wrangle_report

April 21, 2022

0.0.1 Project: Wrangle and Analyze Data (@ WeRateDogs Twitter Archive)

To start with, this project entailed wrangling, analyzing, and visualizing the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog. These ratings almost always have a denominator of 10. The numerators are, however, almost always greater than 10. As in 11/10, 12/10, 13/10, etc. This unique rating system is a big part of the popularity of WeRateDogs. WeRateDogs has over 4 million followers and has received international media coverage.

My data wrangling process for this interesting project was carried out in five steps, namely; - Data Gathering - Data Assessing - Data Cleaning - Storing - Analyses and Visualization.

I carried out the project in its entirety inside the Udacity classroom on the Project Workspace: Complete and Submit Project page using the Jupyter Notebook provided there. I made use of Python and its libraries. Common libraries used include pandas, NumPy, requests, tweepy, json, and IPython.

Data Gathering: In all, the project required three datasets from three different sources in various formats. The methods required to gather each data were different. One dataset, namely, twitter_archive_enhanced.csv was already provided. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively to use in this project. This archive contained basic tweet data (tweet ID, timestamp, text, etc.). For this file, the download was done manually by clicking on a given link. Once downloaded, I uploaded it and read the data into a pandas DataFrame.

The second file image_predictions.tsv file was downloaded programmatically using the Requests library and a given URL. And the last file was acquired from Twitter API. Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. Afterwhich, the file was read line by line into a pandas DataFrame only extracting the tweet ID, retweet count, and favorite count.

Data Assessing: after gathering all three pieces of data, I went ahead to assess them visually and programmatically for quality and tidiness issues. Common assessing codes in this part included; .info(), .describe(), .head(), .tail(), .value_counts(), .unique(), .isnull(), etc And some issues I came across are: 1. presence of outliers in rating values, retweets, missing data, inaccurate dog names and data types, the need to merge some columns and tables, etc.

Cleaning Data: This section was code heavy as it entailed cleaning all of the issues documented while assessing. I started off by making copies of the original data before cleaning. The Cleaning step included merging individual pieces of data according to the rules of tidy data. During cleaning, I used the define-code-test framework.

Storing: The resulting data was then stored as twitter_archive_master.csv.

Analyses and Visualization: Is the last step where I analysed the resulting data and provided insights into my analysis.