Wrangle Report

Introduction

WeRateDogs is a Twitter account that rates participated dogs. In this project, our goal to wrangling a portion of this dataset to discover useful information.

Data Gathering

The project dataset was collected from three different sources. The first data was csv file downloaded directly from the Udacity portal which contains the archive of the WeRateDogs account.

The second data was tsv file extracted programmatically through the provided URL by Python request library.

The third data was txt file downloaded directly from the Udacity portal as JSON file. I was actually excited to get the information through the Twitter API using python's tweepy library but unfortunately, my request to get a Twitter developer account was not approved.

I generally haven't faced any issues to transfer these files to Pandas DataFrame except that in Json file that needs to be read lines one by one.

My plan was assessing and cleaning each dataset separately then combining them and check if it needs to more cleaning.

Data Assessing

Datasets were assessed in two ways:

Visually by using MS Excel and its tools like filter, sort, find, etc.

Programmatically Dataframe and Series pandas attribute like head, tail, sample, info, value_counts, etc.

By using these methods I got more than 10 untidy and messy data columns.

Data Cleaning

In the clean stage, the issues documented in the assessment stage should be high quality and tidy.

The three step process deployed are:

DEFINE, CODE, TEST, STORE

Some methods and functions that used to make cleaning: rename, drop, astype, replace, copy, append, etc.

One of the hardest working in this stage of the assignment was dealing with issues in the text column. Extracting and removing data need a good knowledge of functions like Series.str which contains methods such as contain, extract, match and Python Regex (Regular Expressions).

Finally, a master file was created by combining the three sources which were gathered, assessed, and cleaned. This Dataframe was checked to see if it need extra cleaning.

Analyzing, and Visualizing Data

Finally, there were five insights that analyzed in the master dataset as below:

- Most stages of dog are more popular on the WeRateDogs
- Most dog breeds are commonly on the WeRateDogs
- Distribution of dog ratings out of ten
- Account activity over the year

Conclusion

The cleaning stage of the data was almost consuming the big part of the entire project. The task with little to no guidance was really challenging but in the end, it is very helpful to learn more tactics to solve the problems.