Udacity Data Analyst Nanodegree Program

Wrangle and Analyze Data Project

Wrangle Data Report

Name: *Soyoung Cho*

Introduction

The purpose of this paper is to wrangle and analyze the data of "WeRateDags". The goal of this paper is to wrangle WeRateDogs twitter data and create interesting and trustworthy analyses and visualizations. “WeRateDogs” is the twitter page where people upload dog picture and rate those pictures. I am going to gather, assess and clean the data and pull out some insights.

Gathering Data

I have gathered 3 types of data sets. First, the enhanced twitter archive dataset. This data was downloaded from the Udacity project page. This data includes information of tweet_id, timestamp, source, text, retweeted_status, expanded_urls, rating_numerators, rating_denominator, name, doogo, floofer, pupper and puppo.

Second, the tweet image prediction dataset. This data was downloaded from Udacity server. This data includes information of tweet_id, type of breed and etc.

Third, Twitter API and Twitter json. Twitter json file was downloaded from Udacity project page. And I have collected additional data by accessing the Twitter API key numbers.

Assessing Data

I have observed each datasets in “Gathering Data” section by using; .head(), .info(), and describe(). And some number of issues regarding quality and tidiness came up.

Quality Issues:

1. The number of rows in twitter archived enhanced data (2356) and tweet image prediction data (2075). This seems to occur due to retweets being included.
2. The type of 'timestamp' is 'object'; it has to be changed into correct data type.
3. Some values in the breeds in 'image prediction data', p1, p2,p3 have upper cases.
4. The type of 'tweet_id' has to be changed into correct data type.
5. There are many columns in data frames; not all the information is needed for wrangling and analyzing the data. Some of the columns in twitter_arch should be dropped.

6. There are many columns in data frames; not all the information is needed for wrangling and analyzing the data. Some of the columns in tweet_json should be dropped.
7. Some of the names are wrongly shown. When I have looked at the data in Excel (raw data), additional words such as "a", "an", "the".... etc. makes names mislabeled more.
8. New rating column needed; easier to interpret.
9. Merge p1, p2, p3. Able to see the type of breed in one column.

Tidiness Issues:

1. Merge three datasets used for this project. Easier to observe.
2. Combine each dog stage columns into a one column and named it as "dog stage". I have also considered the issues occurring with multiple do stages and treated them as ‘multiple’ in dog_stage column.

Cleaning Data

I have tried to clean all the issues listed above by using various techniques. I have also double-checked by testing after fixing the issues. The structure of this section is followed as; define the issue – designing code to solve the problem – testing the code whether the issue is solved.

Conclusion

In conclusion, for wrangle data part, I have learned that cleaning and organizing the data before analyzing is important for effective analyze. The process of gathering, assessing, and cleaning the data were not easy but this has developed my skills of using Python and statistical techniques.