# Dimensional Warehousing Report

## Overview

The United States government has tasked me with analyzing socio-economic data to get a better understanding of the nation’s current climate. The project demonstrates integration of multiple datasets into a dimensional database model to be used for querying and analytical analysis. The report will discuss the data collection, data understanding, warehouse framework, ETL, and analysis of a business question query.

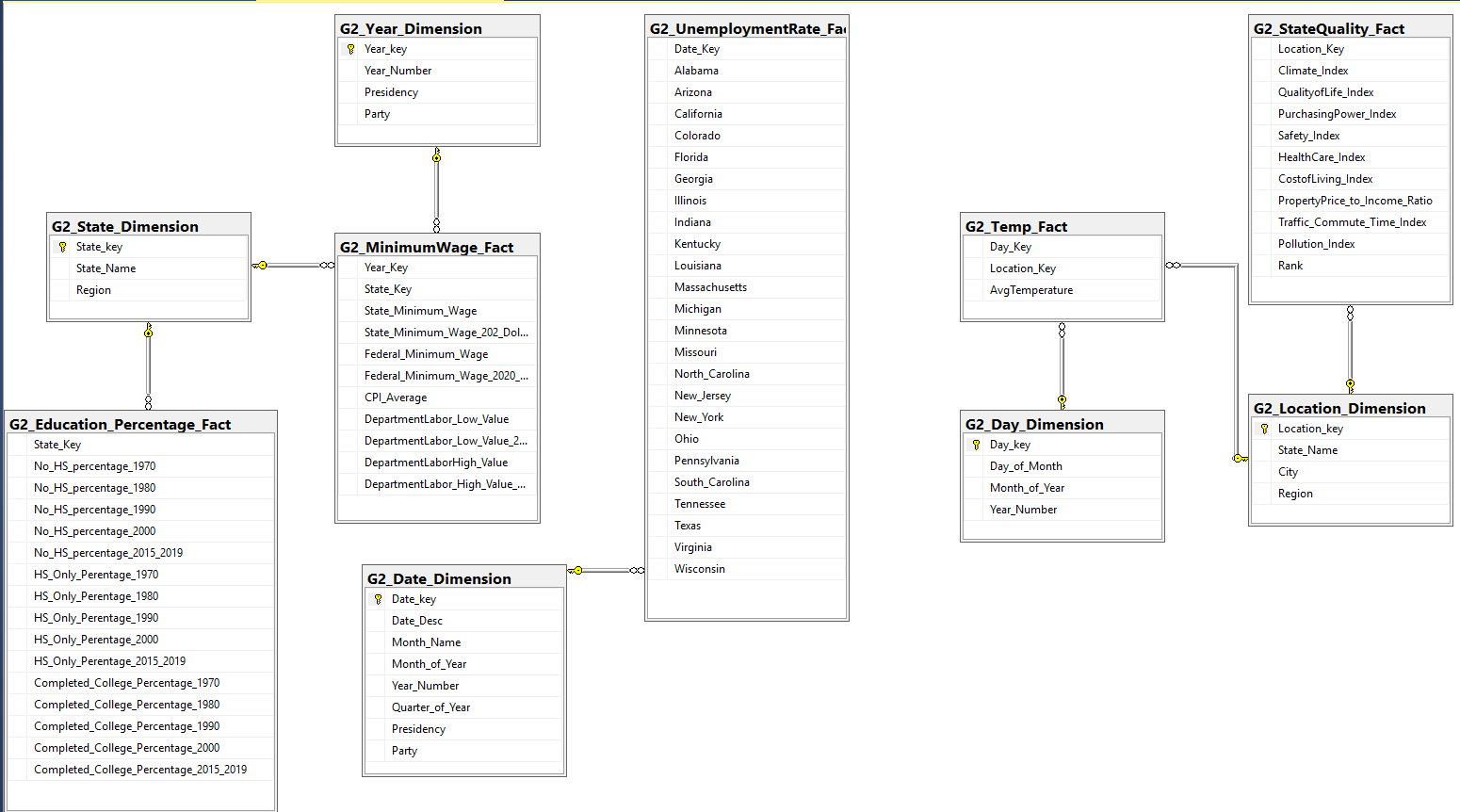
## Data Collection

The final data set encompasses a variety of socio-economic metrics collected from the U.S. The datasets are minimum wage data, U.S. unemployment, education percentage, state quality index, and state temperature. All but one was web scraped from various national databases while state quality index was scraped using python. All datasets were in excel csv form.

