Data Wrangling Report

Mada AlAhmadi (Udacity Data Analyst Nanodegree)

- ## Data Wrangling:

1) **Gather the data:**
Start with download the csv file (twitter-archive-enhanced.csv) that provided by Udacity in the project page, then download programmatically by using requests library the image file (image_predictions.tsv) that contain images information and also provided at the same page, and data from Twitter API such as retweet count, favorite count, and tweet id are written to text file (tweet_json.txt) as json.

2) **Assess the data:**
I use methods such as duplicated(), value_counts, info(), sample(), and describe()..etc., to assess the data and find following:
First: Quality
    1- The timestamp must be datetime
    2- The retweeted_status_timestamp must be datetime
    3- Rewrite the tweet source, from iphone ,web...etc
    4- Remove the tweets without images and only keep the tweets with images
    5- Remove retweets
    6- Some incorrect dog names
    7- Retweeted data included in df (Twitter archive data)
    8- There is some missing values in name columns 'None'
    9- IDs should be sting type
    10- Replies are included df df (Twitter archive data)

    Second: Tidiness

    1- Merge df2 and image_predictions to df (Twitter archive data)
    2- There is must be (Dog Stages) coulmn that include (doggo, floofer, pupper, and puppo) as a single column

3) **Clean the data:**
It's the most important part of this project to clean the data and store it in new and proper form to make it useable. Here I programmatically clean the data to change some types, drop some columns that is not useful, correct some errors in data, and merge data frames.

4) **Store the data:**
I store all the cleaned data into csv file (witter_archive_master.csv)

5) **Analysis:**
I analysis the cleaned data through histogram plot to see the retweet_count & favorite_count also, I use two plots to show the retweet count and favorites count over the

time to know what is the most time that tweets receive highest number of retweets and favorites this kind of analysis can help any account to know what time to tweet specific tweet or what type of tweets that can receive a lot of retweets and favorites.