Ahmed Aldayel – Wrangle Report

Wrangle and analyze data project

Project details:

We have the following tasks in this project:

•Gathering data

•Assessing data

•Cleaning data

Gathering data:

In this project we can obtain the data from three datasets, so we have different dataset as following:
•Twitter archive file: the twitter archive enhanced.csv was provided by Udacity that can be downloaded.

•The tweet image predictions: what kind of the image which will be presented in each tweet according to a neural network. This file (image predictions.tsv) is hosted by Udacity's servers and can be downloaded programmatically using the Requests library and URL information.

.•Twitter API: Read (tweet-json.txt) file line by line to create date frame.

Assessing data:

After looking to the collecting data and applying programming methods such as:

1-head()

2-info()

3-Value_counts()

4-isnull().sum()

I have founded the following issues on quality and tidiness:

Quality issues have been discovered:

- in the name of the dog in 'df_clean' have some issues sometimes is written (a, an ,the) so we should change it to null value.
- Change 'None' in 'name',doggo','floofer','pupper','puppo' colums to null in df_clean
- The type of timestamp need to change from object to datetime in df_clean.
- we need to remove the following columns(`retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp`) because it is unuseful in df_clean.
- Drop the tweet that has tweet_id = '778027034220126208' because it does not have any rating in the tweet in df_clean.
- p1,p2 and p3 have inconsistent capital words in image_clean, so it should be modifieded
- p1_dog, p2_dog and p3_dog have unnecessary underscore so we should put space in image_clean.
- p1_conf, p2_conf and p3_conf have unobvious understanding so we should rename it.(images_clean)

Tidness issues I have discovered:

- melt the 4 colums [doggo','floofer','pupper','puppo'] to one colum for each row with the name Dog_stage..
- we need to combine tweet_clean with df_clean enhanced table.
- we need to combine image_clean with df_clean enhanced table.

Cleaning data:

Cleaning data steps:

1- define
2- code
3- Test

First, we should copy the data frame before cleaning and then start cleaning process, so In the "define" type: the issue has been defined, then I identified the dataset that the issues come from.

In the next step: I have put some special "code" (replace(),drop(),rename()), which will help us in the issues cleaning.

In the last step: the cleaning ability of the code has been tested, to check the output whether if it is corrected or not.