COMP8430 Project
U5883475 word count:1900 words

1. Data sets, problem description and overall strategy

Airbnb is one of the most popular platform business in recent years and it facilitate exchange between producers and consumers. By merging Airbnb data and other city datasets, the end use of merged dataset is to examine the relationship between rent price and inner attributes and outer attributes. Inner attributes means the factors about apartment/house itself such as number of bedrooms or cancellation policy. Outer attributes means the environment around apartment/house such as transportation convenience or criminal rate and so on.

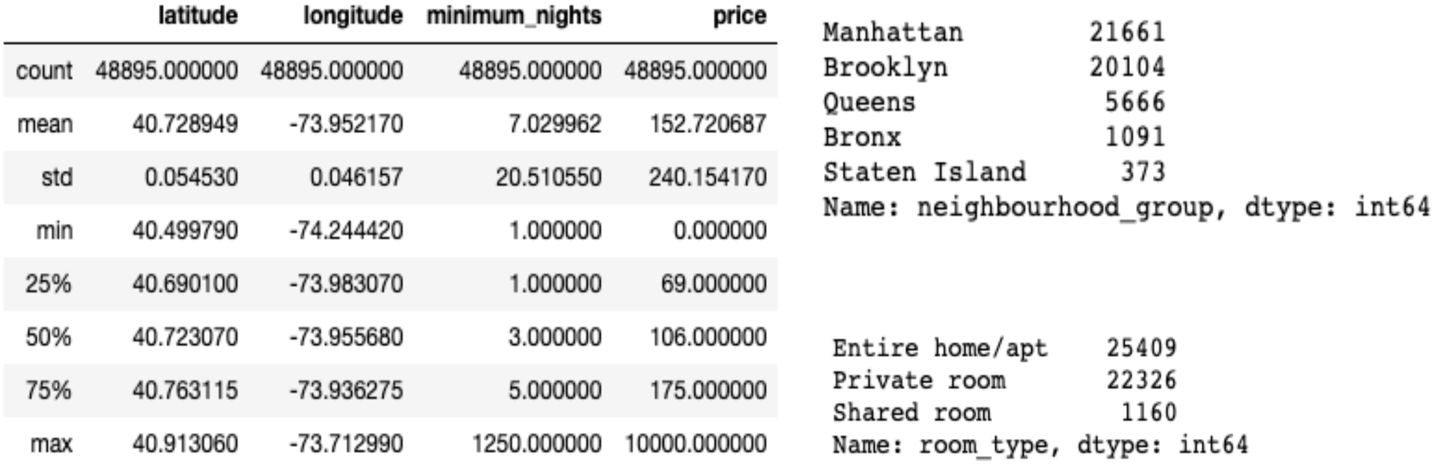
This project use five datasets about New York City from different sources and merge them together to meet the end-use. The overall strategy is merging datasets (Dataset1 and Dataset2) with inner attributes together by id and then merging with other three datasets(Dataset 3, Dataset 4, Dataset 5) with outer attributes by geographic information.

2. Data description and data exploration

Dataset 1: AB_NYC_2019.csv

URLs: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
Purpose: This dataset was collected to make predictions and draw conclusions about Airbnb hosts by using geographical availability and necessary metrics.
Who collected: Dgomonov (Kaggle user)
When collected: 13/08/2019 (2 months ago)

Data exploration: This dataset has 48895 observations with 16 columns and it is a mix between categorical and numeric values. The relevant attributes are id, neighborhood group, latitude, longitude, room type, minimum nights, and price. Table 1 shows statistics about all relevant attributes. latitude and longitude have small standard deviation and mean is very close to median, but minimum nights and price have large standard deviation. In minimum nights, the mean is 7 and median is 3, which suggests the distribution is strongly skewed. Similarly price has strongly skewed distribution as well. The right table shows the frequency of categorical variable. The boxplots and histogram in Figure 1 shows price and minimum nights have left skewed distribution and many outliers.



