

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.

C	D	E	F	G	H
host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude
2787	John	Brooklyn - **Borough**		40.64749	
2845	Jennifer	Manhattan		40.75362	
4632	Elisabeth	Manhattan		40.80902N	
4869	LisaRoxanne	Brooklyn		40.68514N	
7192	Laura	Manhattan		40.79851	
7322	Chris	Manhattan		40.74767	
7356	Garon	Brooklyn		40.68688	
7490	MaryEllen	Manhattan		40.80178	
7549	Ben	Manhattan		40.71344	-73.99037~∞ W
7702	Lena	Manhattan		40.80316	
7989	Kate	Manhattan		40.76076	
11975	Alina	Manhattan		40.7353	
15991	Allen & Irina	Brooklyn		40.70837N	
17571	Jane	Brooklyn		40.69169	
18946	"Doti"	Manhattan		40.74192	
20950	Adam And Ci	Brooklyn		40.67592	
17985	Sing	Manhattan		40.79685	
21207	Chaya	Brooklyn		40.71842	
22486	Lisel	Brooklyn		40.68069	-73.97706~∞ W
22486	Lisel	Brooklyn		40.67989	-73.97798~∞ W
22486	Lisel	Brooklyn		40.68001	
26394	Claude & So	Manhattan		40.86754N	
30193	Tommi	Manhattan		40.76715	
21904	Dana	Manhattan		40.7292	
32045	Teri	Manhattan		40.81305	-73.95466~∞ W
32169	Andrea	Brooklyn		40.72219	
32294	Ssameer Or	Manhattan		40.8213	
7355	Vt	Brooklyn		40.68876	
44145	Tyrome	Brooklyn		40.70186N	
45445	Harriet	Brooklyn		40.63702	
7549	Ben	Manhattan		40.71401	
46978	Edward	Manhattan		40.7229	
47610	Abdul	Brooklyn		40.66278	
47618	Yolande	Brooklyn	South Slope	40.69673	
16800	Cyn	Manhattan	Fort Greene	40.79009	
49670	Rana	Brooklyn	Upper West Side	40.65944	
50124	Oreto	Queens	Prospect-Lefferts Gardens	40.74771	

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_borough”, 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 “Private-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 preceding “\$” 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.