Abelmoneim Salah

Data Analysis Nanodegree Student

31/10/2020

Cairo, Egypt

Wrangle and Analyze Data

Introduction:

In this project i worked in wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations, but data didn’t clean so my role was gathering data from three sources, assessment this data, cleaning it and finally make some analysis and data Visualizations.

Steps:

1. **Gathering Data**
    - I know that I should gather data from three sources, first source was csv file I use pd.read_csv to read it in my jupyter notebook.
    - Second source was gating data from url and save it in my juptyer as tsv file, I make this code using Requests.get() library , then read it as tsv file.
    - Final source was using twitter API by tweepy library, first step was making twitter developer account to get my key and token to access to twitter API, after that I get data include favorite count, retweet count and Data time from twitter using this API, after getting data making it as dataframe then save as txt file.

2. **Assessing Data**
    - After gathering data, I assessing this data to know quality and tidiness problems in it.
    - First I made some visual assessment using pandas like .head() , .sample() and Excel.
    - Then I made some programmatic assessment like .info() , .value_counts() .

- **Tidiness problems that cleaned**

  we have 3 tables I merged 3 tables to one dataset.

  Combine 4 columns (doggo,floofer,pupper,puppo) to one column called dog_type
- ***Quality problems that cleaned***
  - *Change datatype of timestamp to datetime.*
  - Make source column easy to read.
  - correct rating_denominator columns all values should be 10.
  - in rating_numerator column there are wrong ratings that different about the rating in text column.
  - remove retweets data but if retweeted_status_id equal NaN I kept it
  - remove retweeted_status_user_id ,retweeted_status_id and retweeted_status_timestamp.
  - name column has values that are None instead of NaN.
  - some names in column name are wrong names so I will make it Nan.
  - change datatype of some columns.
  - we have timestamp and Date_time both are the same so I droped Data_time

3. Cleaning Data

- After assessment I clean problems that I found using the sequence of Define, Code and Test.
- I use the following techniques
  - reduce that from functools to merge 3 tables
  - re.findall()
  - .drop()
  - fillna()
  - pd.to_datetime
  - .replace()
  - .astype()

Summary:

After great effort I gathering data from 3 sources , assessing this data and cleaning it , I spend a lot of time for understanding each step in this project but I learned new and good methods that will support me in my career future, finally I make some analysis and data visualizations that I will mention it in act report.