Wrangle Report

1. Gather:

In that project we gathered data from three different sources.

● tweet_archive data downloaded from Udacity servers.

● json_df data extracted from json.text that we got from the Twitter page called ‘WeRateDog’ through API Key

● Image_prediction data downloaded by using the request library, url was provided by Udacity.

● The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitteruser @dog_rates, also known as WeRateDogs

2. Assess:

In the data assessment, we inspect our data’s quality issue(i.e content issue) and lack of tidiness issue(structural issue). I assessed my data both In excel and jupyter notebook as well.

I used pandas read (), head (), tail(), info(), describe(), value_counts() functions, checking nullvaluesused isnull() function, duplicate values find out by pandas duplicated() method. To check the column name used tweet_archive.columns() method we also can use the list(tweet_archive) function too.

➢ Assessment report for Quality:-

1. After checking the rating section in twitter_archive, I noticed that denominator wasn’t 10 all the time.

2. There were various Null Values in some columns like in_reply_to_status_id in_reply_to_user_id retweeted_status_id retweeted_status_user_id .

3. There were retweets in the text which were actually causing the data to be repetitive.

4. Many datatypes were incorrect.

5. Columns were needed to be renamed with more easy names.

6. The names of dogs weren’t correct at all.

7. The Denominator can’t be Zero

8. Under the Image Predictions Table, p1,p2,p3 name were un evenly cased, i.e. Some were first letter Capitalized while some were not. Moreover, the names under image predictions table were having underscores, with some uneven spaces.

9. Under the rating numerator Sections some of the ratings were float too, which obviously wasn’t included.

❖ Assessment report for Tidiness :-

1. The dog stages were shown differently in different columns i.e. doggo, pupper, floofer, puppo. They needed to be put into one.

2. Timestamp Column In twitter archive contained both Date and Time. So there was a need to separate them both.

3. The source text was mere a link. However the source could be extracted from it.

3. Cleaning Efforts :-

As I worked on these issues, the dataset was getting neater and tidy eventually. I gave every section of code some index alphabetically such as A, B, C and more on. You can inspect the report steps with the notebook and can walk correspondently with it. Here are some efforts of mine:-

(A) Removed all the string values from the text section of the twitter_archive_enhanced dataframe. After seeing the data frame visually, I realized every text which was retweeted has @RT in front of it. So after learning from it, I used the str.contain function to drop all the texts which had retweets element.

(B) Some of the ratings were incorrect and some of them were missing. But after discovering the fact that they were also present in the tweet texts, I extracted all the ratings present in the text with str.extract regex. Afterwards I split them with the split function of pandas into two new columns.

(C) As I told some of the ratings were incorrect, same goes with base denominator. Not every denominator was 10. So I ran the code which will make all the denominator values equal to 10

(D) Few ratings were in the form of decimals. So I used the str.extract regex to extract all the decimal values to the rating_numerator. This way it will be easy to work on all of the float values.

(E) Not every name written under the name section of twitter_archive was correct. So I extracted the names from the text column by using the str.extract regex. And then overwrote them as names of dogs.

(F) As we know that the doggo, floofer, pupper and puppo are the four stages of the dogs. But somehow they are all differently categorised in the different columns so we need to extract them all into single column named stage. I first replaced all the none values to empty string. Then combined all of them into single column named stage of dog. As there were some cases, where there were more than one stages per dog, So I put them in the separate column named 'dog_stage'

(G) Timestamp columns is readable but with a bit of difficulty. So I thought it will be a good idea to extract time and date from the Timestamp column and make them both as separate column. I used simple string slicing technique to extract Time and date into separate columns named respectively.

(H) The source from which the tweets were made were some links. Readability was bit difficult. So I replaced all the source values with some easy and readable values. First I checked what the distinct values in the source series are. Found out that there were only 4. I replaced each and every one with the easy and readable source.

(I) The text in the tweets sections was combined with the URL. So I used the basic String slicing technique to extract all the URL and removed it.

(J) Uneven Cases of Letters in p1, p2 and p3 & Replacement of '_' with ' ' & Removal of Uneven Spaces. I used the str.lower functions to lower case each of the name of dog to maintain consistency. Moreover I used str.to_replace to replace the underscores with the spaces. It made data look clean.

(K) I removed all the values which weren't confident regarding the p1 category of dogs.

(L) I changed the names for p1 and p1_conf to dog_breed_prediction and prediction_confidence respectively. Nothing much, just to provide joy to the reader.

(M) Since I decided to use only P1 values for the whole Analysis I removed all the other predictions category plus some non-wanted values.

(N) Now as I cleaned most of the data. It was the time to merge all the data into a single data frame named as master_df. I also changed the ratings to float. I first joined two dataset then joined the third with the previously merged dataset.

(O) It was the one last time try to check for any wrong data types and I found out that The tweet_ids were int data types. I converted them into str format.

(P) Now the last step was to simply save the all efforts into a single 'twitter_archive_master.csv'file.\

4. Conclusion:-

After all the efforts, I have finally cleaned some of the data and made it readable and easy