INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Laboratory works No.1.
Dataset Processing and Analysis

Simona Eneva
simona.eneva@ktu.edu

For instructions on how to run the .py file, please go to the end of the file.

1. Introduction

For this research and demonstration of statistical analysis, I'm using a dataset containing open data for New York City Airbnb's in 2019. The dataset contains eight numerical attributes: coordinates, price, minimum nights, etc., and 4 categorical attributes, such as host name, neighbourhood group and room type.

attribute	description
name	Name of listing on Airbnb website
host_name	Name of the host in listing on the Airbnb website
neighbourhood_group	One of the five main neighbourhood groups in NYC
room_type	One of the three room types in the Airbnb system
latitude	Latitude of property
longitude	Longitude of property
price	Price per night for a property
minimum_nights	Minimum number of nights to book a property
number_of_reviews	Number of reviews for a listing
reviews_per_month	Number of reviews per month for a listing
calculated_host_listings_count	Numer of published listings by host
availability_365	Property availability out of 365 days

Table.1. *Dataset attributes and their brief description*

To gather some useful information about this dataset, I started by performing quality analysis in order to find if there are some inconsistencies, such as missing data or surprising values in any of the attributes.

2. Performing a quality analysis of the data

2.1. QA of numerical data

attribute	total number of values	% missing values	cardinality	min	max	1st quartile	3rd quartile	average median	standard deviation
latitude	48895	0	19048	40.49979	40.91306	40.6901	40.763115	40.72307	0.054530078
longitude	48895	0	14718	-74.2444	-73.713	-73.98307	-73.936275	-73.95568	0.046156736
price	48895	0	674	0	10000	69	175	106	240.1541697
minimum_nights	48895	0	109	1	1250	1	5	3	20.51054953
number_of_reviews	48895	0	394	0	629	1	24	5	44.55058227
reviews_per_month	48895	0	938	0.01	58.5	0.19	2.02	0.72	1.680441995
Calculated_host listings_count	48895	0	47	1	327	1	2	1	32.95251885
availability_365	48895	0	366	0	365	0	227	45	131.6222889

Table.2.1. *Quality analysis of numeric data*

2.2. QA of categorical data

attribute	total number of values	percentage of missing values	cardinality	mode	frequency of mode	percentage value of mode	mode 2	frequency of mode 2	percentage value of mode 2
name	48895	0	47906	Hillside Hotel	18	0.00037	Home away from home	17	0.00035
host_name	48895	0	11453	Michael	417	0.00853	David	403	0.00824
neighbourhood_group	48895	0	5	Manhattan	21661	0.44301	Brooklyn	20104	0.41117
room_type	48895	0	3	Entire home/apt	25409	0.51966	Private room	22326	0.45661

Table.2.2. *Quality analysis of categorical data*

After the analysis, I conclude that there aren't any inconsistencies with the dataset. There are no missing values, the cardinality values are as expected and there aren't min/max values out of the expected limit.

3. Drawing histograms of attributes

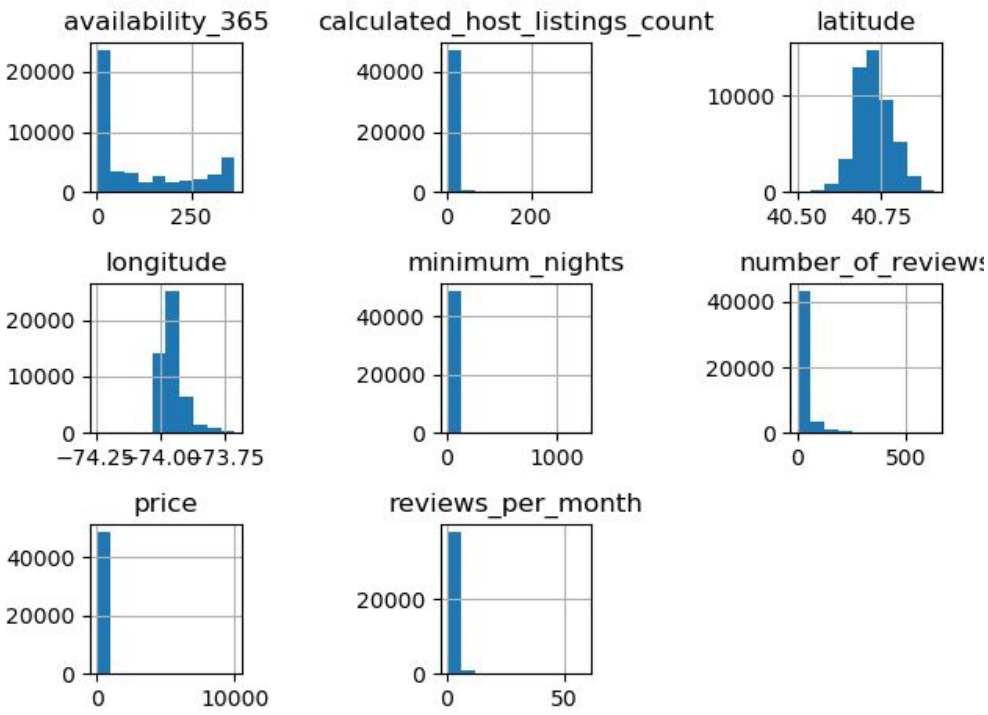

Fig.1 *Histogram of numerical attributes*

In regards to the distribution of our data, we're currently observing normal distribution in column "latitude" and "longitude". As there appear to be some extreme values, **Fig.1** isn't able to show us a detailed distribution of the other attributes, that's why I used a logarithmic scale on **Fig.2** to display the data.

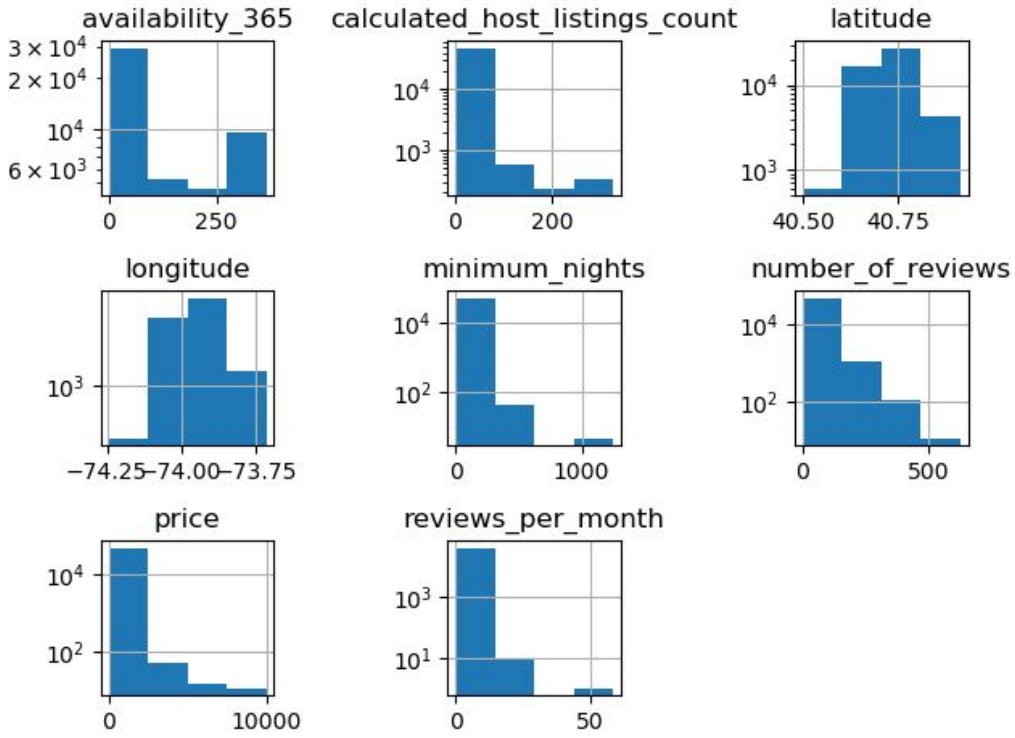

Fig.2. Histogram of numeric attributes, using a logarithmic scale.

With **Fig.2.** It is easier to see that most of our data have some kind of right-skewed distribution.

