Project: Wrangle and Analyze Data

Student: Sherif Shawkat

UDACITY

Wrangle Report

Introduction:

This report is part of the requirements needed to complete the project "Wrangle and Analyze Data" for Udacity Data Analyst nanodegree. The project aims at analyzing dog tweets on the social media giant Twitter. We are collecting tweets, cleaning the received data then analyzing and visualizing them.

Gathering data:

The data is gathered from 3 different sources:

1- A .csv file containing the tweets with the text of the tweet and some basic extractions from this text
2- A .tsv file containing an already run neural network on the pictures of the tweet to identify the type of the dog
3- Finally a .txt file that contains retweets & favorites counts for each tweet.

I took all the 3 files ready from Udacity servers as I have failed to use the twitter API although I sent to Twitter to apply for developer privileges.

Assessing data:

1- Visual Assessment: A quick look on the 3 dataframes was needed to identify our columns and their datatypes, as well as what each column mean.
2- Programmatic Assessment: An in-depth look on each dataframe, so that we can identify our tidiness & quality issues and checking for any duplicated data.

Tidiness Issues (3):

- Merge all dataframes in 1 dataframe for simplicity in our analysis
- Merge the 4 dog stages (doggo, floofer, pupper and puppo) into 1 column "dog_stage". Note: this column value may have multiple types (ex: "doggo, pupper").
- Drop unnecessary columns after fixing the quality issues

Quality Issues (9):

- Replacing names having length < 2 with NaN.
- Remove tweets having rating denominator not equal 10
- Remove tweets that have no images attached
- Clean the “source” column from html tags
- Remove all retweets, and only analyze the basic tweets
- Remove tweets that has the 3 image predictions as “Not dog”
- Adjust ratings to be extracted from the text itself, as some decimal numerators caused problems in the values extracted. And also add a “rating” column that divided numerator/denominator.
- Adjust dog type to lower case letters
- Change format of some columns (ex: timestamp “object → timestamp”)

Conclusion:

Any data received from any source must have tidiness & quality issues that have to be cleaned before starting our analysis. I guess after cleaning our data referring to the issues above, our data is ready to be analyzed and visualized.