## 

## Download Locator Location Map Free Clipart HQ Clipart PNG Free | FreePngClipart

## 

## 

## CIS 9440 Datawarehousing Final Project

## Group 6

## RemoteU: A Data Warehouse To Find the Best Location for US Remote Workers

## 

## 

Andreina Chimi Chong Zheng

Iyob Gebremariam

Raad Khan

Shun Yi Sze

Table of Contents

| Content | Page |
| --- | --- |
| Introduction | 3 |
| Data Sources | 4 |
| Dimensional Model | 5 |
| ETL Process | 6 |
| Analytics and Visualization | 13 |
| Conclusions | 18 |

## 

## 

## Introduction

Due to the Covid-19 pandemic, going to the office is not the norm of working in the U.S. anymore. Instead, there is a growing number of people working from home. According to a study from IBM, 54% of workers would like to remain remote after the pandemic[[1]](#footnote-0). A report from Upwork shows that 14 to 23 million Americans are planning to move due to remote work[[2]](#footnote-1). Because of the remote work mode, more and more people are planning to relocate to other cities or counties where the housing is less expensive, the population is smaller, or the healthcare system is better. This decision however is seldom data driven and most tend to relocate without careful thought. The aim of this project is to build a BI application, RemoteU, which helps remote workers to make a data driven decision of an ideal location to live. There is a lot of information that we should consider, such as healthcare, education, cost of living, political leaning, etc. Therefore, the data for our analysis is obtained from various data sources.

*Key Performance Indicators*

* Graduation rates by degrees
* Health access & quality
* Cost per meal
* Median rent costs by county
* Population by races
* Political leaning by county

*Tools*

| Database | PostgreSQL |
| --- | --- |
| ETL | Pentaho & Python |
| Visualization | Tableau |

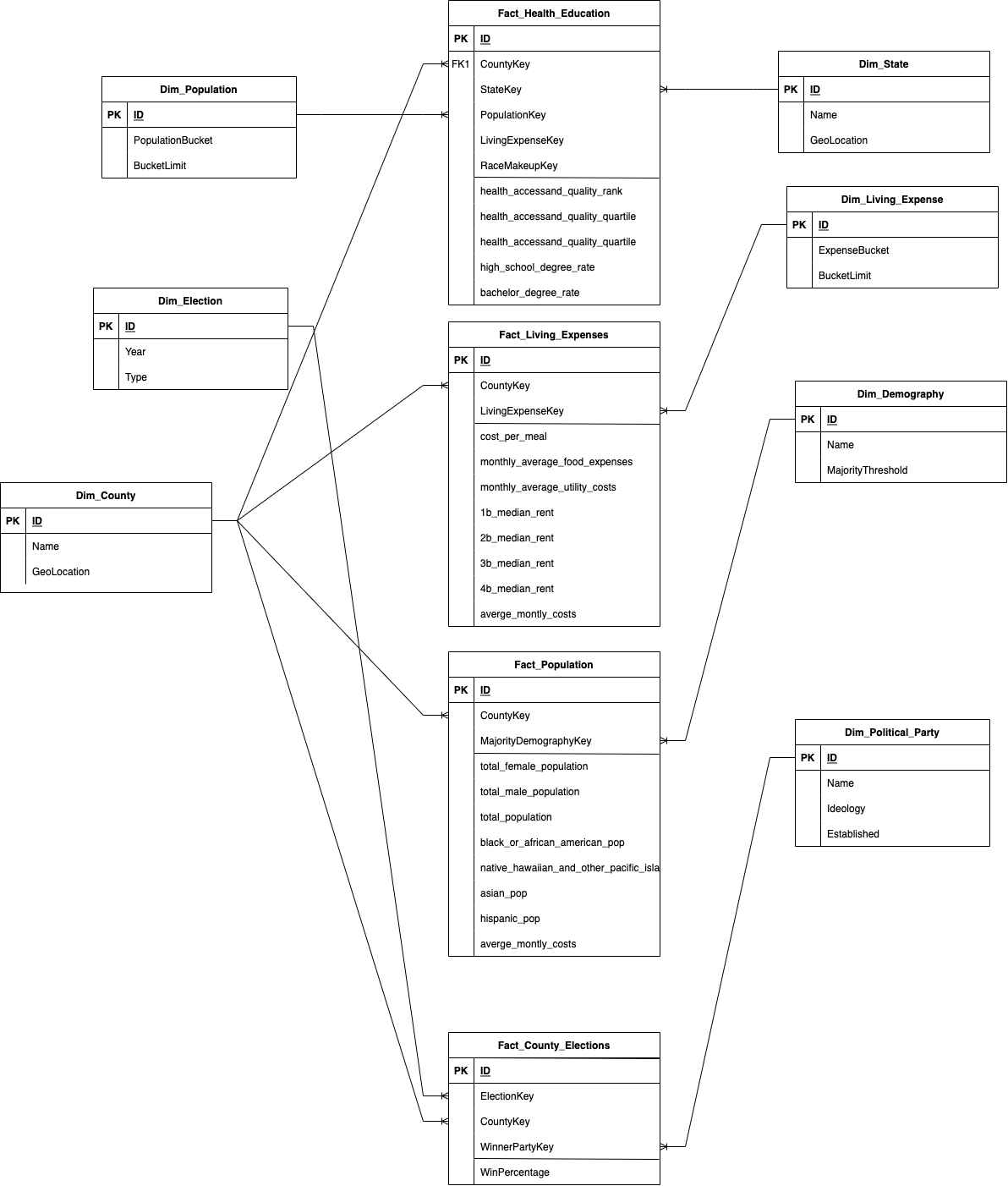
## Data Sources

We looked at various open data sources on US counties to provide the most comprehensive picture of the “livability” of the many US cities and counties. The key dimensions we considered are health and education, living expenses, population makeup and economic activity. The following table summarizes the data sets we used, and a brief description for each.

