Data Wrangling – REPORT

TABLE OF CONTENTS:-

Introduction:-

We Rate Dogs is a popular twitter handle which has over 8million followers and rates the dogs based on various characteristics and upload their photos on it's twitter handle.

In this project we will filter the twitter handle's tweets and make it basis of the analysis.We will wrangle the data by gathering assesing and cleaning the dataset and convert it into a tidy dataset to form visualisation.

Gathering:-

In this first step we will collect the data from three different sources :- import from csv file,import from url with requests library and import the data from the tweepy api.

We will import the necessary modules,load the csv file provided to us in the dataframe and display a few lines of the dataframe.

We will import the tweepy api with the help of the key and secret token provided to us.

We will also import the json file with all the details from the Twitter page and parse the following set of datas from the file to a newly created list which will in turn be used to create a new dataframe.

We will load the tsv file into the dataframe by using the 'tab' as separator and print few lines of the dataframe

We have now gathered all the required and necessary data and now proceed to the assessing part.

ASSESSING:-

We will check the number of columns present in the dataframe and find out which coloumns have missing rows and the type of data in the rows.
.
We observed that the expanded_urls column has few missing values in the dataframe.We will print all those rows.

We will print the type of timestamp as we have already observed that it should be in datetime format but instead it is in string format.

We will check for all the values which are beyond the normal values..

We will check for any duplicate values present in the dataframe.

We noticed that there are some values in the name column of the database which are not names but articles and other english words such as a,the,such,etc.

One more observation was made that in each case of original names the first letter is always capitalised,So we can print all those rows which do not start with capitalised letters using regular expressions.

From further observation we noticed that in the text field there is &amp used instead of & symbol.We will print all such rows.

We will check if there are any duplicated rows present in the dataframe.

We will check if any of the image url is duplicated or not in the dataframe.

We will check for any duplicated rows present in the dataframe and try to understand the data more by checking the values present in it.

We have finally assessed all the dataframes and now we will proceed to the cleaning part.

Assessment Summary

Quality Issues

DF1:-

1.> The type of the timestamp column is string instead of datetime it should be converted to datetime datatype.

2.> We observed that there are few missing rows in the expanded_urls column and we will fill the missing values appropriately.

3.> There are rows where the numerator and denominator are out of range and we will try to fix that.

4.> We observed that the source column has links in it while it should be present in more understandable way which a human can easily read.

5.> The text column contains a few rows where & is written instead of & symbol,these should be replaced.

6.> The name column has a few rows where instead of names articles such as a,the,such,an are used thewse names should be replaced with Not available option.

DF3:-

1.> The type of the tweet_id column is object it shouyld be converted to the int datatype.

Images:-

1.> The predictions have _ joining them which should be replaced with a whitespace character and the first word of each prediction should be capitalized.

2.> The column names should be more understandable and hence we should rename the columns to more simpler names.

Tidiness:-

DF1:-

1.> There are 4 columns for the dog stages namely doggo, floofer, pupper and puppo. These 4 columns should be substituted for one variable.

2.> We are only interested in the tweet id so we should drop the columns with retweet ionformation like retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

DF3:-

1.> The table with the json data should be combined with the tweet data based on the tweet_id column.

CLEANING:-

In this step first we will define the problem then write the code and run it and in the end test the code to see the functionality whether the issue has been resolved or not.
We will now start off by correcting the first dataframe which contained the tweet data. We will copy the dataframe and create a new one and try to analyse the data types and data again.

As we observed that the timestamp datatype is string we will convert it into datetime format.

We will drop all the unneccessary columns not required as we have found that our analysis has no use of the retweet id's.

As we have observed from the examination that the expanded url consists of the https://twitter.com/dog_rates/status/ follwowed by it's tweet id so we will fill the column like this.

We will correct the denominator and numerator manually after checking the correct values from the text column. We will drop the rest of rows where the values can not be corrected.

As we saw above that there are 4 coloumns instead of 1 we will replace the none values with whitespace character and then concatinate all of these to form 1 column.

We have seen that there are few rows where the stage has multiple values we will replace those values and make them one.

We have noticed that the values are in the form of links so we should convert them into more easily readable format.

We will replace the &amp in the text column with & symbol.

We will replace the extra names present in the names column with the value with Not available.

We have observed that the tweet id datatype of the tweet_data is int but ideally it should be a string as we are not supposed to perform mathematical operations on them and they need to be very large numbers so we will convert the tweet_id datatype to string.

We will now merge both the dataframes into one on tweet id.

We will now replace the underscores(_) with the whitespace character and capitalize the first letter of each prediction from the image prediction dataframe.

In the next step we will change the names of the columns of the table to a more understandable form.

We have observed that the tweet id datatype of the image_prediction is int but ideally it should be a string as we are not supposed to perform mathematical operations on them and they need to be very large numbers so we will convert the tweet_id datatype to string.

