Wrangle Report

Data Wrangling Steps (Gathering,Assessing and Cleaning)

Wrangle Report:

```
The dataset wrangle in the project is the tweet archive of Twitter user
WeRateDogs, using TWitter"s API permission to obtain the user's Json data,
while being also provided with the image prediction table and tweet archive by
udacity. WeRateDogs is a twitter account with over 9 million followers, that
rates people’s dogs with humorous comment about the dog.
```

Gathering Data

Data sources are :-

- Scraping Data using TWitter API Access
- Using Python Requests Library to download data from internet
- Importing CSV file LocaL Computer and reading vis Pandas library

3 main Data Sources:

Data using the Twitter API Access

Using Twitter API access to query for each User (WeRateDogs) tweet's using Python's Tweepy library and then storing all in JSON data type file (tweet_json.txt file), with knowledge of python programming, a block of code is written and executed so that Each tweet's JSON data is written on a new line, thereby creating a list of dictionary which is then converted to "twitter_data" dataframe using the pandas library.which contains the columns :-'tweet_id', 'retweet_count', 'favorite_count', 'followers_count'.

Enhanced Twitter Archive

This "twitter_archive_enhanced.csv" which contains basic tweet data for all 5000+ of their tweets, was provided via a http link to me by Udacity.

Image Predictions File

using the request library to directly obtain/ programmatically download the tweet image predictions (image_predictions.tsv) file from Udacity, and it contains image predictions of dogs breed

Assessing data

Having 3 data frames after previously Gathering, its was then individually loaded using Pandas data frames for assessment. Each of the data frames(image_predictions.tsv, twitter-archive-enhanced.csv, twitter_data) which is visualy assessed, programmatically using codes for each dataframe like : .info(), .describe, .head, . sample(), etc. by this we could visually see that there are various errors such as: missing data, inaccurate data, wrong data types, wrong and somtimes impossible data such as ridiculous names,etc.

Cleaning Data

After Assessing the data and spotting several issues, we note it down, in groups such as structural, quality and tidiness issues, then we proceed to programmatically clean the dat. Below are various quality and tidiness issues encountered during the Data Assessing Phase The following quality and tidiness issues were:

twitter_archive_enhanced Quality Issue:

```
 ▪ missing data in the following columns [in_reply_to_status_id,
in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,
retweeted_status_timestamp, expanded_urls.]
▪ dog names: 'None', 'a', 'an'.[validity]
▪ tweet_id is an int (applies to all tables)
• duplicated data and Empty Rows (retweets, retweeted_status_id,
retweeted_status_user_id and retweeted_status_timestamp).
• Accuracy in retweeted_status_timestamp as an object while other dataframes
are as floats
▪ Time-stamp is an object
• Consistency (HTML tags)
```

image_predictions.tsv:

```
• Validity (p1, p2 and p3 columns having invalid data a dog photo as a
starfish, boathouse mailbox.)
• Consistency (p1, p2 and p3 columns have multi-word dog breeds and dog breeds
listed is all lowercase, Sentence Case.
```

twitter_data(tweet_json):

```
• Completeness (Missing Data)
```

cleaning action:

```
• Delete retweets columns alongside unuseful columns not needed for analysis,
Correct the dog types.
• Creating a new dog_breed column using the image prediction data.
• Create one column for the various dog types: doggo, floofer, pupper, puppo
Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id,
retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
• Change the timestamp to correct datetime format.
• Change tweet_id from an integer to a string.
• Correct naming issues and Standardize dog ratings.
•  Finally Merging all 3 clean Dataframes.
```

• Finally copying data frames and using the new copy for visualization.

In [ ]: