Wrangle Report

Introduction:

In this project we will gather, assess and clean data. There will be visualization and analysis to help us reach to the best form of completing this project. The data will be provided offline from Udacity, because there is not enough time for me to contact Twitter.

Gathering Data:

I gathered data from three different file and insert them into three tables in matter to help me with analyzing and creating this report:

- twitter-archive-enhanced.csv
- image-predictions.tsv
- tweet-json.text

Assessing Data:

Here I tried to show all the possible information regarding these three tables. I used pandas:

- info
- head
- describe

After that I addressed data quality and tidiness issues:

Data Quality

1- Name is sometimes not an actual name.

2- Missing data in df file.

3- missing some expanded_urls.

4- Tweets with no images.

5- Text column contains the shortened URL to the tweet

6- In df in_reply_to_status_id & in_reply_to_status_user_id are duplicated.

7- In PixPred values are 2075 out of 2356.

8- In tweetstored had to change id to tweet_id to match other two tables.

Tidiness

1- Adding all the tables to one table.

2- Delete all the columns that won't be needed for this analysis.

3- Use melt for datatypes (doggo, floofer, pupper and puppo columns) to have columns dogs and dog_stage, then, drop dogs and sort dog_stage and remove duplicates.

Cleaning Data:

Tidiness was addressed here, by using:

- melt from panda to make one column from doggo, floofer, pupper and puppo columns
- Join all the three tables to one table
- Delete all the columns that won’t be necessary for this process.

made one clean dataset to be presented. Export new the result of dataset to a new csv file:

final_dogs_data.csv.

Visualization:

Three visualizations from our cleaned data, you will find them in the act report pdf:

- Most popular dog breeds
- Most popular dogs
- Retweet vs. Favorite