ETL Project

Life Expectancy compared to Median Income [U.S.]

January 2020

ISSUED BY

Harvard Extension School

REPRESENTATIVES

Julie Demusz
Soukaina Boutaieb
Ashley Dodge

Introduction & Background

This project our goal was to find and combine data sources using the ETL method. We selected data addressing Life Expectancy and Median Household Income for the U.S. from 2010 to 2015. We were curious if there was any correlation between these two datasets.

Submission Guidelines & Requirements

This document contains guidelines, requirements, and suggestions for Project 1.

Team Effort

Due to the short timeline, teamwork will be crucial to the success of this project! Work closely with your team through all phases of the project to ensure that there are no surprises at the end of the week.

Working in a group enables you to tackle more difficult problems than you'd be able to working alone. In other words, working in a group allows you to work smart and dream big. Take advantage of it!

Project Proposal

Finding Data

Your project must use 2 or more sources of data. We recommend the following sites to use as sources of data:

- data.world
- Kaggle

You can also use APIs or data scraped from the web. However, get approval from your instructor first. Again, there is only a week to complete this!

Data Cleanup & Analysis

Once you have identified your datasets, perform ETL on the data. Make sure to plan and document the following:

- The sources of data that you will extract from.
- The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).
- The type of final production database to load the data into (relational or non-relational).
- The final tables or collections that will be used in the production database.

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

Project Report

At the end of the week, your team will submit a Final Report that describes the following:

- **Extract**: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
- **Transform**: what data cleaning or transformation was required.
- **Load**: the final database, tables/collections, and why this was chosen.
- Please upload the report to Github and submit a link to Bootcampspot.

Data

- Jupyter Notebook
- Query.sql
- Resources/Data

Project Report

Day 1 we outline the layout, time frame and goals for this project. We began brainstorming various topics and areas of interest. We then had to narrow it down based on accessible data. Ultimately we narrowed our search down to a comparison between life expectancy and median household income in the United States. Julia and Souki found the two data sources outlined below:

Data	Source
Life Expectancy	Kaiser Family Foundation
Household Median Income	Data World

Both data sources contained information by state. The Life Expectancy data was only a single column for the Average life expectancy for the time period 2010-2015. The Household Median Income contained a wide range of median incomes from 1984 to 2017 for each U.S. state.

That Ashley began by importing the dependencies pandas and sqlalchemy. The data, which were both .csv files, were then imported into the notebook and converted to dataframes in their natural state. It was discovered that both .csv files had superfluous rows that could be ignored. The miscellaneous rows contained notes or additional headers that were not necessary for the dataframe. To remove those rows Ashley used the header and nrows functions to target the specific data for import.

The next step was to clean the data. To clean the Median Income dataframe (df) only the years from 2010 - 2015 were selected into a new df. A new column calculating the average of the median incomes was then added to this new df. Upon confirming the calculations the year columns were dropped from the df and the remaining columns were relabeled.

The life expectancy data was cleaned by creating a new df that only contained the life expectancy data and not the Footnotes column. Both df were set to have their indexes be 'states'.

It was now time to create the tables in PGadmin using SQL. Two empty tables were outlined in the database mirroring the newly created dataframes. Back in our Jupyter Notebook we were now able to connect it to our local database, life_income_db. After confirming the existence of our tables we loaded the dataframes into our database. To confirm that our database successfully read our new tables we queried them. Success!

We then joined our two tables in our states to get a single table outlining both the average median income and average life expectancy for each state. Once the tables were joined we calculated the correlation of the two columns and determined that there is a moderate positive correlation between life expectancy and median household income.

