Wrangling and Analyze Data report

The objective of this project is to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then cleaning it using Python and its libraries.

In this project, I will wrangle the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The Data: We have the following three datasets.

- Enhanced Twitter Archive:

The WeRateDogs Twitter archive contains basic tweet data. I downloaded the **(twitter-archive-enhanced.csv)** file from Udacity website.

- Additional Data via the Twitter API:

I downloaded the JSON API files which was uploaded by Udacity. **(tweet_json.txt)** i read this file line by line into a pandas DataFrame, Which give is the resulting data from twitter_api.py.

- Image predictions file :

This file **(image_predictions.tsv)** downloaded programmatically using the Requests library and the following
URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

After gathering the three datasets, I discovered a lot of data quality and tidiness issues during the visual and programmatic assessments, Which needs to be clean for the analysis. So i checked the Missing data,duplicates, invalid values,unwanted columns,etc. After that i fixed the quality and tidiness issues in three steps:

Define how to clean the issue in words
Code convert the definitions into executable code
Test test data to ensure the code was implemented correctly

Some of the functions I used in cleaning the dataset:

- **drop()** drop unwanted columns and missing values.
- **dropna()** drop the null values from expanded_urls column.
- **replace()** replacing incorrect names.
- **astype()** to change the datatype to str.
- **pd.to_datetime()** to convert object datatype to date.
- **rename()** renaming columns to be more descriptive.
- **apply()** to apply a function to the dataset.
- **sort_value()** to sort the dataset.
- **islower()** check if column values is in lowercase.
- **merge()** to merge the three dataset.

Then it was time for analysis. Analyzed the three dataset and generated six insights, Three of them visualized.