**吴伟伟 2020Z8017782004**

1. Clearly describe the term “Data science” and reasons why Data scientists are been sourced for across different fields/industries. Briefly discuss at least four areas/fields where the concept of Data science has been applied in real life.

Answer:

A. Data science is set of methods and processes that unify, clean and analysis massive data to extract useful information. It’s an inter-disciplinary field that requires mathematics, statistics, computer science, domain knowledge and information science.

B. Almost all the fields/industries now days can generate lots of raw data, which can potentially contain important information that not being clearly shown directly. Data scientists can use some techniques to transform theses raw data into some readable formats or build tools to instruct us to find unknown problems, improve producing or make prediction.

C. Areas/fields DS has been applied:

1. DS helps cameras to recognize object quickly, like human face focus, by machine learning from thousands of human face pictures.

2. DS helps meteorology department to predict weather precisely by learning from lots historical data.

3. Music App can recommend new songs to us by recommendation system trained from lots of users’ data.

4. Siri or other voice assistant can understand us better by learning from more and more users’ data.

1. By using a case-study, describe the stages in a Data science process highlighting the core importance of each stage in the process. List five top Data science programming languages and discuss their advantages and disadvantages possibly using real-life scenarios.

Answer:

1. Process stages with Airbnb NYC 2019 dataset case
   1. **Understanding** the business problem: assume we need to recommend a price when a new room is available according to historical data.
   2. Data gathering: we can get the data from **multiple sources** like Airbnb website or open dataset from Kaggle.
   3. Data processing: we need to process and **align data from different sources to make sure they are in the same format** like date format should be all ‘YYYY-MM-DD’ if we need to.
   4. Data Analysis: we can use max/avg/histogram etc. functions to show statistics data from raw data to **summarize their main characteristics**. Like maximum/average room price of a certain district.
   5. Data Cleaning: Data cleaning is the process of **removing unwanted or inaccurate records** from a table or a dataset. Some columns of raw may missing like no price information or no locations. We can fill a average price into price-missing cells and just drop those rows without locations because they are invalid for us.
   6. Data Visualization: it’s a graphical representation of the information from the data. It allows us to **observe patterns** in our dataset. Like the room location distribution in a city map.
   7. Modeling and Testing: A model is a **mathematical construct that finds patterns** when we feed raw data into it. We have to choose an appropriate algorithm like Linear Regression model with several full connected layers and start training it with a subset of the dataset that we have. Then use another subset of dataset to test the model to see the how it works.
2. Languages:

|  |  |  |
| --- | --- | --- |
| **Languages** | **Pros** | **Cons** |
| Python | Sufficient libraries, lots of users, easy, cross-platforms, free, open source, object-oriented. | Not that powerful for statistics. Depends on lots libraries to be great. |
| R | Can fully support statistic analysis functionalities, free, open source. | Not readable because of no object-oriented supporting. |
| SAS | Good at Statistical Analysis, professional support. | Business use only, not free, not suitable for individual. |
| Julia | Dynamic, concurrent threading, free, open source, compatible with C and Fortran. | Small user group, too new to be powerful, not much library support. |
| Scala | Based on JVM so compatible with Java, object-oriented | Small user group, too new to be powerful, not much library support, complicated. |

1. Data acquisition/collection represent one of the stages in of a Data science process. Discuss at least four forms of gathering dataset when carrying out a Data science project. If possible, support your description with real-life examples. Also, describe at least three different kind of data formats that a dataset can assume per time.

Answer:

1. Forms of data gathering:
   1. Private Repository: e.g. I can download purchase records data in JD if I am an authorized inner employee of JD company.
   2. Public Repository: e.g. Some research organization can provide public data for students/researchers, like CIFAR dataset: <http://www.cs.toronto.edu/~kriz/cifar.html>.
   3. Database: e.g. I can query student’s information records from the database in school information system if I am an authorized inner employee of the school.
   4. Scrap data from webpages: e.g. If I want to analysis the comments of a certain good in JD, but I can’t get the data from their database or inner file. I can use crawler to gather comment data.
2. Different data formats: e.g. For a good information in JD.
   1. JSON:

{

“id”: 123,

“name”: “XXX shampoo”,

“category”: “shampoo”,

“price”: 50.99

}

* 1. XML:

<good id=123 name=”XXX shampoo” category=”shampoo” price=50.99 />

* 1. CSV: good.csv, comma divided

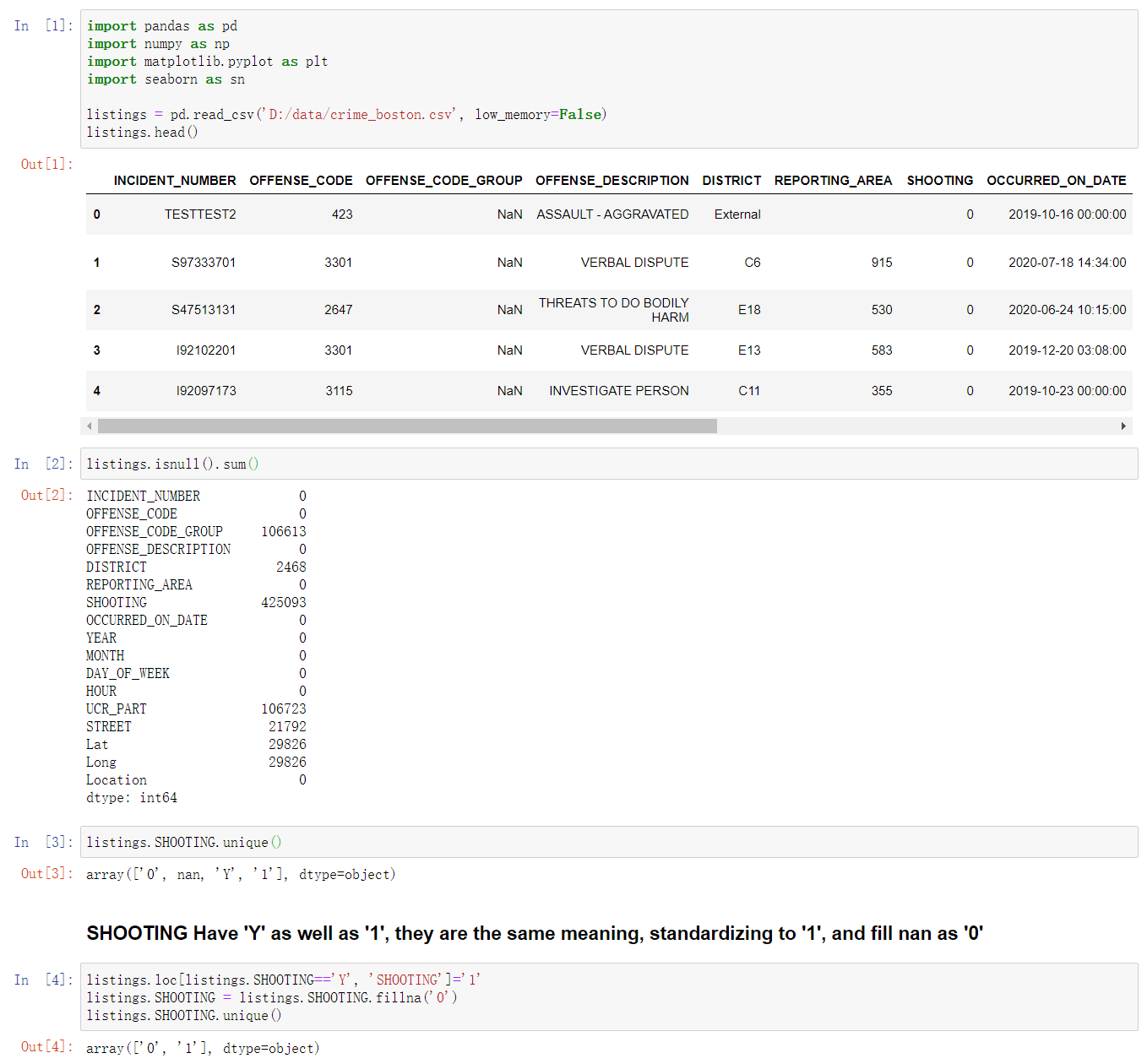
Id, name, category, price

123, XXX shampoo, shampoo, 50.99

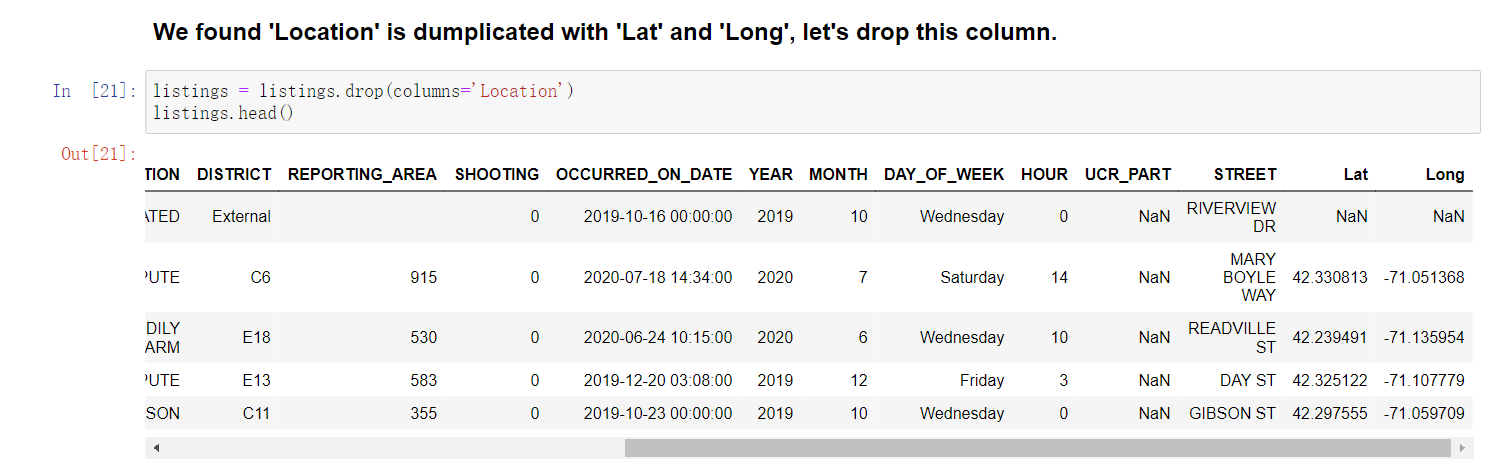
1. Data cleaning/cleansing is one of the stages of Data science process in which issues such as missing values, unit conversion, misspellings, duplicate roles, inconsistent format, and unspecified units are resolved to put the dataset in a good shape for the subsequent process. With this understanding and what you have learned in class, you are to work with the “Crime Incidence Report” dataset which I handed over to you several weeks ago, performing four different data cleaning operations/tasks. Afterwards, you are expected to clearly report your results (screenshots of the outcome of the cleaning tasks and codes used). The implementation can be done using tools/programming languages that we used in class (Python, Jupiter Notebook) or other tools of your choice.

Answer:

