WRANGLE REPORT

INTRODUCTION

The purpose of this project is to put in practice the wrangling steps which consist of 3 phases. They are:

1) Gathering,
2) Assessing,
3) Cleaning.

The dataset that is wrangled 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. This report briefly describes my wrangling efforts.

GATHERING DATA

We gathered data from three different sources in three different formats. The three data pieces of information were the following:

1) **WeRateDogs Twitter Archive**-

This file was provided to us. We downloaded it manually and fetched it in our project. This file was in .csv format and the name of the file was **twitter_archive_enhanced.csv**. It contained basic information about the tweets like tweet ids, timestamp of tweets, sources, names of dog etc.

2) **Tweet Image Predictions File**:

This file was about breeds of dogs i.e. what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file was in .tsv format and the name of the file was **image_predictions.tsv**. It was hosted on Udacity server. We downloaded it programmatically using the Requests library from the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3) **Retweet and Favorite Count File**:

The **twitter_archive_enhanced.csv** file did not include two important data attributes which were retweet and favorite counts of a tweet. So in order to get these two attributes, we used Python's Tweepy library. We used tweet IDs in the WeRateDogs Twitter archive, queried the Twitter API for each tweet's JSON data and stored each tweet's entire set of JSON data in a file named **tweet_json.txt** file. Each tweet's JSON data was written to its own line. Then we read this .txt file line by line into a pandas dataframe with (at minimum) tweet ID, retweet count, and favorite count.

ASSESSING DATA

Once the three dataframes were obtained, we assessed the data by two methods:

a) Visually: We observed contents of the dataframes by printing them in Jupyter notebook and also by opening associated file of the dataframes in Microsoft Excel software.

b) Programmatically: We observed the dataframes by using different dataframe methods like dataframe.info(), dataframe.value_counts(), dataframe.describe() etc.

Then we classified the issues present in dataframes in two categories:

a) Quality Issues: Issues pertaining to the data.
b) Tidiness Issues: Issues pertaining to the structure of the data.

CLEANING DATA

Before cleaning the data, we created copies of our dataframes so that if we mess up with dataframes while cleaning, we will still have our original dataframes with us. After creating copies of the dataframes, we made 3 sections for each issue. They were:

a) **Define**: In this section, we wrote the way by which issue would be solved.

b) **Code**: In this section, we performed our programming procedure to resolve the issue.

c) **Test**: In this section, we tested for the changes i.e. if the changes got reflected or not in the data. In other words, whether the issue got resolved or not.

STORE DATA

After performing cleaning, we saved the dataframe so that in the future, if we need the same data for the analysis, we won’t be in need to go through all the wrangling phases again.

CONCLUSION

For wrangling, we divided our task into four phases: gathering, assessing, cleaning and storing. We went through each phase and at the end, we obtained a clean dataset, ready to be used for analysis/exploration phase.