Wrangle_Report

Project objectives:

The project main objectives were:

1. Perform data wrangling (gathering, assessing and cleaning) on the provided sources of data.
2. Store, analyze, and visualize the wrangled data.
3. Reporting on:

    1. data wrangling efforts.

    2. Data cleaning and combining.

    3. data analyses and visualization.

Step 1: Gathering Data:

In this phase, the three pieces of data were gathered and represented as pandas data frames:

- The first data set to gather data from was named “WeRateDogs Twitter archive” and was a csv, so I used pandas to read it naming it df_arch (file was handed, manual download of 'twitter- archiveenhanced.csv' from udacity website).

- The second data set to gather data from was named “ tweet image predictions” and was a tsv, so I used pandas to read it naming it df_pred (file on hand, manual download of 'image-predictions.csv' from udacity website).

- The Third data set to gather data from was named “ tweet json” and was a txt, so I used pandas to read it naming it df_pred (file on hand, manual download of 'tweet_json.txt' from udacity website).

Step 2 and 3: Assessing and Cleaning Data:

While working with data, a number of observations were made.the observations were solved in actions taken in the Cleaning Step. the observations that were made is there is a number of problems In each data set. These problems was two kinds (Quality, Tidiness) and the quality issues were:

1. Inconsistent naming conventions. This problem I solved by capitalizing every first letter in the names columns so all be the same.
2. Missing values in the name column. This problem was solved by filling

the missing values in name column with unknown.

3. Missing or incomplete information in df_arch. This problem was because there was columns which is a combination between decimal and integer values. So this was solved by making them all one kind of values.
4. We use the tweet ONLY not the retweet there for we should remove those from the table. This was solved by dropping any retweet related columns.
5. We use the tweet ONLY not the reply to the original tweet therefor we should remove those from the table. This was solved by dropping any reply related columns.
6. Inconsistent formatting. This problem was because some data had spaces between words and some had "_". This problem was solved by making any "_" a space in every column.
7. The data set provides no context about the images, such as their source, resolution, or any additional meta data.
8. Some rows have missing or incomplete information in df_pred. This problem was solved by dropping unnecessary things and filling some things.

The Tidiness Issues were:

1. doggo, floofer, pupper and puppo are all dog stages and should be combined to a unique column.
2. All datasets need to be combined.

Step 4: Storing data:

In this step I combined the three data sets into one and gave it a name which is "df" as it has all the data frames in it and store it into a csv file named "twitter_archive_master.csv"

Step 5: Analyzing and Visualizing Data:

In this step I analyzed the cleaned data to get informations about the new data frames and the insights were:

1. Skewed distribution: The distribution is highly skewed to the right.
2. Most tweets have less than 10,000 retweets.
3. High confidence in positive predictions.
4. Low confidence in negative predictions.
5. A significant portion of tweets have a high retweet count.

And the visualization concluded that:

1. the majority of dogs in the dataset have a rating numerator between 10 and 15. There are a significant number of dogs with a rating numerator of 12 and 13.
2. a smaller number of dogs with rating numerators below 10 and above 15.
3. the majority of dogs in the dataset are rated as good dogs, but there is some variation in ratings.