AirBNB Dataset: Data Cleaning

Presentation by:

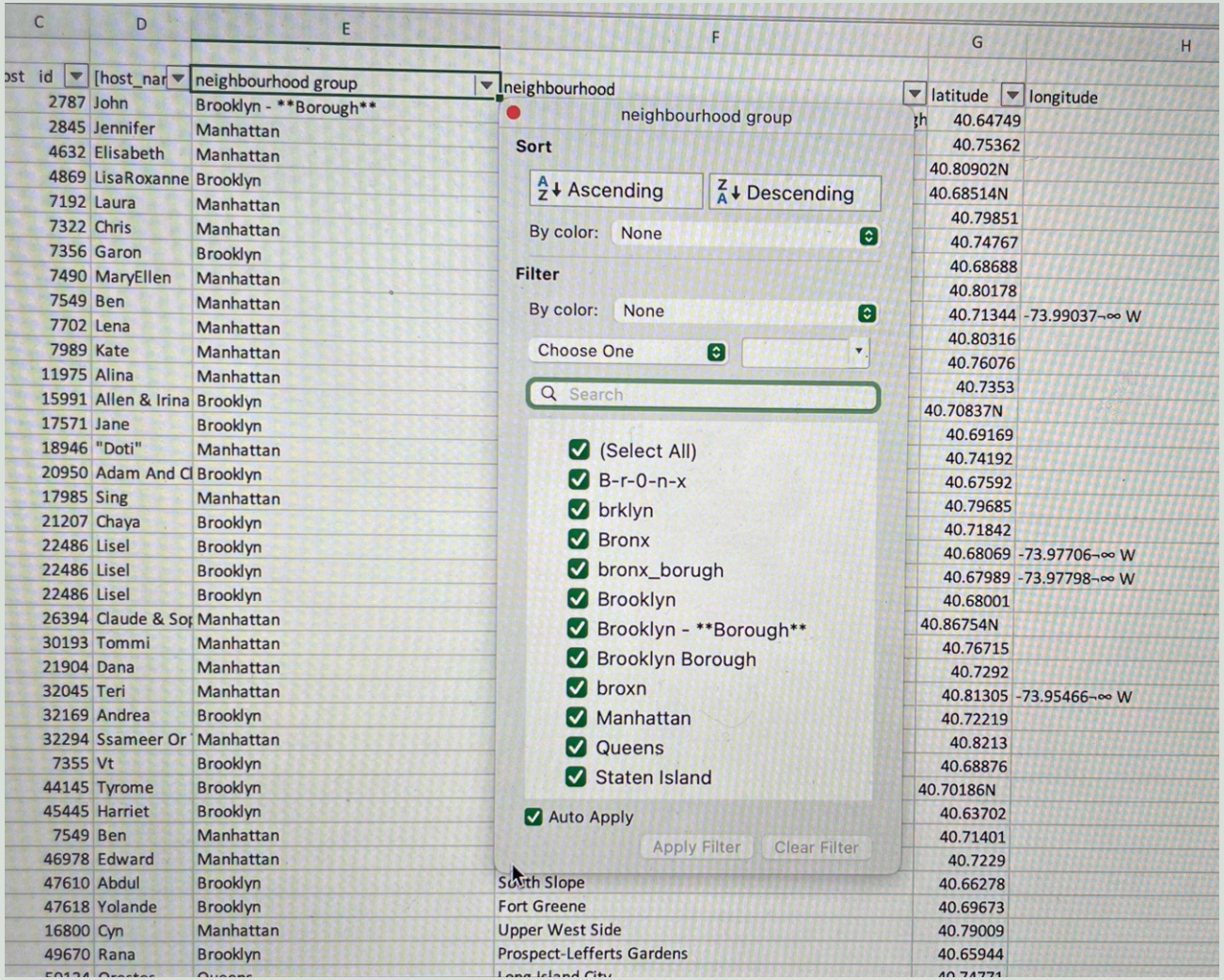
Muskan Burman

Data Science Capstone

Identifying the problems

- I started by visually evaluating the data – a proper lookover, just to get an idea of what the dataset contains and where the data might need some cleaning. I was able to identify the following problems immediately:
  - 1) The header (title rows) were not present.
  - 2) A column full of only NA values was present.

- I then made use of Excel's filter function on each of the columns to see the different values/tuples of the columns, and whether they need cleaning.



For example, here, we can see that the "neighbourhood group" column has unnecessary values that need replacing.

- Using the above-mentioned steps, I was to make a list of all the modifications I needed to make to properly clean the provided dataset:

1. Make Headers (Title Rows).
2. Remove the column full of just NA values.
3. Clean the “[host_name]” column name.
4. Clean “id” values.
5. Clean “host id” values.
6. Clean “neighbourhood group” values.
7. Clean “neighbourhood” values.
8. Clean the latitude and longitude values.
9. Clean the “room and type” values.
10. Clean the “minimum nights” values.
11. Replace the NA values in the “reviews per month” column.
12. Clean the “floor” values.
13. Clean the “noise.dB.” values.
14. Clean the “price” values.

Overcoming the problems

- My goal during this assignment was to preserve as much data as I could. Therefore, I made sure to refrain from directly deleting any outliers or columns with above-mentioned problems (eg: NA values). Instead, I tried to inspect each problem and find ways to clean the dataset without losing any precious data.
- To work on the dataset, I used python as my programming language and made use of the Jupyter Notebooks IDE by importing my CSV file into Jupyter Notebooks, and made use of Pandas and Numpy.

Make Headers (Title Rows)

- I created the headers for my dataframe using iloc.

Remove the column full of just NA values

- For this I used the “dropna” function.

Clean the “[host_name]” column name

- To do this, I renamed my column name using the “rename” function.

Clean “id” and “host id” values

- While using the “filter” function on Excel, I noticed that the “id” column had a random value :“52$$$38”. So, I replaced it with its cleaner version: 5238. Similarly, in the “host id” column, I replaced “2,,,118778” with “2118778”.

Clean “neighbourhood group” ,“neighbourhood” and “room and type” values

- Here, the “filter” function yet again helped me identify the various errors in the tuples of the ”neighbourhood group” column. I used the ”replace” function of python to replace all “B-r-0-n-x”, “bronx_borugh”, and “broxn” values with “Bronx” ; and all “brklyn”, “Brooklyn - **Borough**”, and “Brooklyn Borough” values with “Brooklyn”. Similarly, I updated the values of the ”neighbourhood” column as well. I did the same thing with “room and type” column, replacing all “Privatè-Room” and “Room Type Private” with ”Private room”.

Clean the “latitude” ,“longitude”, “minimum nights” and “noise.dB.” values

- I removed all the trailing unnecessary characters from all these columns: “N” from latitude values, “°W” from longitude values, “nights” from minimum nights, and “dB” from noise.dB.

Replace the NA values

- I took the mean of values belonging to different categories of “room and type” and used that to fill in missing review per month values of the same category. This way I did not delete any tuples containing NAs.

Clean “price” values

- I removed all the trailing unnecessary characters from the price column and also removed the preceeding “$” signs.

Clean ”floor” values

- I converted all values to numerical form: First floor to 1, etc.

Can be imported to R!

- I realized that hostname and name columns had no value in our dataset, however I did not remove them to maintain my goal of preserving data. I will not be using them for Data Exploration and prediction model purposes.
- My cleaned dataset can now be imported to R studio for further investigation.