Project\_2

Resolutionist

Group Members: Nasra Ahmed, Kokila Janarthanan, Huntley Bodden

Project description - This is an Extract Transform Load (ETL) project where the team will use the ETL process to compare two different types of data that contain information about New Year resolutions. One source of information is a web page where the team will use web scraping to scrape in data and transform it into a pandas data using jupyter notebook. the second source of data is a .CSV file (found on Data World) that contains tweets of various twitter account holders’ New Year’s resolution tweets.

Project purpose - The ultimate goal is to 1) demonstrate the execution of the ETL process, 2) merge two different types of data sources into one data comprehensive data set, and 3) identify and catalogue the similarities between the two datasets.

The data sources the team explored are listed below. Of the three sources listed, the first two were used only because they contained more information.

Data Sources:

1) Life Hack - <https://www.lifehack.org/articles/communication/50-new-years-resolution-ideas-and-how-achieve-each-them.html>

2) Data World - <https://data.world/crowdflower/2015-new-years-resolutions>

3) Chamber of Commerce -<https://www.chamberofcommerce.org/top-new-years-resolutions>

Process:

The Life Hack website contains a list, from one to fifty, of the most popular New Year’s resolutions. While the purpose of the page is to provide an extensive list of advice, tips, and tricks to help people see their resolutions through and make life changes, the team wanted to see how the list measured up to the resolutions mentioned in the twitter-verse. Using web scraping, the team read in the html file to a pandas notebook, scraped the list of 50 resolutions from the website, cleared the html formatting and transformed the data into a Pandas data frame.

After the web file was prepared, the team then moved on to importing the Data World .csv file. The data was reviewed, cleansed and converted into a Pandas data frame.

Once both files were in the proper format, the group proceeded to merge the data frames into one, trouble shooting along the way. The ERD data model for showing how the data was merged is displayed below.

[Graphical user interface, text, application, chat or text message

Description automatically generated](file:///C:\GeorgiaTech\GT-VIRT-DATA-PT-09-2022-U-LOLC\project_2\Project2_ERD.png)

Project 2: Data Model

The next step was to load the data into PostgreSQL, the SQL relationship structure for the data set is displayed below.

Text

Description automatically generated

Project 2: SQL Structure