Data wrangling

Jelena Mitrovic

Gathering Data

Gathering data was done in a couple of steps.

Firstly, I read the file (twitter-archive_enhhanced.csv) that was given by the instructor using the pandas .read_csv() function.

Secondly, I downloaded the content of the image predictions URL and then read the tsv file. In that way, I gathered the image_predictions file.

Thirdly, I collected tweet JSON file by using tweepy and JSON libraries and then read it using .read_json() function.

Assessing Data

When I prepared all necessary dataframes for this project, I focused on detecting quality and tidiness issues among them in order to be able to fix them in the next part and prepare dataframes for the data analysis process.

By using different pandas functions such as .info(), .head(), .unique(), etc. I came up with detecting eight quality issues and two tideness issues.

Eight quality issues:

- Column Source in twitter_archive has all links the same.
- In twitter_archive, the column timestamp is a string instead of being datetime.
- In twitter_archive, the column Text has unnecessary information such as ratings and links.
- In twitter_archive, the column Name has some string values that do not represent the names of dogs. For example: such, quite, a.
- In twitter_archive, the columns Rating_numerator and Rating_denominator have some wrong values.
- Missing values in a couple of columns in twitter_archive
- In image_predictions, the columns p1, p2, and p3 have some capitalized string values, some not.
- In image_predictions, the columns p1, p2, and p3 have some string values with an underscore and some of them not.

Two tidiness issues:

- The last four columns in twitter_archive: doggo, floofer, pupper, and puppo should be organized as one column whit those 4 categories.
- Tweet_json_full dataframe should be combined with twitter_archive.

Cleaning Data

Once I detected the issues in dataframes in this project, I started the cleaning process.

When it comes to eight quality issues I did the following:

- The column Source in twitter_archive was dropped since all the values are the same.
- The data type of column Timestamp in twitter_archive was changed from object (string) to datetime.
- From column Name in twitter_archive string values that do not represent names were deleted. Those include: such, quite, a, an.
- The wrong values in the columns Rating_nominator and Rating_denominator in twitter_archive were changed according to their ratings in Text column.
- Unnecessary information (ratings and links) in the column Text in twitter_archive was removed.
- The columns in twitter_archive that had a lot of missing values were dropped.
- The columns p1, p2, and p3 in image_predictions were fixed to have the first letter capitalized.
- The columns p1, p2, and p3 in image_predictions were fixed not to have underscores in some rows.

When it comes to tidiness issues, I did the following:

- The last four columns in twitter_archive: doggo, floofer, pupper, were organized as one column whit those 4 categories.
- Tweet_json_full dataframe was combined with twitter_archive.

Data Analysis and Visualization

When I cleaned all dataframes and combined them into one, I performed a couple of descriptive analysis in order to get a better insight into the relations between some variables.

After grouping the results based on the type of dog, we could see that the average rating score is the highest for floofer and puppo types of dog. The type of dog that has the highest average number of retweeted posts is doggo. Also, the most common names of dogs are Cooper, Charlie, Tucker, and Oliver. Moreover, we could observe that the profile we were analyzing is gaining popularity throughout time.