

**ADVANCES IN DATA SCIENCE/ARCHITECTURE**

**FINAL PROJECT**

**REPORT**

**Submitted By:**

**MEGHA SINGH**

**SNEHA MALSHETTI**

Table of Contents

1. Problem statement 3
2. Overview3

1. Objective Flow of the Project5
2. Data Ingestion ,Exploratory Data Analysis and Data Wrangling11

Steps for Data cleansing and EDA15

1. Logging and Threads 11
2. Feature Selection and Modeling 11
3. Deploying, Dockerizing and Pipelining 11
4. Tableau in and IPython Notebook in web application11
5. Citation4

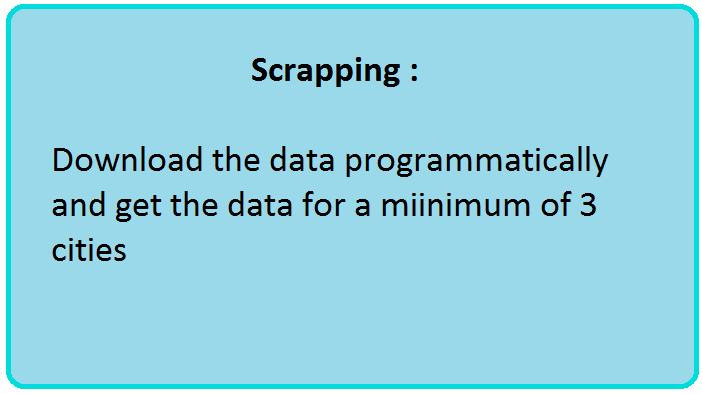
**Problem Stetment :**

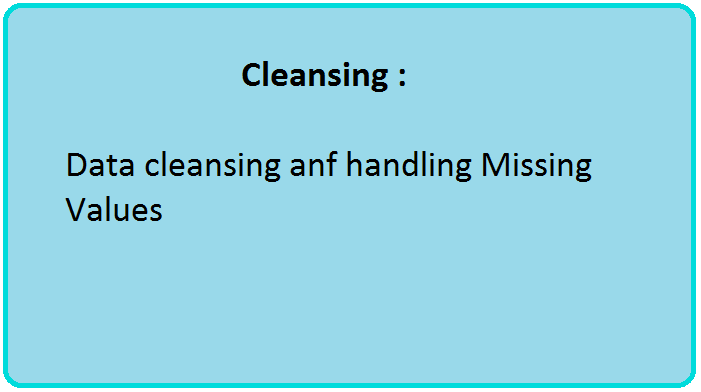
**“** Airbnb wants it analysis and prediction for prices available to their customers to make right decisions while choosing venues for providing good customer experience.

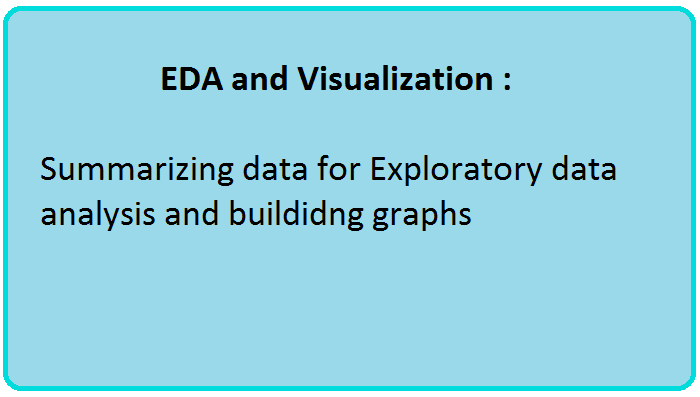
Airbnb also to extend its services to owners or hosts by providing their feedback on how they can improve business.

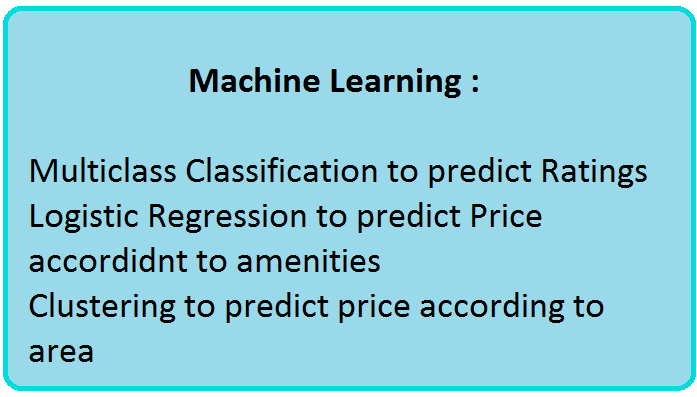
Also, for investors they decide to help them with searching for good investment opportunities based on location. Based on best model evaluation outcome they wish to deploy the services on cloud. **“**

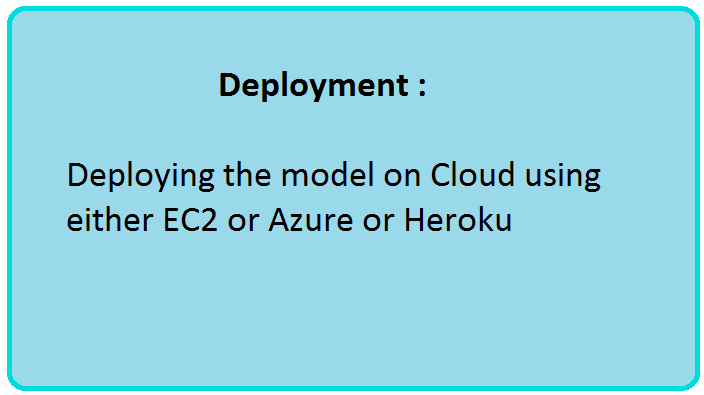
**Overview:**











Objective Flow

**Data Ingestion:**

**Dataset :**

**The link for the dataset of Airbnb is as below :**

<http://insideairbnb.com/get-the-data.html>

The Airbnb daaset consists of 3 files namely,

“listings.csv.gz”

“reviews.csv.gz”

“calendar.csv.gz”

The Listings dataset:

The listings dataset consists of all current the listings in the Airbnb database for a particular city.

The Reviews dataset:

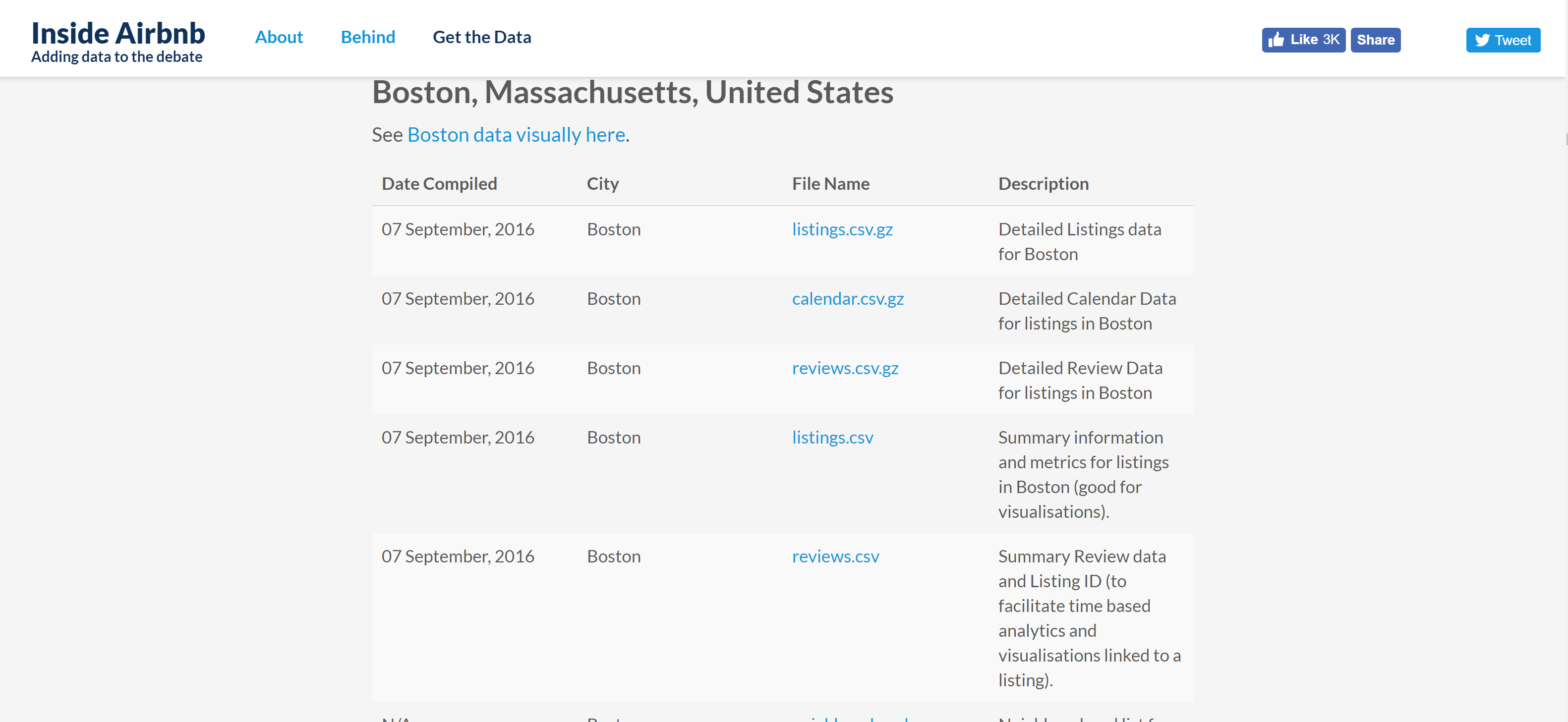
Consists of all the reviews for the respective Listings

The Calendar dataset:

Consists of all the listings availability for each day of the year

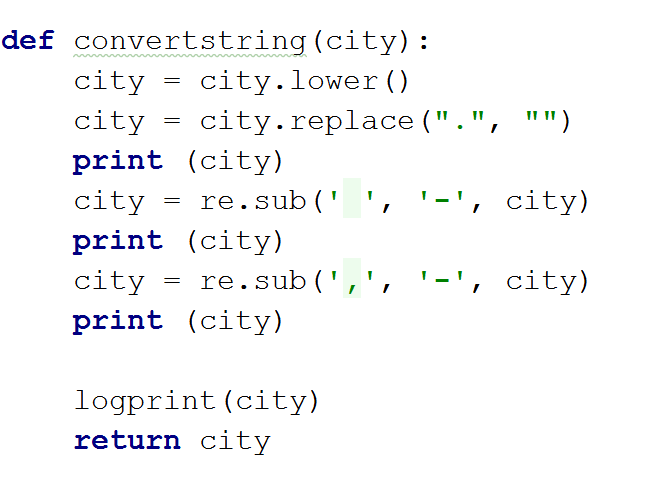
These files were present for each city and there were archived files for the same as well which were scrapped at different points in the year.

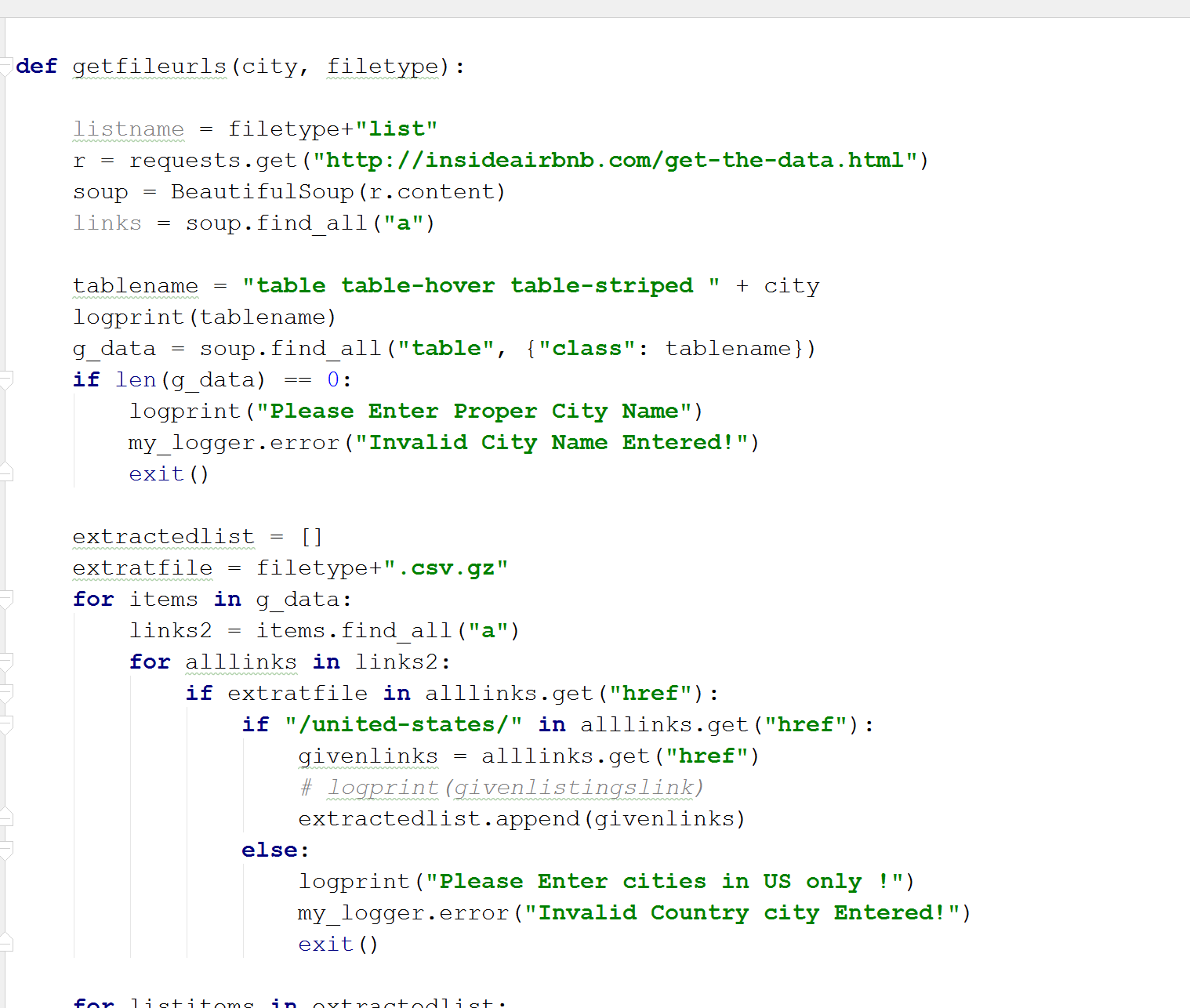
We have only picked the latest files for analysis.



