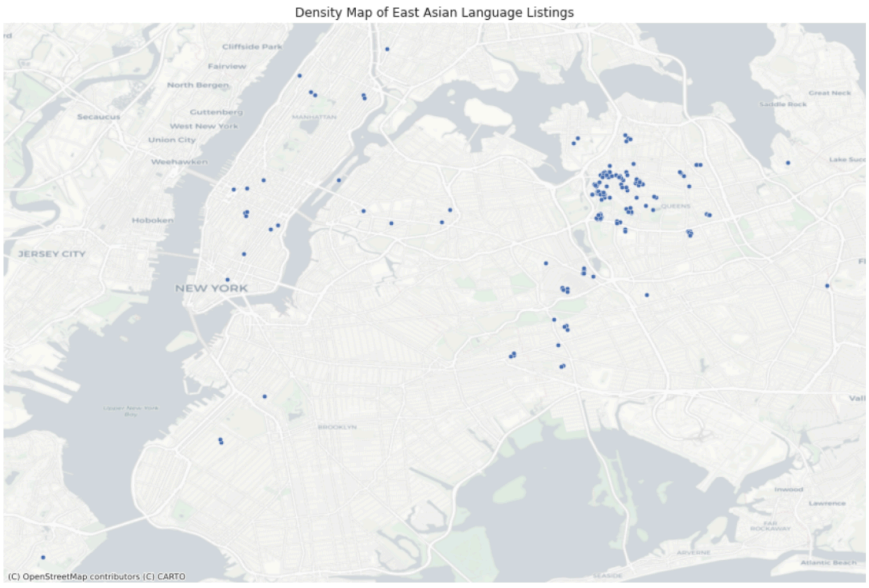
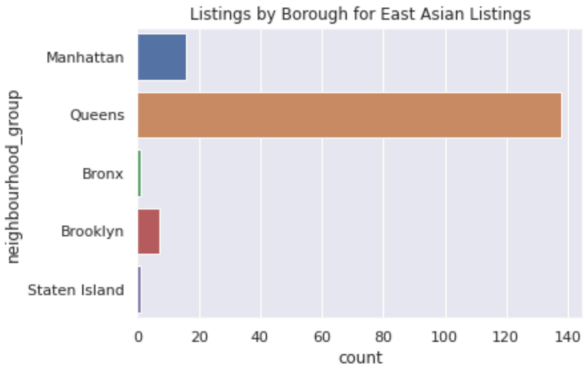


FIG. 30. (Density Map of East Asian Listings).



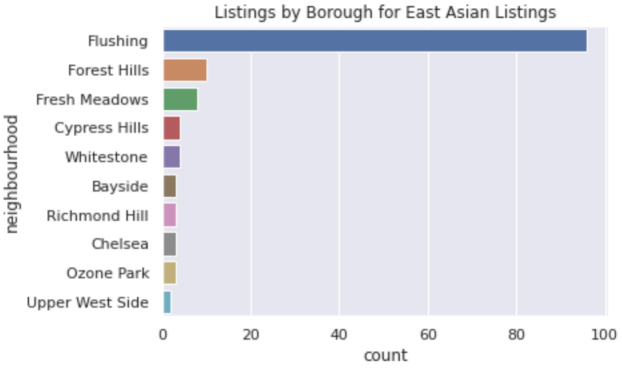


FIG. 31. (East Asian Listings by Borough). FIG. 32. (East Asian Listings by Neighborhood).

From the density map and bar charts, we can clearly see a concentration of non-English listings in Flushing, Queens. This confirms our hypothesis that many of these non-English listings are geographically correlated to areas with large populations of specific East Asian cultural groups, in this case, the Chinese community in Flushing Chinatown.

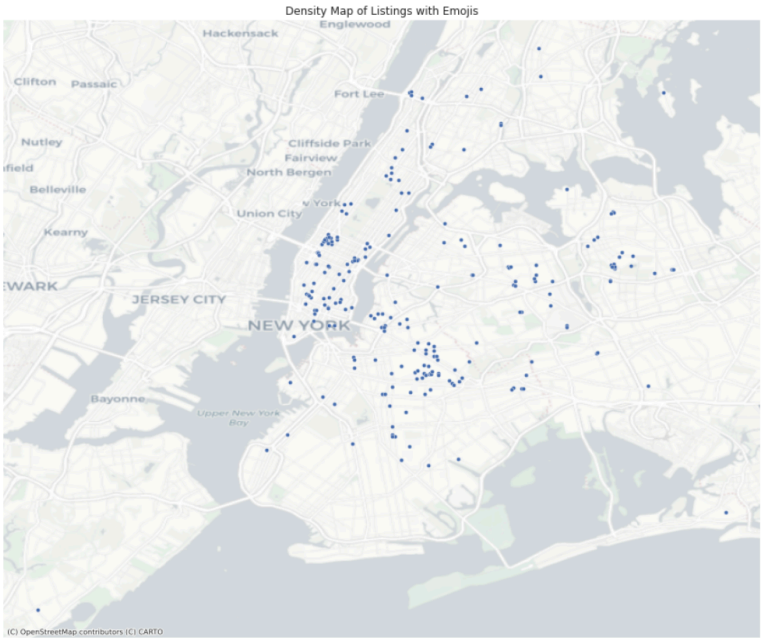


FIG. 33. (Density Map of East Asian Listings).

Unlike non-English listings, there doesn't seem to be a geographical correlation between listing names with emojis, apart from a slight clustering in Hell's Kitchen. Further analysis of the

Hell's Kitchen listings show that this is purely coincidental, as there aren't any single hosts with multiple emoji listings in Hell's Kitchen driving up the concentration apart from Louis (Host ID: 209405908) with 2 listings in Hell's Kitchen.

V. CONCLUSION

In our report, we have found that most New York City Airbnb listings are either in Brooklyn and Manhattan, with private properties and rooms being the most popular listing type in 2019. There is a very wide range of listing prices, but a majority of listings fall below $330. There is also a wide range of minimum night requirements among listings, indicating the availability of many short-term and long-term rental options.

Through NLP, we gained further insight of how Airbnb hosts word their listing names to appeal to customers. Many hosts include technical terms that state the layout of their properties, geographical indicators of proximity, and positive descriptors invoking homeyness. Additionally, these word choices change according to how expensive their listings are. Hosts with non-English listing names, such as Chinese and other East Asian languages, attempt to appeal to corresponding cultural groups based on the cultural demographic of the listings' geographic location. However, there does not seem to be any immediate correlation between the use of emojis and other non-ASCII characters with geography and other variables.

Some limitations that we encountered included the inability to accurately distinguish non-English listings with English listings. Our python package only was able to differentiate between listings with non-English characters. Additionally, many listing names had a lot of typos that resulted in the package recognizing it as a foreign language. A suggestion for further analysis could be looking into correlation of other languages with our variables apart from East Asian languages once this roadblock is addressed.

APPENDIX

REFERENCES (BIBLIOGRAPHIC)

"Creating a GeoDataFrame from a DataFrame with Coordinates." Creating a GeoDataFrame from a DataFrame with Coordinates - GeoPandas 0.9.0 Documentation, geopandas.org/gallery/create_geopandas_from_pandas.html.

Nick-Ulle. "Nick-Ulle/Teaching-Notes." GitHub, github.com/nick-ulle/teaching-notes/blob/master/sta141b/2018/discussion09.ipynb.

"SpaCy 101: Everything You Need to Know · SpaCy Usage Documentation." SpaCy 101: Everything You Need to Know, spacy.io/usage/spacy-101.