	latitude	longitude	minimum_nights	price
count	48895.000000	48895.000000	48895.000000	48895.000000
mean	40.728949	-73.952170	7.029962	152.720687
std	0.054530	0.046157	20.510550	240.154170
min	40.499790	-74.244420	1.000000	0.000000
25%	40.690100	-73.983070	1.000000	69.000000
50%	40.723070	-73.955680	3.000000	106.000000
75%	40.763115	-73.936275	5.000000	175.000000
max	40.913060	-73.712990	1250.000000	10000.000000

```
Manhattan        21661
Brooklyn         20104
Queens            5666
Bronx             1091
Staten Island      373
Name: neighbourhood_group, dtype: int64
```

```
Entire home/apt    25409
Private room       22326
Shared room         1160
Name: room_type, dtype: int64
```

Table 1

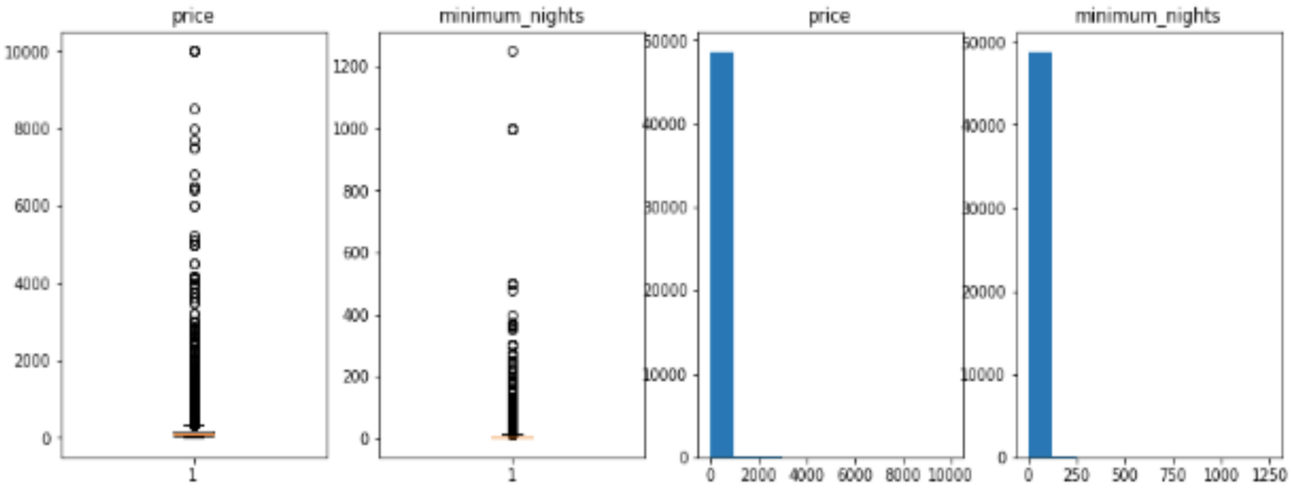
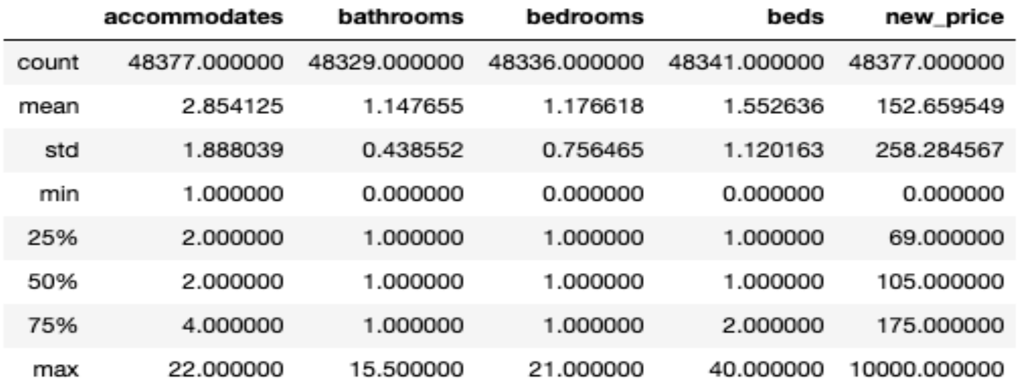