**Web Scraping :**

We have used Beautiful soup library to scrape all the required city urls and then downloaded the data for the same.





We pull the required city name from a config file and then use that city name to get the files in the above code.

The above code only fetchess the list of urls . The code below downloads the csv and gives us a merged file



Each file is converted to csv and sent to S3 after merging and The execution of all the functions is as below :



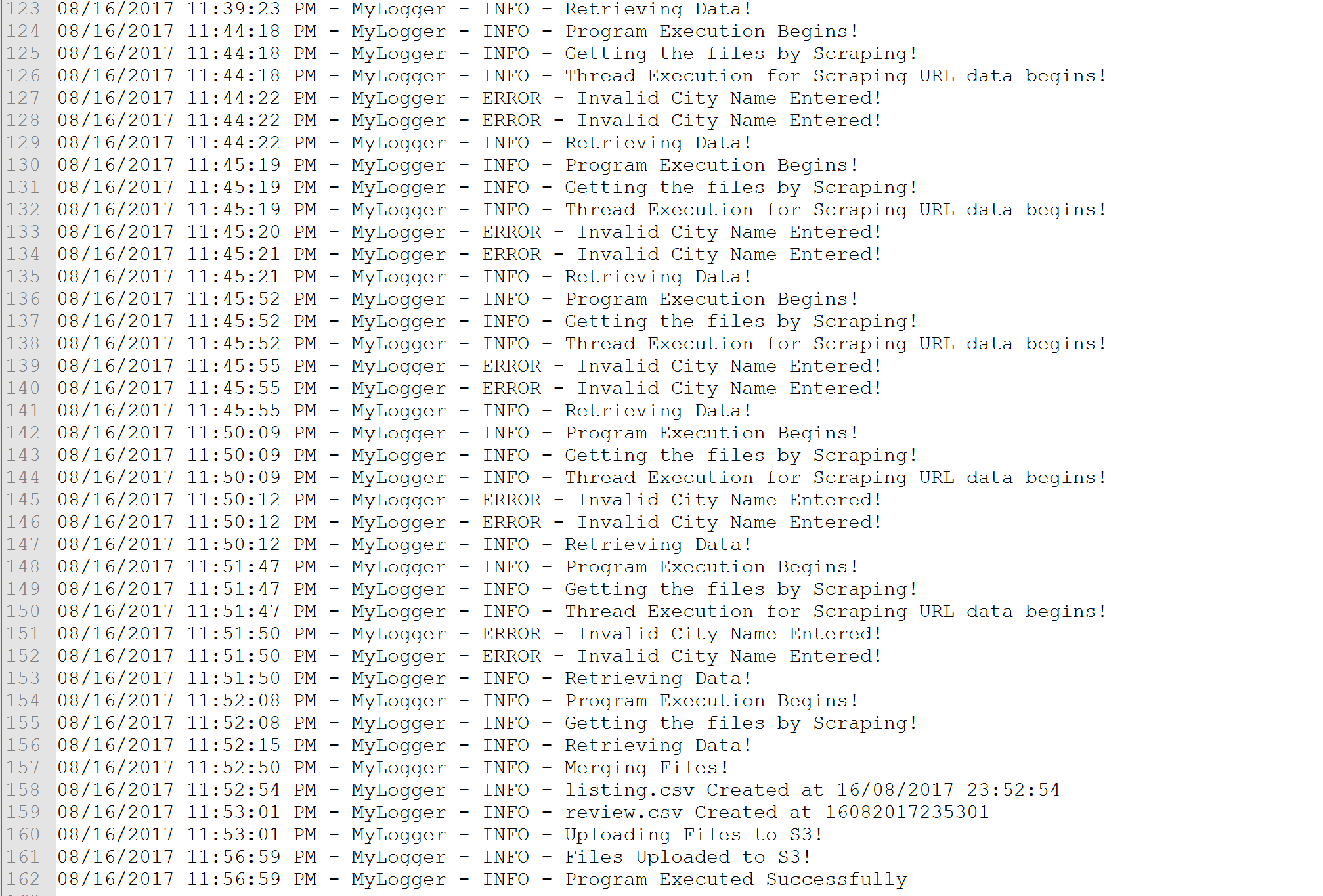
We have also put in the logging functionality in place and the code for it is s below:



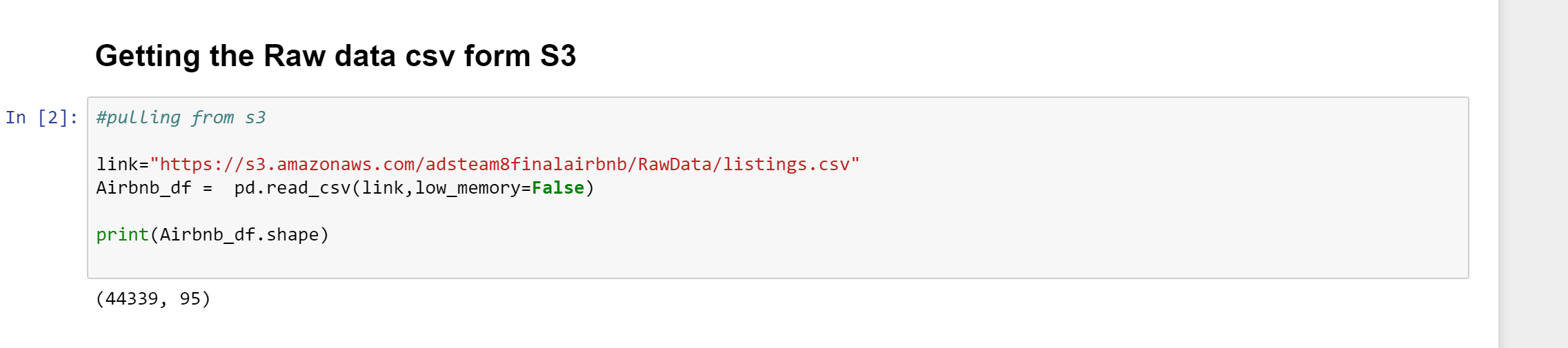
A log file gets generated everyday to track the entire day’s activities.

The log file logs the activities to the greatest detail in time to the second .

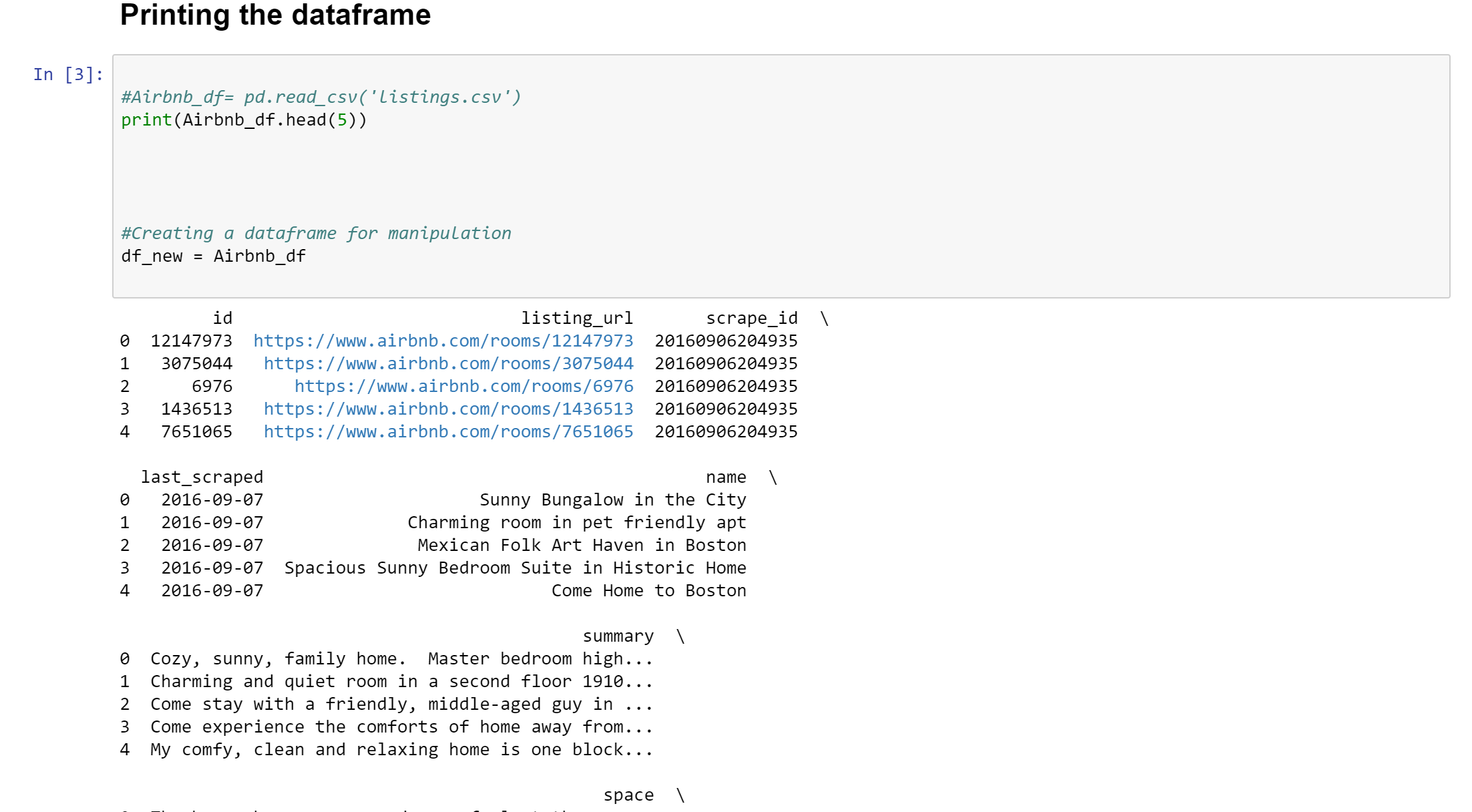
The enteries in the log file look as shown below.

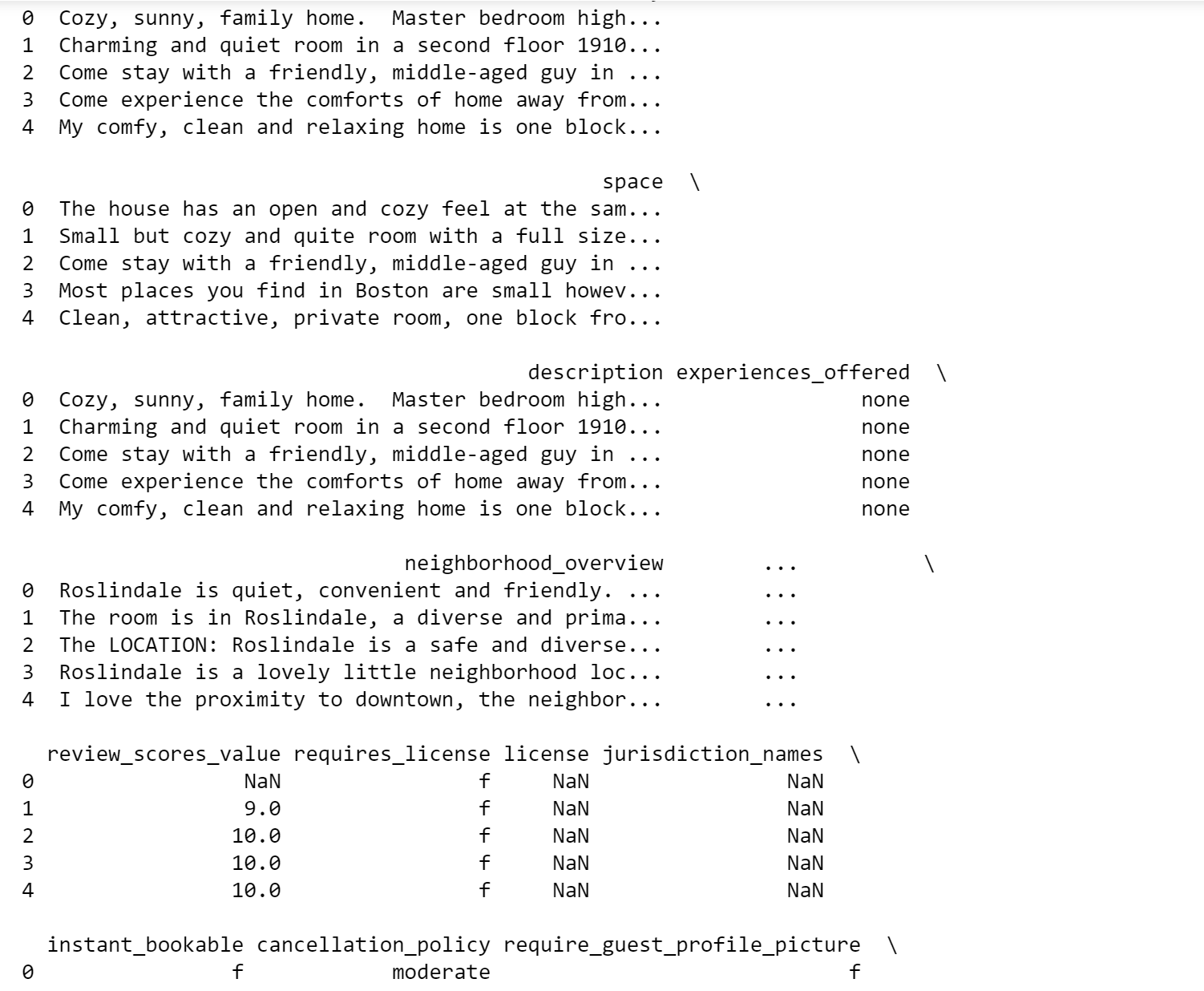


Data Wrangling :

