Data Analyst NanoDegree Program

Project #4: Wrangle and Analyse Data

by Samuel Jiménez Sanabria

Wrangle Report

For the WeRateDogs Tweeter analysis I have first gathered data from three different data sources. Then I have assessed the data both visually and programmatically to clean it and to be able to get insights and to create visualisations.

The process has been as follows:

1.- Gathering data: The data has been gathered from three different data sources. First I imported data from a .csv file called **twitter_archive_enhanced.csv**. This archive contains basic tweet data for the WeRateDogs account tweets. The second data source imported was the **image_predictions.tsv** file. This file contains additional information and was imported using the 'requests' library. Additional data was gathered via the Tweeter API, saving .json data using Tweepy to a .txt file that was then used to create a dataset. In total three datasets were created from the sources: Tweeter archive: **tweet_df**, Image predictions: **pred_df**, and Twitter query : **qr_df**

2.- Assessing data: The data was visually assessed using the **.head()** function on all datasets. Other functions like **.info(), is_null(), .notnull(), .sum(), nunique(), .duplicated(), value_counts(), or .shape** were used to assess the data programatically. After assessing the data the quality issues that I chose to work with were the following:

Issue listing

Quality Issues

Enhanced Twitter Archive (tweet_df)
* Some numerators are incorrectly extracted
* Some column types are wrong

* Dog names wrong or missing
* Some tweets doesn't have pictures (expanded_urls)
* Some tweets are replies (in_reply_to_status_id)
* Some tweets are retweets (retweeted_status_id)
* Source columns very dirty
* Some columns are not needed for the analysis

Image Predictions File (pred_df)
* Fix column datatypes
* Names on P1, P2, P3 columns are inconsistent

Twitter API Query (qr_df)
* Tweet ID column not named tweet_id as in the other two dataframes

Tidyness Issues

Enhanced Twitter Archive (tweet_df)
* Dog stages are stored in separate columns
Data is spread across different datasets.

3.- Cleaning Data: Before cleaning the datasets, copies of the original datasets were created: Tweeter archive: **tweet_df_clean**, Image predictions: **pred_df_clean**, and Twitter query : **qr_df_clean**

Each dataset was then cleaned individually using the Define, Code and Test workflow procedure

4.- Saving processed data: The first step here was to merge all three datasets into a **merged_final** dataset. Then, after an adjustment in the columns order, this dataset was saved to disk as a .csv file called **twitter_archive_master.csv.** A database file **twitter_archive_master.db** was also created using SQLAlchemy.

5.- Visualising data: Some visualisations were created querying the database file and saving the queries as small datasets. Visualisations were created using Matplotlib and Seaborn. Other visualisations ere created directly from the **master_df** dataset.

6.- Regression Model: Simple linear regression model was created using the Statsmodels Python Module.