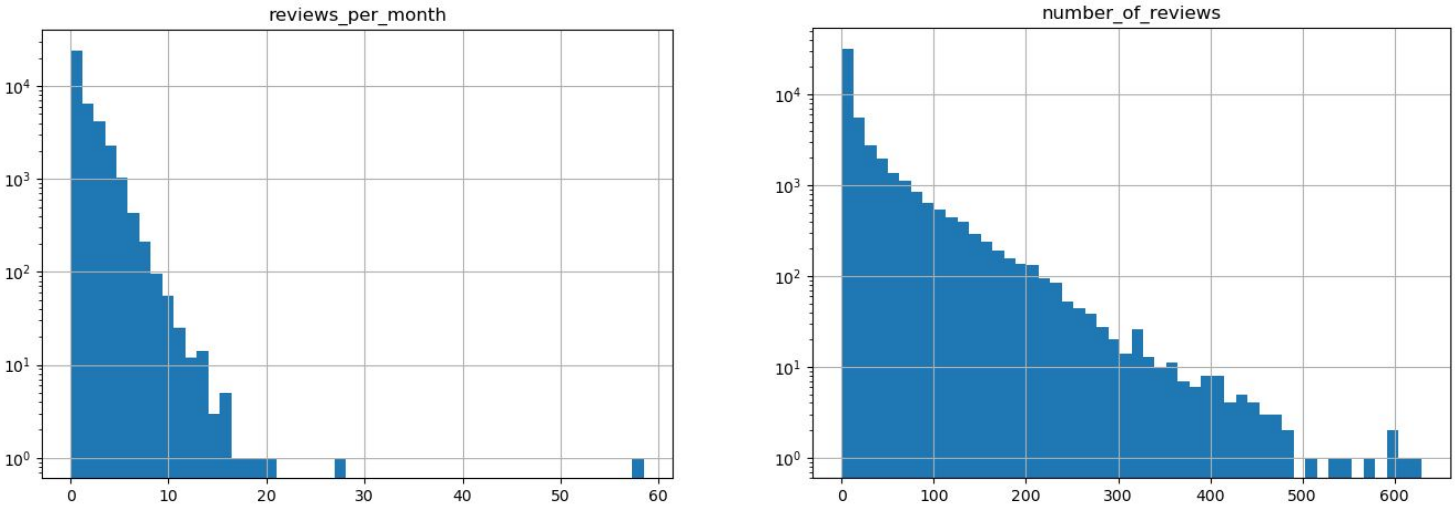

Fig.3. *Histograms of attributes "reviews per month" and "number of reviews"*

As suspected, there are attributes that show skewed distribution and a higher tendency towards smaller values. On **Fig.3.** We can observe show a right-skewed distribution with outstanding values on the far right in “reviews per month”.

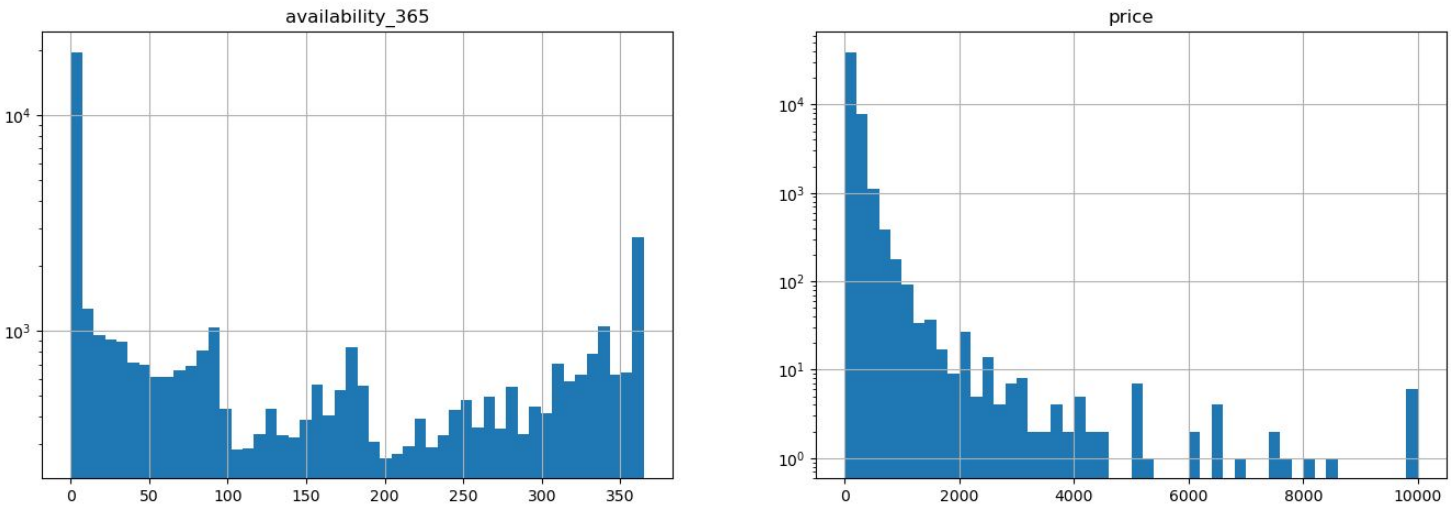

Fig.4. *Histograms of attributes “availability 365” and “price”*

On **Fig.4.** We are able to see a bimodal distribution for column “availability 365” and exponential/right-skewed distribution for “price”. The bimodal distribution shows us that there are tendencies towards two different values. In our case they are opposite.

4. Identifying data quality problems

After completing the numerical quality analysis and the categorical quality analysis, and drawing histograms from the data we received, we gather some information about our data that can help us decide whether there are some unexpected occurrences.

The most noticeable issue with our model is that a few of our columns hold outstanding values that are affecting the distribution of the rest of the data. After the quality analysis of our data, there were no issues found with missing data and cardinality. Both sets of data types (categorical and numerical) show 0.00% of missing data and cardinality values that aren’t alarming.

5. Plan for resolving data issues

1. Remove zero values from some of the columns

Some of our data hold zero values that represent inactive listings. For example, we have properties that are listed for 0/365 days a year, which means they don't offer accommodation and would give us false data. Another attribute that can use this correction is the price attribute. There in order to appear on searches with lower prices, some listings show to be $0, which can alter the average price for a neighborhood, for example.

This resolution will be applied to "availability 365", "calculated host listings count", and "price".

Even though there are some outstanding values, they still represent accurately some of our data. For that reason, they will not be corrected.

6. Establishing relationships between attributes

6.1. Numeric type attributes

Using visualization techniques such as scatter plots, SPLOM and bar charts, I am going to represent the relationships between attributes.