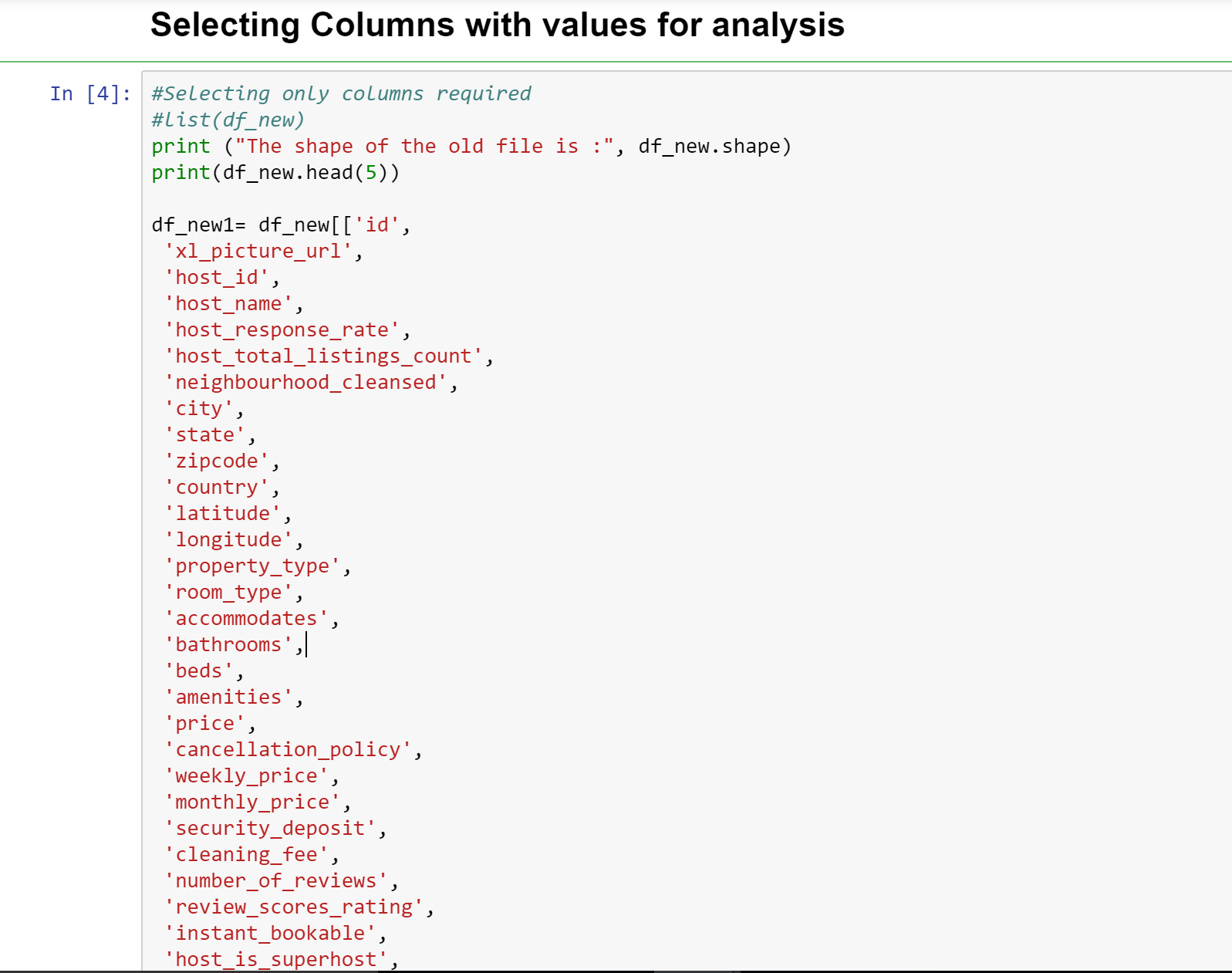


The dataframe looks as shown below with 95 columns

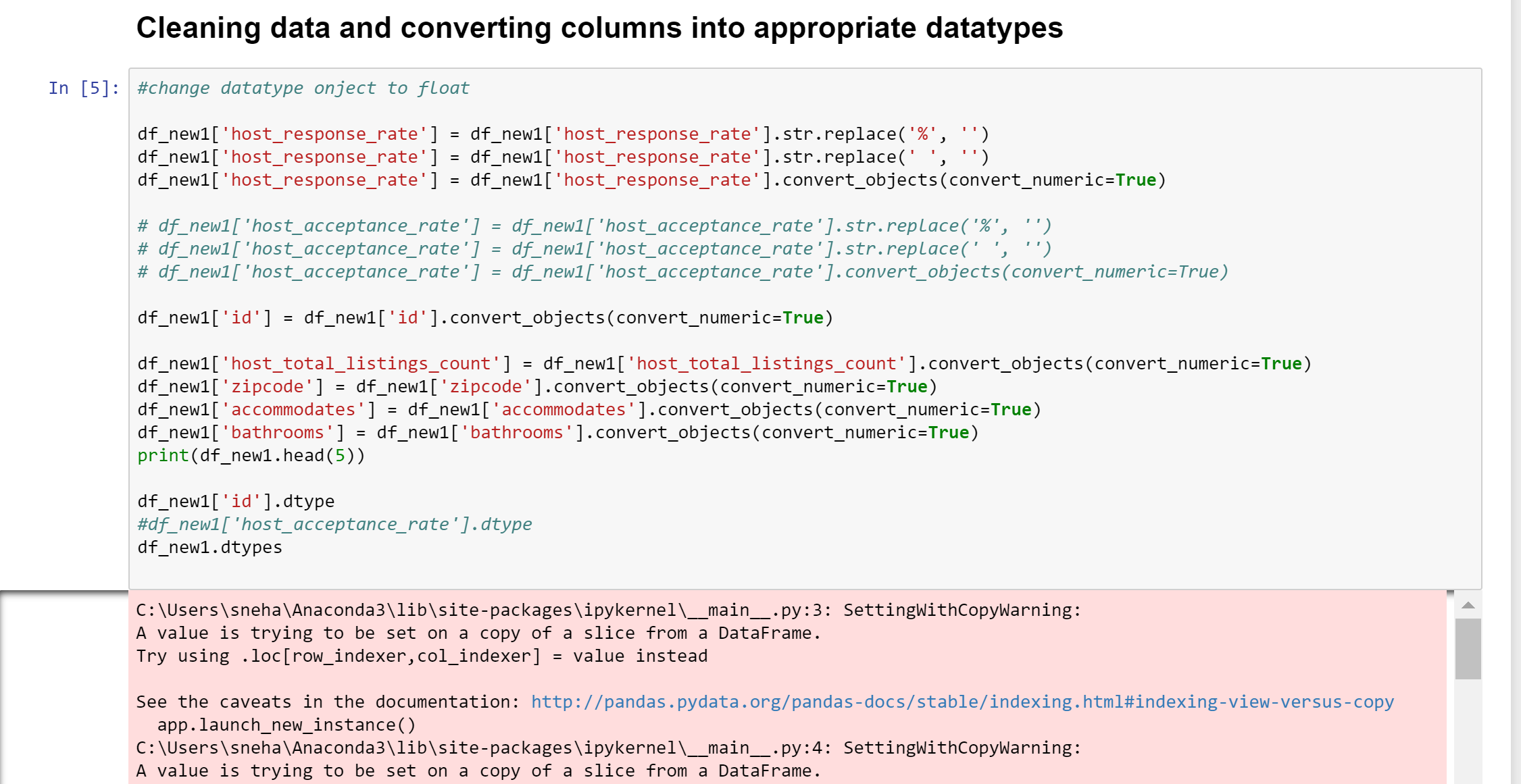




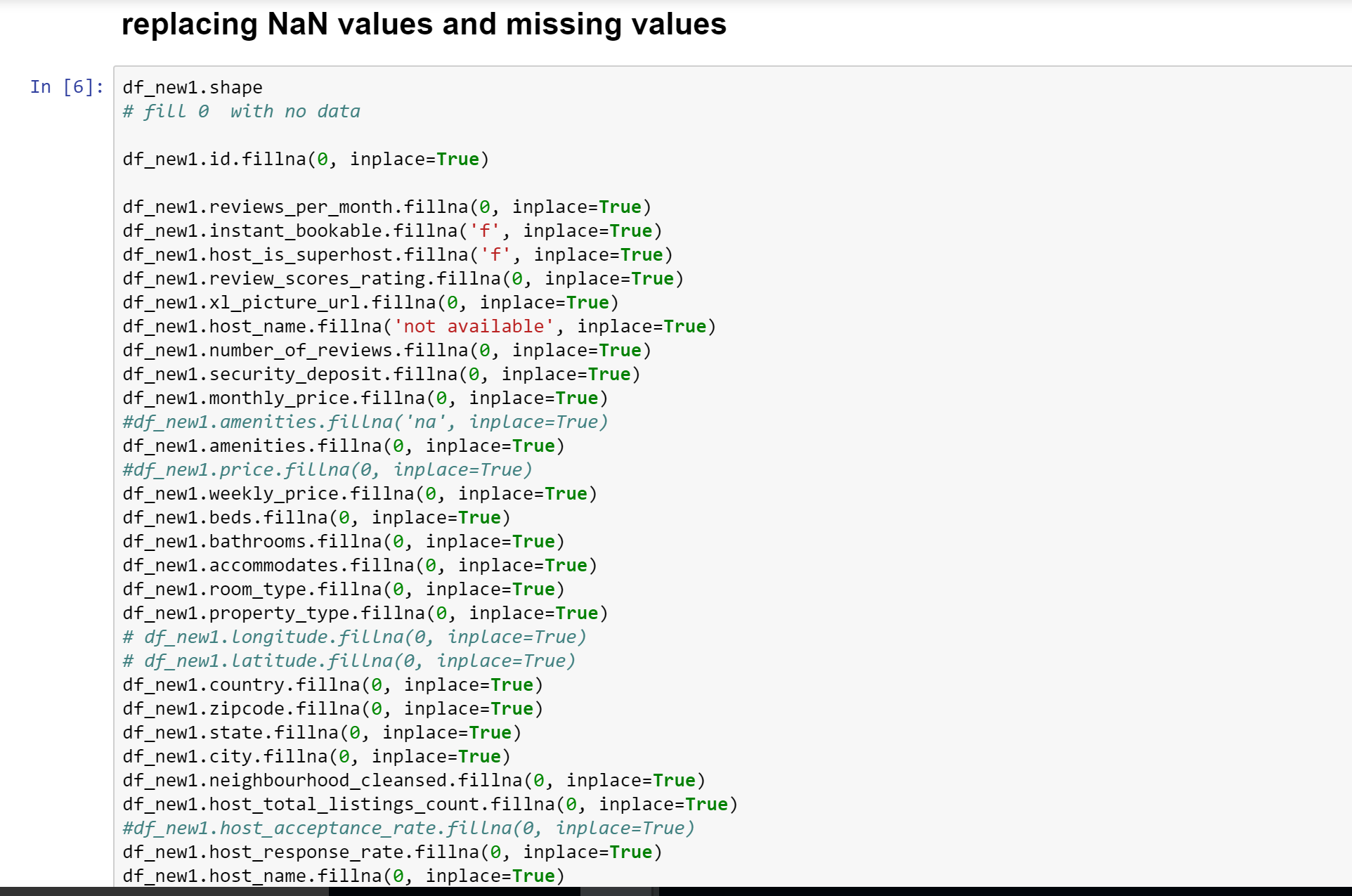
Now we select only the required columns for analysis and delete the rest of them :



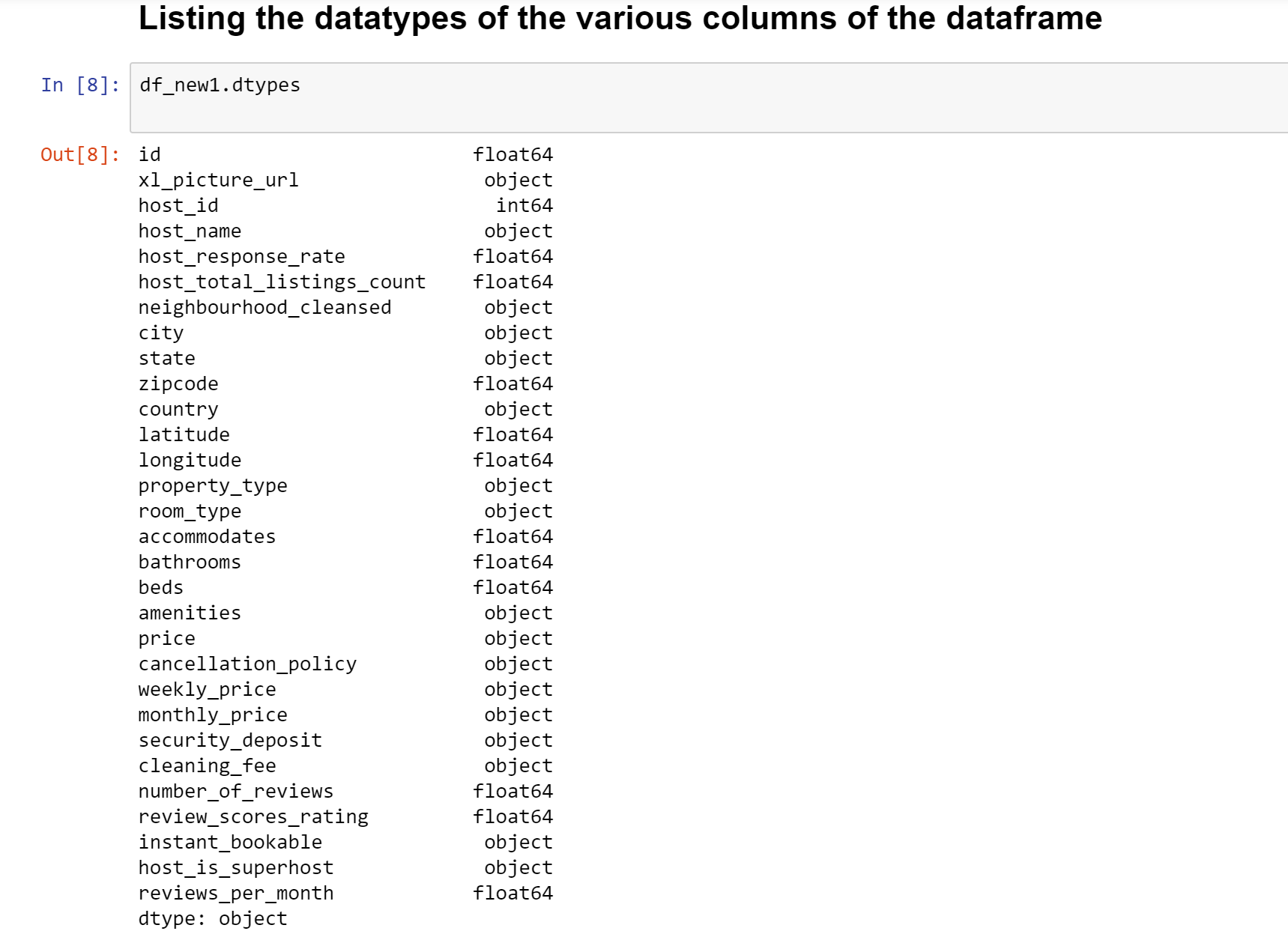
The selected columns are cleaned and converted to appropriate datatypes to perfor further preprocessing on them :



