Data Analyst Nanodegree

Project Report: Wrangle WeRateDogs

Introduction

In this project, I did extensive work of data wrangling, analysis and visualization. I was glad that I have the opportunity to complete this project which tremendously sharpens my data wrangling skills. My project is structured as the following manner:

- Data gathering
- Data assessing
- Data wrangling
- Data analysis and visualization

Data Gathering

I gathered data from multiple sources. First, I read in the archived tweets csv file as a data frame. Next, I programmatically downloaded the neural network predictions of dog breed from an URL and loaded in as a data frame. Lastly, I used the tweepy library to query the Twitter API, storing the necessary information as a json file. I extracted the twitter ID, likes, retweets of the json file, imported them into a list and created a new data frame out of it.

Data Assessing

Prior to conducting any wrangling works, I proceeded to identify issues with the dataset after assessing the three data frames individually and programmatically. I discovered quality issues such as incorrect data types, rating denominators that are not 10, the presence of retweets, and so on. I have also spotted tidiness issues such as that all of the three data frames should be merged into one. All of the issues had been documented in the notebook to guide my wrangling process.

Data Wrangling

Here, I will guide you through my wrangling journey and briefly described each step I took in my wrangling notebook:

- I removed the retweet rows by resetting the data frame to include the rows that have null values in the '*retweet_status_id'* column.
- I merged the four columns titled *'doggo', 'floofer', 'pupper'* and *'puppo'* into one column titled *'dog_stage'* since these four columns represent the same variable. Due to not all of the rows having a *dog_stage* value documented, I set rows without a dog_stage to null. I also renamed the *dog_stage* values that have multiple labels.
- I addressed the tidiness issue by merging all three data frames into one, by using the pandas's inner join command on *tweet_id*, since all three data frames have the same corresponding *tweet_id*.
- I proceeded to correct the data types of the *'tweet_id'* and *'timestamp'* column of the merged data frame.
- I then dedupe the data frame, specifically the *'jpg_url'* column. However, I found that merging the data frames likely resolved this issue.
- I then adjusted the *'rating_denominator'* column by setting them all to 10.
- Next, I extracted the image prediction of the dog breed by selecting the label with the highest probability for each tweet. If the neural network failed to predict a label of dog breed for a tweet, I then set its *dog_breed* value to be 'No Prediction'.
- I have also fixed invalid names such as 'a' by converting these names to null.
- Lastly, I removed unnecessary columns prior to analysis just to make the data frame more concise.

Data Analysis and Visualization

Lastly, I conducted analysis and visualization of the merged data frame. I discussed questions that I had posed earlier and extracted key insights of the dataset. For instance, I found the 10 most frequently predicted dog breeds, the number of dogs grouped by stage, etc.