INTRODUCTION

The project “Wrangle and Analyze Data “ involves wrangling of data from various sources associated with tweets from theTwitter user @dog_rates, also known as WeRateDogs.WeRateDogs rate’s pictures of people’s dogs in a humorous manner, most often giving ratings higher than 10/10. After scraping the data, quality and tidiness issues were assessed and then cleaned.

The wrangling process performed in this dataset are:

- Gathering
- Accessing and
- Cleaning

GATHERING DATA

WeRateDogs data was gathered from 3 different sources:

1)The enhanced twitter archive file was provided and downloaded manually. This file includes various variables for each tweet including tweet id, timestamp, text, rating numerator and denominator, name, etc.

2)The tweet Image prediction file which was provided for Udacity Students.

3)For some weird reasons, I was unable to get the Twitter API keys so I downloaded the ‘tweet_json.txt’ file which was provided by Udacity for those that were unable to set up a Twitter developer account.

ACCESSING DATA

After gathering the data, I was able to access the data visually and programmatically with some python function like;

- .head() : This was used to access the first five records of each data.
- .info() : This was used to access the information of rows and columns of each data e.g number of rows, number of columns, and data type of each column.
- .sample() : This was used to access random records of each data.
- .value_counts(): This was used to count the values in each column.
- .duplicated() : This was used to check for duplicates.
- .isnull(): This was used to check for null values.

During Data Assessment, I found out that the data has some quality issues and tidiness issues like: missing values, duplicate rows,invalid data e.t.c.

CLEANING DATA

I was able to clean the tidiness and quality issues using some functions

- Deleting retweets using ‘.drop function’
- Removing columns that were not needed using “.drop()’ function.
- Change timestamp to correct date format using pd.to_datetime() function.
- Separating timestamp into day, month, year columns using datetime.
- Create one column for the different dog stages using pd.melt() function.
- Merge the clean version of twitter archive, image prediction and tweet file together using pd. Merge function.
- Correct naming Issues using ‘.replace()’ function.

3. The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers. Using machine learning techniques, the breed of dog was predicted based on the picture.

ASSESSING DATA

After the data was gathered, assessment was performed using the following methods:

- .head()
- .sample()
- .info()
- .value_counts()

Tidiness issues that were cleaned:

- Combining all dataframes together as they all contained information about the same tweets
- Combining 4 variables about dog type into 1 column: "dog_stage"

Quality issues that were cleaned:

- Data contained retweets
- Tweet id was the incorrect data type
- Timestamp was the incorrect datatype
- Name contained the string "None" instead of a NaN
- Name contained various inaccuracies which were regular lowercase words
- The name O'Malley was incorrectly extracted as "O"
- Rating numerators which contained decimals were incorrected reported
- Ratings are unstandardized
- Undesired columns present