ETL Project Final Report

Team: Cesar, Yen, Roma, Bill

Introduction

The ultimate objective of the analysis is to explore correlations between suicide and other factors such as age, gender, geographic location, GDP and HDI (Human Development Index). According to United Nations Development Programme, HDI measures quality of life based on longevity, education, and living standards. The hope is to identify risk factors in an effort to prevent suicide.

This project is intended to demonstrate our ability to Extract Transform and Load datasets utilizing Python and either SQL or NoSQL databases.

From this data we plan to answer the following questions:

1. What are the Top 10 Countries for highest suicide rates for males in 2016?
2. What are the Top 10 Countries for highest suicide rates for females in 2016?
3. What are the suicide rates for males in countries with the lowest GDP growth in 2016?
4. What are the suicide rates for males and females in countries with the lowest HDI in 2016?
5. What are the global averages for suicides rates for males and females?
6. What are the suicide rates for males and females in countries with the highest inflation in 2016?

Extract

Data was collected from online sources in the form of csv files. The data files were reviewed to determine a schema for our database. All of the files included data of multiple countries over multiple years allowing us to compare the data from each source by country and year. Since the data in each dataset is related we decided to use PostgreSQL as our database.

The layout of the database was visualized with an entity relationship diagram (ERD) designed using the website Quick Database Diagrams (https://app.quickdatabasediagrams.com) and exported as a .sql file. A new database called ETL_db was created in postgres and the schema was designed using the exported ERD .sql file.

Transform & Load

We scoured the web for free data sets from reliable sources. After rejecting several data sets due to incomplete or limited information, we were able to find three data sets which we felt we could join with one another based on common factors such as year and country. Our next and biggest challenge was to manipulate the data so that we can join on overlapping variables. The datasets we found were in a wide format with repeating data making it not ideal for querying. After loading to Pandas, we utilized the melt() method to convert the wide format dataset into a long dataset. We then utilized the pivot() method to sort data by country and year. This was completed for each dataset to standardize the table format before loading it into tables in our database. Each data set was loaded into its own table in ETL_db using the sqlalchemy library.

In addition to rearranging the data, another pain point we experienced was that the data was inconsistent in the country names being used. (For example, USA vs United States vs The United States of America.) We manually created a .csv file with “universal” country codes which we used to replace all Country values to be identical between all three data sets. Once we overcame these challenges, the three tables were connected using the join method on the “Year” and “Country”.

Conclusion

The most important part of a project like this is to provide good clean data as an output. Understanding the end use of the data will help determine how to build a warehouse to meet the needs of the customer.

The data source plays a major role in how much munging is required. Collecting data yourself vs utilizing available datasets online can determine how intensive the munging process will be.

Appendix

Github repo: https://github.com/UncleBacon/ETL_Project