UDACITY
Data Analyst Nanodegree Program

PROJECT DETAILS: Wrangle and Analyze Data

INTRODUCTION:

The Wrangle and Analyze Data project is a part of Udacity's Data Analyst Nanodegree Program. This project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. The WeRateDogs rate's pictures of people's dog in a humourous manner, most often by giving them ratings higher than 10/10. After scraping together the data, quality and tidiness issues were accessed and then cleaned.

Gathering Data:

In the above project data was gathered from 3 different sources:

1. The twitter_archive file that was provided and downloaded programmatically. It was the dataset provided from the WeRateDogs account user and included information's like tweet_id, timestamp,rating_numerator,rating_denominator,name,etc.
2. Additional data including favourite and retweet count were gathered from twitter API.
3. The image_prediction file was downloaded programmatically using the requests library in python from the udacity servers. It included information on the breed of dog that was predicted by a machine learning algorithm.

Accessing Data:

After the data was gathered, it was accessed programmatically and manually in google spreadsheets. Programmatic assessment was performed using the following methods:

- .head()
- .sample()
- .info()
- .value_coounts()

Data Quality Issues:

1. Delete columns that won't be used for analysis
2. Separate timestamp into day - month - year (3 columns)

3. Correct numerators with decimals
4. Correct denominators other than 10.
5. Name has values that are string "None" instead of NaN
6. Looking visually in Excel,some names are inaccurate such as "a", "an", "the", "very", "by", etc
7. It is also found that name of a dog being "O" instead of "O'Malley"
8. Drop duplicated jpg_urls
9. Delete columns that won't be used for analysis
10. Keep original tweets only

Tidiness Issues:

1. Change tweet_id to type int64 in order to merge with the other 2 tables
2. All tables should be part of one dataset

Cleaning Data:

The issues found during the assessment process were cleaned and tested using the following methods provided by pandas library iin python:

- Merge()
- .drop()
- .astype()
- .to_datetime()
- .islower()
- .replace()
- .loc[]
- .value_counts()
- .info()
- .head()

Conclusion:

From this project we were familiarize that all the data needed for a project need not come from a single source and already tidy. We did gathered data from various sources like downloading from the servers and scraping from the net using api’s, etc. This project also emphasized on using python to access various quality and tidiness issues in the dataset before any analysis can be performed.