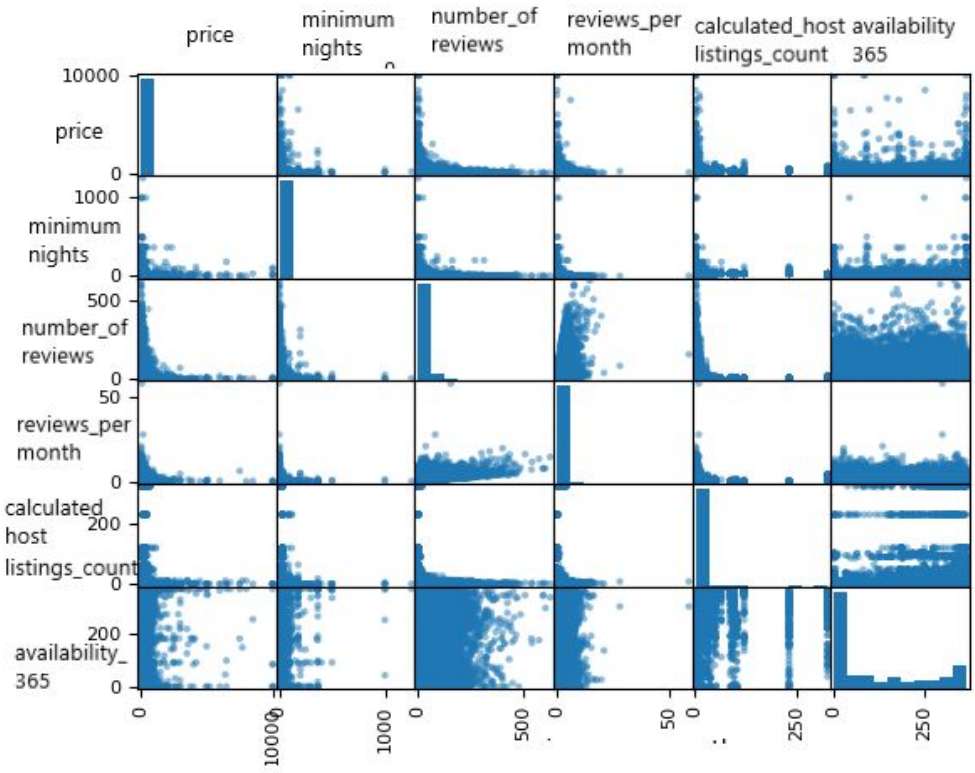

Fig.5. *SPLOM diagram of numeric attributes*

From what we see in this SPLOM diagram (**Fig.5.**), our data isn't strongly correlated. There might be some slight correlation between the total number of reviews and reviews per month but other than that, the data is non-correlated. From **Table.3.** this observation is confirmed by the correlation values. The strongest correlation we have is between number_of_reviews and reviews_per_month - 0.549867506. To review some of the attributes I considered are worth checking for correlations, we can observe **Fig.6.** where a group of scatter plot displays the correlation between 3 couples of attributes. Nevertheless, this group doesn't represent any correlation in our data.