| Dataset | Source | Description |
| --- | --- | --- |
| The County Health Rankings & Roadmaps data[[3]](#footnote-2) | University of Wisconsin’s Population Health Institute | Includes comprehensive rankings of each county based on a health ranking model. The model includes health behaviour (30%), clinical care (20%), social and economic factors(40%) and physical environment (10%).[[4]](#footnote-3) |
| Education attainment county data[[5]](#footnote-4) | United States Census Bureau | Includes highschool, bachelor’s and advanced degree attainment levels for each county in the US. |
| Populations[[6]](#footnote-5) | United States Census Bureau | Includes total populations in each county also broken down by gender and race |
| Cost of food[[7]](#footnote-6) | Map the Meal Gap 2021 | Includes cost per meal and average monthly food expenses |
| Median rent costs by county[[8]](#footnote-7) | US Department of Housing office of policy development and research | 1 to 3 bedroom fair market rent prices for each county in the US |
| Utility costs[[9]](#footnote-8) | Move.org | Includes monthly average utility costs for each county |
| Election and political leanings[[10]](#footnote-9) | Harvard University | Election outcomes of the past five election cycles for each county |

## Dimensional Model

Having assembled the datasets, we designed the following dimensional model. As we were able to gather a large collection of datasets we broke down our model into 4 facts and 7 dimensions as shown below.

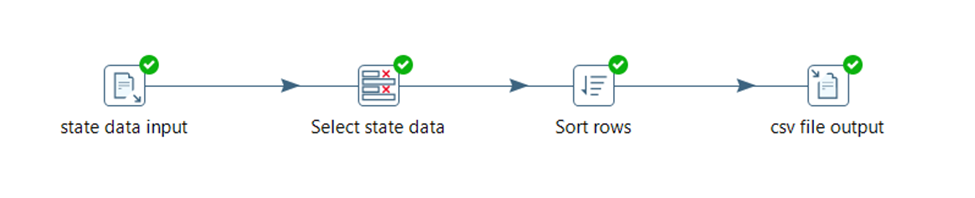


*Fig 1. Dimensional Model*

## ETL Process

As part of the ETL process, we utilized Pentaho to clean data and Python to transform and load data into the database.

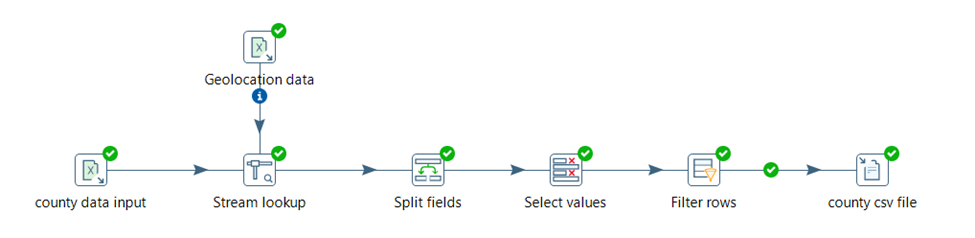
**State data**



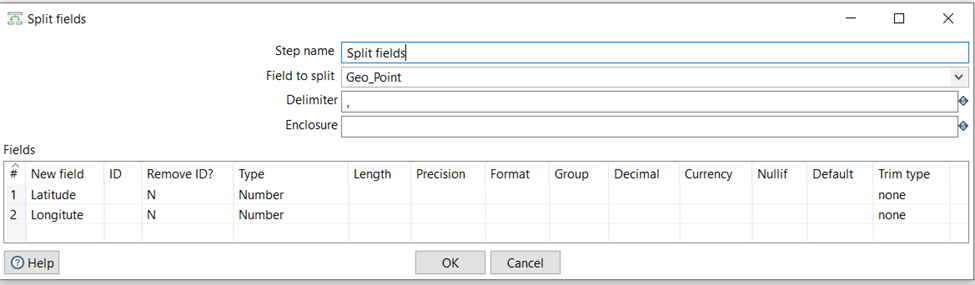
*Figure 2*

For state data, we extracted a csv file(state data input) and selected only state data(state fips, state abbreviation and state name) from it. Then, we sorted the state name in ascending order. Finally, we loaded it into a csv file for further transformation.

