Wrangle and Analyze Data Project

Wrangle report

By: Abdulellah Al-hudaithy

Introduction:

In this project I will use three steps to wrangle WeRateDog account data WeRateDog is an account in Twitter.

- Gathering
- Assessing
- Cleaning

Gathering: in this step I will gather the date from multiple resources such as CSV, downloading file from Internet and Twitter API

- From CSV: file contain archive tweet from WeRateDog account, and I have used read_csv function to extract it. This file was given by Udacity.
- Downloading from internet: file contain detailed data about images in each tweet. I have used Request library to extract it and save it in TSV format.
- From Twitter API: file that count the number of favorite (like) and retweet in each tweet. I will use tweepy library to extract the file but unfortunately I didn't get the authorization from Twitter, so I read the JSON file using JSON library.
-

Assessing: in this step I have go through the data and try to find out quality and tidiness issues either visually or programmatically.

Quality issues:

- Missing values.
- Incorrect dogs name.
- Duplicated tweet (count the retweet and retweet with comment)
- Incorrect data types.
- Column has a None instead of NAN (nulls)
- Include the HTML tag in the source column.
- Correct input.
- Remove unnecessary row or attributes that will not affect the analysis and visualization.

Tidiness issues:

- in Twitter-archive table should have one column for the dog type instead of having four columns for each.
- Merge all the dataframes into a master table.

Cleaning Data: in this step I have cleaned all the issues above using Pandas and numpy libraries. And I have used the methodology “Define, Code and Test” for each issue to make the process organized.