We will now merge the previously merged dataframe to the new dataframe on the basis of tweet id and create a final dataframe.

Storing:-

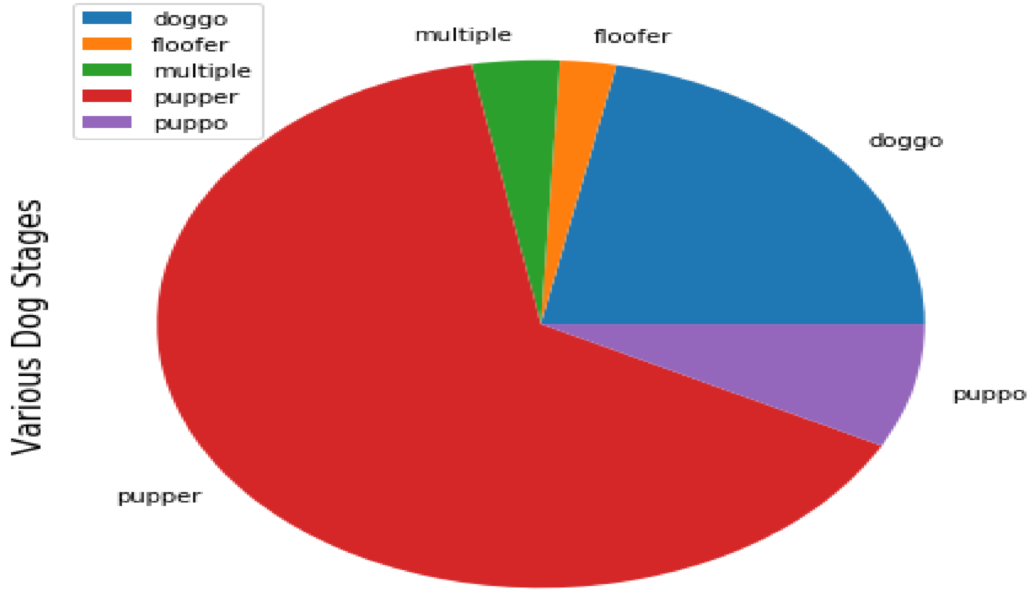
We will store the single cleaned and merged dataframe into a new csv file.

Analysing the data:-

We will now use the dataframe created to answer some questions through visual analysis.

Percentage of different stages of dogs:-

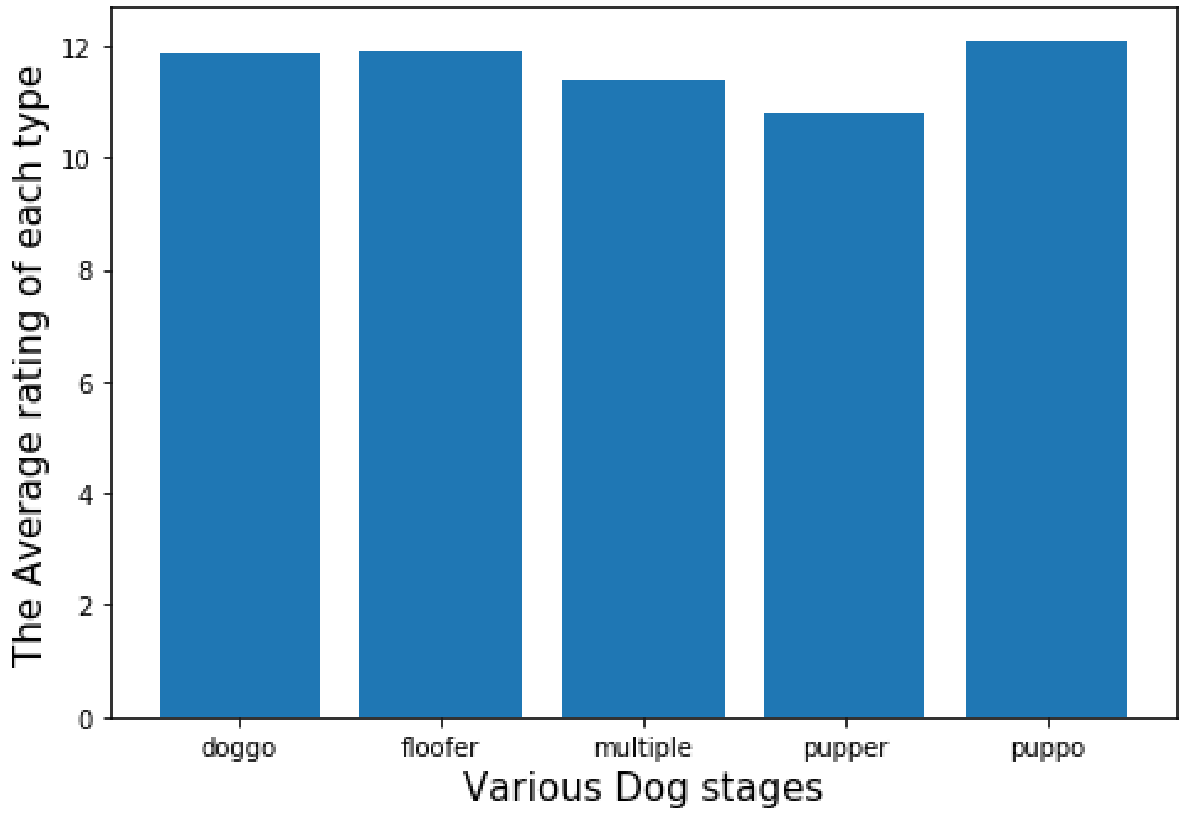
We will find the number of different stages of dogs available in the final dataset created after cleaning and plot the pie chart based on the count of each kinds of dogs available.