**County data**



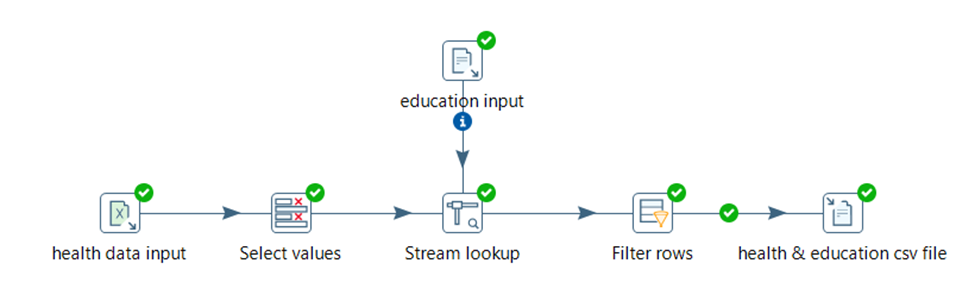
*Figure 3*



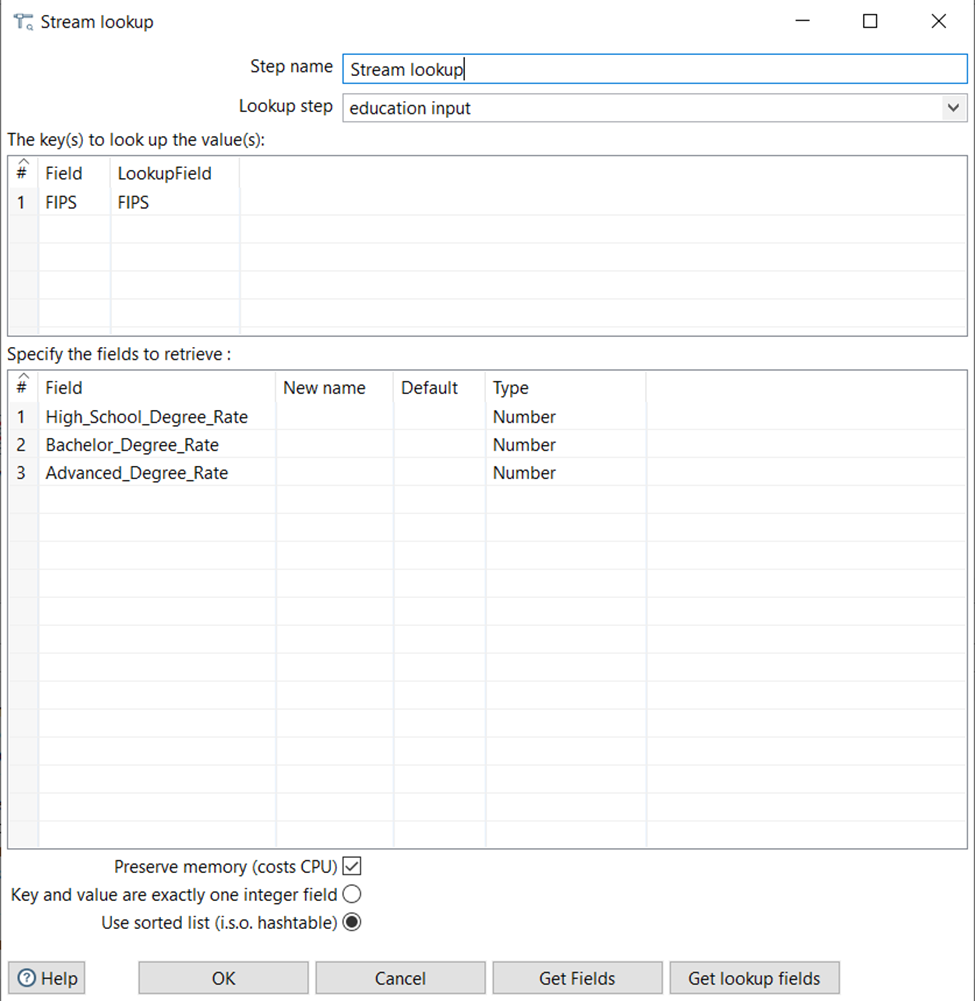
*Figure 4*

For county data, we extracted data from an excel file(county data input). Then, we added geolocation to county data. The Geo\_Point column had the latitude and longitude, so we split it into two fields which were Latitude and Longitude. After “Split fields”, we removed the Geo\_Point field. There were some rows that the county value was null because they were state data, which were unnecessary for us. Therefore, we filtered rows that the county value was not null. Finally, we loaded it into a csv file(county csv file) for further transformation.

