WRANGLE REPORT

Udacity Data Analyst Nanodegree

By: Deepshikha Patni

Introduction

The Wrangle and Analyze Data is part of Udacity's Data Analyst Nanodegree. This project involves wrangling of data from different sources associated with tweets from the user @dog_rates, also known as WeRateDogs. After scraping data, quality and tidiness issues were addressed. Finally, some visualizations and insights are captured in the act_report.pdf document.

It includes three parts,

1) Gathering Data
2) Assessing Data
3) Cleaning Data

Gathering Data

Data is gathered from three different sources in a Jupyter notebook titled **wrangle_act.ipynb**.

1) The enhanced twitter archive file (**twitter_archive_enhanced.csv**) was provided and downloaded manually. This file includes various variables for each tweet like tweet_id, timestamp, text, rating, numerator, denominator etc. This was used to create twitter_archive data frame.
2) Twitter API and Python's Tweepy library were used to gather each tweet's retweet count and favorite ("like") count. Using the tweet IDs in the WeRateDogs Twitter archive, we query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. We read tweet_json.txt file to create **tweet_json** dataframe.
3) The tweet image predictions, i.e. what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file is hosted on Udacity's servers and was downloaded programmatically using the Requests library. We used this to create **images** dataframe.

Assessing Data

Data assessing was done in **wrangle_act.ipynb** both visually and programmatically by using functions like *head(), value_counts(), info(), duplicated(), describe()*

I detected and documented quality and tidiness issues as the following,

Quality Issues:

- Completeness:
  - Missing data in some columns like in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
  - Incorrect data type for tweet_id
- Validity:
  - Dataset includes retweets (i.e. duplicated data)
- Accuracy:
  - Incorrect data type Timestamp
- Consistency:
  - Ratings are unstandardized.
  - Undesired columns present.

Tidiness Issues:

- Correct naming issues.
- All dataset should be merged into one dataset.

Cleaning Data

We performed data cleaning by using below functions,

1. merge()
2. extract()
3. drop()
4. value_counts()
5. sort()
6. head()
7. info()
8. Regular Expressions

Finally, the cleaned data which is used further in data analysis process is stored in a csv file named **twitter_archive_master.csv**

Conclusion

This project helped to enhance the data wrangling skills when the source data is to be scraped from multiple data sources. It also helped to learn the importance of addressing data quality and tidiness issues before any data analysis could be performed.