Udacity - DAND

Project 7

Wrangle and Analyze Data

OVERVIEW

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "**they're good dogs Brent**." WeRateDogs has over 4 million followers and has received international media coverage.

Your goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "*Wow!*"-worthy analyses and visualizations.

Data wrangling, which consists of:

- Gathering data
- Assessing data
- Cleaning data

Wrangling

1. Gathering

Wrangling this dataset begun with downloading the data from the twitter using the Twitter API. A Twitter app was created for this project and its keys were used to authenticate and download the data. This took a fairly large amount of time to download. There were a few exceptions during the first run of the API. This could have been due to the URL's unavailability at that moment. An **Exception list** was made and that list was re-run. This time, there was only one instance failure in the exception list implying that the tweet was unavailable permanently. I then downloaded the **'image-predictions.tsv'** provided by Udacity.

2. Assessing

The data was assessed with pandas. There were many unwanted junk in the data. Many columns had missing values. NULL values were present in columns that were important for analysis. Data types of key data were messed up. Redundant columns were present and had to be removed. The datasets were now merged with **TweetID** as the **Foreign Key**.

3. Cleaning

QUALITY

1: This column was simply a replication of the index in floating point representation format. This also has NULL values hence using this might bring us confusion to our analysis, hence it had to be removed.

2: The given timestamp might cause confusions when analyzing since they had inconsistent format and many NULL values. Hence they had to be converted to pandas date-time format, which is the standard for analysis.

3: Getting rid of the rows that can affect our analysis on retweets. There are many rows with NULL value in the retweet column

4: Unnecessary columns are removed from our dataset. E.g.: ALL TYPES OF URLS.

5: ‘in_reply_to_status_id’ and ‘in_reply_to_user_id’ are found to be totally inconsistent and they are of no use to the analysis. Hence there removal was performed.

6: Dropping the four ‘dog stages’ columns that were processed for tidiness.

7: Removing the ‘image-predictions.tsv’ data (predictions 2, 3 and 4) as only the first prediction was taken into account in a separate column.

8: To remove null values and useless data by dropping the rows

TIDINESS

TIDINESS 1: Condensing the four dog type columns into one column helps us to analyze better and easier.

TIDINESS 2: Condensing the dog breed prediction columns, by considering only the 1st and best prediction. This helps us to analyze better and easier.

Conclusion

Real World Data is almost never comes clean. The Twitter scrapped data-set has now been wrangled and is free from null values, missing data and inconsistency. It is now ready for Analysis.