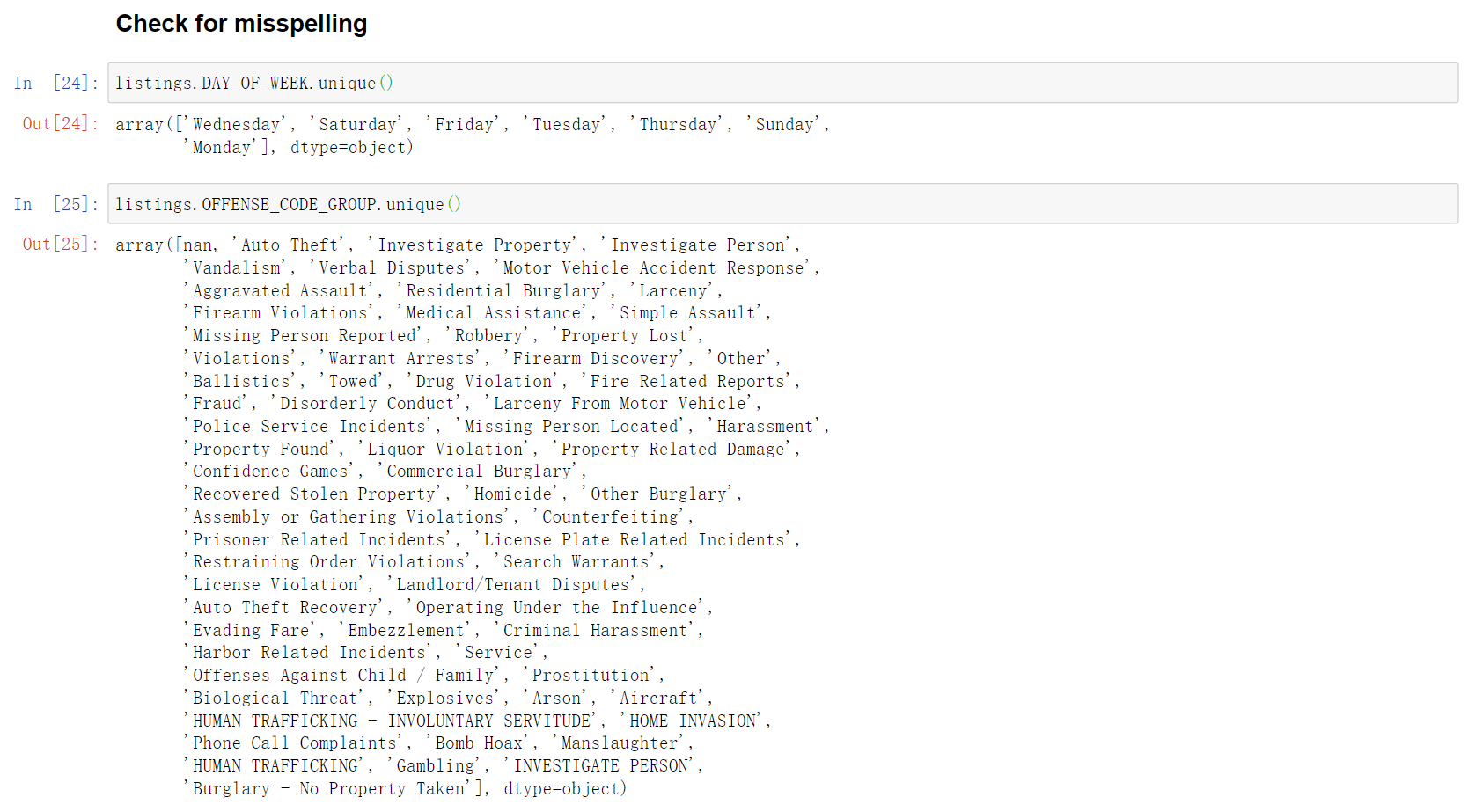
**1. Inconsistent formats standardization** and **2. Fill missing values** for SHOOTING column of Boston crime data set.



