Data Wrangling Report

By: Amani Abbas

Introduction:

This project is part of Udacity Data Analysis Nanodegree, and is focusing on data wrangling steps, which comes before analysis and visualization. The project will gather data that are related to WeRateDogs Twitter account, this account rates people's dogs with humorous comments, the rating in most of tweets is having a denominator of 10, however, the numerator in most of tweets is bigger than 10.

The wrangling process:

The wrangling process includes three main steps, and each step includes mini-steps.

First step: Gathering:

In this project I gathered data from 3 sources:

1- Source: File on Hand (twitter-archive-enhanced); this file has been giving to us from Udacity instructors. It is a CSV file that contains WeRateDogs twitter archive.
2- Source: Downloading Files from the Internet (image_predictions); this file is a TSV file that I have downloaded using this link; it contains three image predictions for each tweet id.
3- Source: Twitter API (tweet-json.txt); this file's data has been gathered using twitter API and saved in a JSON.txt file, then I only extracted two columns, tweet_count and favorite_count, from it and saved it in a data frame.

Second step: Assessing:

In this step I have assessed the data frames to look for Quality and tidiness issues, and followed two assessing ways:

1- Visually, opened each data frame in Jupyter Notebook and scroll over the tables to look for any quality or tidiness issue.
2- Programmatically: where I used Pandas methods to assess data, like: .info() .describe() .value_counts() .. etc.

The majority of issues were found in twitter archive data frame (df_tarchive).

I documented the issues to get back to it when I start the cleaning process.

Third step: Cleaning:

In this step I started to clean my data referring back to the issues I have documented in the assessing step.

The cleaning process includes three steps:

1- **Define:** to define each cleaning step; what am I going to do.
2- **Code:** to do the cleaning process using code.
3- **Test:** to test if the cleaning process went well and I have got the result I wanted.

Finishing:

At the end after having three cleaned data frames, I combined them into one data frame called (twitter_archive_master), then converted it to a CSV file and saved it.