Wrangling report

By: *Shimaa M. Badawy*

After gathering the data from the resources given in the course and investigating the data it is found that the data has many issues. ***Tweepy API*** is used to get the data from Twitter. And ***Request library*** is used to get the data that is hosted by Udacity. During the wrangling process, these issues are found in the data.

- First issue, is that there are missing values in the given dataset.
- Second issue is in the ***'timestamp'*** column which has **object** as a type whereas it should have **date** type, and the same issue is found for the column ***'retweeted status timestamp'***.
  - These two columns are handled by converting them to ***datetime*** type.
- The ***'timestamp'*** column and the ***'retweeted status timestamp '*** columns also need to be separated to a date column and a time column.
  - This was handled by making a separate column for **time** and another one for **date**.
  - For ***'timestamp'***, it was separated to ***'tweet_date'*** column and ***'tweet time'*** column and then **dropped** the original column.***'timestamp'***.
  - For ***'retweeted status timestamp'***, it was separated to **'retweeted status date'** column and ***'retweeted status time'*** column and then **dropped** the original column ***'retweeted status timestamp'***.
- Another issue is that some columns need to have a more clear name.

- For example, **‘p1’** column, **‘p1_conf’** column and **‘p1_dog’** column . Same for **‘p2’** columns and **‘p3’** columns.
- This is handled by the following:
  A. ***‘p1’*** is converted to ***‘prediction_class_1’***
  B. ***‘p1_conf’*** is converted to ***‘confident_of_class_1’***
  C. ***‘p1_dog’*** is converted to ***is_class_1_dog***.
- Same is done for the **‘p2’** column and **‘p3’** column.

- Another issue is that Number of records in the tweetInfoDF is **2331** whereas it should be **2356**.
- And the same for the tweetInfoDF, the number of records in it is **2075** whereas it should be **2356**.
- Also one of the issues is that some images are not classified correctly.
  - For example, some entries have the **p1_dog** value ,**p2_dog** value and **p3_dog** value is equal to **False**.
- One another issue that the dog stages columns need to be merged in one column.
- Also the retweets related columns and rows need to be removed.