wrangle_report

June 30, 2020

0.1 Data Wrangling, Analysis and Visvualization Project Report

This project is about wrangling the dataset from the twitter account of WeRateDogs. Wrangling means gathering data from different sources and cleaning them by looking for quality and tidiness issues. Without cleaning the data the data analysis can lead misleading results. So it is very import to wrangle data correcting using right techniques. For this project, I have wrangled the data provided by Twitter API for WeRateDogs account. This account rate dogs on a ery different scale with most commonly numerator being higher than denominator because dogs deserves it. There are total of three steps involed in this process which includes data gathering, accessing data and cleaning data.

0.2 Step 1

0.2.1 Data Gathering

In this project, there were three different source from where i had to gather data. The three different sources were local, from url and from api. The first local file I obtained was from the udacity server named 'twitter-archive_enhanced.csv'. Moving forward, I used the url provided by the Udacity to acces the second file name 'image_predictions.tsv' which I downloaded programmatically. The final file was downloaded using Twitter API. For which I had to make an developer account on twitter. Later, I had to create an app to get credentials to access the data. The credentials was inputed in the code to obtain the data.

0.3 Step 2

0.3.1 Accessing Data

In this step, the downloaded data from above is assessed both visvually and programmatically to look for any issues in quality and tidiness. On observing the data, I found the following issues which had to be solved. I named the twitter archieve as tw_df_clean, image prediction as p_pic_clean and twitter database as twitter_df_clean.

Quality Issues

1. tw_df_clean has columns to be removed which are not useful for analysis like in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

2. tw_df_clean has tweet_id which needs to be changed to string, timestamp to todate.

3. p_pic_clean has tweet_id which needs to be changed to string.

4. tw_df_clean has retweets rows which needs to be removed as we do not need retweets data.
5. tw_df_clean name column has various names which are not capitalize which needs to be replaced with Nan.
6. tw_df_clean name column has various names which are one one letter which needs to be replaced with Nan.
7. tw_df_clean has numerator ratings which does not match the text which needs to be corrected.
8. twitter_df_clean retweet_count and favorite_count has to be changed to int.
9. td_df_clean has source which is not readable which needs to be extracted.

Tidiness Issues

1. tw_df_clean Melting the columns doggo,pupper, puppo, floofer to one row.
2. p_pic_clean Reduce the number of columns p, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3 to two columns dog_breed and conf_test.
3. p_pic_clean The dog_breed column created above has name which are not capitalized and needs format change
4. Extra columns which to be removed like p, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3 after converting it into two columns
5. Merging all the three dataframes into one.

0.4 Step 3

0.4.1 Cleaning Data

In this step, the data is cleaning tacking one issue at a time to make the dataframe as clean as possible so that the analysis is done properly. Also, the data is checked after tackling each issue to check if the issue is solved.

0.5 Conclusion

Data Wrangling is very important for anyone who handles data. Data wranging in python is seemed to be most effienct than any other tools present in the market because of its versatile libraries and ease of coding practing.