WRANGLE REPORT

Udacity DAND: Wrangle and Analyze Data Project

By: Vatsal Prakash

INTRODUCTION

This project called Wrangle and Analyze Data is part of the Udacity's Data Analyst Nanodegree program. The project involves gathering, assessing and cleaning data from various sources. All the data is associated with the tweets from WeRateDogs. The ratings for WeRateDogs are often higher than 10/10 which is a part of their popularity. Once the data was scraped from various sources, it was visually and programmatically assessed for quality and tidiness issues. These issues were then cleaned. Finally, some visualizations were created, and insights was gathered about the data.

GATHERING DATA

For our purposes, data was gathered from 3 different sources:

1) The enhanced twitter archive file was provided and downloaded manually. This file includes various variables for each tweet including tweet id, name, timestamp, ratings for denominator and numerator etc.

2) The second one was downloaded programmatically using the Requests library from Udacity's servers which was saved into a tsv file.

3) Additional data, including favorite count and retweet count, were gathered using the Twitter API and saved in a JSON file.

ASSESSING DATA

First visual assessment was performed on the gathered data using the following methods:

- .head()
- .tail()
- .shape

Then programmatic assessment was performed on the data using following methods:

- .info()
- .value_counts()
- .duplicated()
- .isnull()

Tidiness issues that were addressed and cleaned:

- Combining the three dataframes together as they contained the same tweet information
- Combining the 4 variables ('doggo', 'floofer', 'pupper' , 'puppo') into 1 dog_stage column

Quality issues that were addressed and cleaned:

- doggo floofer pupper puppo columns have None instead of NaN
- There are some retweets in the dataframe as evident from the retweeted_status_id
- tweet_id column is not a string type
- timestamp column is a string type instead of datetime object
- Many names in the 'name' column are invalid and also have 'None' instead of NaN
- Source has extra text making it unreadable
- numerator ratings columns have missing decimal values and was incorrectly extracted
- extended_urls column has some missing values

CLEANING DATA

All the issues found during the assessment were cleaned programmatically and then tested using the following methods and techniques:

- .astype()
- .merge()
- .value_counts()
- .extract()
- .drop()
- .to_datetime()
- .islower()
- .replace()
- .query()
- .loc[]
- Loops
- Regular expressions

CONCLUSION

This project emphasized on the use of Python from scraping data from various sources and then cleaning it using Pandas library. We got a glimpse that rarely the data in real-world is clean and that we need to wrangle and clean it and make it tidy before we can start our analysis.