Data Wrangle Report

Before the data wrangling process we need to import the necessary libraries such as pandas, numpy, json, requests, os.

Gathering Data:

There are three types of data being used in the project. First type of data is a CSV file and this file is directly access from Udacity's local drive and is read using the pands.read_csv() command. The second type of file is a tsv file and it needs to be downloaded programmatically from the internet. The url was given and we need to use the requests.get command to get the url and then open the url and save if as a file in our local drive for use to use pandas.read_csv to open the file and we need to use \t as the separator here because the file is tsv. The third type of data is a text file and we need to read the txt file line by line into a pandas DataFrame with tweet ID, retweet count, favorite count, and retweeted status. To generate the text file we had the option to use tweep API to download the texts or just directly access the text file. I prefer not to set up a twitter account due to personal reasons and thus I will use the additional twitter data provided ('tweet-json.txt) in the project and I will convert this txt file into a data frame. Once all the data files are read into the jupyter notebook, we are ready to assess the data.

To comment on the level of gathering difficulty, the first csv file is gathered and read most easily, the second tsv file added some additional effort because it has to be downloaded programmatically and saved locally before it can be read. The third file had more added difficulty because it is a text file which as block of texts and it is harder to read. It must be opened first and then each of the four data component to be extracted and saved into dictionary lists and then this list will need to be converted into a DataFrame.

Assessing the Data:

There are two types of assessment being used here 1) visual assessment and 2) programmatic assessment.

Visual Assessment

The purpose of the visual assessment is to display the data in jupyter notebook for us to look through the dataframe to see if there are any obvious issues. The visual assessment is a quick way to get familiarized with the data and to spot any obvious data issues. Some observations from visual assessment are the size of the data frame, whether there are typos or null values. More observations as follows:

Programmatic assessment

The majority of the data issues are assessed programmatically. Here are some of common pandas functions and methods used to assess the data in its entirety:

- info

- describe

- sample(5)

- value_counts

From these functions, we are able to get a detailed examination of the data frame. We are able to get the number of duplicates, null values, maximum, minimum and certain values that look invalid.

Here is an issue summary from the data assessment:

Quality issues

- Timestamp should be type datetime not object

- Rating numerator and denominator change to type float

- Data contains retweets, indicated by retweeted_status_id, retweeted_status_user_id, there are 181 of them

- Data contains reply tweets, indicated by in_reply_to_status_id, in_reply_to_user_id, there are 78 of them

- There are 59 missing records of expanded urls

- The source column is long and hard to read, replace with simple text

- Invalid rating denominator with values greater than 15 and needs to be verified

- Invalid rating numerator with values greater than 15 and needs to be verified

- duplicated image and prediction results as a result of the inclusion of retweets, 66 instances

- Dog breed prediction p1 has uppercase for some records and lowercase for some other records, this is not following the consistency aspect of data quality dimension

- Records where all three predictions are all false

Tidiness Issues

- There are four columns for the four dog categories:doggo, floofer, pupper, puppo. According to the rule of tidiness, each variable is one column, therefore these should all be in one column instead of four different columns

- the retweeted status are all false, it is not necessary to include this column

- this table contains the important information which should be included in the main archive table as well

Cleaning Data

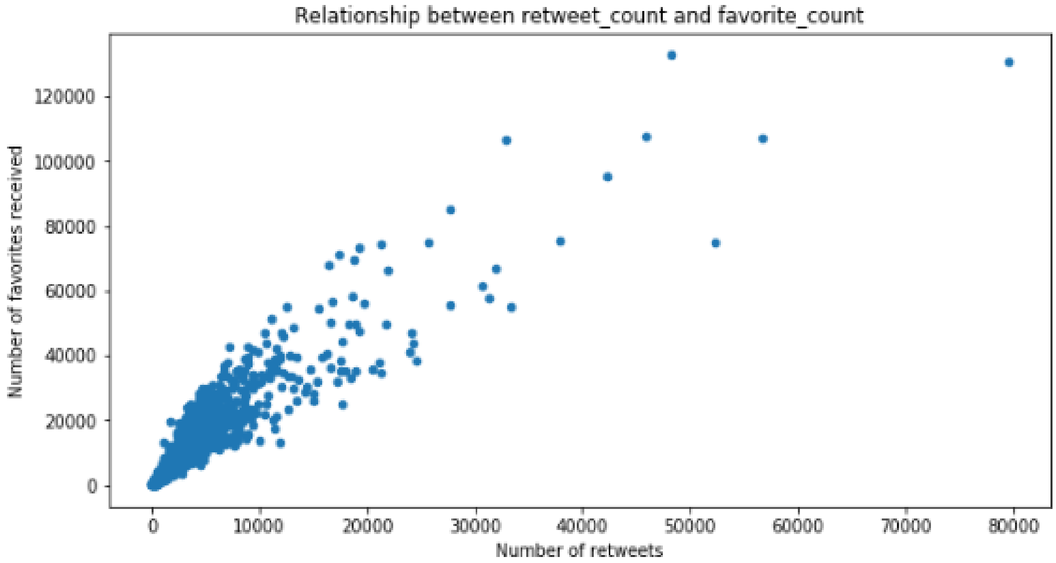
The data data cleaning portion of the project used the majority of the time and this section of the project is used to resolve the issues found in the data assessment section. The data cleaning uses a Define – Code – Test structure which first defines the issue, write the code to fix the issue, then test the fix. Some of the main technique used for data cleaning are .drop(), .query(), .str.replace(), .astype(), .at().

