WRANGLING (WeRateDogs) TWITTER DATA.

JUNE 15 2022

1 DATA WRANGLING REPORT

During this analysis I used the following software: Jupyter Notebook which helped me to have a better understanding of the documentation. Then I used the following packages (libraries) :Pandas,NumPy,Requests,Tweepy,Json,re,Seaborn,Matplotlib.pyplot,Datetime,Timeit.

1.1 1. Gathering Data

The Dataset(s)

1. The first data set on this project is the dataset to be wrangled in this project is the WeRateDogs which is the tweet archive of Twitter user @dog_rates. This data set includes 2356 observations and 17 features, which consist of the data observation from November, 2015 to August, 2017. I downloaded the Twitter archive from a CSV file manually using the link given to me by Udacity(twitter_archive_enhanced.csv).which I later imported/read into my Jupyter Notebook using the pandas library with the read_csv function.

2. The second data set is the tweet image predictions, which consist of the breed of dog present in each tweet alongside each tweet ID, image URL. It has 2075 entries and 12 columns without any missing values. downloaded the file (image_predictions.tsv) manually also using the link given to me by Udacity class. which I later imported/read into my Jupyter Notebook using the pandas library with the read_csv function.

3. Lastly, I needed additional data. So I queried Twitter API Using the tweet IDs in the WeRateDogs Twitter archive for each tweet's JSON data using Python's Tweepy library then I stored each item into JSON data in a file called tweet_json.txt file. I read each tweet's JSON data in its own line one after the other. I went ahead to then read the txt file line by line into a pandas dataframe. The rtwt_df dataframe which I imported from this JSON includes the tweet_id, hashtag, retweet_count & favorite_count columns.

Assessing Data

After gathering all 3 datasets mentioned above, I went ahead to assess all 3 visually(scrolling through the data in your with my eyes on the preferred software application like excel,google sheet) and programmatically(by using code to view specific portions and summaries of the data set) for quality and tidiness issues. To be able to achieve that I had to assess all 3 data sets I gathered earlier. For my Visual Assessment I used google sheet to open the 3 dataset and scrolled through them to find both quality and tidiness issues. Then I also use the programmatic method from the pandas library such as info(),head() etc as well to assess the 3 data sets. During my assessment process I detected a total of 11 quality issues and 5 tidiness issues. Drop unuseful columns in the image data set and then delete the columns that won't be used for the analysis.

Cleaning Data

In this section, I clean all of the issues I documented above during data assessment. And before I started the cleaning process, I made a copy of the original data before cleaning using the pandas copy()function .In the data cleaning aspect,the programmatic cleaning process of the quality/tidiness issue took place in 3 stages which are the Define, Code & Test stages. I defined each identified issue,wrote each code to clean all and went ahead to test to check if the correction had been made. In order to obey the law of tidiness during the cleaning process, I merged all useful columns in each dataset to one :nnew_df.

Storing Data

After the completion of all stages which includes gathering,assessing and the cleaning process, the nnew_df DataFrame has been stored and saved in twitter_archive_master.csv file.