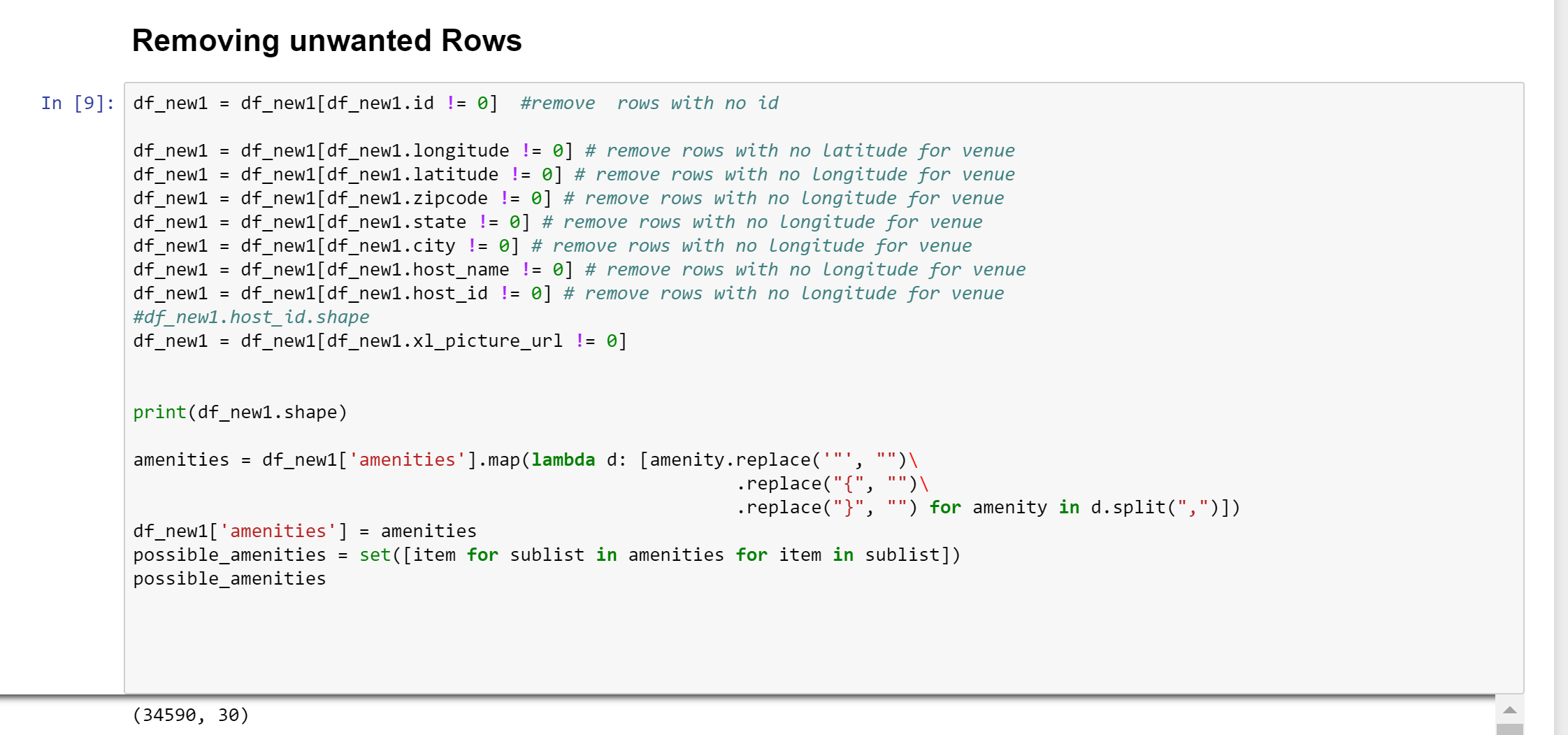
Replacing all NaN values we get ,



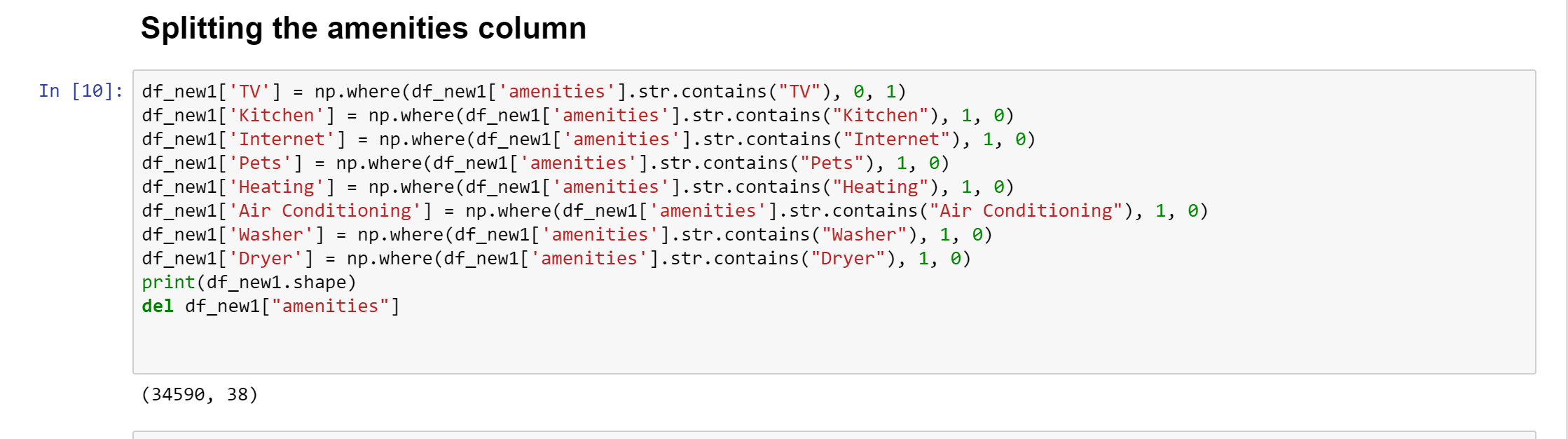
The data set is reduced to 30 columns for further cleaning



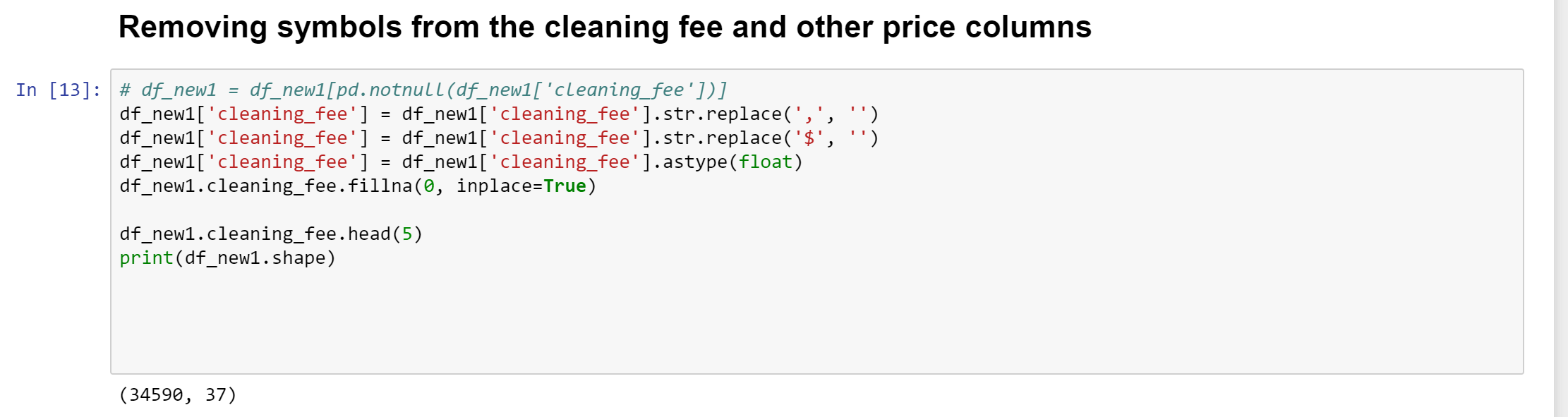
Now we remove the unwanted rows or the rows with NaN values :



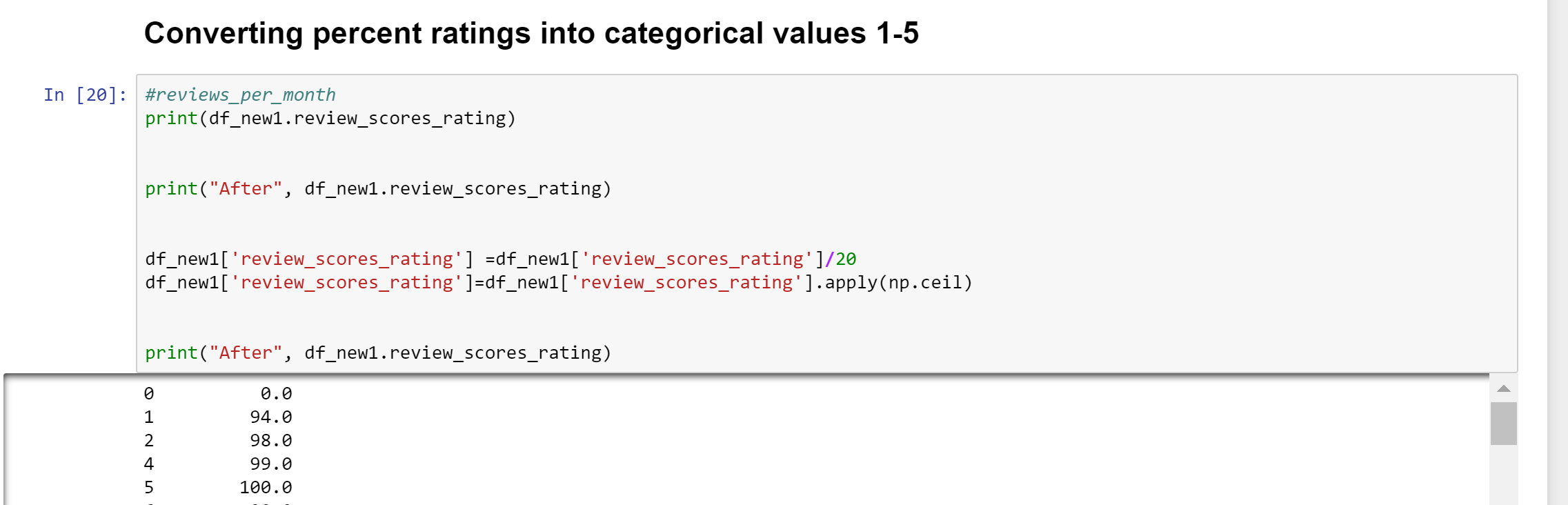
The amenities column in the dataset consists of various values and needs to be split into different columns with 0 and 1 as their output.



