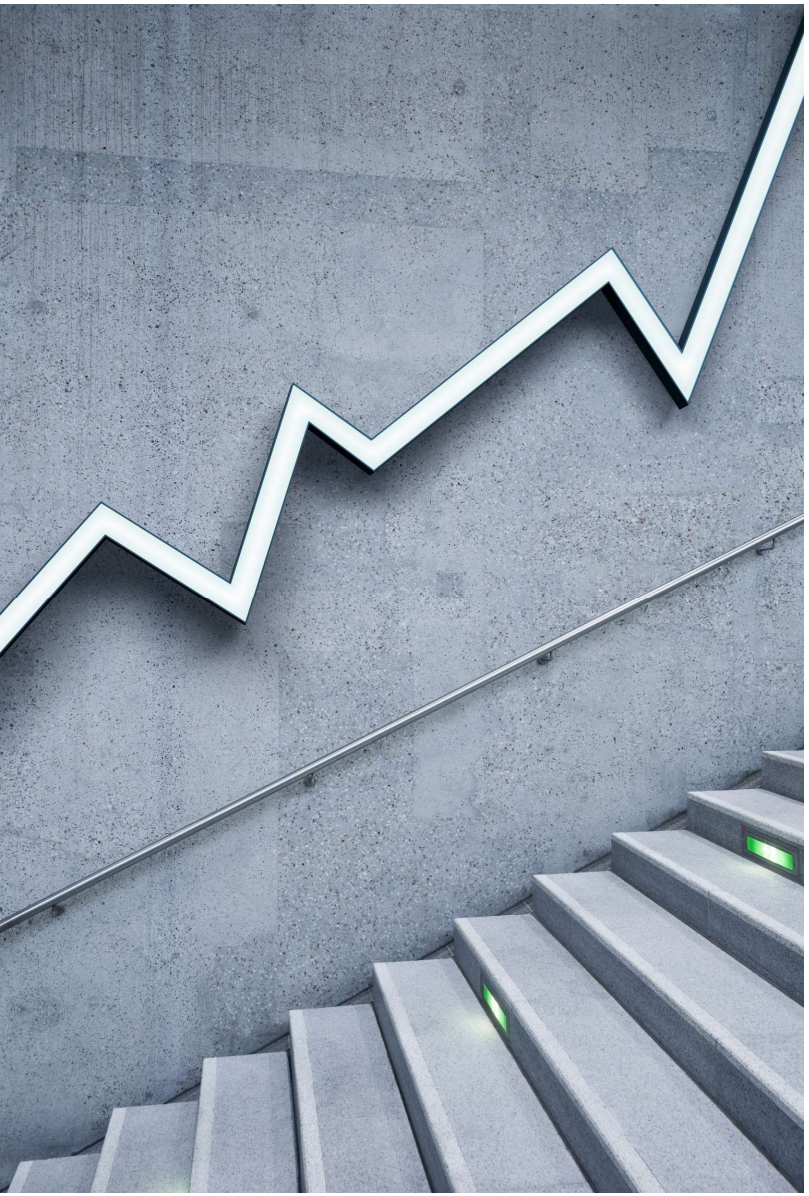


We Rate Dogs - Data Wrangling Analysis

MICHAEL STEPHENSON

UDACITY DATA ANALYST NANODEGREE

Introduction

This project represents the procedural steps of a data wrangling task for data analysis. The project data consists of a data set pulled from a Twitter account called “We Rate Dogs” via its archive of tweets saved to a csv file. The Tweepy Python Library is also used to make accessing the Twitter services possible. There are four stages to the study beginning with Gathering data in three diffent collections of information. This includes the tweet archive, an image predictions data set, and JSON api information pulled from Twitter.com.

The Assessing stage explores the data sets to gain an understanding of the information present in each collection, tidiness and usefulness of information is noted. If changes to the data set are required, the actions are noted at this phase. This is followed by the Cleaning phase which now rearrange information, make changes to table information and edit column names for easy understanding. Each part of the cleaning step is defined, coded, and tested for conclusions and results. Finally, the Analyzing phase will place the data into meaningful visualizations using .

Gathering

The necessary Python Libraries are imported, the Twitter Achieve data is read into the Jupyter Notebook along with the Image Perditions file and JSON api information collected about the “We Rate Dogs” account. The information is checked for correctness after loading.