**3. Drop duplicated columns:**



**4. Check for misspelling, no misspelling found, nothing to do:**

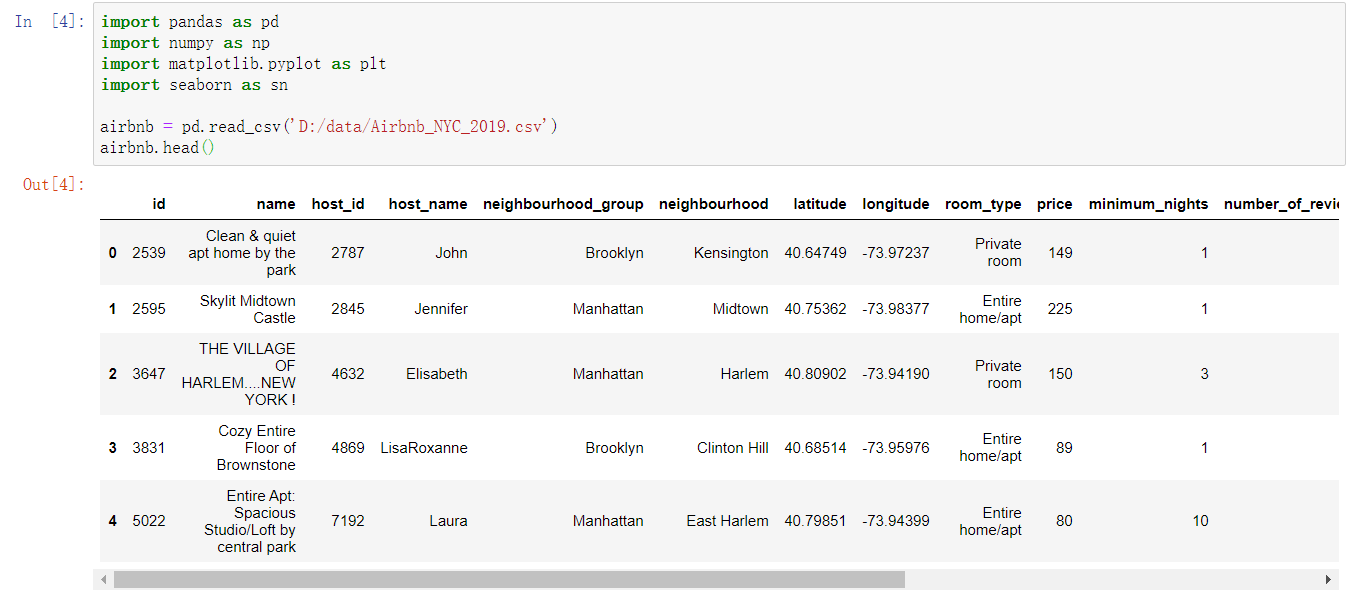


1. Briefly discuss why Data visualization is an important stage of Data science process. Using the “New York City Airbnb Open Data” which I shared with you a few weeks ago, perform at least four different kind of Data visualization tasks by exploring the ***Matplotlib*** and ***Seaborn*** libraries and their functions. Also, clearly report your results (screenshots of the outcome of the visualization tasks and codes used).

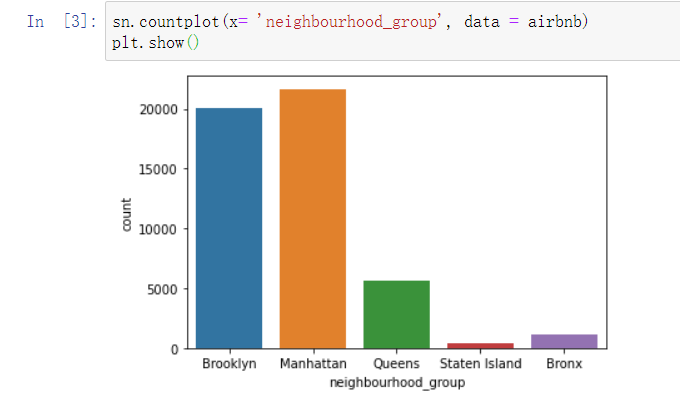
Answer:

1. Data visualization is an important stage in DS process because: raw data are usually simple records one by one and the information behind them is not clear. Visualization uses graphical/pictorial representation to present information/data clearer. And it allows us to observe trends/patterns in our data set. Just looking at a large set of numbers in tables may not be that helpful in aiding understanding of inherent trends.
2. Data visualization for Airbnb data set:

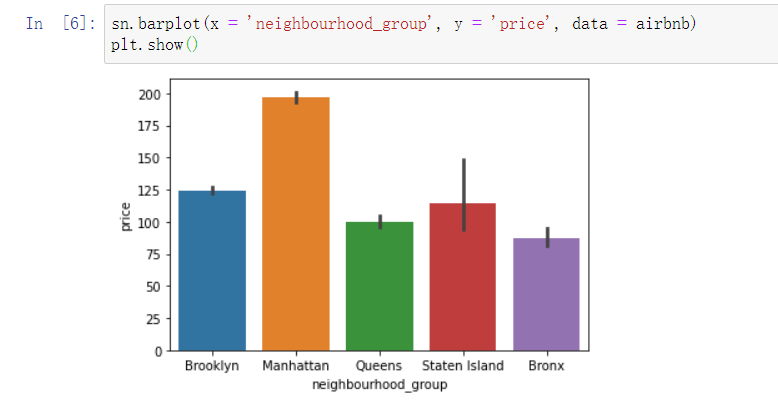
Import libraries and load data, check first several rows of the data set:



