Report: Wrangle and Analyze Data

Introduction:

For this project, the goal was to work through this notebook and wrangle(gather, assess, clean) tweet archive of Twitter user @dog_rates, popularly known as 'WeRateDogs' to create interesting and insightful analysis and visualizations.

The flow of the process is as follows:

1. Data Gathering
2. Data Assessing
3. Data Cleaning
4. Storing cleaned Dataset
5. Gain insights and visualization

1. Data Gathering:

For this project, I have gathered data from 3 different sources and file formats.

I. I downloaded Twitter archive file manually from udacity website. The file was in .csv format.

II. I used requests library to download Image Prediction data from udacity server and stored in dataframe.

III. I used twitter API to collect additional data required for project. This was in json file.

2. Data Assessing:

After gathering data, I started assessing data. For visual assessment I used Microsoft excel along with jupyter notebook. After visual assessment I followed programmatic assessment to perform deeper assessment. After assessing I documented the quality and tidiness issues which I came across.

3. Data Cleaning:

I created copy of all three datasets. This step was necessary because if something goes wrong it should not affect the original gathered data.

This step of data cleaning was subdivided into three steps which are define, code and test. The define section was to define the issue I was cleaning. Code section was included code for cleaning operation. Test section was to confirm that issue has been addressed accurately.

I firstly addressed tidiness issues followed by quality issues.

4. Storing Cleaned Data:

After cleaning the dataset, I stored the cleaned dataset into new file for further processes.

5. Insights and Visualization:

Here we explored the data to gain insight and visualized to communicate our findings. This step is explained in detail in ‘act_report’.

6. Conclusions:

The real-world data mostly come from different sources and in many file formats. This data comes with many quality issues which needs to be addressed. Real world data is rarely tidy. Before doing exploration on data we need to thoroughly assess and clean data to gain better and reliable insights.