Image predictions

In [3]:

```
#Downloading URL programatically
url = "https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv"
response = requests.get(url)

with open('image-predictions.tsv', mode ='wb') as file:
    file.write(response.content)

#Reading TSV file
image_prediction = pd.read_csv('image-predictions.tsv', sep='\t' )
```

Twitter API and JSON

In [4]:

```
# Query Twitter API
'''
auth = tweepy.OAuthHandler('unYL4mx7PPU1DlAKasamDsrIW', 'JLfXGfswKX3RMbjCWdG73mmFVggyCrBxGqXNGlUayXr69dRQ1H')
auth.set_access_token('1349844283818434565-Uo1l7h1DEe1wGJ2Bm0yVpy41B4A9R8', 'tfNU5oDKIoVc2FmqVZFwEPbJ7Oxe7Sa8LrisE6JhPMfsu')
api = tweepy.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=True)


test=[]
```

Assessing

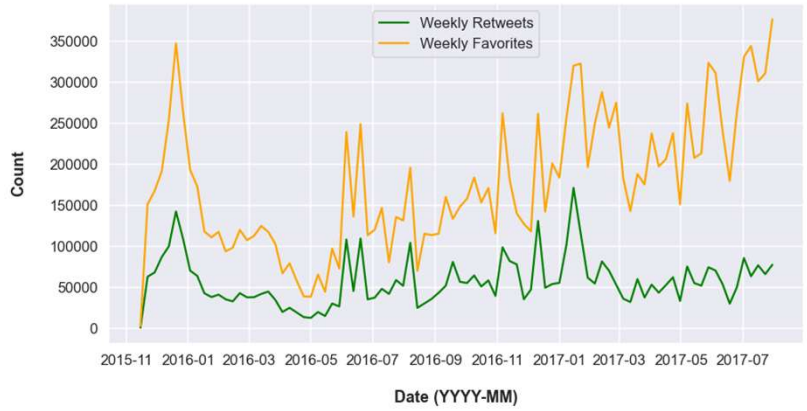
Once loaded into the project it the Assess section looks at the .head(), .tail(), and .description() information. This is required to identify useful information contained within each data set. It will be important to know which information to keep and which columns will not be needed or combined into new column data. The protentional for new columns can be assess also, however this section is only about looking at the variables before acting upon them in further sets.

Cleaning

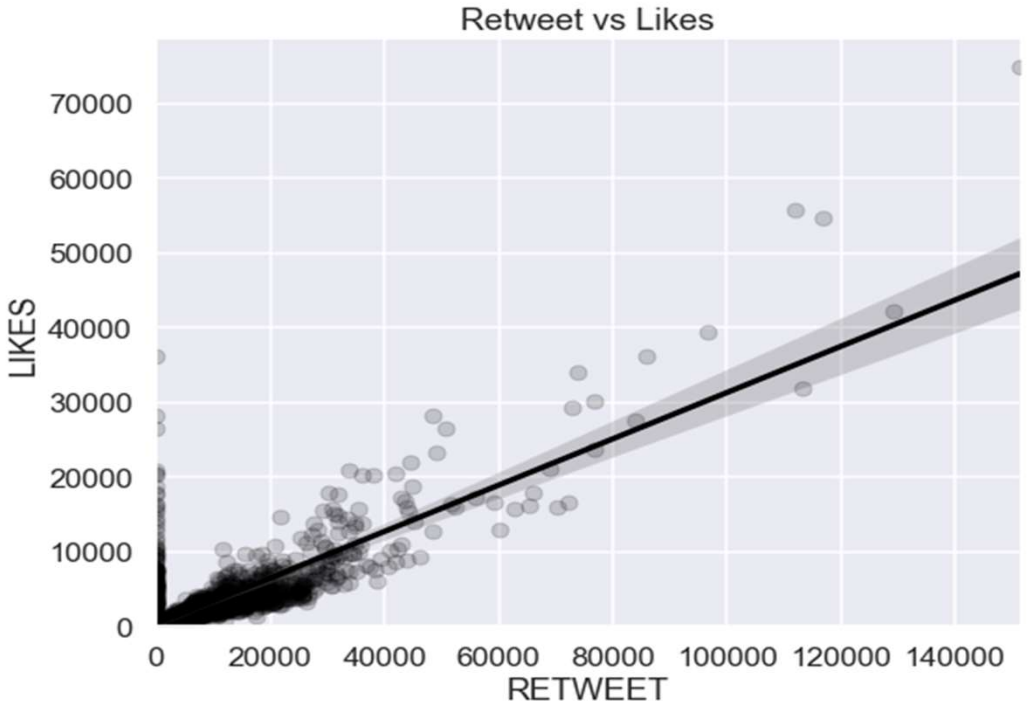
This phase on the wrangling project is probably the most involved because it seeks to change the structure of the data to something preferable for study and visualization. The data will be formatted for inclusion in various analysis tools. The required parameters for Python will be made accessible cleaning steps. The table data is first copied then each set is operated on for uniqueness in the data. Columns that are not required are removed. Since the table will be merged the “twitter_id” column is converted to string data as the merge point. Each cleaning step contains a definition of the step, the code used to complete the step, and the test of results after completion in order to be consistent and display results as changes are made.

