UDACITY WeRateDog DATA WRANGLING PROJECT

Introduction

- This is a Udacity project aimed at implementing Data Wrangling part of Analysis. We will utilize data from 3 sources. It will cover the process of gather, assess and clean the data.

Data Gathering

1) **twitter_archive**: It is the WeRateDogs Twitter archive provided by Udacity and I have downloaded on local machine.

2) **image_predictions**: It is the predictions of the tweets about what is breed of the dog using a neural network. It has been downloaded using a URL provided by Udacity using the request library.

3) **tweet_1** : using the twitter API and tweet ids, we have made a JSON file storing the tweets.

Data Assessing

Now we will proceed to cleaning the data. I am mentioning a brief plan on how will we go about cleaning all our 3 datasets. Since we have seen above that tweet_1 data has a completed data and is something that we have formed, we will not have to explicitly clean it.

1. Tidiness Problems
   - We can also think about merging the tweet_1 with twitter_archive table to get twitter_archive_master table.
   - We then merge the twitter_archive_master with image_predictions

2. Quality Problems

twitter_archive
- Remove all the retweet data and replies
- Removing tweets that don't have an image
- Removing data without the image
- Categories of dog are mentioned in 4 different columns, we need to melt them to form a new column which has all 4 categories in it.
- Reducing 3 columns of confidence to only one with that of correct prediction confidence.
- The urls are very long and not really human readable.
- The columns 'rating_denominator' should have standard value of 10 like a 5/10 or 6/10.
- Converting numerator rating to decimal type.
- The column names p1 and p2 are not intuitive.
- The predicted dog breeds have both upper and lower case for first letters.

A good practice for cleaning data is making a copy and then trying to amend that new table so that we can have a backup.

Data Cleaning

In this section, we will try and provide programmatic solution to each of the problems mentioned above in Data Assessing section. During implementing the code to solve the issues, I have also tested for the same to verify what is being implemented.

Analyzing the Visuals

- We can infer that the dog level called as 'pupper' (a small or young dog) is the most popular dog stage. It is followed by 'doggo' level and 'puppo' level but are significantly less. This can be attributed to the fact that people like dogs more when they are cute and really small/young. Although the visual gives us some idea about the situation, but due to lots of data missing, we cannot confirm this finding.
- we can tell that there is a single source among all the sources that is majority of the source which is 'Twitter for iPhone'.
- A fair insight or let's say hypothesis can be stated as 'the more popular tweet gets more retweets' and which can be deduced to be true in general. Even the coorelation coefficient is 0.8 which indicates the same.
- The pie chart reveals that nearly in 2 out of 3 cases, the predictions are correct. The results are low considering a deep learning model. Also the confidence level for the algorithm is considerably high.