ETL-Project

Unemployment and Diversity's Impact on Wages

Introduction

This project is designed to conduct a presentation of Median Income, Diversity, and wages from a 5-yr window of data (2010-2015).

The purpose of this project was to build a database that demonstrates the relationships between these variables and understand how unemployment and the diversity index relate.

Data Extraction

In this project we extracted, transformed, and loaded 5 years of data from the following:

- Diversity Index from Kaggle.
- Unemployment from Kaggle.
- Median Income by county from Data World.

For the Median Income Dataset (Figure 1), unnecessary columns were dropped - the columns County-State & State were removed. The State Code column was changed to "State".

Location columns were split from Diversity table into County & State to make it easier to merge and group with the other 2 datasets.

The unemployment dataset was grouped by State and County and averaged for each County. Then Unemployment dataset and Median Income dataset were merged on State and County, using an inner join. The Diversity dataset was then merged on that by State and County again to create consistent formatting before final output.

Implementation

The last step was to transfer the final output using MySql. The database and respective tables were created to match the columns from the final Panda's Data Frame using MYSQL and then connected to the database using SQLAlchemy and loaded the result.