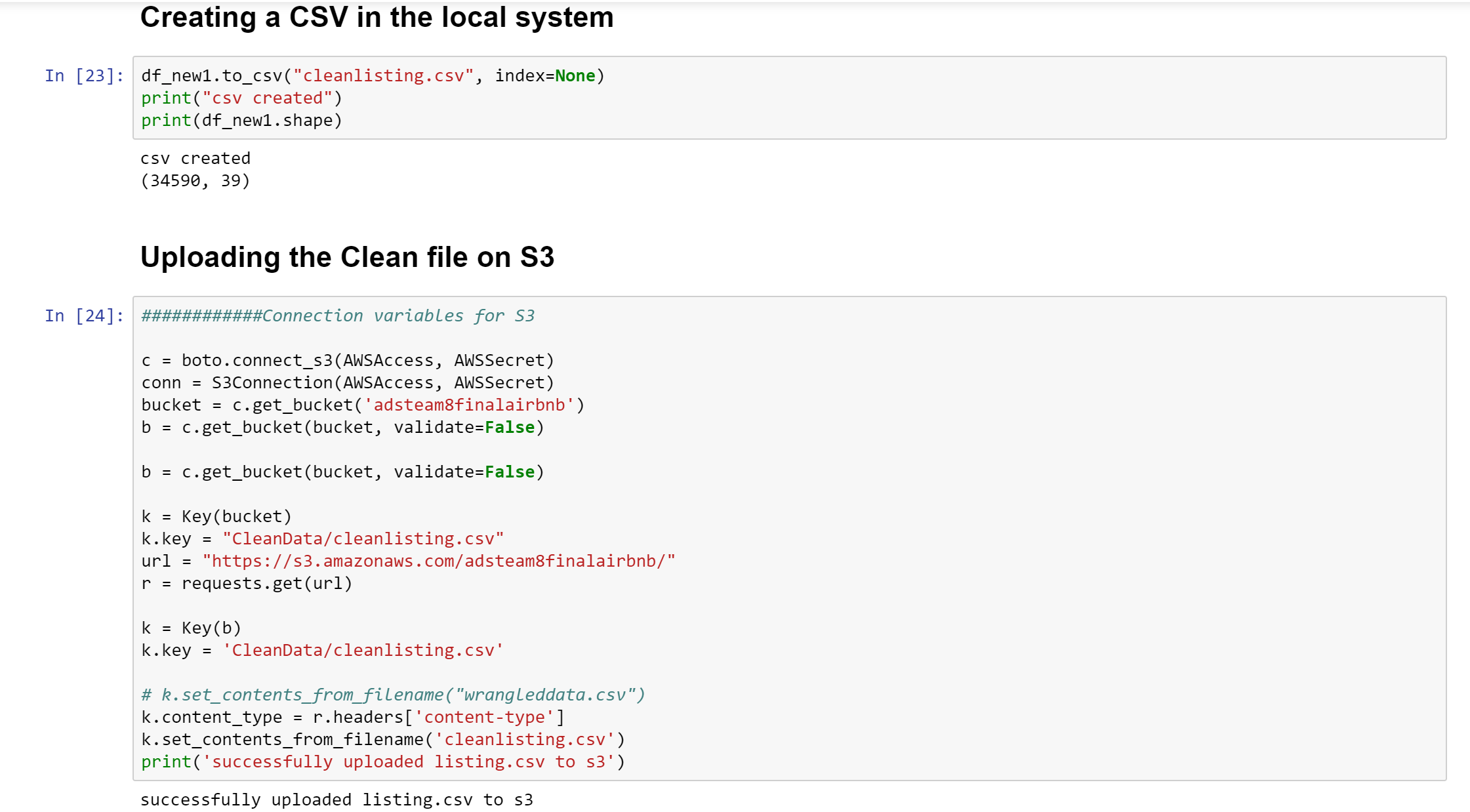
Some columns in the dataset have special chracters like “$” , which have to be removed



Converting Score Reviews into Ratings, we get values from 0-5



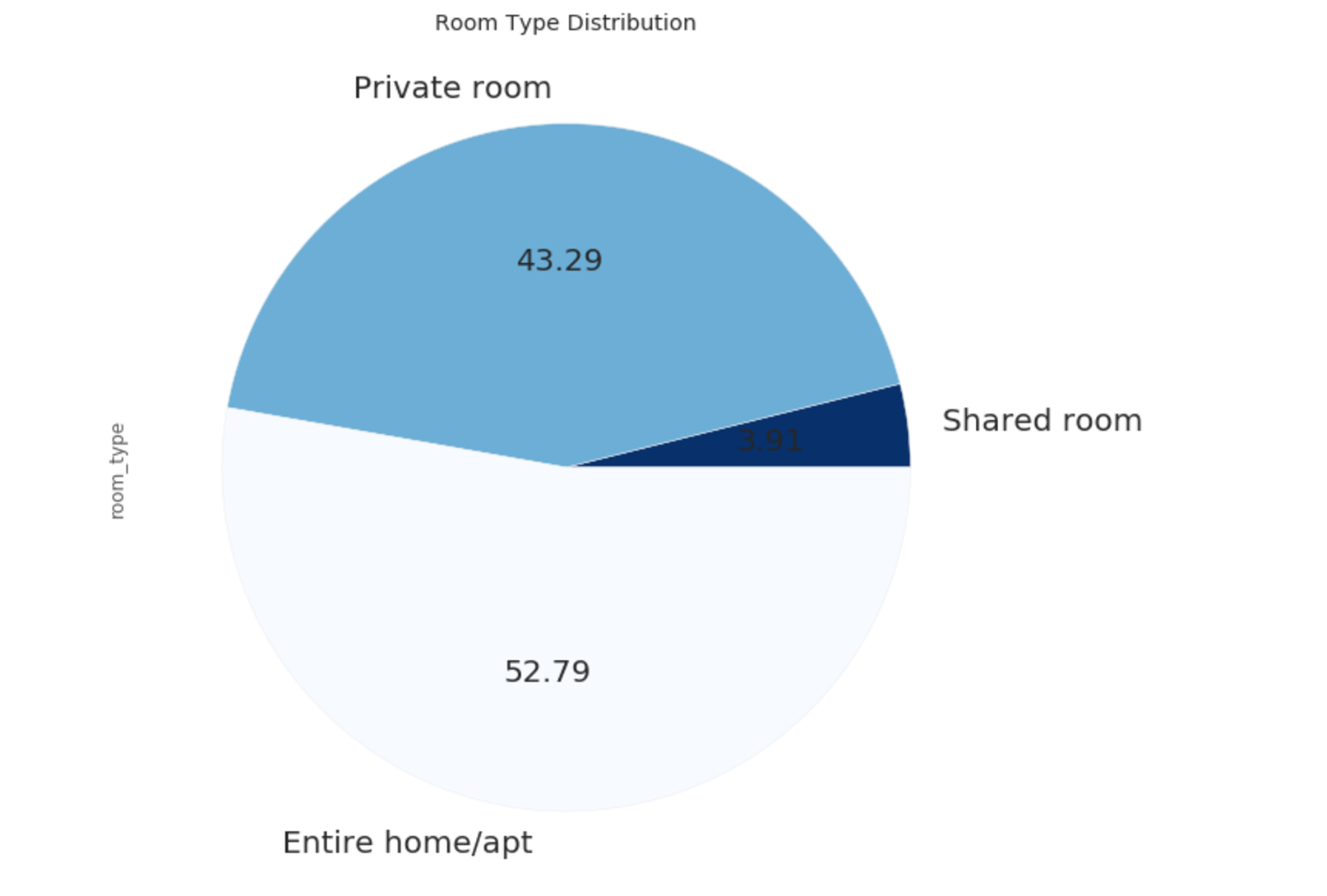
The dataframe is converted to csv and sent to s3 to a folder named “CleanData”



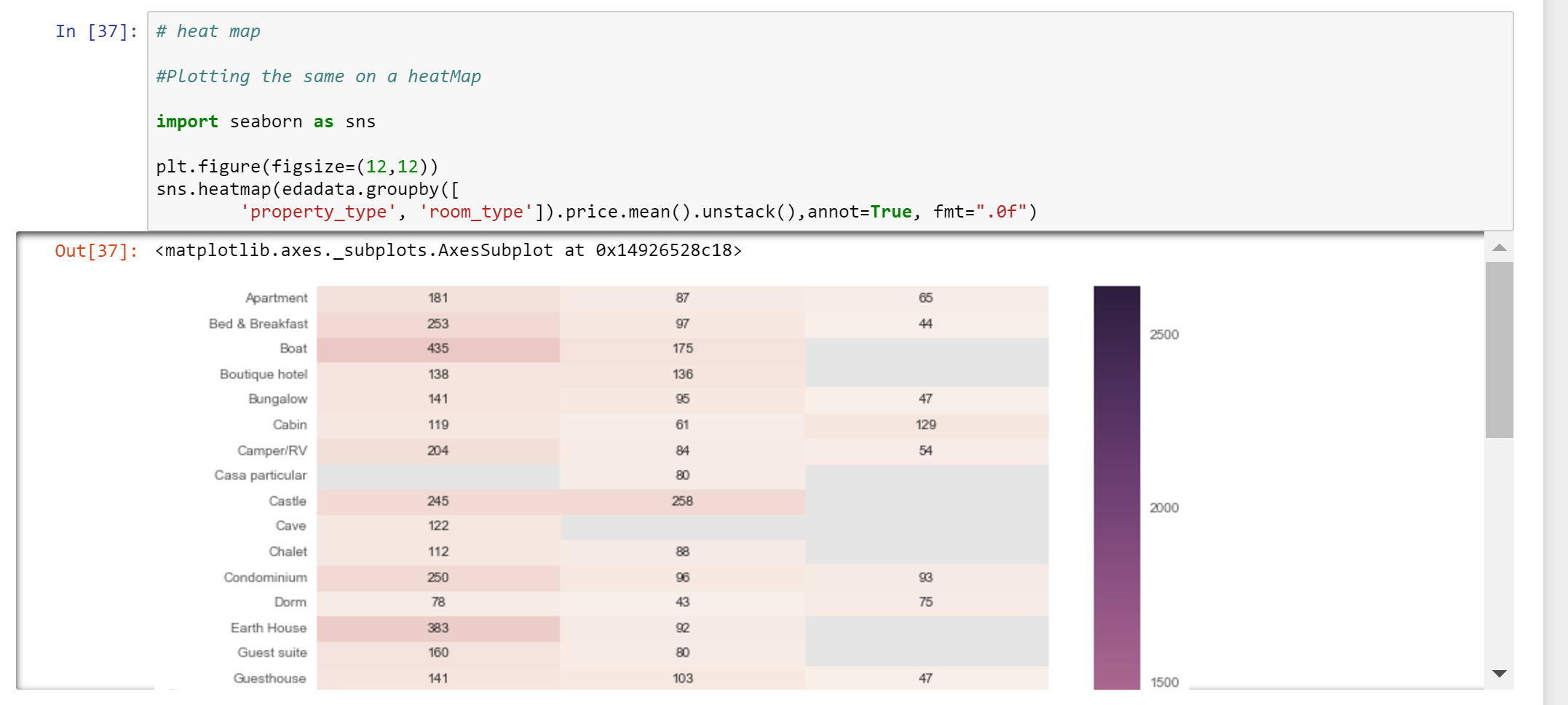
EDA

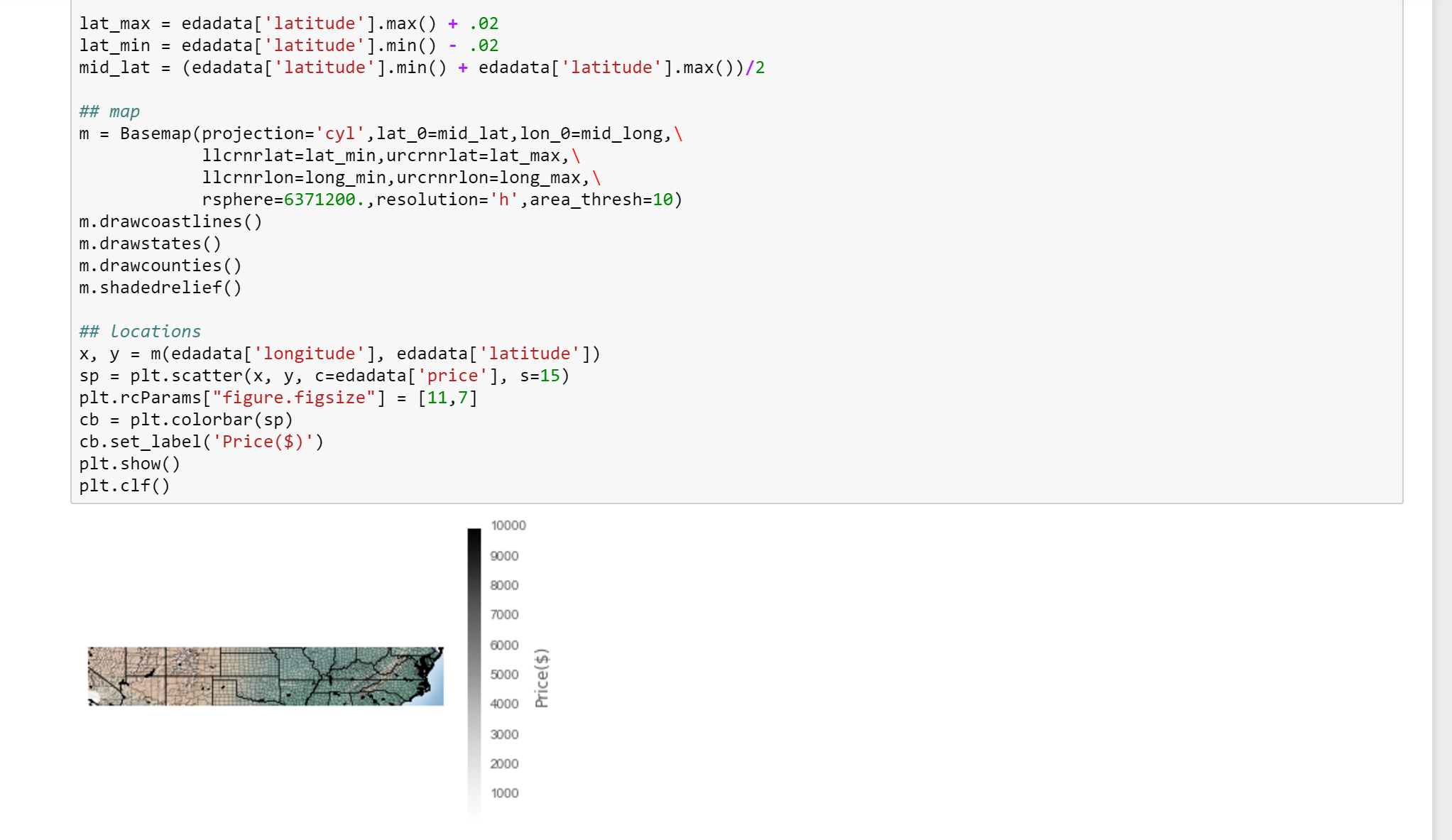
The following graph and code gives the different roomtypes present in the dataset.