Figure 1

Dataset 2: listings.csv
URLs: http://insideairbnb.com/get-the-data.html
Purpose: This dataset was collected to describe the detailed listings data for New York City on Airbnb.
Who collected: Inside Airbnb
When collected: 12/09/2019
Data exploration:

This dataset has 48377 observations with 106 columns and it is a mix between categorical and numeric values. The relevant attributes are id, neighborhood group cleansed, zip code, latitude, longitude, accommodates, bathrooms, bedrooms, beds, price, instant bookable, cancellation policy. Same as before, Table 2 and Figure 2 shows accommodates, bathrooms, bedrooms, beds, price has a strongly skewed distribution. Also, Table 2 suggests there are some missing value in bathrooms, bedrooms and beds.



	accommodates	bathrooms	bedrooms	beds	new_price
count	48377.000000	48329.000000	48336.000000	48341.000000	48377.000000
mean	2.854125	1.147655	1.176618	1.552636	152.659549
std	1.888039	0.438552	0.756465	1.120163	258.284567
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	1.000000	1.000000	1.000000	69.000000
50%	2.000000	1.000000	1.000000	1.000000	105.000000
75%	4.000000	1.000000	1.000000	2.000000	175.000000
max	22.000000	15.500000	21.000000	40.000000	10000.000000

Table 2

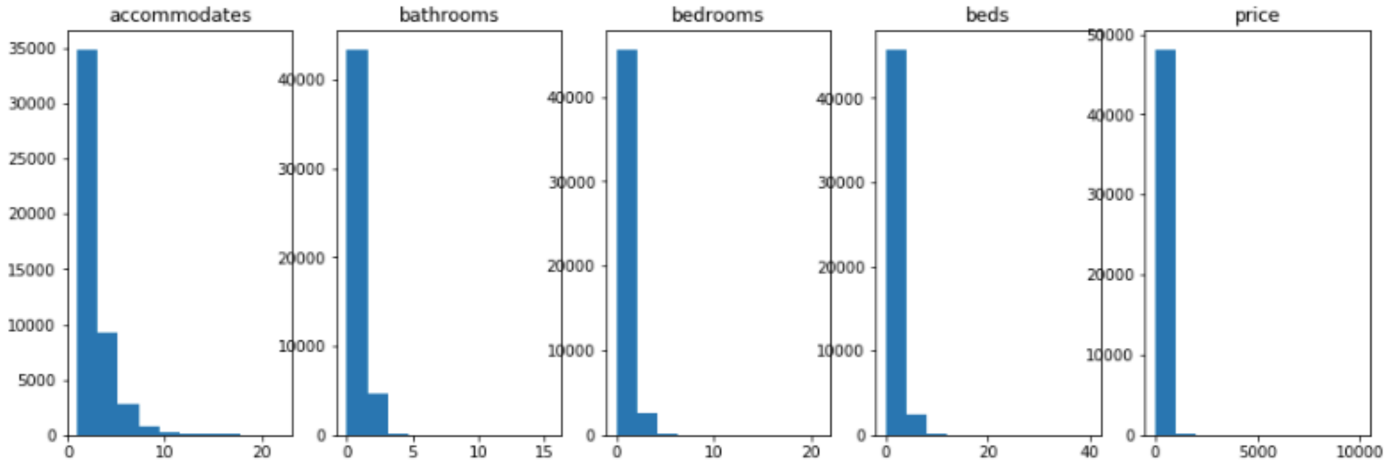


Figure 2

Dataset 3: SUBWAY_ENTRANCE.csv
URLs: https://data.cityofnewyork.us/d/drex-xx56?category=Transportation&view_name=Subway-Entrances
Purpose: This dataset was collected to engage people in the information that is produced and used by city government and benefit every people
Who collected: Metropolitan Transportation Authority (MTA), NYC OpenData
When collected: 11/09/2018
Data exploration: This dataset has around 1928 rows with 7 columns and it is a mix between categorical and numeric values. The relevant attributes are the_geom (including longitude and latitude) and LINE. There is some outliers in longitude and latitude but this is not a series problem.

