Wrangle Report

Introduction

This project is a real case of how to gather, assess, clean, and analyze the data in the real world. The database used as an example is about the WeRateDogs Twitter user archive, this account has more than 7,572,000 followers, 9,500 tweets, and 141,000 likes.

The Data Gathering process tackle three different tasks, the first one download file from URL and later loading to the Jupyter Notebook, which requires a manual step, the second downloading a file programmatically, and the third gathering data from the Twitter API. This step has also required to save these data in a local machine.

Based on the data gathered, I have assessed the most evident issues and documented it to create a record of modifications.

Later, in Data Cleaning process I have fixed all identified issues, and I have also merged (the two downloaded files from the Data Gathering process) into one and added some missing values (from the archive downloaded from the Twitter API). The final data frame was stored as twitter_archive_master.csv.

In the Data Analysis and Visualization, which I have interpreted as Exploratory Analysis, I have tried to find some insights and patterns in my data.

1- Data Gathering:

- I have gathered the first file twitter_archive_enhanced.csv by downloading it manually then read it into my jupyter notebook using pandas library method read_csv.
- The second file the image_predictions.tsv, I have gathered it using the request package and the given url to it at Udacity server.

  Although, these two files have almost all the information from the WeRateDogs user, but:

- I gathered some missed variable, like (favourite_count & retweet_count), using the tweepy package.

2- Data Assessing:

Both visually and programmatically, I have assessed my data for quality and tidiness issues.

Systematically, I have documented problematic issues to clean them in the next step.

3- Data Cleaning:

I have tackled quality issues first, then the tidiness ones.

Quality issues, related to missing values, inaccurate names, wrong data types

Tidiness issues, I have merged information from img_prediction table and json_data table to the main twitter archive table as the all related together, dropping any unrelated access data.