**Health and education data**



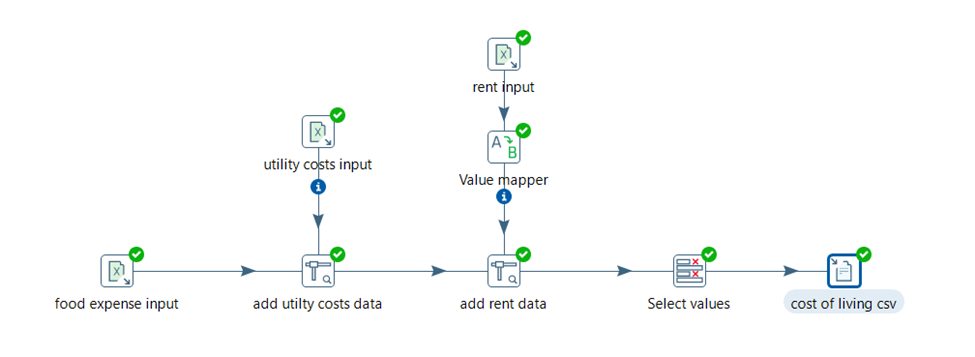
*Figure 5*



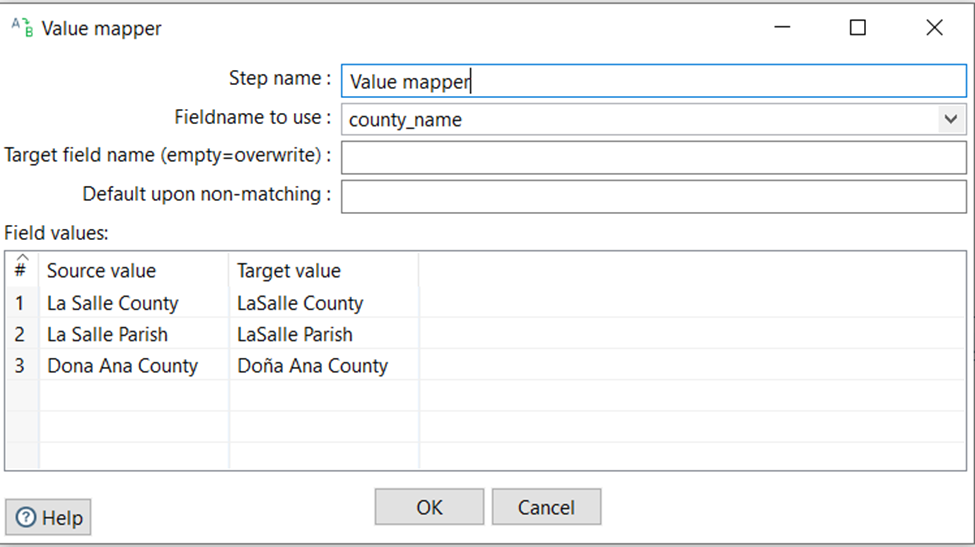
*Figure 6*

For health and education data, we extracted health data from an excel file and we removed unnecessary data. Then, education data was added. Lastly, we filtered rows that county value was not null and we loaded the data to a csv file(health & education csv file).

**Cost of living data**



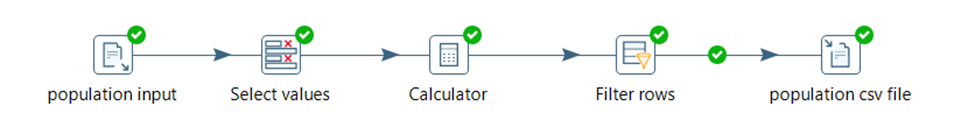
*Figure 7*



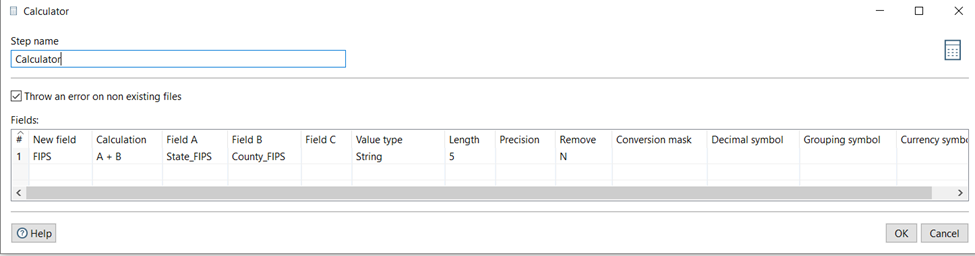
*Figure 8*

For cost of living data, we extracted food expense data from a csv file(food expense input). Also, we added utility costs data and rent data to cost of living data. Because some county names have different expressions, we used “Value mapper” to change those county names for consistency. Then we removed unnecessary columns and loaded the data to a csv file(cost of living csv).

**Population data**



*Figure 9*



*Figure 10*

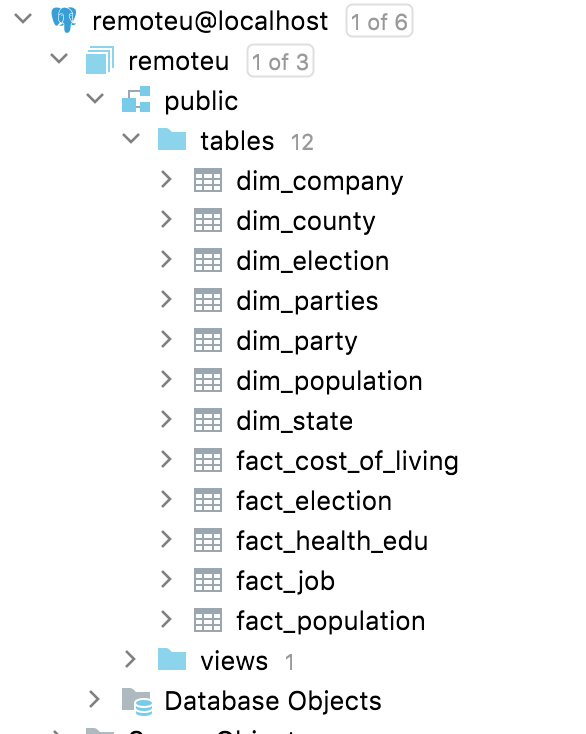
For population data, we extracted data from a csv file(population input). Because there was a lot of data in the csv file, we only selected data that was necessary. Then, we used “Calculator” to add State\_FIPS and County\_FIPS to a new column called FIPS to uniquely identify each county. After that, we used “Filter rows” to select rows that we needed. Finally, we loaded the data into a csv file(population csv file).

