**Wrangle and Analyze Data**

**-Introduction:**

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog\_rates, also known as WeRateDogs.It is a twitter account which rate peoples dogs with a humorous comments about the dog.

**Project Details:**

Wrangling process has mainly three stages

* Data Gathering
* Data Assessment
* Data Cleaning
* Storing analysing and visualising the data.

**Data Gathering:**

Data for Analysis gathered from three sources:

* Twitter API
* URL
* Manualy Downloaded Data (.csv)

**Assessing Data:**

After gathering this data devided this data into three dataframes depending on the sources they were gathered. Data Assessment is a process where we observe the data visually and programmatically for quality and tidiness issues.

**Observations :**

**Quality issues in twitter\_df dataframe:**

1) in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id , retweeted\_status\_user\_id, retweeted\_status\_timestamp columns have missing values and does not contain any useful data so remove them.

2) Doggo,pupper.floofer and puppo columns has missing values and they are denoted as None instead of NaN.

3) Timestamp and retweeted\_status\_timestamp columns must be datetime instead of object.

4) Remove unwanted columns from dataset.

5) Dogs name column contains random names.

6) denominator contains value=10 for 2333 results so considering 10 for other conditions as well.

**Quality issues in image\_predictions dataframe:**

1) p1, p2, p3: dog breed names are not all in lowercase.

2) p1, p2 and p3 should be categoral datatype.

**Quality issues in df\_api\_data dataframe:**

1) user\_favourites value is same for all rows, which seems incorrect.

2) date\_time should be of datetime datatype instead of object.

**Tidiness:**

1) doggo, floofer, pupper and puppo are different dog types which should be in a same column rather than different.

2) retweet\_counts and fav\_count should be combined twitter\_df instead of a separate table of tweet\_id.

**Cleaning :**

This part of data wrangling is divided into three sections: Define, Code and then Test the code.Data cleaning is necessary process as we have to make changes in the dataset for better analysing the data.In this process we need to make some important changes and what is required for the proper analysis.

**Changes have made to clean data:**

* in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id , retweeted\_status\_user\_id, retweeted\_status\_timestamp columns have missing values and does not contain any useful data so remove them.
* Doggo,pupper.floofer and puppo columns has missing values and they are denoted as None instead of NaN.
* Timestamp and retweeted\_status\_timestamp columns must be datetime instead of object.
* Remove unwanted columns from dataset.
* denominator contains value=10 for 2333 results so considering 10 for other conditions as well.
* p1, p2, p3: dog breed names are not all in lowercase.
* p1, p2 and p3 should be categoral datatype.
* user\_favourites value is same for all rows, which seems incorrect.
* date\_time should be of datetime datatype instead of object.