Project # 2.0, Wrangle and Analyze Data

Introduction

tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs" has over 4 million followers and has received international media coverage

After gathering data from different sources for the required analysis and do the required cleaning of data , this report describe wrangling efforts to clean the data as much as possible

1. Data Gathering

We gathering data from 3 different sources as the following: -

- ✓ **Gathering data from downloadable file** in the Resources tab which hosted at udacity server, file name twitter_archive_enhanced.csv, this file imported in tabular data using pandas' package at data frame named df_archive_clean
- ✓ **Gathering data from the following URL** https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv using requests library, and the data imported into tabular data using pandas' package at data frame named df_image_predictions
- ✓ **Gathering data from Twitter API** for each tweet's JSON data using Python's Tweepy library, in fact till now my twitter developer account don't approved, therefore I used the JASON file which hosted at UDACITY server, and the data imported into tabular data using pandas' package at data frame named df-api

Project # 2.0, Wrangle and Analyze Data

2. Data Assessing

After gathering data, we made assessment visually and programmatically for quality and tidiness issues to detect quality issues tidiness issues as the following: -

- ✓ **Visual assessment**

We made quick visual assessment for the three data frames, using the pandas library for example we use the head ( ) , tail ( ) , sample ( ) ……etc for each data frame to detect the issues

- ✓ **programmatically assessment**

We made programmatically assessment for the three data frames, using the pandas library for example we use value_counts ( ) , nunique( ) , dtypes( ) ….etc For each column to detect the issues , then clean all issues as per the below table which describe the issue and the action done as a solution

Project # 2.0, Wrangle and Analyze Data

3. Data Cleaning

We do the best effort for cleaning all issues which appeared in the assessment step, we found 10 quality issues and 2 tidiness issues as mentioned in the below table

Serial	Issue	Type of issue	Data frame	Solution
1	Column 'name' has 745 cell None instead of NAN	Quality	df_archive_clean	we replaced all None vlaues by Nan to be compitable in analysis
2	Data type of column 'timestamp' is object instead of datetime	Quality		we convert the type of data in column to timestamp instead of object using built in function
3	Column 'doggo' has 2259 cells , column 'floofer' has 2346 cells , column 'pupper' has 2099 cells and column 'puppo' has 2326 cells None instead of NAN	Quality		we replaced all None vlaues by Nan to be compitable in analysis
4	rating_numerator column data type must be float not int64	Quality	df_archive_clean	we convert the type of data in column to float instead of int64 using built in function to conclude the right rating from the text column

Project # 2.0, Wrangle and Analyze Data

5	rating_numerator column for dog Almost always greater than 10 , but there are some outlier observation, for exmaple 666 , 143 , 182 , 204 , 961 , 1776 , 172 .. this values may affect the analysis	Quality		we extracted the correct rating_numerator from 'text' column using python functions
6	rating_denominator for dog have values less than 10 and values more than 10 for some observations , we will replace all values at this column by 10	Quality		we changed all rating_denominator to 10
7	column 'text' has a combined content of tweet text , link and ratings will not used in analysis , therefore we can delete it	Quality		we dropped the column from the data frame after extracting the correct rating_numerator
8	As per Key points in Project Motivation , we need only original ratings (no retweets) that have images , consequently we can delete all rows which mean for Retweet or Reply..... for rows which have Nan in 'in_reply_to_status_id', 'in_reply_to_user_id' , 'retweeted_status_id', 'retweeted_status_user_id' , 'retweeted_status_timestamp ', will considered original tweet	Quality	df_archive_clean	we dropped the columns('in_reply_to_status_id', 'in_reply_to_user_id' , 'retweeted_status_id', 'retweeted_status_user_id' , 'retweeted_status_timestamp ') from the data frame after querying original tweets

Project # 2.0, Wrangle and Analyze Data

9	expanded_urls columns have 59 NAN values before deletion of some rows in the previous cleaning steps , we can drop the remaining rows from data fram	Quality		After cleaninig data from serial 1 to 8 , onlt two rows with NaN at expanded_urls column, we dropped this rows from data frame
10	Column 'img_num' not important at analysis	Quality	df_image_predictions	we dropped the column from the data framw due to the useless in analysis
11	name of column 'id' in df_api is diffrent than the two other data frame , the name for the same column in df_archine and df_image_predictions is tweet_id	Quality	df-api	we changed the name of column 'id' in df_api to "tweet_id" , to be the same name in the other data frame to be the same column name in merging data frame step
12	Types of dogs during their age in columns 'doggo' , 'floofer' , 'pupper' , 'puppo' may be combined in one column	Tidness	df_archive	we add new column called dog_type and quarying the values from the four columns, then delete the four columns
13	we need only three columns in the analysis 'id' , 'favorite_count' and ' retweet_count' only	Tidness	df-api	we modified the data frame to include only the required rows ('id' , 'favorite_count' and ' retweet_count')