We observe from the above pie chart that the number of pupper is highest among all the different dog stages and occupies the highest percentage followed by doggo while the number of floofers are the lowest.

The Average rating of each stages of dogs:-

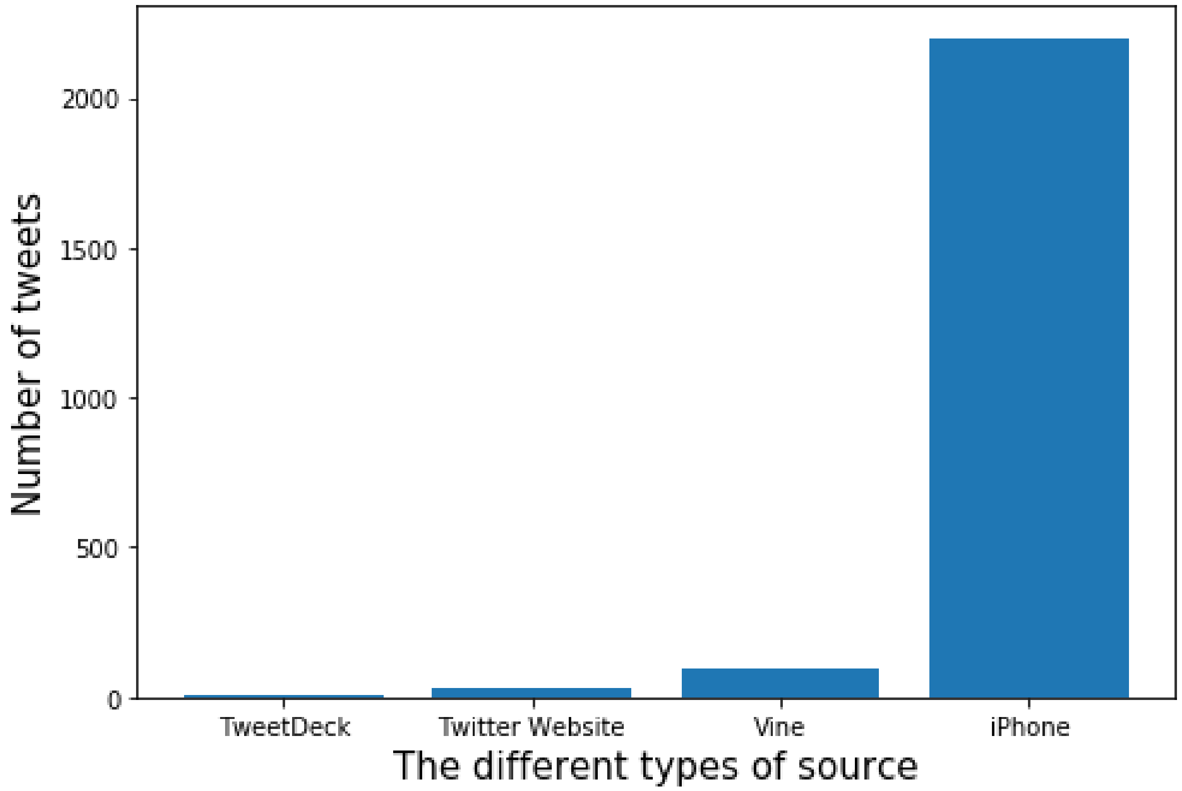
We will find the mean rating of each stage of dogs and plot a bar graph to see the differences in the mean rating of the various stages.



We observe from the above chart that the mean rating of almost all stages of dogs are nearly the same but still the pupper stage has the lowest mean rating among all the other stages while puppo stage has the highest mean rating above all the stages.

The Source distribution:-

We will find the number of each types of sources from which the tweet was posted and plot a bar graph to see the differences in the count of the posts done from the various sources.



We can observe from the above bar chart that the highest number of tweets are posted from the iPhone followed by Vine-Make a Scene app while the lowest number of tweets are made from the TweetDeck. Although the Vine app is at 2nd position but the difference between iPhone and vine is very much.