Sometimes simply adding the columns and then use the replace function is quite useful and is simpler than using functions such as melt or concat. I used this method to add the columns (doggo, puppo, floofer, pupper) and then did a str.replace to simplify the columns.

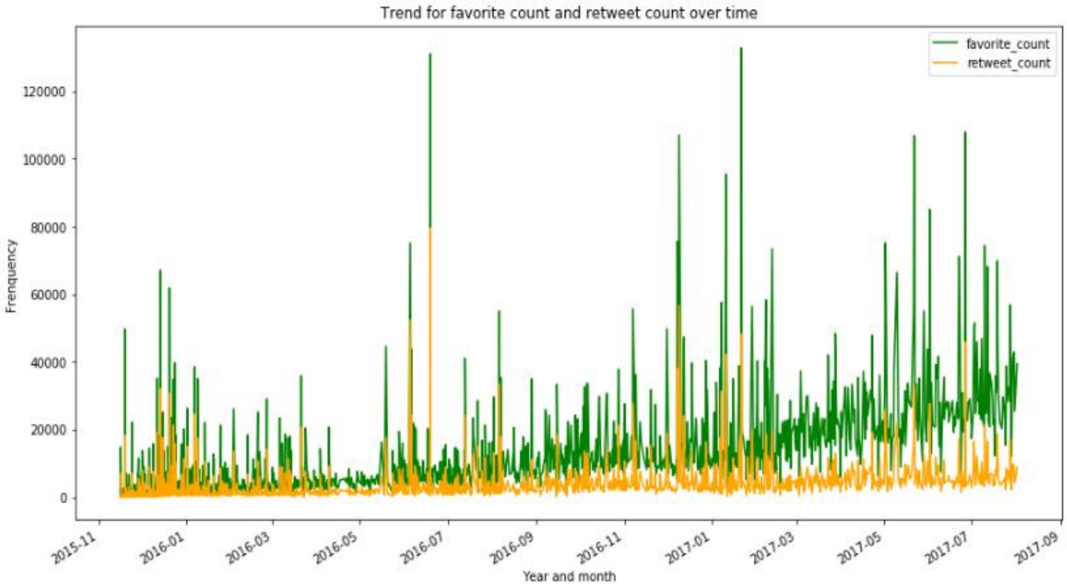
Data Analysis

After the data are cleaned and restructured, I saved the master data into a csv file. The cleaned data are then analyzed to generate some useful insights about the twitter account, here are some of the interesting findings. Please refer to the act_report.pdf for more interesting insights.

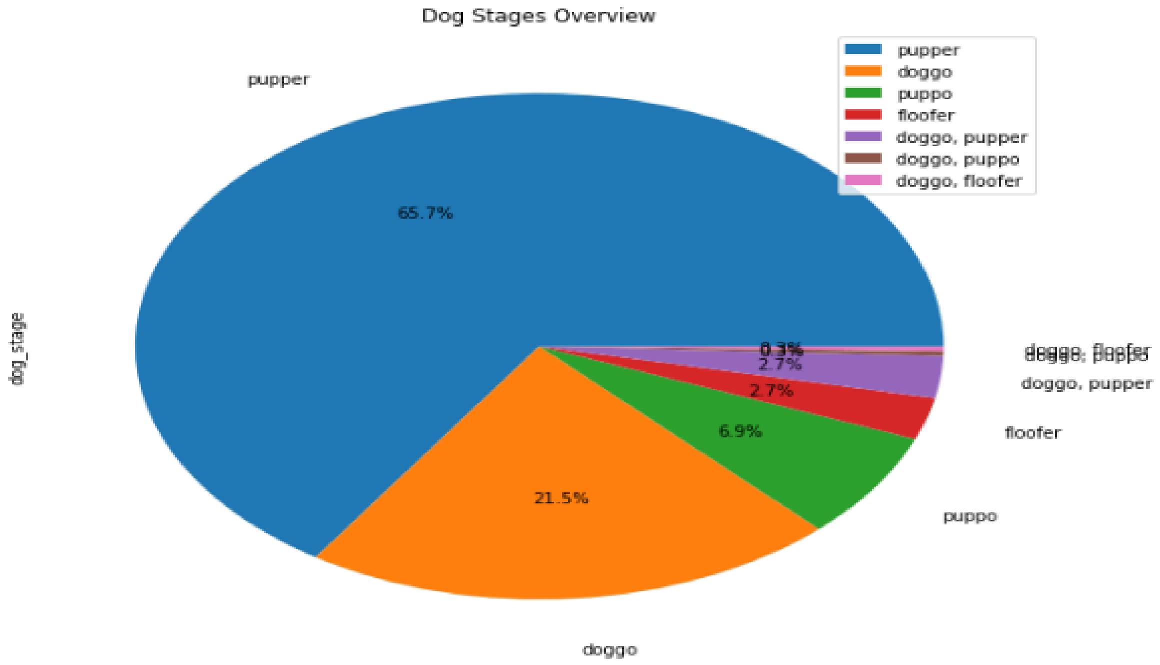
1) There is a positive correlation between number of retweets and number of likes received as shown in the scatter plot below:



2) The twitter account is getting more and more popular over time as indicated by the increasing trends for the retweet and favorite counts. Please see the line graph below:



3) The most frequent dog stage being rated is pupper, as indicated in the pie chart below:



Dog Stages Overview
pupper
doggo
puppo
floofer
doggo, pupper
doggo, puppo
doggo, floofer
dog_stage
pupper
65.7%
doggo
21.5%
puppo
6.9%
floofer
2.7%
doggo, pupper
2.7%
doggo, puppo
0.3%
doggo, floofer
0.3%