

Wrangle Report

Wrangle and Analyze Data Project

Preparations

The first thing I did was import the necessary Python Libraries that I would use in this project. They include but are not limited to Pandas, NumPy, SeaBorn, and Matplotlib.

Data Gathering

The project requires three files. twitter-archive-enhanced.csv, Data that is obtained through Twitter API, and an image predictions file.

twitter-archive-enhanced.csv was originally downloaded programmatically by the course instructor. I went ahead and downloaded it manually from Udacity Classroom. I then loaded this data into a pandas dataframe using `read_csv()`. The project requires a student to extract Twitter API data using Twitter API. However, my request to access the API was denied therefore I used the provided python code to extract the data. I then read the JSON file provided line by line into a Pandas dataset. The third dataset (image predictions) was provided as a txt file and all I did was to import it into the dataframe.

Assessing the data

After successfully loading the data into pandas Dataframes, I merged them to create a single dataset. I went ahead and assessed the data visually and also programmatically. Some of the Pandas methods I used to assess the dataset programmatically include `pandas.dataframe.info()`, `Pandas.dataframe.isnull().sum()`, `Pandas.dataframe.duplicated()` and many others as described in the comments in my Jupyter Notebook.

Cleaning the Data

After carefully assessing the data I went ahead and typed out the observations I had made. Some of the observations include; missing values, incorrect data types, dogs that have multiple stages assigned to them and inconsistent alphabetic cases in a given Pandas series. From the observations made, I went ahead to decide which methods I would be using to handle them.

I went ahead to remove columns that have numerous missing values using the Pandas.drop() method. incorrect data types were corrected by changing the data types.

Conclusion

The wrangling process was generally good and challenging at the same time.