WRANGLE REPORT

Dalal AlAli

Introduction:

In this project we are going to wrangle and analyze the tweet archive of Twitter user @dog_rates. This account @dog_rates is a Twitter account that rates dogs. The account opened in 2015 by a man who called Matt Nelson, and has received international attention. By Using Python and its libraries,I will gather data from many sources then,I will assess these data and its quality and tidiness issues, then,I will clean it. This is called data wrangling. And I will document my wrangling efforts in a Jupyter Notebook.

The wrangling processes:

1- gather the data
2- assess the data
3- clean the data

Step 1: Gathering The Data

First, I downloaded and read file (We Rate Dogs Twitter archive) which contains the fundamental tweets info for all the tweets.

Second, by using the request library I requested the tweet image predictions file from the provided link in Udacity's servers. Which is https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.

Third, by using Python's Tweepy library I used the provided file from Udacity Because Twitter did not let me access to developer account.

Step 2: Assessing The Data

Quality issues:

1-the doggo, floofer ,pupper and puppo supposed to be merged in one column.

2-Replace the names 'DoggoPupper" to "Doggo", 'DoggoPuppo', 'Puppo', 'DoggoFloofer', 'Floofer'.

Tidiness issues:

1-There are some tweets that are retweeted and supposed to be excluded.

2-There are some tweets that are replies and supposed to be excluded.

3-tweet id is int and it is supposed to be string by using astype method.

4-the p1,p2,p3 supposed to start with capital letter.

5-Timestamp has wrong data type which is string and it is supposed to be datetime.

6-the unneeded columns in twitter_archive supposed to be dropped.

7-the unneeded columns in image_predictions supposed to be dropped.

8-change id name to tweet_id in tweets_json_clean.

Step 3: Cleaning The Data

1-I merged the doggo, floofer ,pupper and puppo in one column.

2- I Replaced the names 'DoggoPupper" to "Doggo", 'DoggoPuppo', 'Puppo', 'DoggoFloofer', 'Floofer'.

3-I changed tweet id from int to string by using astype method.

4-I changed the p1,p2,p3 to start with capital letter.

5-I changed Timestamp data type which is string to be datetime.

6-I dropped the unneeded columns in twitter_archive.

7-I dropped the unneeded columns in image_predictions.

8-I changed id name to tweet_id in tweets_json_clean.

9-I excluded some tweets that are replies and supposed to be excluded.

10-I excluded some tweets that are retweeted and supposed to be excluded.