Wrangle and analyze data

WeRateDogs Twitter archive
Wrangling a dataset

Introduction

I will discuss in this report how I managed to collect data provided - by different methods - for WeRateDogs Twitter account, and wrangle it in order to have a final result that is greatly refined and ready for analysis.

I am using python and its libraries (mainly pandas) as a tool for my wrangling efforts.

Body

Gather

I am provided:

- "The WeRateDogs Twitter archive" as a '.csv' file ready to use. Which is a dataset of most of the data available for the account.
- "The tweet image predictions" as an online '.tsv' file that I need to download inorder to use.
- Access to Twitter API in order to collect more helpful data (like and retweet counts)

The first step is to download the needed '.tsv' file, and query Twitter API using Tweepy package to collect needed data (which was saved in a local file).

assess

A full assessment is then performed for the three datasets available to identify what needs to be cleaned which led to - in short - these issues:

- Incomplete records
- Wrong entries
- Erroneous data types
- Redundant columns
- Messy grouping of values and entries

Clean

After identifying what needs to be done, I started cleaning with the main dataframe and then jumping into the other two so that I end up with one master data frame that includes all I need for my analysis.

I use different Pandas functions and techniques such as:

- Dataframe subsets: to select specific qualities of the dataframe
- Pd.melt: to separate and rearrange values and variables
- Pd.merge and pd.concat: to join different dataframes together either vertically or horizontally
- Pd.astype: to change columns dtypes
- Pd.loc and pd.replace : to locate and correct wrong values
- Pd.value_counts: which comes in very handy to group and subset dataframes as well

I test all changes made to the dataframes, this is where a lot of other functions are used such as pd.info, pd.sample, pd.isnull, and pd.duplicated.

Conclusion

Data wrangling is crucial to enhance the quality of the dataframe we are using and to open up a lot of potential for analysis. After I finish this step, I have a high quality and very tidy dataframe that is ready for exploratory analysis to answer all the research questions posed.