**Database and DDLs**

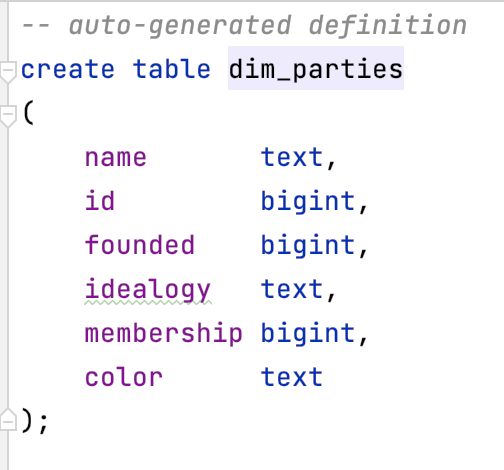
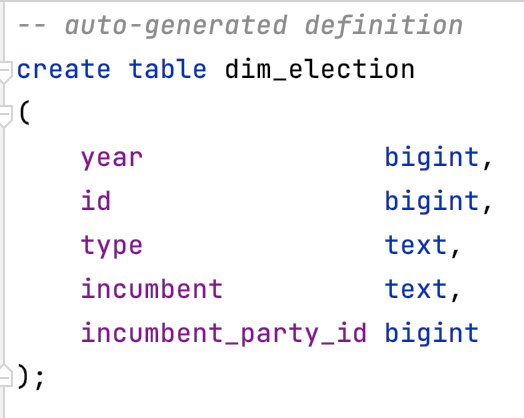
Once CSV files were cleaned through Pentaho, we used Python to load the data into Postgres for easy querying and plugging into Tableau. We used pandas (a python library) to automatically generate DDLs and insert the data in bulk.



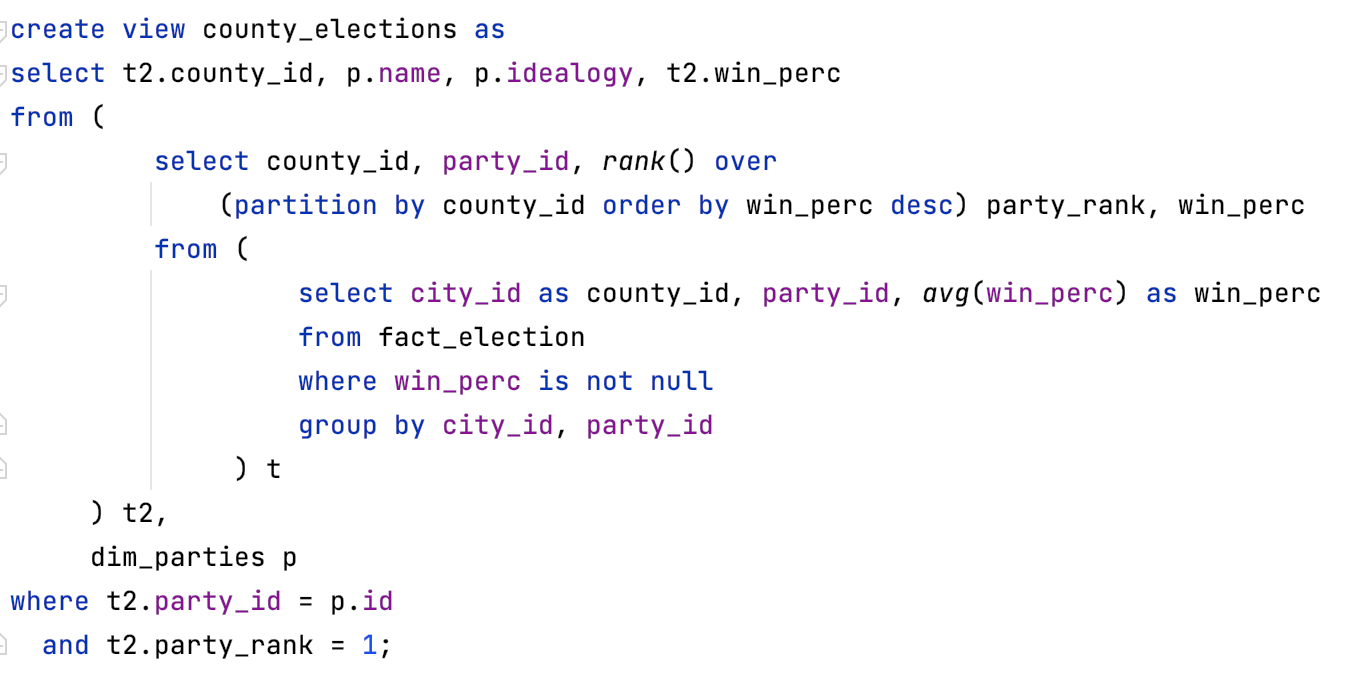
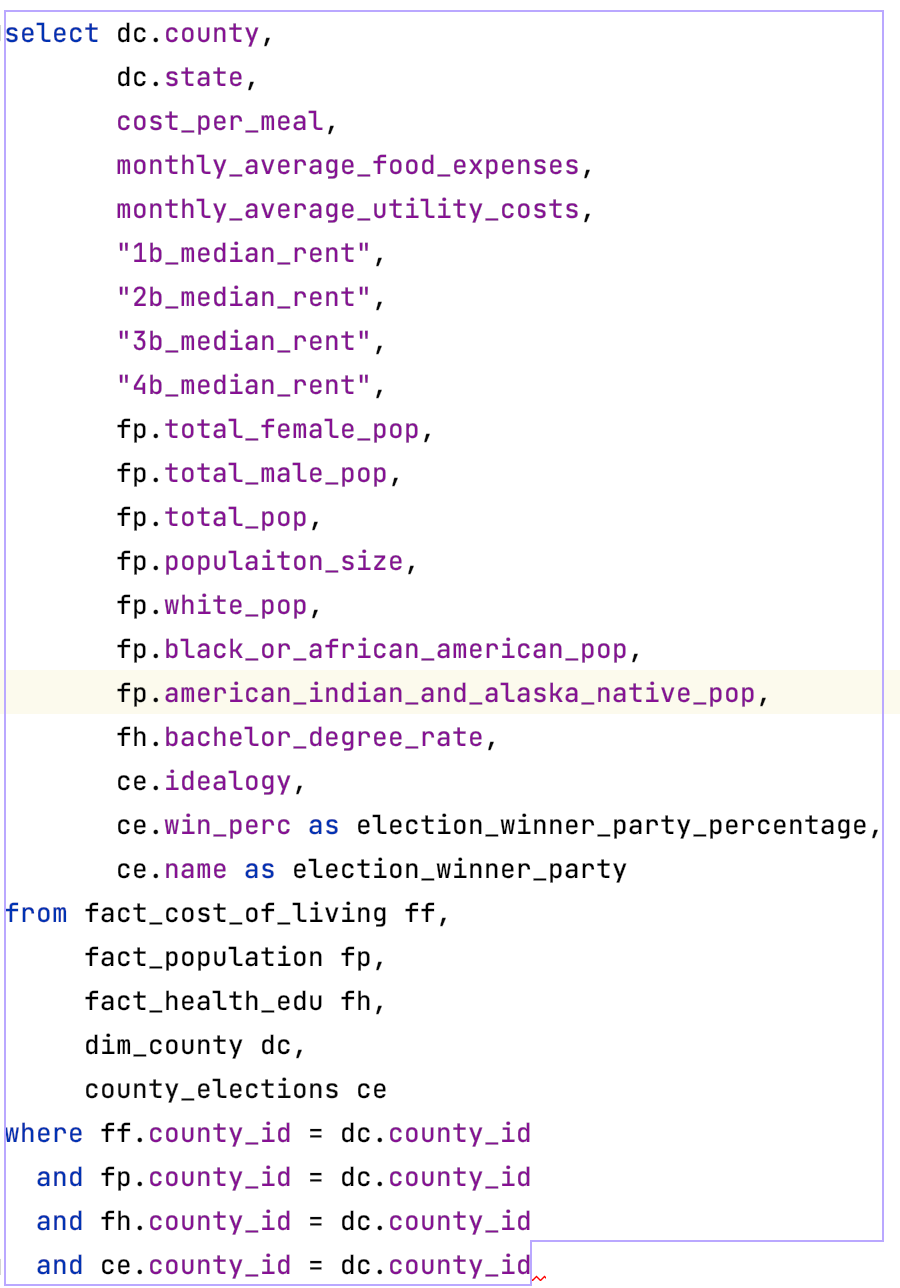
*Fig. 10 Loading CSVs into Postgres in bulk*