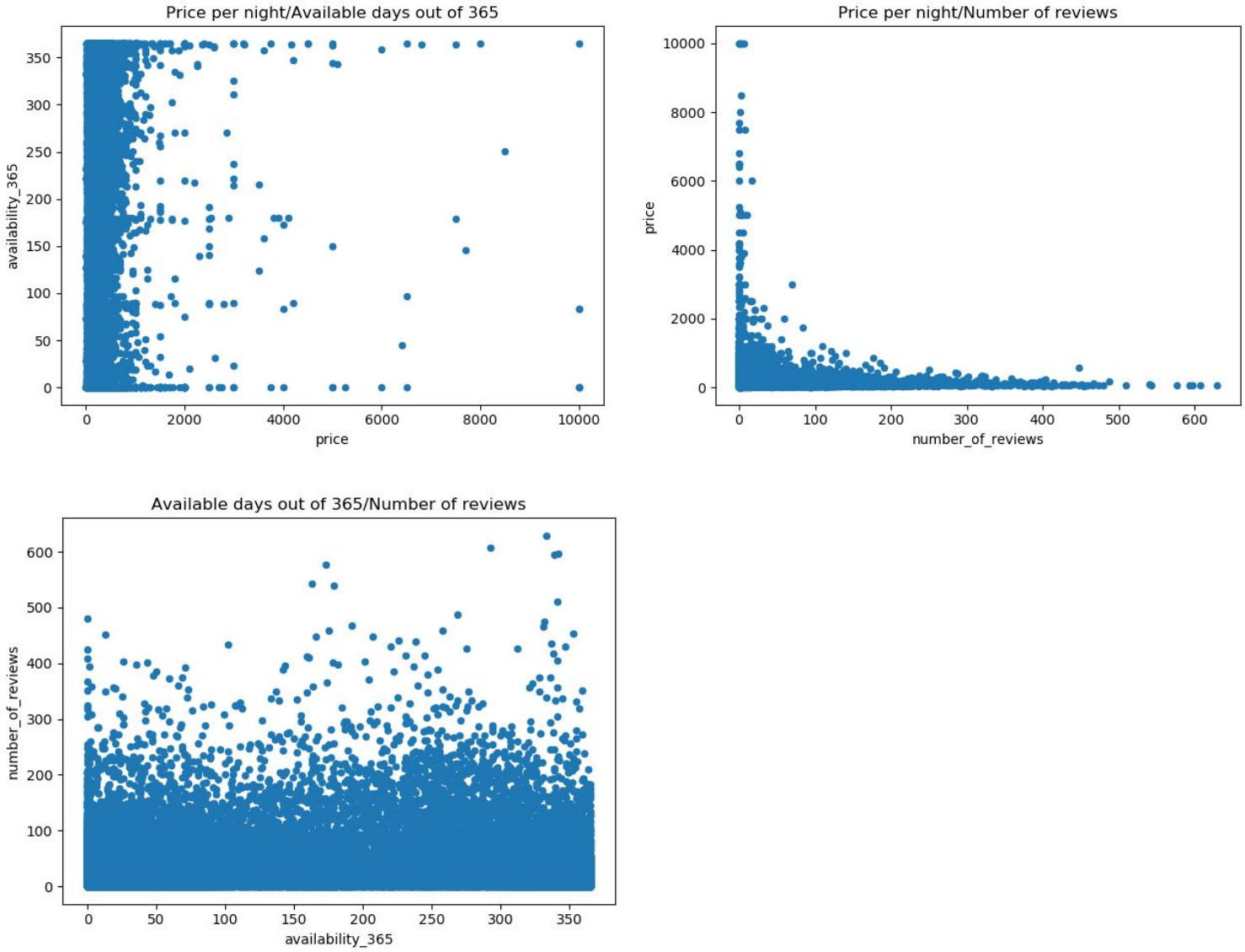

Fig.6. *Group of scatter plots*

6.2. Categorical type attributes

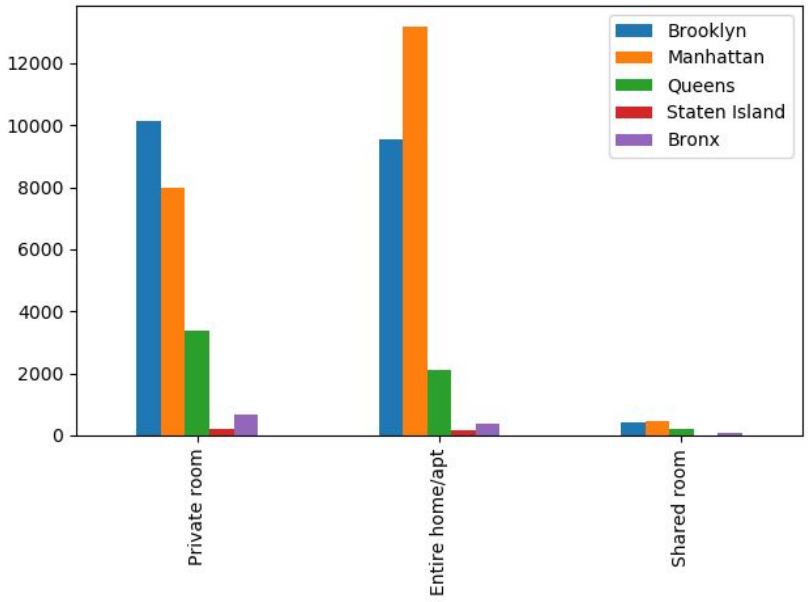

Fig.7. *Bar chart with categorical data*

In **Fig.7.** it is easier to extract some information from the given graphic. We find that most of the properties are either a private room or an entire home/apartment. Between the two, private rooms are mostly offered in Brooklyn and the least in Staten Island, while entire homes are mostly offered in Manhattan and, again, the least in Staten Island. With **Fig.8.** we can see sorted visualization of data distribution in our categorical attributes.

