ETL-Project

By: Tara Madhyasta

It is no secret that Donald Trump wields enormous power through his Tweets. The objective of this project was to correlate between the tweets made by Donald Trump and the change in Dow Jones Index for those days. But since Trump lost his Twitter power, the project focused on correlation between major news headlines about Trump and the Dow Jones Index for those days. It also includes the crime data for the 4 years from Trump's 2016 election campaign through his Presidency years.

Extraction

3 datasets were used from various platforms. The data ranges over the years from 2016 to 2020. The sources for the datasets are as follows:

* <https://finance.yahoo.com/quote/YM%3DF/history?p=YM%3DF>
* <https://www.cnn.com/article/sitemap-2016-1.html>
* <https://crime-data-explorer.fr.cloud.gov/api>

Dow Jones data was pulled from Yahoo Finance website from their historical data section. The data was downloaded in CSV format. read\_csv function was used to read the csv file and convert it into Pandas dataframe.

Major news headlines was extracted from CNN.com website using BeautifulSoup. The extracted headlines and dates list from BeautifulSoup was converted into Pandas dataframe.

Crime data was downloaded using API from Federal Bureau of Investigation agency website. The desired data for analysis was extracted from the JSON file and normalized into a Pandas dataframe.

Transformation

The first few steps in cleaning up the datasets involved dropping unwanted and redundant data. Further, the following steps were taken to clean up the data for analysis:

Used Pandas functions in Jupyter Notebook to load CSV file.

Used BeautifulSoup to get the extracted data in desired form.

Transformed the CSV, JSON files into data frames.

Converted the web extracted data into data frames.

Dropped the extra columns which were not relevant to the focus of this study.

Calculated the change is Dow Jones Index.

Converted the datatypes of dates from string to datetime format.

Further cleaned the dates column by extracting only the date portion and removing the timestamp.

Extracted the sought results from JSON file and converted into a dataframe.

Rearranged the columns in dataframe for easier analysis.

Sorted the data in ascending order for easy comparison and analysis.

Load

Next task was to transfer the cleaned and sorted dataframes into a database. A PostgreSQL database was created using PGAdmin. A connection was established with the PostgreSQL database and 3 tables were generated from Pandas Jupyter Notebook, one each for the Dow Jones Index, news headlines, and the crime data. The final step was to load the data from Pandas to PostgrSQL using the connection.