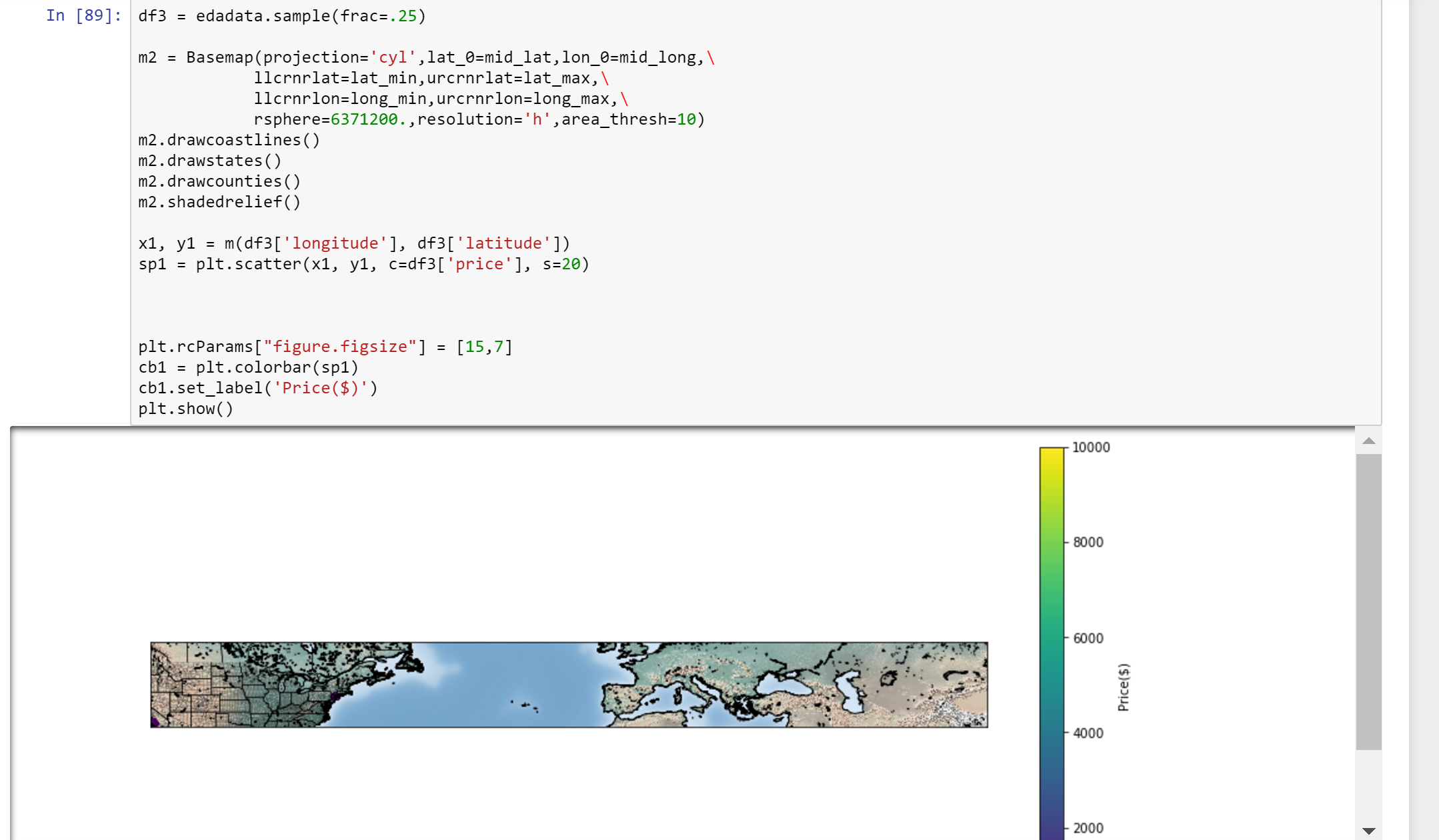


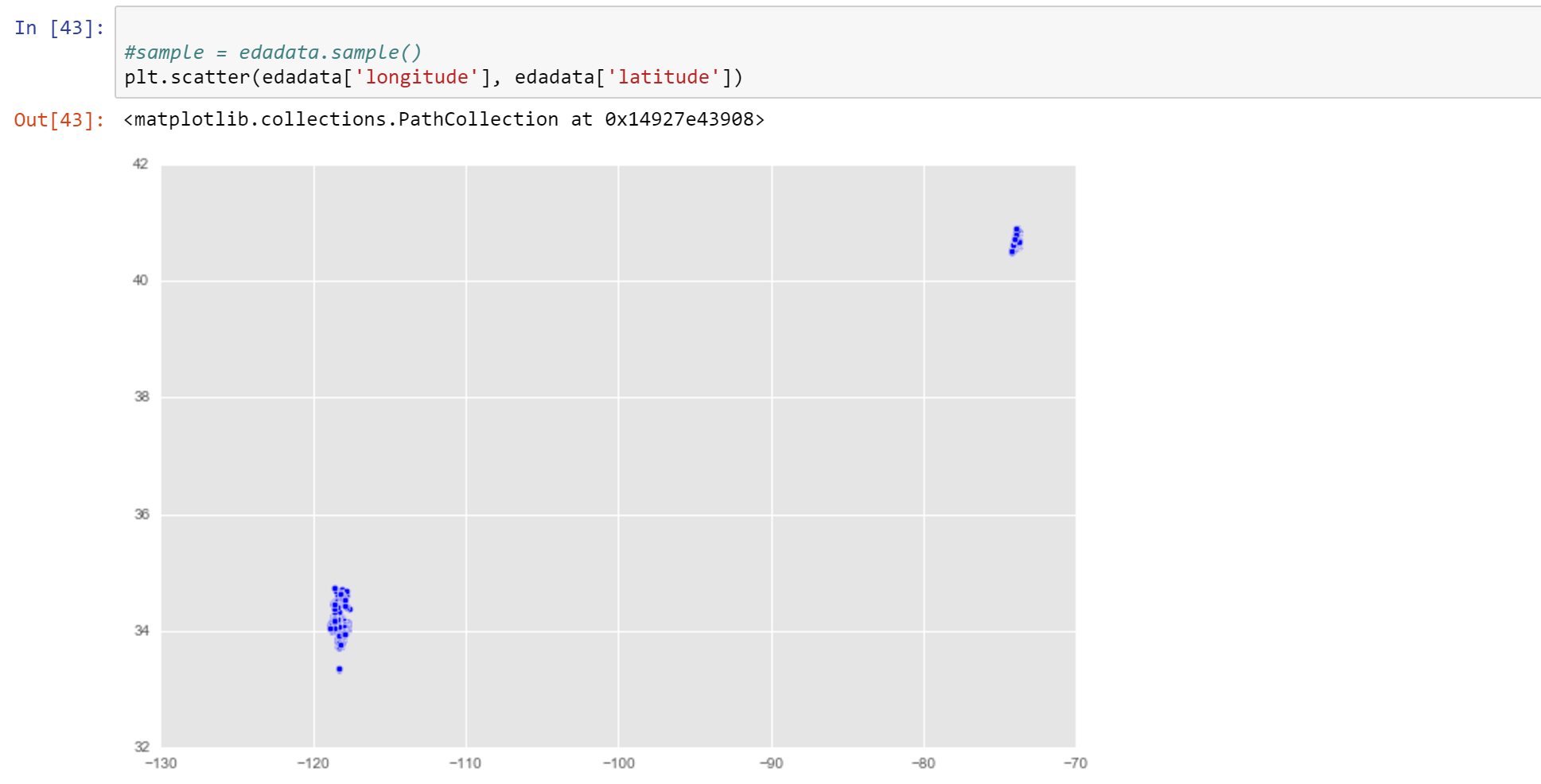
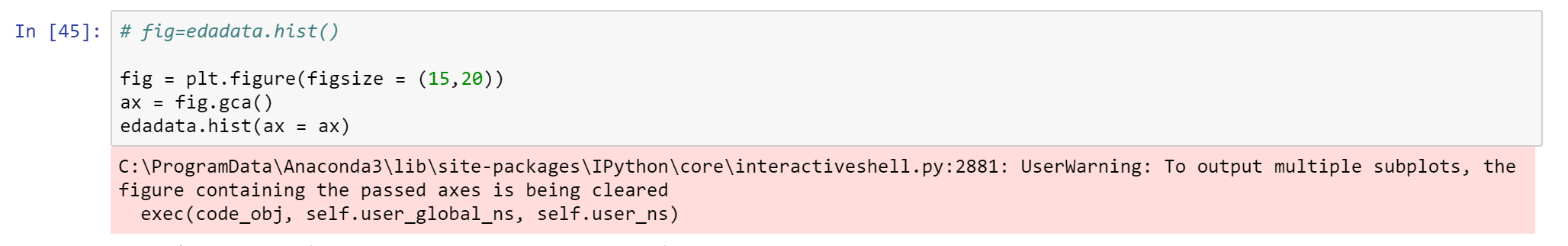


