WRANGLING REPORT

INTRODUCTION

The project shows the data wrangling process for the tweet archive of the user @dog_rates which is also known as WeRateDogs. The account rates people's dogs, giving comments and various captions. This report covers the description of data wrangling process which includes gathering, accessing and cleaning techniques on the twitter_archive dataset provides, the image prediction dataset and the twitter Api dataset.

Data Gathering

- From WeRateDogs twitter archive a csv dataset was provided which consist of tweet data for 2300+ tweet from WeRateDogs which was loaded into the work environment using pandas
- The image prediction dataset consists of dog breed predictions in columns p1, p2 and p3 and their respective confidence level. The dataset also includes the image url and the number of images per tweet. All these were gathered using the request library gotten from the image prediction url provided for the project
- My twitter Api dataset was gotten form the text file provided for the project. Using the json library and latter appending the necessary column to my empty list previously created.

Accessing

After loading the three files into my workspace, I proceeded to visually and programmatically access my data. Visually by scrolling through the dataset and programmatically by using the .info and .describe function. Later proceeded in checking for null and duplicated values in suspected columns

Quality Issues

- Missing values in twitter_archive dataset
- in-accurate name like "a" and "an" in the name column of the twitter_archive dataset and image prediction dataset
- wrong data types
- Most dog stages are recorded as "None"
- Duplicated values in the jpg_url column"?

Tidiness issue

- Incomplete dataset which resulted to gather more data form twitter which will require merging the three dataset for this project
- Merging the dog breed column

Cleaning

The quality and tidiness issue were further cleaning programmatically using functions and according to the structure provided.

- Dropping unnecessary columns in the twitter archive clean data set
- Replace the None string to NA in the dog stages column and further dropping it
- Change the datatype of the timestamp column
- Correct the invalid name issue
- Drop duplicates to uniquely identify tweet_ids to remove retweets
- Created a new dog breed column with respective confidence level
- Merge all three datasets together using the pd.concat function