Wrangle Report

Introduction

This report briefly explains the steps used in data wrangling for twitter dataset. The main reason for project is to practice skills learnt from Udacity Nanodegree program under the topic wrangle and analyze data. Dataset used for this analysis is for the tweet archive Twitter account **@dog_rates** and username **WeRateDogs**. This account rates people's dogs with humorous comments.

The steps involved in the wrangle and analysis for this dataset includes:

- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data

This step involved gathering data from 3 different sources.

1. ***twitter_archive_enhanced.csv:*** This dataset which is provided by Udacity and is available for downloading it manually by clicking the link provided.
2. ***image_predictions.tsv:*** This dataset is hosted in Udacity servers and is available for downloading it programmatically using the Requests Python Library. The dataset presents dogs name predicted using neural network.
3. Data from the Twitter API:
   To query this data from Twitter, a twitter developer account with elevated access is required and then create an API object using Tweepy Library
   Then, using the using the API object query the JSON data and store in ***tweet_json.txt.*** Then create a data frame by reading line by line from the txt file.

Assessing Data

Now, having gather the 3 datasets, the next was to assess data for quality and tidiness issues. This includes both assessing the datasets visually as well as programmatically.

For visual assessment, the technic used was to print the dataset on jupyter notebook and manually checking for quality and tidiness issues.

Then programmatic assessment, involved using different methods such as ***describe, info, isnull*** etc.

Later, documenting the quality and tidiness for cleaning later.

Cleaning Data

The steps involved creating copies of the 3 datasets and in this stage is cleaning the earlier noted quality and tidiness issues using the framework provided on Udacity classroom, which is **Define, Code** and **Test**.

Conclusion

Some notes for the twitter data wrangling and analysis: this step is a back-and-forth strategy, that is even after assessing and cleaning, during analysis and visualizing the data you may note some quality issues which prompts you to go back to cleaning.