*Fig 11. Postgres Tables*



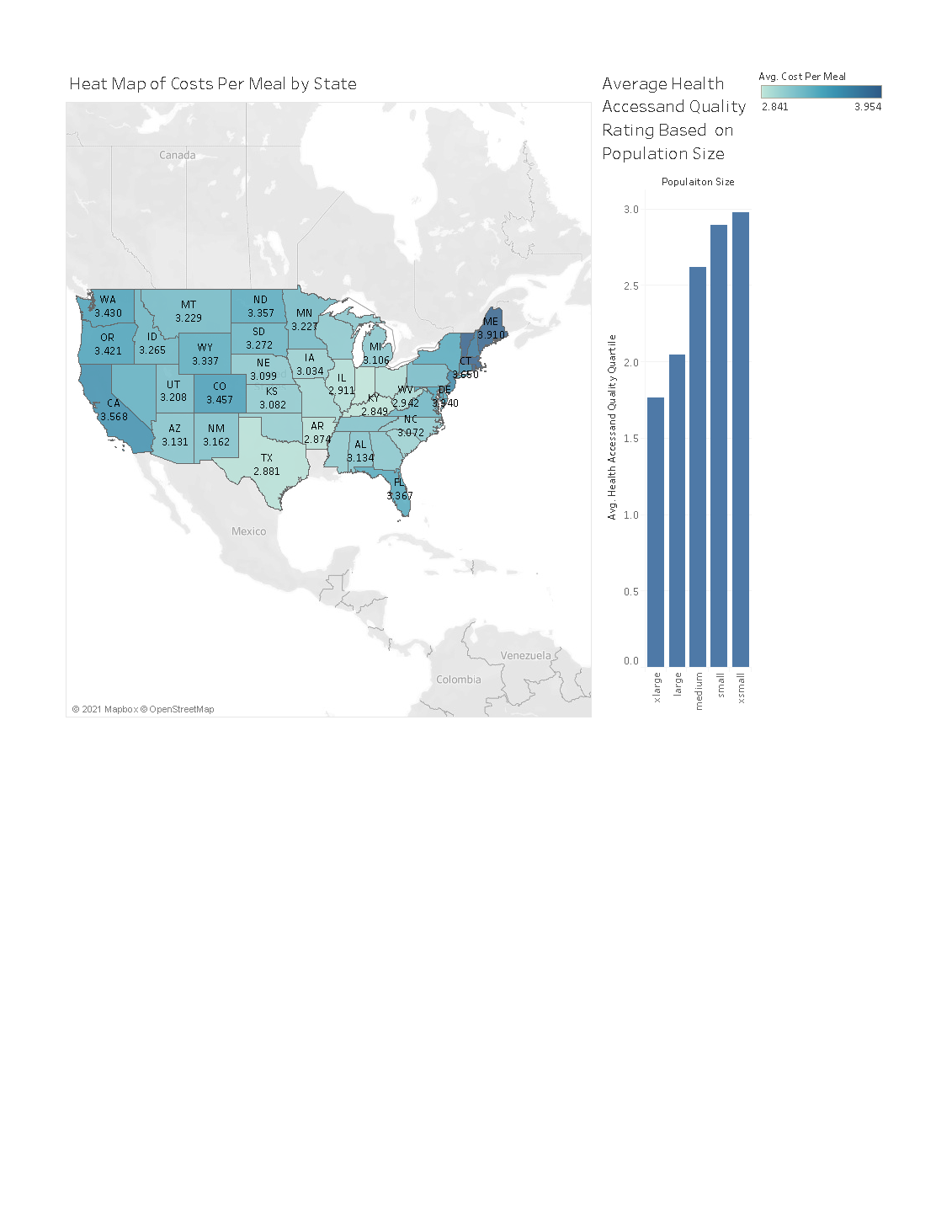
*Fig 12. Example of generated DDLs*



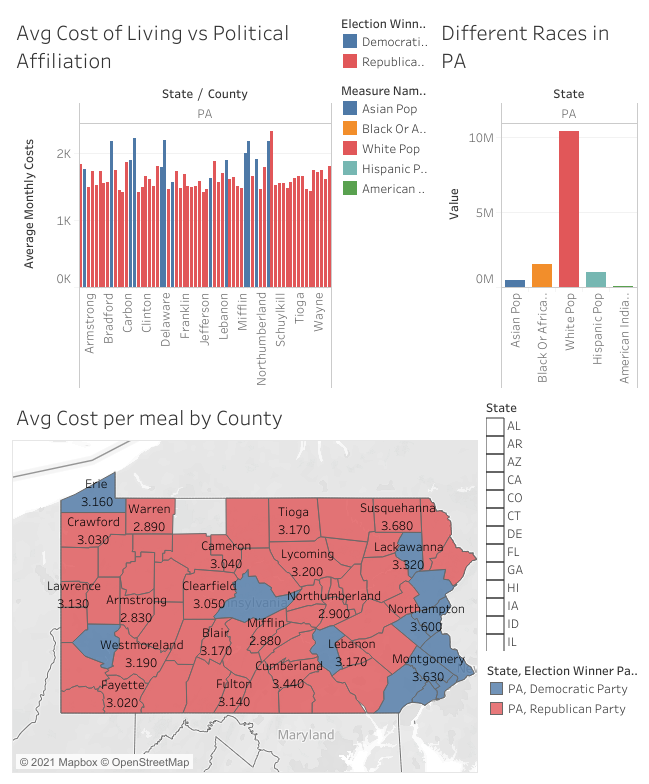
*Fig 12. Example of joins for visualization*

## Analytics and Visualization

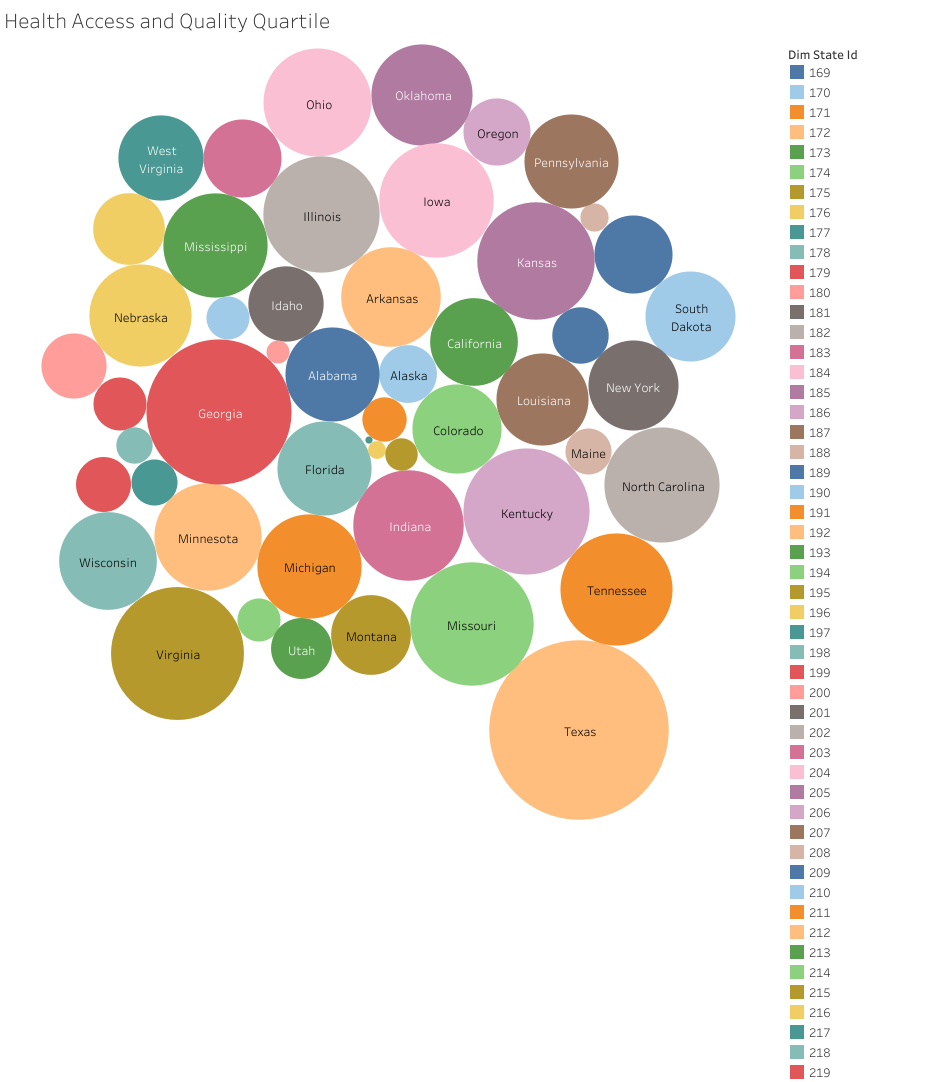
We performed various exploratory visualizations on the data we collectioned. We used Tableau for all visualizations. The data was joined and merged to make visualization and exploration easier via queries like the ones shown in Fig 12. The following are some examples of visualizations based off of our data warehouse.



