wrangle_report

July 25, 2021

1 Gathering Data

I used the following packages to gather data from a downloaded CSV file, the Twitter API, and a Udacity server.

- Pandas
  - I used the Pandas 'read_csv' function to load the downloaded CSV data into a dataframe.
- Requests
  - I used the Requests 'get' function to retrieve the data (payload) from the Udacity server. I saved the payload to a TSV file using the 'to_csv' function, and subsequently loaded the data into a Pandas dataframe with the 'read_csv' function.
- Tweepy
  - I instantiated a Tweepy API object using a config object from the Hydra framework. I then used the 'get_status' function from the API object to retrieve tweets, save them to a JSON file, and load them into a Pandas dataframe.

2 Assessing Data

2.1 Visual Assessment

I used the Pandas package and Microsoft Excel to visually assess the data. This showed obvious data quality and tidiness issues, such as columns with mostly NULL values and column headers that should be column values.

- Pandas
  - head
  - tail
- Microsoft Excel

2.2 Programmatic Assessment

I used the Pandas package to programmatically assess the data. This confirmed my assumptions that most of the values for several columns were null, and some outliers were present in certain columns.

- Pandas
  - info
  - describe
  - shape

- value_counts

3 Cleaning Data

I used the define, code and test method to list the specify the data issues, code the solution, and verify the results. I successfully corrected data quality and tidiness issues and joined the three tables into a master dataset.

4 Storing Data and Insights

I stored the cleaned data to a CSV file and to a SQLite database for analysis. This meant I could write SQL queries and return the results as a dataframe using the Pandas 'read_sql_query' function. After this it was easy to create plots from the aggregated results using the Pandas 'plot' function.