Data Wrangling Report

Goal of the project:

wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

**Gather the data**

Here I have gathered 3 dataset from different sources one of them using an API provided by Udacity.

Files as follow :

1. twitter-archive-enhanced.csv
2. tweet-json.txt
3. image-predictions-3.tsv

**Assess Phase**

First of all, I have mainly worked on tweet-json.txt dataset as I have saw it has way more problem than others.

* Visual Assessment:

1. user column contain json object
2. retweeted, possibly\_sensitive\_appealable and truncated columns only contains False value
3. lang column has 9 languages
4. id and id\_str has the same value
5. extended\_entities conatin the image of the dogs in json object
6. entities contains json object
7. most columns are string
8. source have tags
9. we need to drop most of the columns as they don't have any useful values

* Programmatic Assessment:

Here I have used multiple methods to get an idea if there is any problem with the data I have found

1. Date is consider as string which is wrong
2. there are 300 missing values in extended\_entities
3. some tweets doesn't contain rating we need to Drop them
4. id should be string as we will not be having any computation on it
5. lang all of tweets in english doesn’t make sense to differentiate based on language we need to drop the column

Here is the list of the quality issue I have noticed :

**Quality issue :¶**

1- there are 300 missing values in extended\_entities

2- convert created\_at from String type to datetime type

3- there is retweets in full text remove them

4- id should be string as we will not be having any computation on it

5- some tweets doesn't contain rating we need to Drop them

6- get the specie name of the dog from full text column

7- no need for tags in source column

8- there is error in extracting rating for some rows that have decimal we need to correct them

9- ratings is int convert it into float

**Tidiness issue:**

1- extended\_entities have multiple values in each row, here we need to create a new dataframe for it and then get the data from it

2- User column has the same issue (later i realized we don't need this column)

3- Full text has multiple varible so it's not adhere to the Each variable forms a column. it has the name of the dog and the type of the dog and rating

4- there are 4 columns of dog stages, doggo, puppo, pupper and floofer. These violate the rule 1 of the tidy data. So merging them into 1 column will be a tidiness issue

5-The tweet\_id column should be named same in all the DataFrames and it's datatype should be same in all the tables

**Cleaning Phase:**

1- we need to get extended\_entities url for image of the dogs

2- we need to remove rows that don't have extended entites

3- we need to convert created\_at to date from string

4- we need to transform full text and extract from it the names

5- we need to drop any row that doesn't conatin image

6- we need to remove the tags in source column

7- Convert id column from int to string

8- The tweet\_id column should be named same in all the DataFrames and it's datatype should be same in all the tables

9- rating\_numerator and rating\_denominator both have int convert it to float

10- some rating which has decimal has wrong values we need to fix it

**Conclusion:**

What I have learned in this project that data cleaning is one of the most important phases in any data analyst, it is not an easy task to do as I have thought before. I have worked in this project a lot and there is still too much things to do bring it up to the best clean data it can be.