WRANGLE REPORT

1. INTRODUCTION

About Data set:

- WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention for its popularity.
- WeRateDogs asks people to send photos of their dogs, then tweets selected photos rating and a humorous comment. Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as "13/10".

DATA GATHERING

Data will be gathered from three resources:

1. The twitter_archive_enhanced.csv file containing data from WeRateDogs Twitter archive.
2. Image-predictions-3.tsv: The tweet image predictions, i.e., the breed of dog (or other object, animal, etc.) that is present in each tweet according to a neural network.
3. Twitter API and Tweepy library to gather each tweet's retweet count and minimum favorite("like") count, and any additional data of interest.

Note: I downloaded tweet-json.txt from Udacity classroom because of a problem in my twitter account.

DATA ASSESSMENT

I used jupyter notebook and python data science tools for thorough assessment of datasets. Some methods that I used are :

1. df.info()
2. df.isnull().sum()
3. df.describe()
4. df[‘column’].value_counts()
5. df.sample() and many more.

While assessing the data I came across various issues in datasets. I classified these issues into two types: quality issues and tidiness issues.

Quality Issues:

1. For Twitter Archive Dataset

- The datatype of timestamp is object(string) instead of datetime object.
- Some rows still have the HTML tags in source values.
- tweet_id is integer which is why it is being rendered in scientific notation. Not clear to read and compare.
- There are retweets in the dataset(as for some rows retweeted_status_id and retweeted_status_user_id are populated with numbers instead of NaN).
- Dog Ratings are wrong at some places(wrong values for numerator and denominator). Ratings contain twitter links and fractional ratings were not stored properly.
- Many names for pets were peculiar(like 'are', 'and', 'mad').

- The dataset looks cluttered due to the large number of rows. Many rows will be not required for later analysis. We'll drop these columns after all above corrections.
- Parts of text in some tweets have been interpreted as retweet status ID.

2. For Image Prediction Dataset

- Many jpg_url values occurring twice. It implies there are many duplicate rows and the same images were analysed because of retweets.
- Many tweets contain animals which are not dogs or objects like ice lolly and basketball.

Tidiness Issues:

1. All tables should be combined in a single dataset.
2. A single column should be created for all 4 dog types namely: doggo, floofer, pupper and puppo.

DATA CLEANING

The Data Cleaning Phase is divided in three parts:

1. Define
2. Code
3. Test

For each of the above listed issues with the dataset, these methods were applied. Correcting the numerators and denominators was very difficult because it had many quality issues. There were many rows for re-tweets which had to be handled in all datasets. There were datatype issues.

After thorough cleaning all the unnecessary columns were dropped and three datasets were merged and saved in a CSV file.

CONCLUSION

This project is a great example of why data wrangling and cleaning is important. Data rarely comes in clean and tidy form. Hence data wrangling is important before any statistical analysis can be performed or a machine learning/deep learning model can be applied.

Although data wrangling is a tedious process, conducting it will ensure that the data collected is not outdated or irrelevant. Therefore, data wrangling provides credibility to data analytics. It picks the right data required in order to provide the necessary solutions to a problem.