Dataset 4: NYPD_Arrest_Data__Year_to_Date_.csv
URLs: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc
Purpose: This dataset was collected to explore the nature of police enforcement activity by the public.
Who collected: Police Department (NYPD), NYC OpenData
When collected: 19/07/2019
Data exploration: This dataset has around 113651 rows with 18 columns and it is a mix between categorical and numeric values. The relevant attributes are Latitude, Longitude, AGE_GROUP, PERP_SEX. There is no series problem in terms of outliers and unexpected values in all relevant attributes.

Dataset 5: NYC_Free_Public_WiFi_03292017.csv
URLs: https://data.cityofnewyork.us/Social-Services/NYC-Wi-Fi-Hotspot-Locations/a9we-mtpn
Purpose: This dataset was collected to engage people in the information that is produced and used by city government and benefit every people.
Who collected: Department of Information Technology & Telecommunications (DoITT), NYC OpenData
When collected: 27/09/2019
Data exploration: This dataset has around 3319 rows with 29 columns and it is a mix between categorical and numeric values. The relevant attributes are the_geom (including longitude and latitude), POSTCODE, PROVIDER, TYPE. There is no series problem in terms of outliers and unexpected values in all relevant attributes.

3. Data quality assessment

Because some original datasets are very large and include many attributes, here I mainly examine the data quality in attributes that is relevant to end-use. Relevant attributes are selected in last section.

Dataset 1: AB_NYC_2019.csv
Completeness: no missing value
Consistency: Between dataset1 and dataset2, some attributes have inconsistent value. Latitude has 149 inconsistence values, longitude has 152 inconsistence values, price has 2662 inconsistence values.
Uniqueness: no duplicate records
Validity: all value have same format and appropriate type in each attribute, also in the reasonable range though minimum nights and price have extreme value
Accuracy: for convenience I assume all values are accurate otherwise need to external source to verify. This assumption applied to other dataset.
Timeliness: all data are collected in 2019 and up-to-date

Dataset 2: listings.csv
Completeness: 471 missing value in zip code, 48 missing values in bathrooms, 41 missing value in bedrooms, 36 missing value in beds, no missing value for other relevant attributes.
Consistency: same as consistency in dataset 1
Uniqueness: no duplicate records
Validity: all value have same format and appropriate type in most attribute, also in the reasonable range though accommodates, bathrooms, bedrooms, beds and price have extreme value. Some value in zip code has different format.
Accuracy: same as dataset 1
Timeliness: all data are collected in 2019 and up-to-date

Dataset 3, Dataset 4, Dataset 5
Completeness: no missing value
Consistency: can't measure consistency because no same entity across datasets
Uniqueness: no duplicate record in each dataset
Validity: all value have same format and appropriate type in each attribute
Accuracy: same as dataset 1
Timeliness: Dataset 4 and 5 are all updated to 2019, dataset 3 was collected in 2018. Although dataset 3 about subway entrance was collected in 2018 and not up-to-date, the information can be used in 2019 because usually there is no big change of city subway during one year.

Other data quality assessment
Usability: all datasets include relevant information for end-use
Understandability: all relevant attribute in dataset is easy to understand. There are some attribute in WiFi and arrest data that is hard to understand, but they are irrelevant.
Flexibility: all datasets are in csv file and easy to read and load
Volume: most datasets have an appropriate size. However, dataset 2 (listing dataset) has 106 columns so I need to use subset dataset. Also, dataset 4(arrest data) has 113651 rows and I need to summarize the information.
Privacy: dataset 1 and dataset 2 include the host name that maybe the true name of person. For other datasets, they have no issue of privacy.

4. Data integration

Process:

1. Select subset with relevant variables in each datasets.
2. Merge subset of dataset 1 and dataset 2 on unique id. I use inner merge to ensure every record include all relevant attribute from two datasets. This merged dataset includes all inner attribute and also some outer attributes such as zip code, latitude and longitude.
3. Merge subway entrance data with previous merged data. The basic idea is each record in previous merged dataset has latitude and longitude, and each record in subway data has line number, latitude and longitude. Based on latitude and longitude, I count how many subway entrances within 1 kilometer(difference in latitude and longitude are all smaller than 0.007) for each house/apartment. Also, I collect the subway line information. If there is no subway entrance around the house/apartment, then the number and line are 0.
4. Merge WiFi data with previous merged data. Same idea as process 3. Based on latitude and longitude, count the total number of WiFi within 1 kilometer. Also, collect the information of WiFi type and provider. If there is no WiFi around the house/apartment, then the number and provider are 0.
5. Merge arrest data with previous merged data. Same idea as process 3. Based on latitude and longitude, count the total number of arrest within 1 kilometer. Also, collect information of arrest age group and sex. If there is no arrest around the house/apartment, then the number, age group and sex are 0.

Problems in data integration

1. In process 2, there are some inconsistent values in latitude, longitude and price from dataset 1 and dataset 2. These three variables are important to merge other data and is important for end use. For the inconsistent value in latitude and altitude, the difference is very small such as 40.66684 and 40.66604 or -73.95877 and -73.95914. I take the mean of two value as the final value. Also, inconsistent value of price, for example 199 and 215, 350 and 275, I take mean of two values. This is because usually the price fluctuate during different periods, mean value is better to represent the overall price in 2019 year.
2. In process 3 and 4, there are various value under the same attribute. For example, within the 1 km of a house/apartment, there are 20 WiFi spots and 5 different providers. I select the mode of provider as the main provider. Similar to arrest data, I select the mode of age group as the major age group of arrest. Same to major sex of arrest.
3. After merging all dataset, there are 477 missing values in zip code. WiFi dataset including zip code is useful to impute missing value. For each house/apartment has missing value in zip code, find all WiFi point which is close to this house/apartment and select the mode of zip code as imputed value. Close means the latitude and longitude of two points are smaller than 0.01.

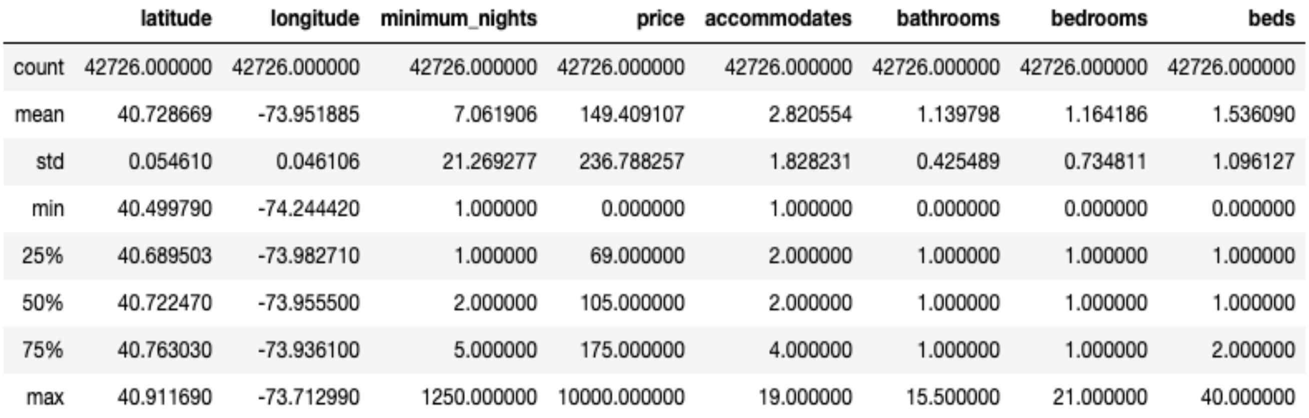
4. There are 47 missing values in bathrooms, 21 missing values in bedrooms and 28 missing values in beds. For these missing value, I just impute them with mode. This is because the distribution of these three variables are heavily skewed distributed. The majority of value are 1 for all three variables.

