We Rate Dogs Data Wrangling Report

Introduction

This assignment I was tasked with examining three datasets from twitter and perfom data wrangling on the datasets. The goal was to use data gathering, accessing, and cleaning abilities we learned in class to perform the data wrangling process after which insights and visualizations were to be drawn from the cleaned data. In this report I will give a brief detail of the steeps I followed to complete the task.

Gathering

I used three data sources in the project namely 'twitter-archive-enhanced.csv' , 'image-predictions.tsv' , and 'weet_json.txt'.

The 'twitter-archive-enhanced.csv' was made available for download on the Udacity web page. This archive contained information which included tweet ID, tweet text, date tweeted, tweet URL,dog ratings numerator and other data points. It was turned into a dataframe using pandas read_csv() function.

The 'image-predictions.tsv' file was provided through a URL. I used *requests* package and the *get()* function to access the file, then I read it into a dataframe using *read_csv* function and specifying the delimiting character using *sep*. The file contained a table of image preditions (top 3), and each corresponding tweet ID, image URL, and the image number that corresponded to the most confident prediction.

The 'tweet_json.txt' is a text file in JSON format which provided on the Udacity webpage. The instruction was to gather the file from the Twitter API using the Tweepy but due to problems encountered while requesting access to the API I opted to use the provided text. The file contained tweet IDs, retweet count, and favorite count among other columns.

Assessing

The data was accessed both visually and programmatically using the pandas functions. Some of the functions that were used were;

describe()

info()

duplicated()

Quality and Tidiness Issues

Dataset	**Issue**
twitter-archive-enhanced.csv	Contained retweets which were not required for analysis
twitter-archive-enhanced.csv	Contained replies which were not required for analysis
twitter-archive-enhanced.csv	Contained unnecessary columns (*in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , source , retweeted_status_id, e.tc*)
twitter-archive-enhanced.csv	Timestamp column was of string type not datetime type
twitter-archive-enhanced.csv	The rating denominator was not standard
twitter-archive-enhanced.csv	Source had HTML tags
twitter-archive-enhanced.csv	Missing values of 'expanded_urls'
twitter-archive-enhanced.csv	Duplicated tweets
twitter-archive-enhanced.csv	Dog Type can be one column with only the type name
All Datasets	Twitter JSON table and image prediction table should be merged to twitter archive table

Cleaning

The following steps were done to clean the data;

1. Remove retweets
2. Remove replies
3. Convert Timestamp to datetime using *to_datetime* pandas function
4. Starndardize rating denominator and make it 10 using *replace* function
5. Remove HTML Tags from *source* using *regex* in a lambda function
6. Remove duplicates(by tweet_id) in Twitter Archive and Tweet JSON Archive using the *drop* functions
7. Drop replies and retweets columns
8. Drop duplicated Tweets
9. Merge Twitter JSON table, Image Prediction Table and Twitter Archive table
10. Create a dog *development_stage* column and drop the *'doggo','floofer','pupper','puppo'*

Storing Data

The merged data was stored into a single twitter archive called ‘twitter_archive_clean.csv'

Conclusion

This project took longer than expected as I had a difficult time with regex and getting access to the twitter API which in the end I had to forgo. I am also a beginner in both python and Pandas, to accomplish the tasks in this assignment I had to read a lot of information from the Pandas documentation and the book by Wes McKinney titled ‘Python for Data Analysis’. I also came to realise that Data Wrangling is a time consuming process.