Wrangle Report

Introduction:

This report is part of the fifth project "We rate dogs – Data Wrangling". In addition, this report should explain the data wrangling efforts I have made in order to apply all requirements needed.

Data Wrangling process consists of data gathering, data assessment, and data cleaning.

Data Gathering:

Data gathering in this project involved obtaining three different dataset in three different formats from three different sources.

The first dataset was downloaded manually in csv file. It contains tweets archive for we rate dogs account. The dataset has large amount of data.

The second dataset was downloaded programmatically using Python Request Library. The file In tsv format (Tab Separated Values) contains dogs images predictions. The file hosted on Udacity's servers and I was able to collect the needed data from requesting the URL given in the project details page.

The third dataset is tweet_json txt file, I have applied for twitter developer account. Unfortunately, they did not approve my request even though I have answered all their questions based on what written on Twitter API section in the project submission. Therefore, I have downloaded the file manually, convert the txt file into json in order to use read_json function and get the three important columns needed which are tweet_id, retweet_count, and favortie_count.

Data Assessment:

After collecting and saving the three datasets into three dataframes, one for tweet_archive, the second for image predictions, and the last one for tweet_count. The main goal here is to assess data quality and tidiness.

Two types of assessment performed in each dataframe:

1- Visual Assessment.
2- Programmatic Assessment.

Visual assessment performed on both Jupyter Notebook and MS Excel. It helps to get an overview based on visualizing the data by eyes. We can see some issues found visually written on the notebook.

Programmatic assessment is by using python and pandas functions to help us assessing and getting accurate assessment. Using for example, .info, .describe, .sample. .duplicated. etc.

Quality issues are those issues related to data. For example, missing data, duplicated rows, meaningless values etc. it is also called dirty data.

Tidiness issues are those issues related to the structure. Where there is a rule of thump to make the data tidy, which is :

1- Each variable forms a column.
2- Each observation forms a row.
3- Each type of observation unit forms a table.

Data Cleaning:

After addressing the issues related to quality and tidiness at the assessment phase, we start clean our data step by step.

The cleaning phase usually consist of three sections:

1- Define the issue.
2- Code the issue in order to clean.
3- Test the code to confirm cleaning.

Conclusion:

After cleaning, we formed a new dataframe called master_df and store it as csv. In addition, we applied our analysis based on master dataframe.