Analysis

The final step in the project is to display the results of the cleaning step in useful visualization in order to gain an insight about the data collection. Since . The success of the account would be a useful metric to understand, the Retweet History is measured against the Favorites history in both a short weekly period and again in a longer yearly average.



Analysis

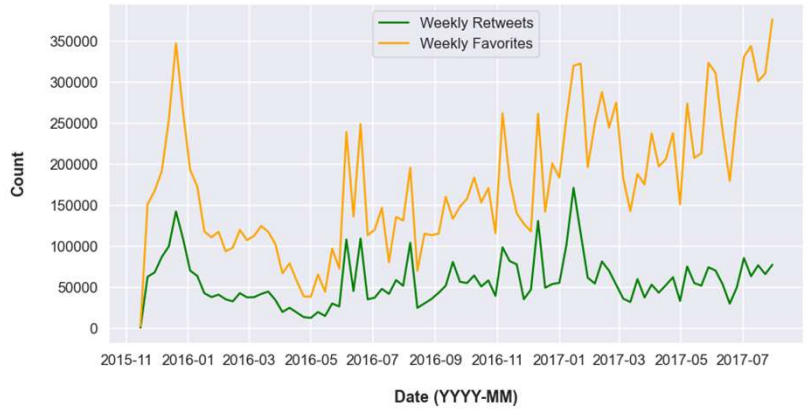


These metrics became identifiable once a trend was identified within a scatter plot of Retweets vs Likes.

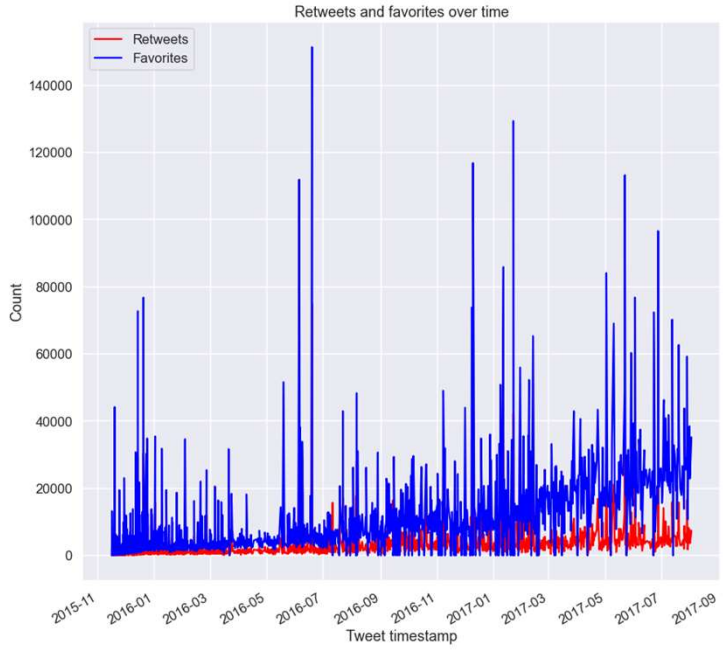
It clearly identifies a relationship between the two variables. Likes increase as Retweets occur.

Analysis Conclusion

In a weekly scale, the frequency of Favorites does have a similar trend line to the Retweets line. Thought the Favorites trend velocity seems to also be influenced by other factors as the heights of the trend lines do not mirror each other,



Analysis Conclusion



When the Yearly trend pattern is observed, though it does contain outliers, seems to reflect the same progression of events as the weekly scale did. The Favorites metric does appear to be more popular than the use of Retweet functionality of the Twitter platform. Though a conclusion might be made about the frequency of Retweets because it appears Favorites occur more often even when Retweets increase in a period.