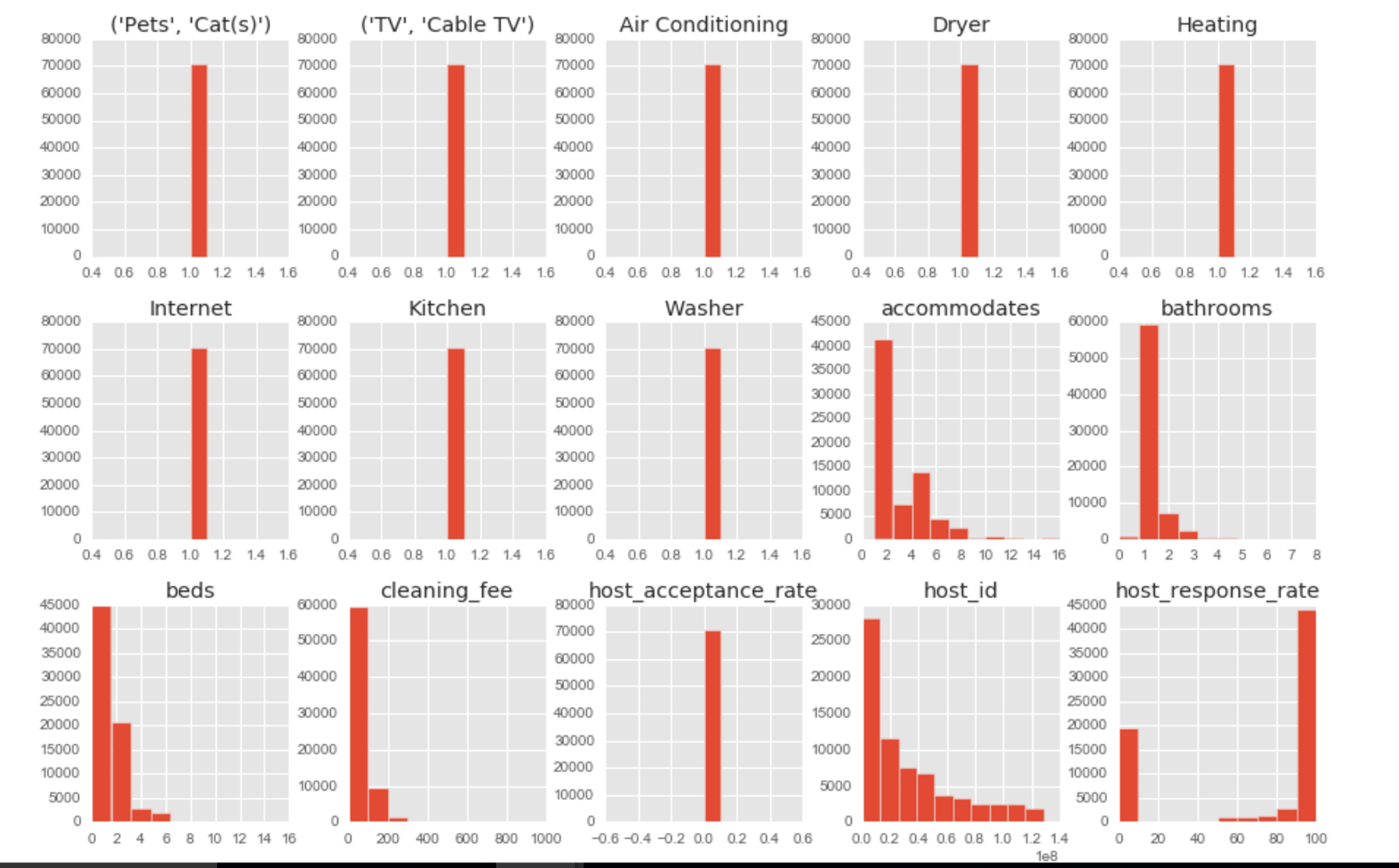




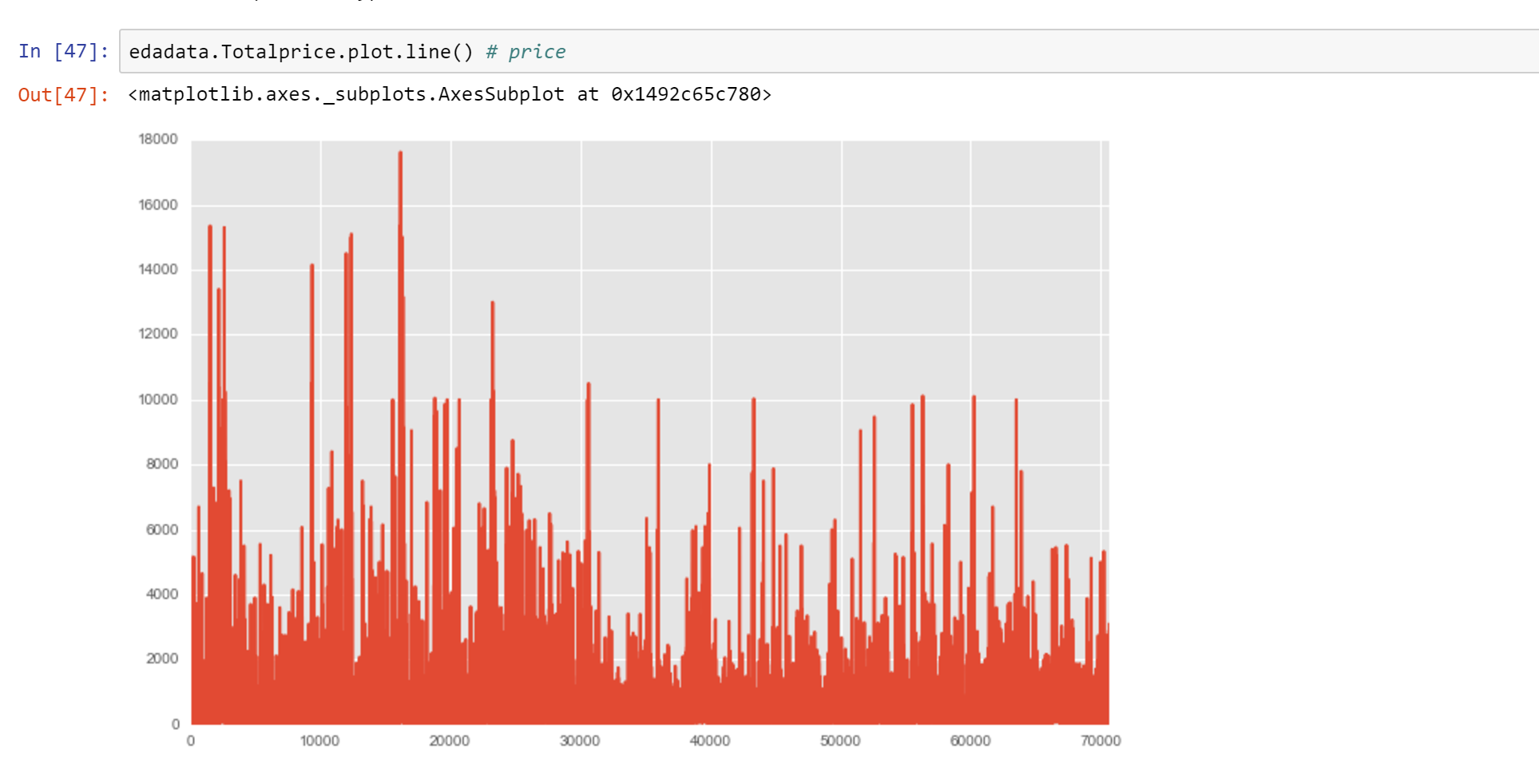


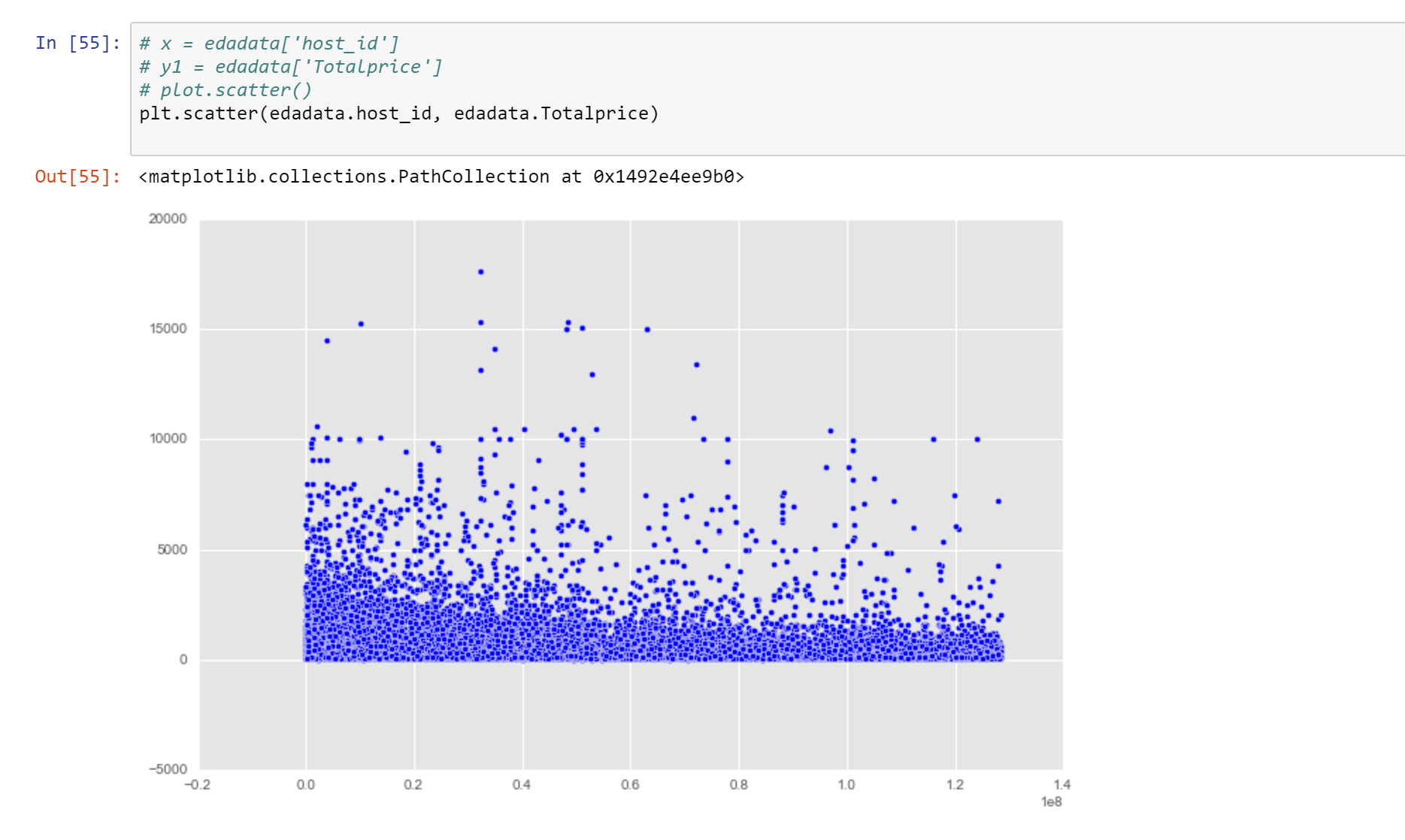




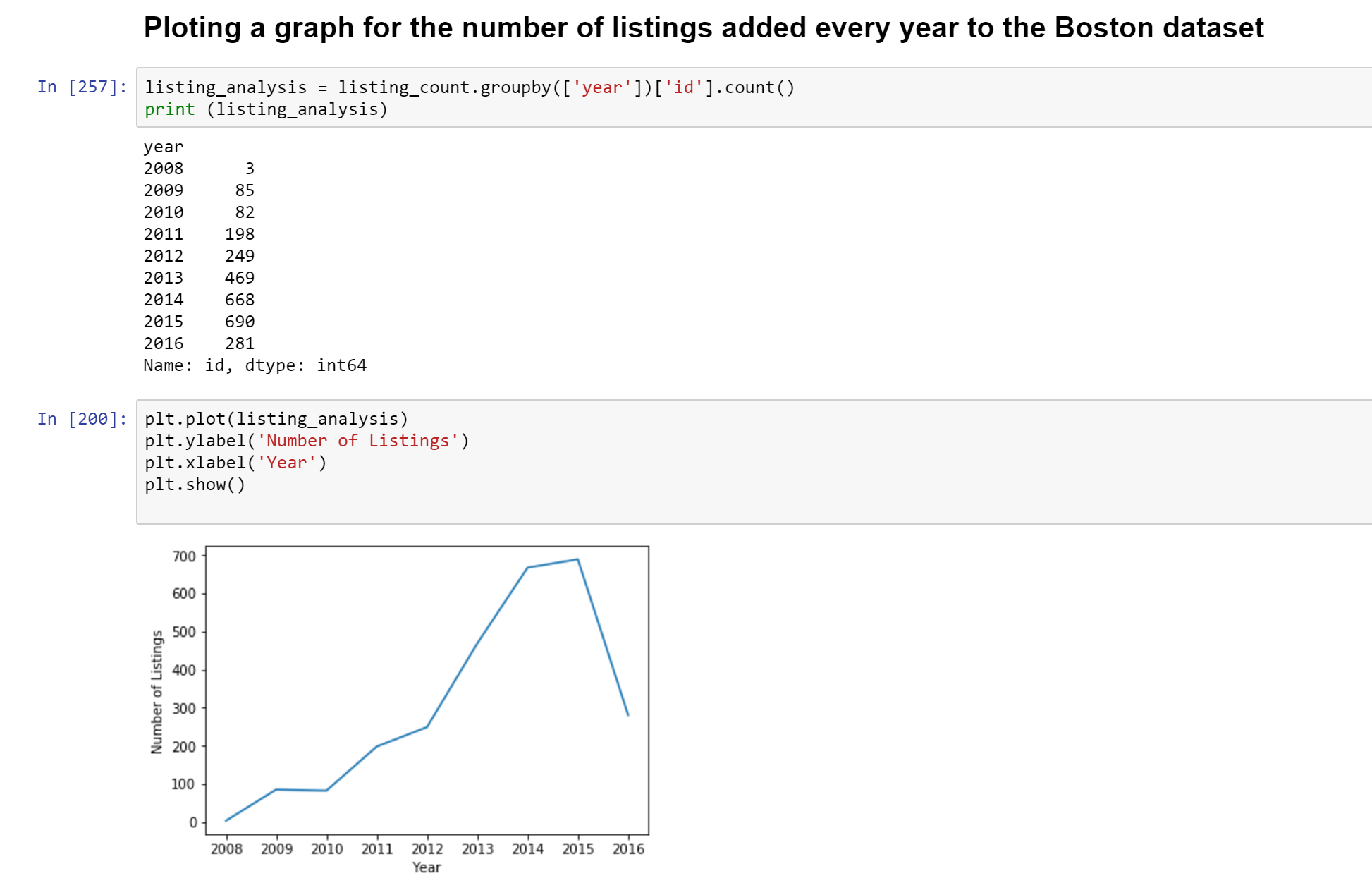






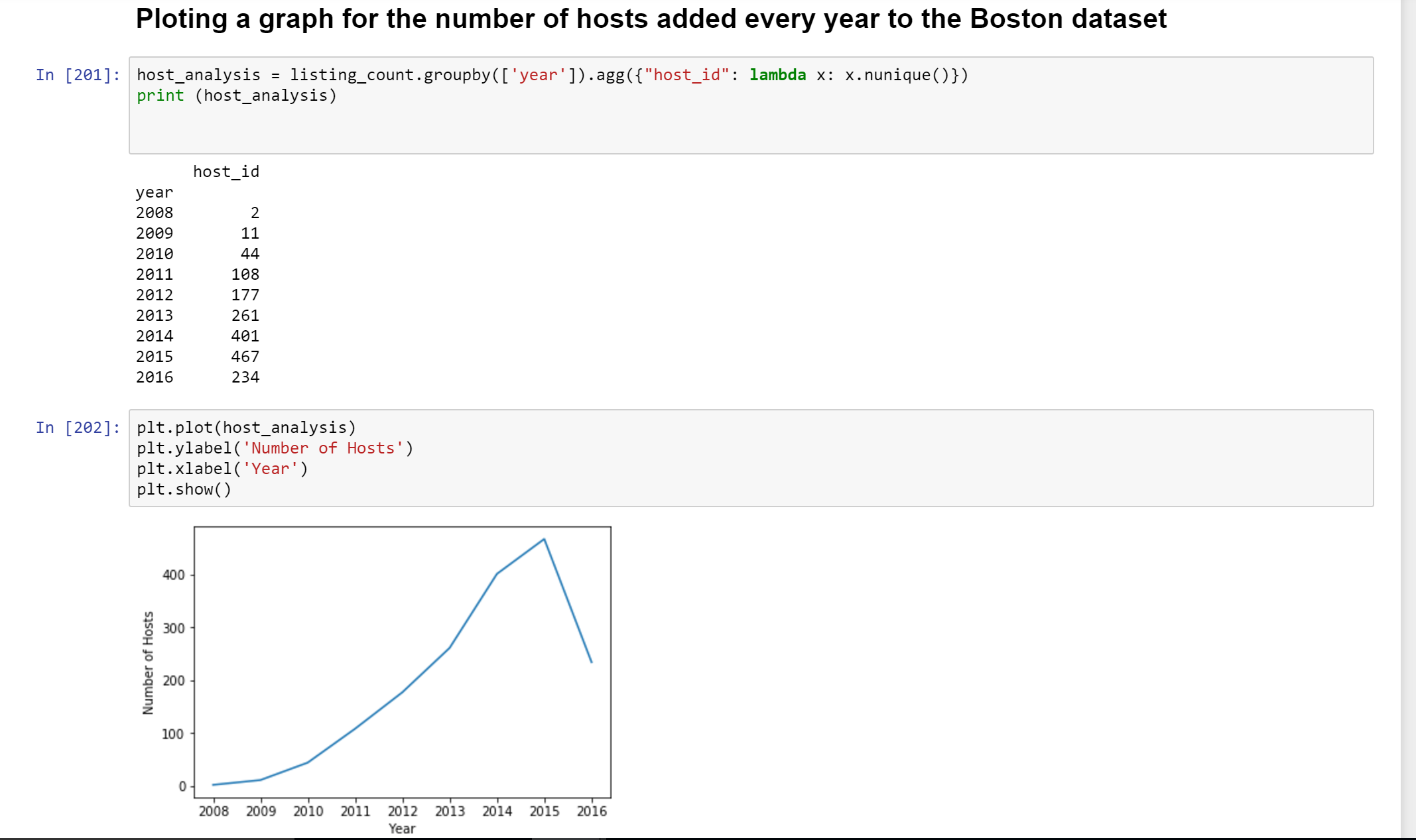
The rest of the Analysis is performed on the Calendar dataset:

