wrangle_report

February 7, 2019

1 Wrangling efforts report / Ayah Ahamdan

The main objective of this project is to apply what we learned from Udacity (Data Wrangling) into a real dataset by performing some key tasks: gathering, assessing and cleaning data. In this project, we used a dataset from the tweet archive of WeRateDogs account. This twitter account rates people's dogs with humorous comments. The tweet contains a rating of the dog with an image. A big part of the popularity of this accounts is that ratings have a unique system, where the numerators can be greater than the denominators. To get insights on the dataset, some wrangling efforts needed to be done to be able to get an analysis.

1.1 Gathering Data

There are three different datasets needed to complete this project:

1. **Twitter archive file:** This CSV file "twitter-archive-enhanced.csv" was downloaded manually from Udacity. It contains the tweets' text which was used to extract rate, dog name and dog stage.
2. **Image predictions file:** This file is hosted on Udacity's servers, and was downloaded programmatically by using requests library and the URL given. It contains information of the images in the tweets (like image url) and predictions of the dog types from a neural network.
3. **Twitter API & JSON:** This data was supposed to be gathered by using twitter API and tweepy library. Unfortunately, Twitter did not reply to my developer account application (my status is still pending). Since I don't have access to the API code, I downloaded the "tweet_json.txt" instead. This file was provided by Udacity. The text file was read line by line to a pandas DataFrame with the following columns extracted: tweet_id, favorites, retweets, followers and retweted_status.

1.2 Assessing Data

After having a look on the three datasets, there were a couple of things that needed to be assessed to have a high quality and tidy data while making sure that the key points are met (like including original tweets only). The quality and tidiness issues were documented to be dealt with in the "Cleaning Data" part.

1.3 Cleaning Data

First, a copy of each dataset was created to make a new clean data frame that I can play around with until I was satisfied with the final structure of the data. Then, the retweets were removed from the dataset to make sure that there were no repeated tweets and only original tweets were included. After that, all columns that were not needed in the analysis were dropped from the datasets. There were a couple of data types that were converted in order to make the data frame more consistent. After that, I deleted the duplicates in the datasets, to ensure accurate analysis. Also, there were some invalid values that were handled, like denominators other than 10 and invalid name inputs.

Finally, the three datasets were combined and stored in a new CSV file, and some questions were asked to start performing some analysis and get insights through visualizations.