Wrangle and Analyze Data Project
Wrangle report

By: Mishal AlKhalifa

29th of August 2020

Introduction

The wrangle and analyze project is part of Udacity Data Analysis Nanodegree program. This project is about wrangling of data from multiple sources associated with tweets from twitter application of users rating dogs' pictures. So, in order to wrangle and analyze the data there are main steps to follow as gathering, assessing and cleaning.

Body

I. Gathering

This step is about data being gathered from multiple sources such as

- The enhanced twitter archive file. It contains important tables that will be needed in the analysis such as tweet id, text and rating.
- Twitter API. This file JSON contains data such as favorite count and retweet count.
- Tweet image predictions file. Contains data of images of each tweet.

II. Assessing

This step is about finding the issues of the data which affect the quality of the data. And tidiness is about the issues that were cleaned.

Quality issues

- Expanded urls has a lot of missing value.
- img_num & source datatype should be categorical.
- Timestamp datatype should be DateTime.
- Some names are inaccurate such as 'this','a' ,'an', 'the', etc.
- Name has values that are the string "None" instead of NaN.
- Removing the values with incorrect rating in the rating denominator & rating numerator.
- The source should't include the full url only the content between it e.g Twitter for iPhone.
- Twitter count retweet & Retweet with comment as new tweet so there are some duplicated tweet start with RT but with same content.
- jpg_url has duplicated rows.
- Merge 1 variable (dogkind) in 4 different columns (doggo, floofer, pupper, and puppo).
- Merge all the dataframes into a master table.
- some missing ids such as reply status id, replay user id, retweet status id, & retweet user id.
- change data type of reply status id, replay user id, retweet status id, & retweet user-id into an integer.
- the source shouldn't include urls!
- retweeted the same tweet in text so duplicated the expanded urls.

- There are many columns in this dataframe IS HARD to read, and some will not be needed for analysis**.**

Tidiness Issues

- Expanded urls has a lot of missing value.
- img_num & source datatype should be categorical.
- Timestamp datatype should be DateTime.
- Some names are inaccurate such as 'this','a' ,'an', 'the', etc.
- Name has values that are the string "None" instead of NaN
- Removing the values with incorrect rating in the rating denominator & rating numerator.
- The source should't include the full url only the content between it e.g Twitter for iPhone.
- Twitter count retweet & Retweet with comment as new tweet so there are some duplicated tweet start with RT but with same content.
- jpg_url has duplicated rows.
- Merge 1 variable (dogkind) in 4 different columns (doggo, floofer, pupper, and puppo).
- Merge all the dataframes into a master table.
- There are many columns in this dataframe IS HARD to read, and some will not be needed for analysis.

III. Cleaning

This step is about enhancing the quality of the data by applying some cleaning to it. Using multiple functions that are powered by libraries that were imported in the begging of the project.

Conclusion

In conclusion, this project involved wrangling data from various sources which implies using different libraries to clean the data in order to become tidy.