Data Analyst Nanodegree

Wrangle and Analyze Data

Name: Ibrahim Riyad Aldekhiyl

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

Data wrangle have 4 main steps:

1- Gathering Data
2- Assessing Data
3- Clean Data
4- Analyze Data (Data Insights and visuals)

In this project I have used data from Twitter account that interesting in rating dogs. (@dog_rates)

Here are the project steps:

1- **Gathering Data:**
   In this step, I've faced some issue in gather the data from Twitter API, then I choose to go with downloading the data directly from Udacity (twitter-archive-enhanced.csv) and I've download (image_predictions.tsv) programmatically.

2- **Assessing Data:**
   In this step, I've exploring the data and finding errors In the data. I've identify around 10 errors:
   <u>Data type issue:</u>
   - twitter: tweet_id,in_,in_reply_to_user_id, reply_to_status_id, retweeted_status_id, retweeted_status_user_id and in_reply_to_user_id shoud be
   - twitter: timestamp should be --> datetime
   - images: tweet_id shoud be --> object
   - tweet: retweet_count and favorite_count should be --> integer

Null issue:

- general: the naming of some columns are vague (not discriptive)
- twitter: names,floofer, puppo, pupper and doggo have none as a value for some records, need to be updated with NaN

Other issue:

- images: p1, p2 and p3 some enries with capital letter and the other with small letter (in inconsisty of the data)
- images: a lot of data are missing (have just 2075 records)
- tweet: some data are missing (have just 2347 records)

Tidiness:

- twitter: doggo, puppo, pupper and floofer can be merged into one column
- twitter, images and tweet: all the three tables should be merged (all of them describe same tweet)

3- Clean Data:
In this step, I've fixed all the findings or errors in the 2nd step.

4- Analyze Data (Data Insights and visuals):
In this step, I've identify some question then i answered then through out some visuals and graphs.