ETL Project

Final Report

Datasets used:

Heritage Foundation - Index of Economic Freedom

Data will be Web Scraped for the years 1995 to 2019

https://www.heritage.org/index/explore?view=by-region-country-year&u=636924242568355010

The World Bank - GDP per Capita

Data will be from an CSV/Excel file from 1960 to 2017

<https://data.worldbank.org/indicator/NY.GDP.PCAP.CD>

When obtaining the data from Heritage.org, we used web scraping for each year. Data from the World Bank was available via download from their website into an Excel file.

As our group consisted people from varying background and interests, we went through a lengthy process of selecting the data to organize. Because of the shared interest of each group member, we selected economic and happiness data that we discovered was readily available from the sites mentioned above.

There was very little cleaning that needed to be completed, with the exception of removing n/a values. The first task was to join the GDP data to the happiness index data. The happiness indices had each year as a column, while the GDP data had years as rows. After much discussion, we decided to unpivot the happiness data and make the years into rows, so that it could be joined with the GDP data. That left us with one dataset for each region of the world. The data from Heritage.org was organized by year, and by region. Because of this, we appended the yearly region data into one large dataset that included all regions, for all years that could be organized by year, country, GPD or happiness index.

After speaking with Sahil and Cam and getting their suggestions on how to use the flat files, we decided to use Microsoft Access and SQL for organizing and storing our data. This made our options of joining and viewing very simple and straightforward.

Regarding the schema, we have a separate dataset for each region with all sets of data included in the datasets (GDP and happiness indices). We also have a separate dataset for the GDP before we discovered a source for GDP per capita. Once we discovered the GDP per capita, we immediately pivoted to using that as our main source of GDP data. That dataset is not used, but we decided to leave it for full transparency. The final dataset if the fully joined dataset from each region that has been detailed above.

There would be many ways to analyze the current data we have collected and cleaned. You could easily do surface level analysis comparing GDP per capita to the freedom index. The more detailed analysis could include info from sources like the Would Health Organization, financial data of companies from the countries in questions, refugee information, climate data, spending data, etc. Because GDP and freedom indices are quite broad, this data can be very helpful when paired with multiple other datasets to create analysis that is very deep, very broad, or somewhere in between.