David Phillips

Vadym Zhukhovytskyy

**ETL Project Report**

Extract

The data we used in this project came in two forms: CSVs and a SQL relational data table from a preexisting database. One source of data we used was the World Health Organization (WHO). We pulled CSVs from this website containing data about suicide rates and alcohol consumption for countries across the globe. We also used data from the MySQL sample database world\_db, which includes tables called Country, City, and CountryLanguage. For this project we used the Country table which includes statistics such as GNP, life expectancy, and government structure. Population data was collected from World Population Review 2019 in the form of a CSV. Happiness data was collected from World Happiness Report 2019 in the form of a CSV. We used this data to decide how we wanted to filter our data as to only select for the top 20 happiest countries from our other datasets. Finally, suicide rate data was collected from the World Health Organization in the form of a CSV. Our goal for this project is to collect this data and transform it to a succinct collection of tables and load them into a SQL Database so they can be utilized for future analysis.

Transform

For this project we wanted to transform the data we collected in three ways:

1. Delete unnecessary data columns from our tables/dataframes
2. Select only the top 20 countries from the World Happiness Report 2018 list of happiest countries.
3. Join multiple dataframes to one more complete table

We were able to accomplish this transformation through a few different methods. For the CSVs we selected the columns and records (countries) we wanted to keep using Pandas in a Jupyter Notebook after converting the CSV’s to a dataframe. We then joined these dataframes together to make one succinct dataframe with appropriate fields and exported this dataframe as a new CSV. For the data the we had in a SQL relational database we performed the transformation of dropping irrelevant columns and selecting for the 20 desired countries using mySQL script in SQL Workbench. After this process we were left with clean, concise data in the form of CSVs and a SQL table.

Load

Since the data we are using is relational we loaded the data to a new SQL database for the final step of the ETL process. We did this in a Jupyter Notebook using SQLAlchemy to connect with our local SQL server and store the data we created using Pandas in the same Jupyter Notebook. The compiled SQL database contains data that would be useful to an analysis of countries that rank amongst the happiest according to the World Happiness Report 2018. Specifically, the data we created in the ETL process contains information such as population, suicide rates for males, females, total suicide rates, and alcohol consumption rates for these countries.