Correlation = 0.7811864031497228

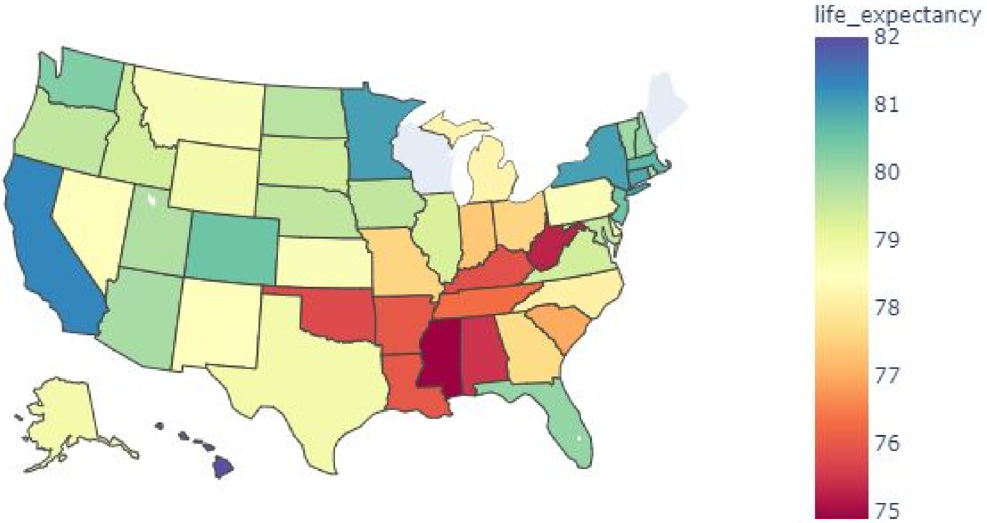
The ETL file was then ready to be pushed to git to be accessed by others.

Souky used the regression analysis on our Data , the scatter plot was linear which means that there is a strong and positive association among 2 variables which is .

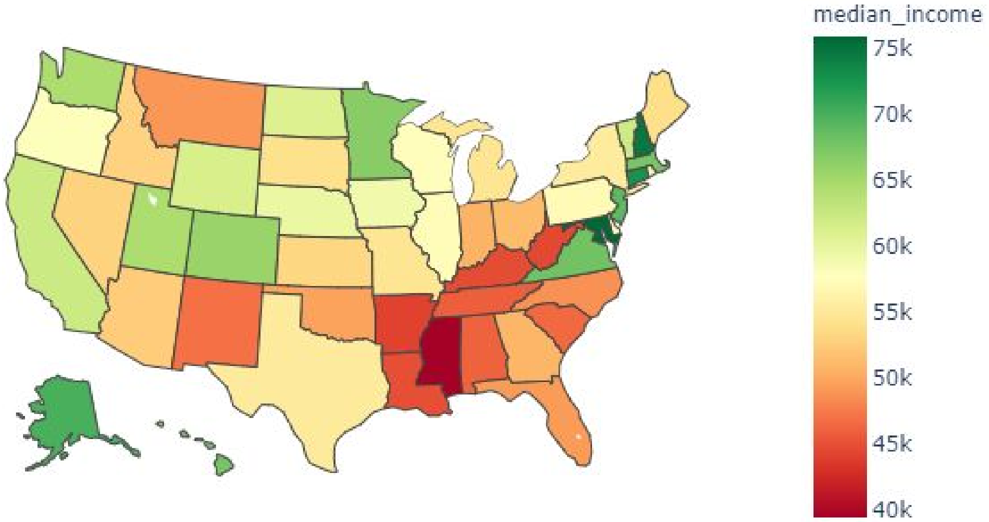
We also discussed the best visual to present our study. Firstly we were considering using heatmap. However, since our data is based on states, latitude and longitude are not the best index in these circumstances.

Therefore, we were looking for alternative visual tools for the best practice. Ashley proposed the Plotly and Julie makes it happen!

Life Expectancy Across the States 2010-2015



Median Income Across the States 2010-2015



Limitation

1. **Timeframe of the Datasets:** The timeframe of the datasets are half a decade ago. We would like more up to day data if we have more time for the project;

2. **2 Units are missing from one of the datasets:** Life Expectancy for Wisconsin and Maine are missing from the life expectancy dataset.

3. **Finding the Data:** Although there are plenty of datasets available for Median Income and Life Expectancy, it took us more than 3 hours to identify and polish the datasets for our study.

Conclusion

With the given time and available datasets, although the correlation is not perfect, however **+0.70 indicates a** strong positive correlation in between Income and Life Expectancy.