DATA WRANGLING - REPORT

Data wrangling process involves three phases. They are
1 - Gathering of Data
2 - Assessing of Data
3 - Cleaning of Data

Gathering of Data:

Data Collection for this project in three types to collect different types of data.

The first dataset (twitter_archive.csv), data that was sent to Udacity by WeRateDogs is downloaded manually by clicking the link provided in the resources section. It is a .csv file and is imported into our programming environment using pandas .read_csv() method.

The second dataset (image_predictions.tsv) is downloaded programatically by using requests library of python and the link that was provided in the resources section. Using requests library request is sent to the specified URL and the response is saved into an .tsv file locally on our system. The .tsv (tab seperated values) file is imported into our programming environment using pandas .read_csv() method thereby assigning its attribute sep to tab (/t).

The third dataset (tweet_json.txt) is downloaded programmatically by using tweepy library of python. Using the tweet_id from twitter_archive dataset, we have downloaded the tweets from twitter with that id in json format and saved it locally as text file. It is then imported into our programming environment using pandas .read_json() method.

Assessing of Data:

Assessing of data can be performed in two ways. They are
1 - Visual Assessment
2 - Programatic Assessment

Visual Assessing of Data:

Visual Assessment of data is done by viewing few rows of data in each dataset. By accessing each dataset, we identified there are many null values in twitter_archive dataset, no major issues were identified in image_predictons dataset and null values are identified in tweets dataset. More programatic assessment should be done to get clear picture about all the issues.

Programatic Assessing of Data:

Programatic Assessment involves observing summary and general information of data using pandas .describe() & .info methods. By going through each column programmatically, we have identified the following issues in each dataset.

Issues identified in twitter_archive dataset:

Quality:

1 - Source Column values should be sliced more, to get the main content.
2 - Missing values(NaN) are found in many columns in four id's, retweeted_status_timestamp, expanded_urls.
3 - Inconsistent values in rating_numerator and rating_denominator columns.
4 - Timestamp and Retweeted_status_timestamp columns are of object datatype. They should be of datetime format.
5 - Some of the names in the name columns seems to be invalid (like a, an, the).
6 - Seperate columns should be there for time and date for data in timestamp column.
7 - Keep only original tweets.
8 - Remove the shortened url at the end of each value in the text column.

Tidiness:

1 - Four columns doggo, floofer,pupper, puppo should be combined into a single column.
2 - tweet_id is of int datatype. It should be changed to string datatype.

3 - Four id's in the dataset are of float datatype. They should be of int datatype.
4 - Some of the columns in the dataset are not useful for analysis. They should be removed.

Issues identified in image_predictions dataset:

Quality:

1 - They are about 66 duplicates values in the image_predictions dataset in the jpg_url column (same jpg_url for different tweet_id's).
2 - 2356 rows of data in df_twitter_archive dataframe, whereas there are 2075 rows of data in the df_image_predictions dataframe.
3 - Condensing dog_predictions and confidence_level into one column each.

Tidiness:

1 - tweet_id is of int datatype. It should be converted to string datatype.
2 - Remove all the columns that are not necessary for our analysis.
3 - Predictions contains values(p1, p2, p3) with capital letters and some are not. They should be made consistent.

Issues identified in tweet_json dataset:

Quality:

1 - Many Null Values are present in the following columns.
contributors, coordinates, extended_entities, retweeted_status, quoted_status_permalink, quoted_status_id_str
quoted_status_id, quoted_status, place, in_reply_to_screen_name, in_reply_status_id, in_reply_to_status_id
in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, geo.
2 - Should change the column name ID to tweet_id.
3 - Extract the values from the user column in the dataset.

Tidiness:

1 - Id is of int datatype. It should be converted to string datatype.

2 - Remove columns that are necessary for our analysis.

Some General Issues:

Combine all the three tables into a single table.

Cleaning of Data:

Cleaning of data is performed by defining each problem, and then cleaning data programmatically, removing unnecessary columns in the table and last testing the data to check the above performed transaction on data.

After addressing every problem and cleaning the data, inner join is performed between these table on tweet_id as checking factor and combined into a single data frame.

Saving of Data:

In the final master data frame, index column is reset by keeping timestamp as index in our new master dataset. Then it is exported into a .csv file to a path in our local system using pandas .to_csv() method.

```
df_final.to_csv('twitter_archive_master.csv')
```

The import the dataset into our programing environment using .read_csv() method to perform further analysis.

Hence Data Wrangling part for our analysis is done. It can be an iterative one, because some time's analysis requires more data to be collected and to be wrangled again even after wrangling part is done, it depends on the question we will be tackling in our analysis.