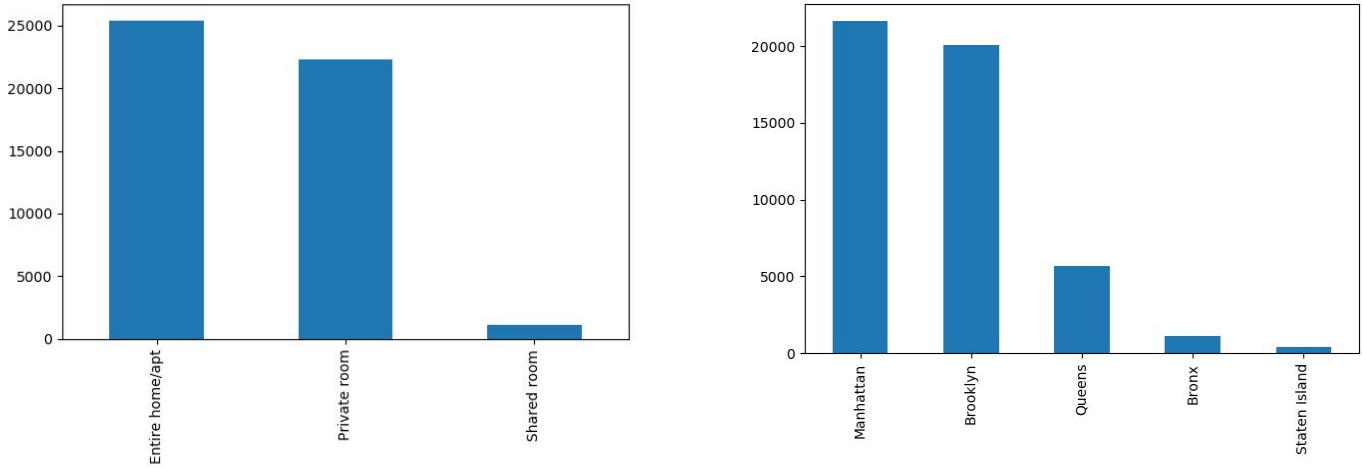

Fig.8. *Bar chart for categorical data*

6.3. Mixed type attributes

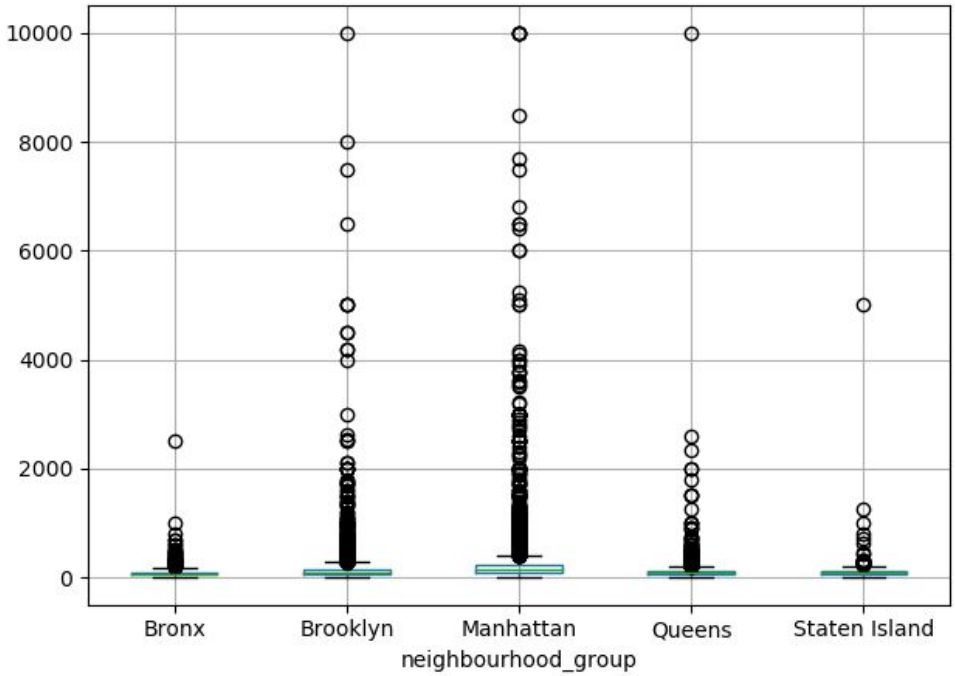

Fig.9. *Box plot of price/neighbourhood attributes*

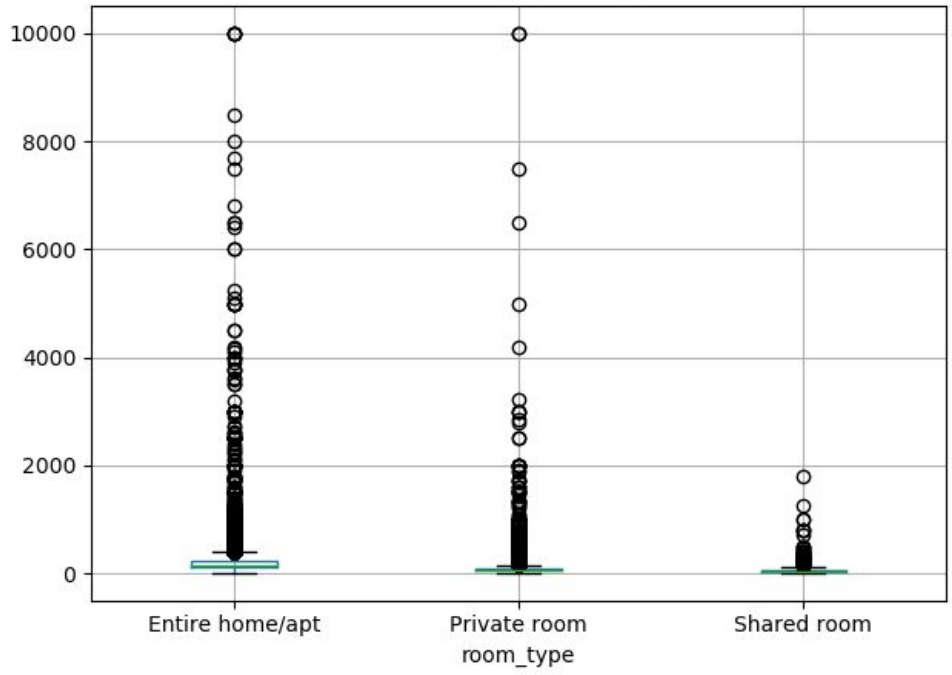

Fig.10. *Box plot of price/room_type attributes*

7. Calculate the covariance and correlation values

	latitude	longitude	price	minimum nights	number of reviews	reviews per month	calculated host listings count	availability 365
latitude	1	0.084788368	0.033938668	0.024869274	-0.015388804	-0.010141595	0.019517351	-0.010983458
longitude	0.084788368	1	-0.15001927	-0.062747114	0.059094288	0.145948027	-0.114712791	0.082730748
price	0.033938668	-0.15001927	1	0.042799334	-0.047954227	-0.030608349	0.057471688	0.081828827
minimum_nights	0.024869274	-0.062747114	0.042799334	1	-0.080116068	-0.121702201	0.127959629	0.144303063
number_of_reviews	-0.015388804	0.059094288	-0.047954227	-0.080116068	1	**0.549867506**	-0.072376061	0.172027581
reviews_per_month	-0.010141595	0.145948027	-0.030608349	-0.121702201	**0.549867506**	1	-0.009421162	0.185790961
calculated_host_listings_count	0.019517351	-0.114712791	0.057471688	0.127959629	-0.072376061	-0.009421162	1	0.225701372
availability_365	-0.010983458	0.082730748	0.081828827	0.144303063	0.172027581	0.185790961	0.225701372	1

Table.3. *Correlation values of numeric/continuous attributes*

	latitude	longitude	price	minimum nights	number of reviews	reviews per month	calculated host listings count	availability 365
latitude	0.002973529	0.000213406	0.444448093	0.027814837	-0.037384742	-0.000937165	0.035070795	-0.078832385
longitude	0.000213406	0.002130444	-1.662923498	-0.059402694	0.121516144	0.011452173	-0.174475943	0.50261041
price	0.444448093	-1.662923498	57674.02525	210.8164231	**-513.0626584**	-10.13000733	**454.812827**	2586.579899
minimum_nights	0.027814837	-0.059402694	210.8164231	420.6826422	-73.2066121	-3.555422743	86.48462115	389.5671116
number_of_reviews	-0.037384742	0.121516144	**-513.0626584**	-73.2066121	1984.75438	44.52519455	-106.2519581	1008.743866
reviews_per_month	-0.000937165	0.011452173	-10.13000733	-3.555422743	44.52519455	2.823885299	-0.416305507	40.44493935
calculated_host_listings_count	0.035070795	-0.174475943	**454.812827**	86.48462115	**-106.2519581**	-0.416305507	1085.868499	978.9313915
availability_365	-0.078832385	0.50261041	**2586.579899**	**389.5671116**	**1008.743866**	40.44493935	**978.9313915**	17324.42692

Table.4. *Covariance values of numeric type attributes*

From **Table.3.**, we can already extract the conclusion that we don't have correlation between our attributes. On the other hand we observe some significant positive covariance between a few of the attributes:

1. *Price - Calculated host listings count*
2. *Price - Availability 365*
3. *Calculated host listings count - Availability 365*

And negative covariance:

1. *Price - Number of reviews*

What this means is, that these couples of attributes vary together and is expected if one changes the other will change with it as well.

8. Instructions

The main executable file is `airbnb_nyc_2019.py` in folder *Lab01*. It is best run from an IDE. The other files contain methods used in the main file or .csv files generated after running some of the methods.