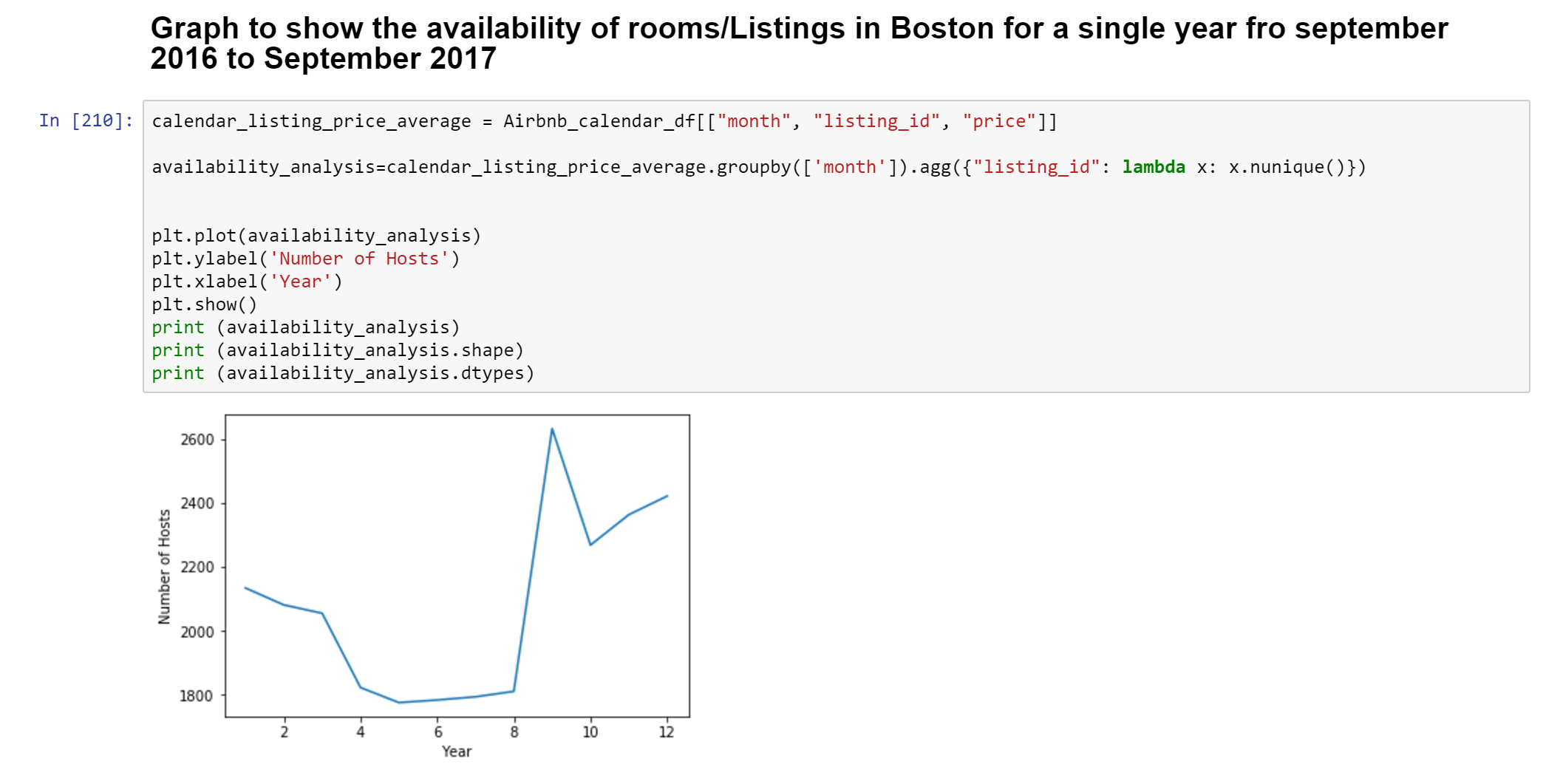
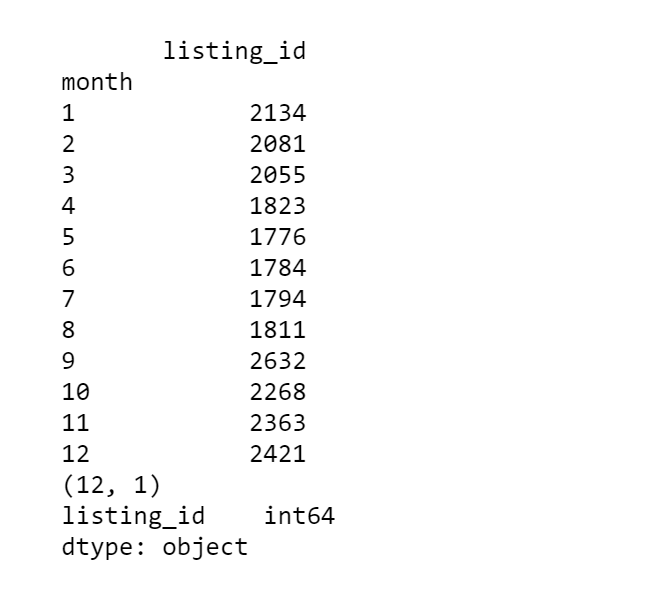
Shown below is the yearwise numbers of listings that are added to Airbnb

The below analysis shows that the number of venues added were at it’s peak in 2015 and reduced during 2016. One reason may be that 2016 dataset is not complete. 

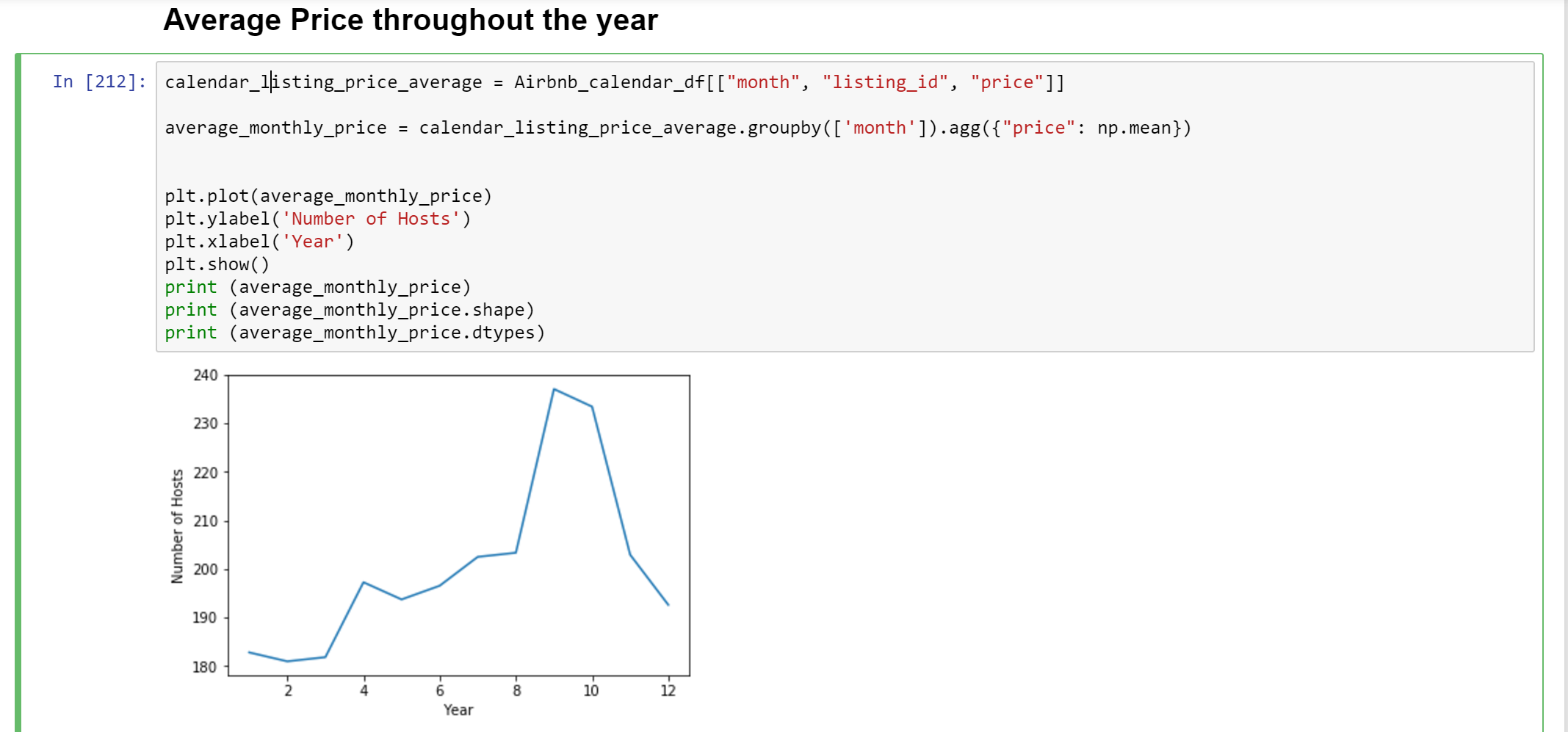
The next gragh shows us the yearwise count of hosts added to Airbnb

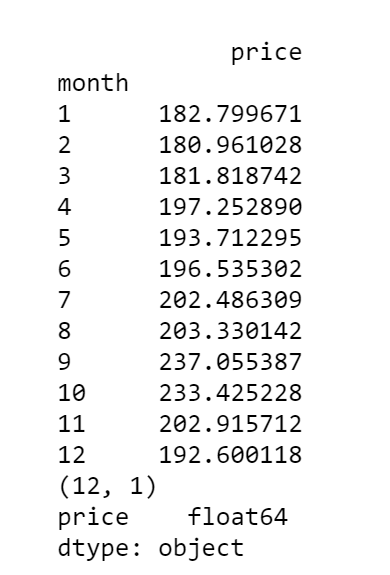
The same analysis goes for the number of hosts that were added to airbnb in boston over the years.



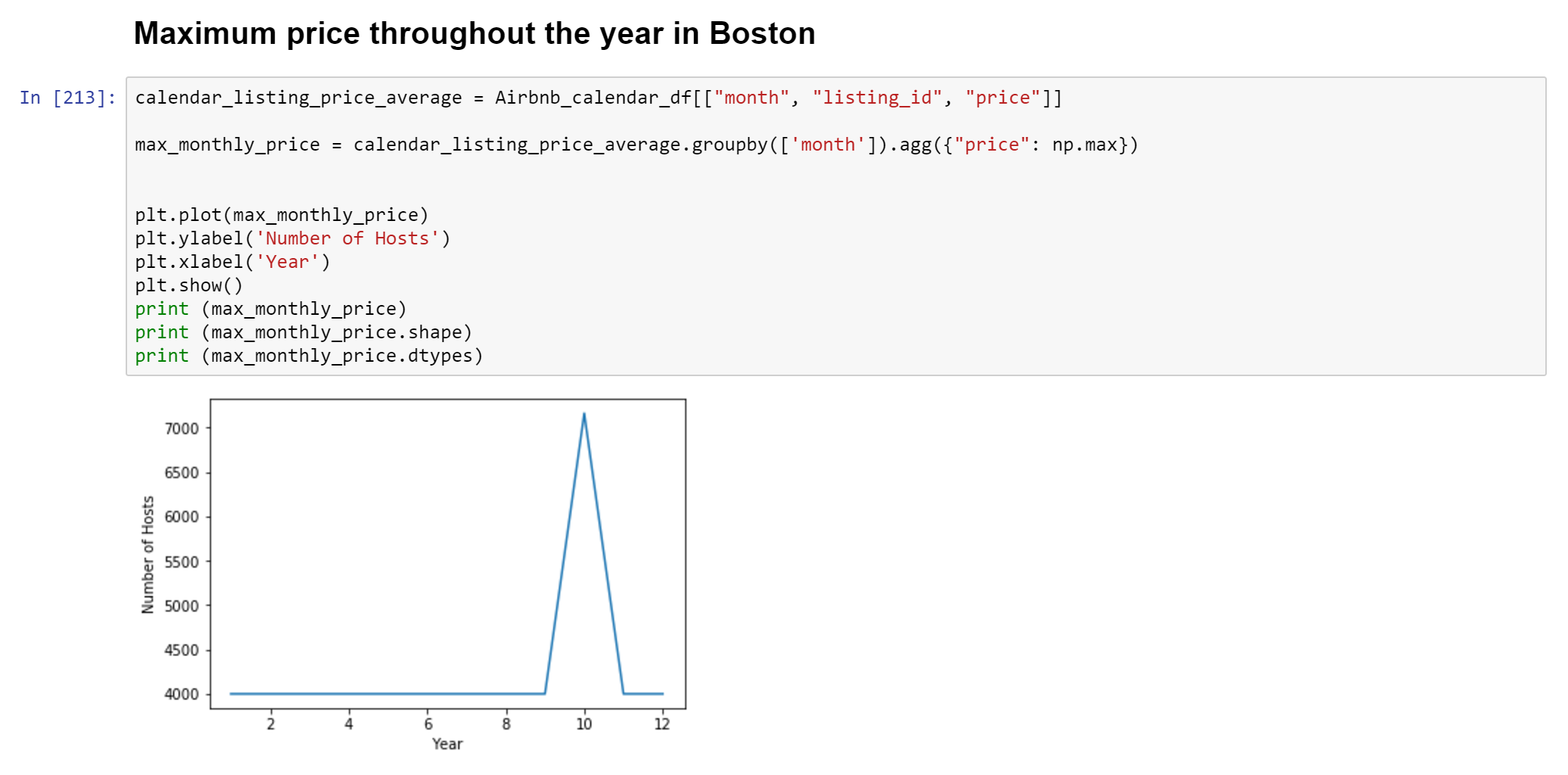
The below graphs show that during October and November most number of listings are available while from April to August the availability is very less  

Since the season has more listings the price also is high during the same time as shown below :





The prices, even the maximum prices are skyrocketing in the month of October which makes it ideal time for hosts but not for guests in Boston.





Also according to the graph below, it is clear that the best time to have a cheap holiday in Boston is between April and July .