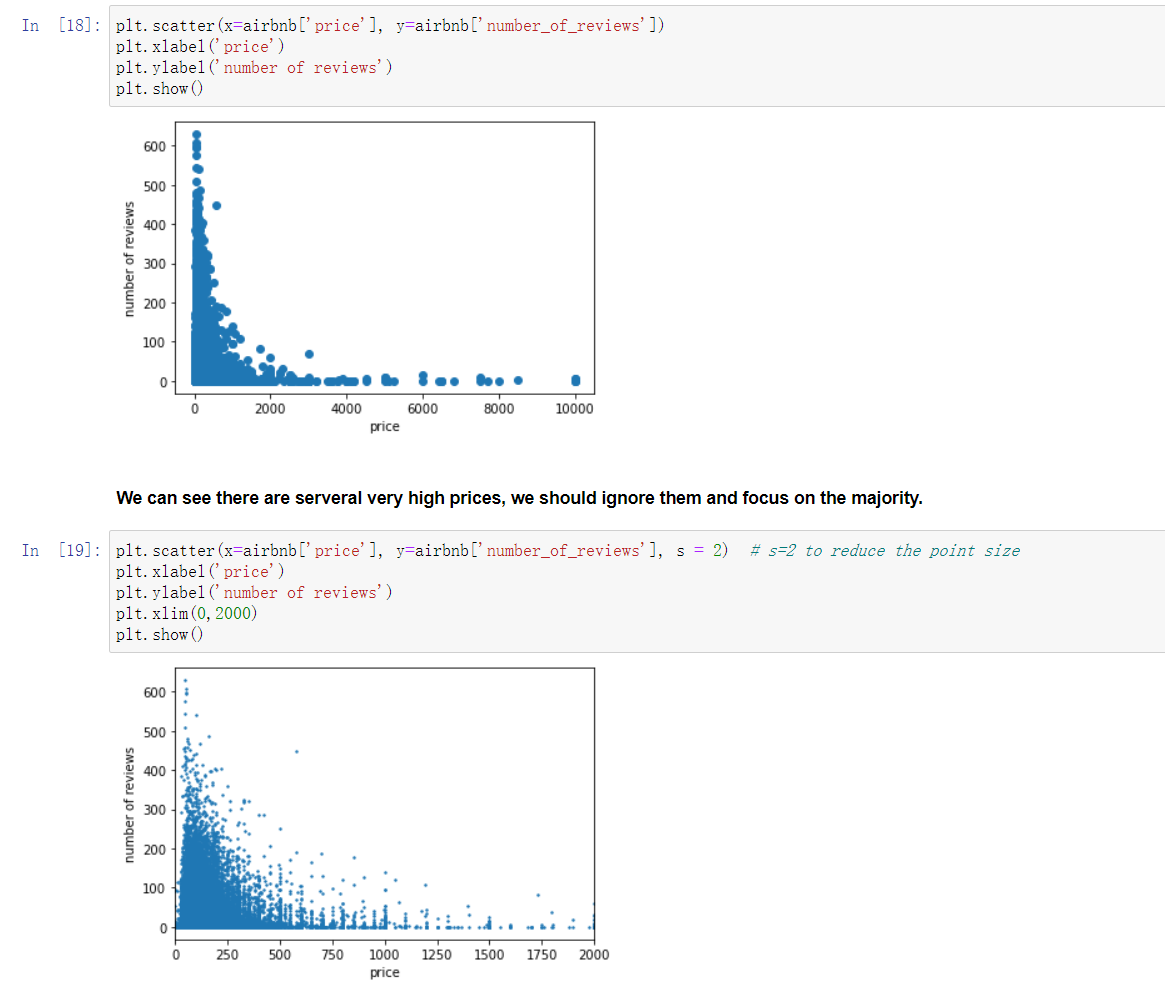
1. Check the counts of all neighborhood groups, the most of the rooms are from Manhattan.



1. Check the price statistics of each neighborhood group, we can see that Manhattan has the highest prices in average.



1. Check the distribution of number of reviews as with price.



1. Check the prices distribution, we can see that most of the rooms have lower prices(0-150).