5.Data preparation and cleaning

1. Data reduction. Before merging, need to select subset dataset in each datasets.
2. Data cleaning. Some value is not in good format and is not easy to compare and calculate. For example, price with value $150 is transformed into numeric value 150 , location value POINT (-73.86835 40.84916) is split and transformed into longitude -73.86835 and latitude 40.84916.
3. Data cleaning. After merging all datasets, zip code value have different type. This is because the original value come from listing dataset has type integer and imputed value come from WiFi dataset has type string, also with other noise such as “11103-3233”, “NY 10005”. I remove the noise and transfer all value into string.

6.Final dataset quality and significance

The final dataset has no missing value in most variables and there are only four missing values in zip code. All values are in appropriate format and reasonable range, and they are pretty accurate because they come from original dataset except few imputed value. The imputed value and adjusted value of inconsistent data would not have big effect on end use. Table 3 and Figure 3 shows the similar statistics and distribution with original value. Also, the 22 variables include all inner attributes and outer attributes information to meet the end use of price prediction. This dataset also can be used for other purpose such as explore the surrounding criminal rate for each hotel/apartment. In conclusion, this dataset has high quality and can be used with confidence.



	latitude	longitude	minimum_nights	price	accommodates	bathrooms	bedrooms	beds
count	42726.000000	42726.000000	42726.000000	42726.000000	42726.000000	42726.000000	42726.000000	42726.000000
mean	40.728669	-73.951885	7.061906	149.409107	2.820554	1.139798	1.164186	1.536090
std	0.054610	0.046106	21.269277	236.788257	1.828231	0.425489	0.734811	1.096127
min	40.499790	-74.244420	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	40.689503	-73.982710	1.000000	69.000000	2.000000	1.000000	1.000000	1.000000
50%	40.722470	-73.955500	2.000000	105.000000	2.000000	1.000000	1.000000	1.000000
75%	40.763030	-73.936100	5.000000	175.000000	4.000000	1.000000	1.000000	2.000000
max	40.911690	-73.712990	1250.000000	10000.000000	19.000000	15.500000	21.000000	40.000000

Table 3

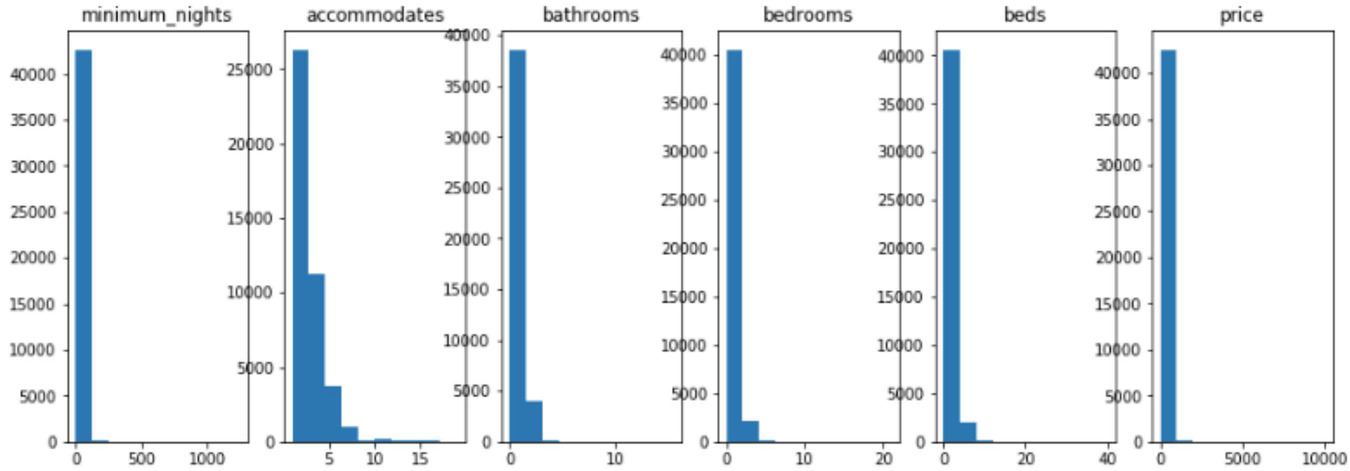


Figure 3