## Data Understanding

Going through each dataset, the minimum wage data includes a variety of variables at the state and region locations dating back to 1968. The U.S. unemployment data includes unemployment rate at the state level (it does not include all states) at a monthly, quarterly and yearly level dating back to 1976. The education percentage data shows levels of education by state binned by decade beginning at 1970. The state quality index data contains is a current snapshot of various metrics regarding quality of life at the state and city locations. And lastly, the state temperature data shows weather temperatures at the city, state, region, and country locations and day, month, year time levels. While the datasets are not all from the same industry we thought it would be interesting to discover if there were relationships between any of these factors.

## Data Warehouse Model

The final dimensional model is shown in figure 1 below. Each dataset has its own fact table showing quantitative data. The minimum wage fact is joined with the state dimension table and year dimension table. The year dimension has year number, presidency, and party affiliation metrics. Education fact is joined with the state table as well. Education fact has a time grain already accounted for with the education percentages (years are binned for each education level category). The state dimension gives education and minimum wage their location grain at the state and region level. The Unemployment fact table has the unemployment rates for each state. The location grain, state, is already accounted for with the unemployment rates. The unemployment fact is joined with a date dimension table. The fact table has a time grain of month of year, quarter of year, and year number. The temperature fact table only has a time grain of day of month, month of year, and year number in the day dimension table. It also has a location grain of city, state, and region in the location dimension table. State quality fact table is a periodic snapshot so it does not have a time grain. It does however share the location dimension with temperature fact. The main challenge in making the model was figuring out how to implement the correct tome and/or location grain to integrate the different fact tables.

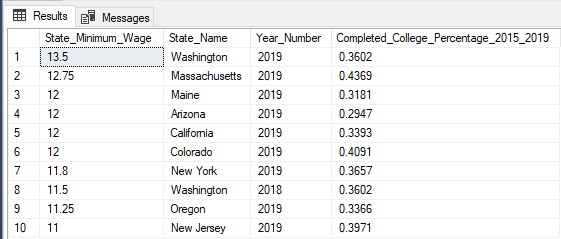
**Fig 1**. Dimension Model

## Extract, Transform, Load

An ETL framework was developed to take the excel data and lad it into the dimensional model skeleton. The SQL data tools SSIS first took the model made and mapped the primary keys and attributes so the model could accept the excel data loaded into it and align the data to the model.

## Analysis

One question I wanted to better understand was: Is there a relationship between state education level and minimum wage rates? The results from the query are below in figure 2. With the most current data, Washington, Massachusetts, and Maine have the highest minimum wages but have a variance in percentage of people that completed college by about 12%. Furthermore, when trying to identify a relationship between the two variables Arizona has the 4th highest minimum wage rate but the lowest percent of people who have completed college out of the top ten state wage rates. The query for this table is illustrated in the appendix.



**Fig 2.** SQL query output

# Appendix

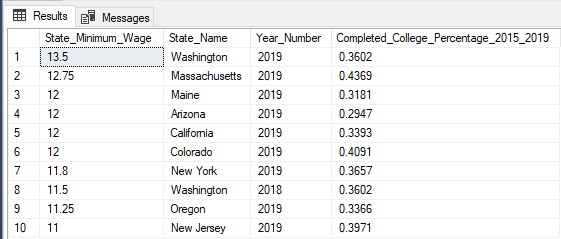
**Select top 10 states with the highest minimum wage rates in 2019 and the percent of people that have a college education.**

SELECT top 10 m.State\_Minimum\_Wage, s.State\_Name, d.Year\_Number, e.Completed\_College\_Percentage\_2015\_2019

FROM G2\_Year\_Dimension d join G2\_MinimumWage\_Fact m on m.Year\_Key = d.Year\_key join G2\_State\_Dimension s on m.State\_Key = s.State\_key join G2\_Education\_Percentage\_Fact e on e.State\_Key = s.State\_key

WHERE Year\_Number <= 2019

ORDER BY m.State\_Minimum\_Wage DESC

;