wrangle_report

January 8, 2023

0.1 Reporting: `wragle_report`.

The dataset that I wrangled was 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.

In the project, I wrangled `WeRateDogs` Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. I then made `additional gathering`, then `assessing` and `cleaning` which was required for **"Wow!"**-worthy analyses and visualizations.

I firstly started by importing the necessary data analysis libraries and packages for the project. I then downloaded the twitter archive data directly from the internet. I then used the `requests` library to download the image prediction file before making use of the twitter API to collect the twitter data to be used in the project. I then proceeded to access the datasets using the methods taught in the classroom including *visual assessment* and *programmatic assessment* methods. In the visual assessment step, I just briefly skirmmed through the dataset while in the programmatic assessment was were the most work was done. I made use of the popular pandas methods including `tail`, `head`, `describe`, `isna`, etc to explore the datasets. Afterwards, I explored the data to check for quality and tidiness issues. I noted down the various issues before proceeding to clean and address those quality and tidiness issues. The **cleaning stage** was where the vast amount of efforts was put in to evaluate the data. It took a lot of time as I ensured I checked all forms of irregularities in the data and them addressed them. In the end, I was finally able to merge all three datasets into a clean and fresh data to now store it into a csv file called `twitter_archive_master`. I then moved on to analyzing and visualizing the data to perform **exploratory data analysis** and **explanatory data analysis**. In this stage, I generated key insights from the data from asking questions such as: - most popular dog names - most common dog stages and dog breeds - average retweet and favorite counts of dog stages - average retweet and favorite counts of dog breeds - average retweet and favorite counts of dog names - most popular source of tweets - dog stages and types that where the highest rated - word cloud of tweet texts - word cloud of dog breeds, etc With my analysis, I was able to perfectly answer the questions I posed.

[ ]: