Wrangle Report

Wrangle and Analyze Data: Disha Mukherjee

This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree. The project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs rate's pictures of people's dogs in a humorous manner, most often giving ratings higher than 10/10. After scraping together, the data, quality and tidiness issues were assessed and then cleaned. Finally, two visualizations were created, and insights can be found in the act_report.pdf document.

GATHERING DATA

Data was gathered from 3 different sources:

- The enhanced twitter archive file was provided and downloaded manually. This file includes various variables for each tweet including tweet id, timestamp, text, rating numerator and denominator, name, etc.
- Additional data, including favorite count and retweet count, were gathered using the Twitter API.
- The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers. Using machine learning techniques, the breed of dog was predicted based on the picture. Following instruction from Udacity project detail page first I downloaded Twitter archive file 'twitter_archive_enhanced.csv' which was provided for download, second file 'image_predictions.tsv' download programmatically from Udacity server using request library and third file which cause me all sort of problem as it was tweet_json.txt file using twitter API for @WeRateDogs as I requested for permission from twitter for almost a week and in end contacted mentor and it was provided me by email so I read this file line by line to create data frame for at least three columns id(tweet) , favorite and retweet counts.

ASSESSING DATA

After the data was gathered, assessment was performed using the following methods:

- .head()
- sample()
- .info()
- .value_counts()

Tidiness issues that were cleaned:

- Combining all dataframes together as they all contained information about the same tweets

- Combining 4 variables about dog type into 1 column “dog_stage”

Quality issues that were cleaned:

- Data contained retweets
- Tweet id was the incorrect data type
- Timestamp was the incorrect datatype
- Name contained the string “None” instead of a NaN
- Name contained various inaccuracies which were regular lowercase words
- The name O’Malley was incorrectly extracted as “O”
- Rating numerators which contained decimals were incorrected exported
- Ratings are unstandardized
- Undesired columns present

CLEANING DATA

The issues found during the assessment process were cleaned and tested using the following methods and techniques:

- merge()
- reduce()
- .extract()
- .drop()
- .isnan
- .astype()
- to_datetime()
- islower()
- .replace()
- .rename()
- set_option()
- .loc[]
- .value_counts()
- .info()
- .head()
- Loops
- Regular expressions

CONCLUSION

As Matt said during course Lectures that rarely does all the data, we want for a project come from 1 source and is already tidy. This project emphasized that you would need to using Python and its various libraries to scrape data from various sources in various formats, and clean various quality and tidiness issues, before any data analysis can be performed. I think only after

learning thoroughly from Udacity platform I have started to grasp coding mindset but as I am changing career so still a lot more to learn.