DATA WRANGLE REPORT

BY: ABDELRAHMAN HANAFY

GATHARING:

As the first step in wrangling gathering is important to get the data needed for analysis. So, this report data gathered from three different sources and with multiple extension. Those sources are downloadable file, online file, and API data. The efforts done to fetch all the data can be concluded in:

1. Downloading a CSV file and loading it into pandas dataframe.
2. Using requests library to fetch a TSV file then load it into pandas dataframe.
3. Using TweetPy to request the tweeter API data and load it into txt file then load the important data into pandas dataframe to save it into CSV file for easy visual assessment.

ASSESSING:

In this step of data wrangling, the data quality and tidiness issues should be found and summarized for cleaning step. The assessment goes through two main rounds: Visual assessment and programmatic assessment.
Visual assessment: In this type on assessment, I used MS excel to investigate the data gathered and have an overview about the problems of data such as unnecessary columns, trash or default records, and extreme or faulty values.
Programmatic assessment: On the other hand, this type uses the pandas dataframes methods to check the data such as duplicated, info, and describe to look at the duplicated data, data types, and the distribution of values.

Summary of assessment

This summary is divided into two main section quality and tidiness.

Quality

Archive dataframe

1. tweet_id data type is int instead of string.
2. timestamp is object but must be datetime.
3. NaN vlaues in the name + last 4 columns represented as 'none' string.
4. the column name has a lot of missing values and trash value such as 'a'.
5. tweets that are tweeted after 1/8/2017 that have no api data.
6. tweets that have no image prediction.
7. tweets that are not original.
8. source and expanded_urls are not used columns.

Image prediction dataframe

9. tweets that are not original.
10. the names of the columns are not representing.
11. the values of how confident the algorithm is not represented in percentages.

Tidiness

Archive and API dataframes

1. the last 4 columns are values in the form of columns.
2. rating Num and Den are stored in 2 columns.
3. tables archive and api are 2 tables for the same observation unit.

Image prediction dataframe

4. p1, p2, p3 are values represented as columns.

CLEANING:

In the last step, pandas tools used to fix the problems found in assessing process following the define code test method in addressing each issue. First, copy all the dataframes to start cleaning process. Second, define how an issue should be solved by stating the steps of the cleaning. Finally, save the cleaned data to use in visualizations.