ACT REPORT

The act report shows the summary of the insights as well as displays the visualization produced from the wrangled data.

In this project, I carried out an exercise on data wrangling making use of three separate datasets from different data sources. The datasets are the; Twitter_archive_enhanced.csv, Image_prediction.tsv and the Tweeter API data. Each of the three dataset tables has a unique column called the ‘tweet_id’. The tweet_id is the primary key which connects the tables of the three datasets together.

Following the completion of the processes of data gathering, assessing, and cleaning, the datasets were merged to form a single master data file named 'twitter_archive_master.csv'. In addition to the earlier mentioned works done, analysis and visualization was also carried out on the master data and below are some of the insights generated from the analysis.

Insights

- It was observed that out of all image predictions recorded in the master data, 15.51% of the image predictions of the dataset was predicted wrongly while 84.49% of the dataset have correct predictions. This showed that a major percentage of the dataset had correction image predictions.

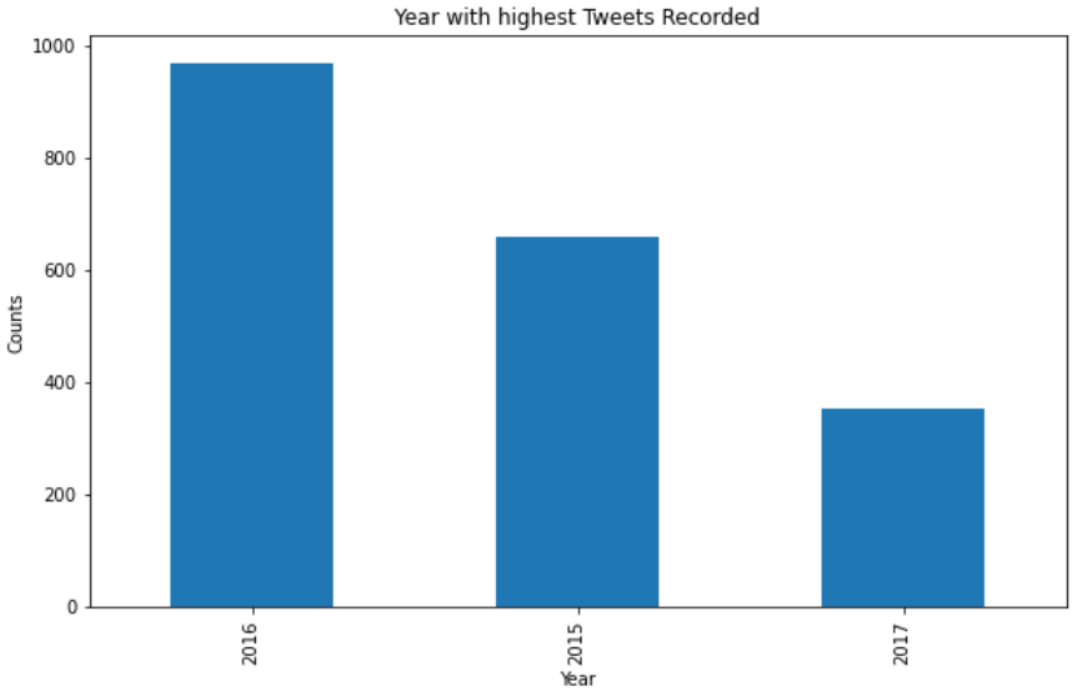
- However, in the analysis and visualization of the year with highest record of tweet, it was also observed that the year with highest number of tweets recorded was 2016 with a total of 971 tweets, followed by year 2015 with a record of 662 and lastly year 2017 with a record of 353.

- Furthermore, it was also observed that amongst all the tweet_ids recorded in the master data table, tweed_id 744234799360030481 recorded the highest number of favorite count with a total of 144235, while tweet_id 666102155909144576 has the least favorite count with a total of 66.

- Finally, in the master data table, it was also observed that the tweet_id 749981277374128128 has the highest rating numerator with a rating total of 1776 while we have two tweet_ids (835152434251116546 and 746906459439529985) with zero (0) rating numerator.

Visualization

The figure below represents visualization of years with highest record of tweets.



The figure above shows that the highest number of tweet was recorded in the year 2016.

Conclusion: This project gave me a complete and better understanding of the processes involved data wrangling.