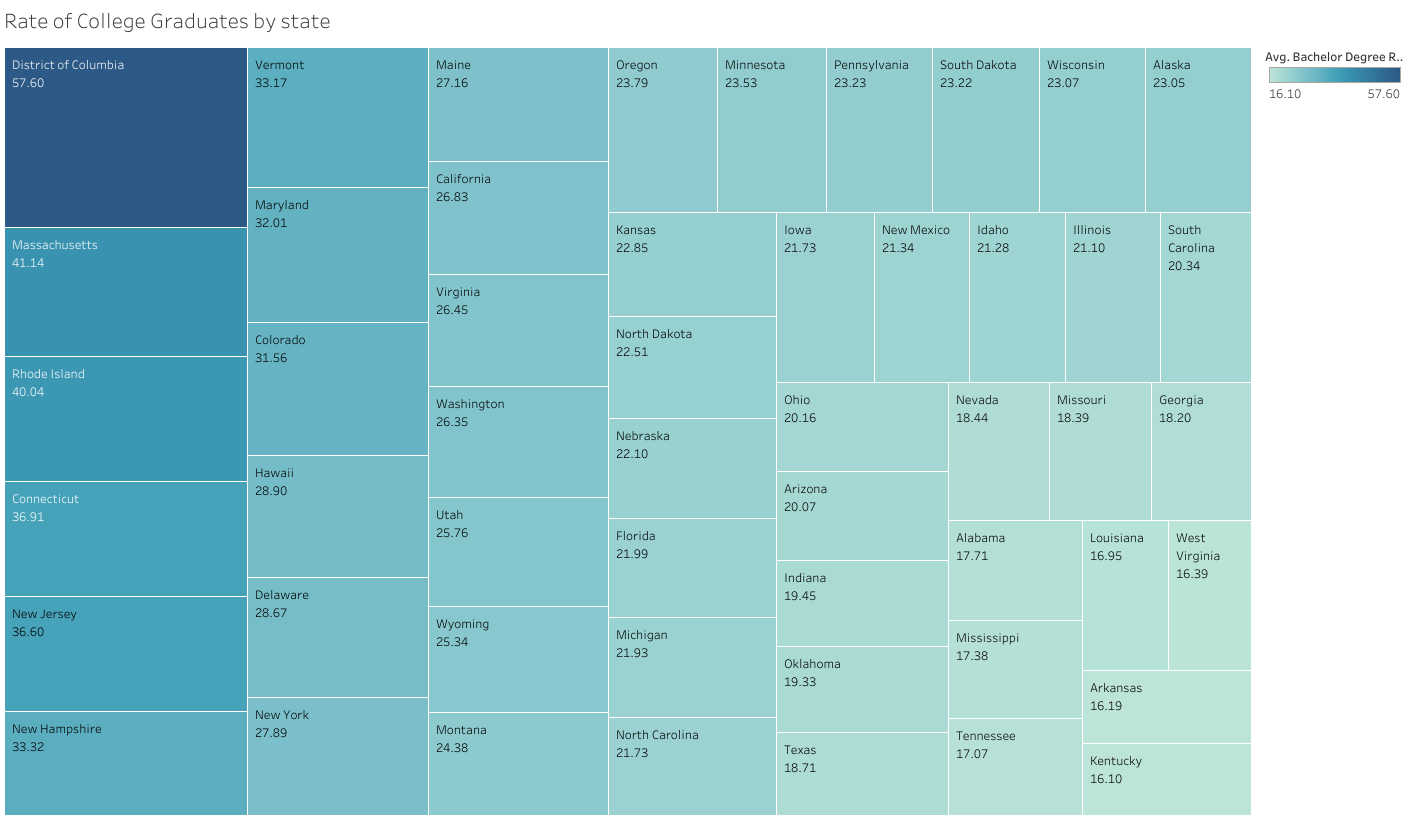
The heat map above shows the spread of the average cost of a meal by State. The darker the state map, the more expensive it is. Maine has the highest with $3.90 per meal. This will allow the users to get an idea of moving into states where the cost of surviving and food could be based on their budget. Our visualization allows us to see the average Health Accessand Quality Quartile by population size of counties. The higher the value, the better health facilities these areas have. It seems that the smaller the population of a county, the better healthcare facilities they provide.



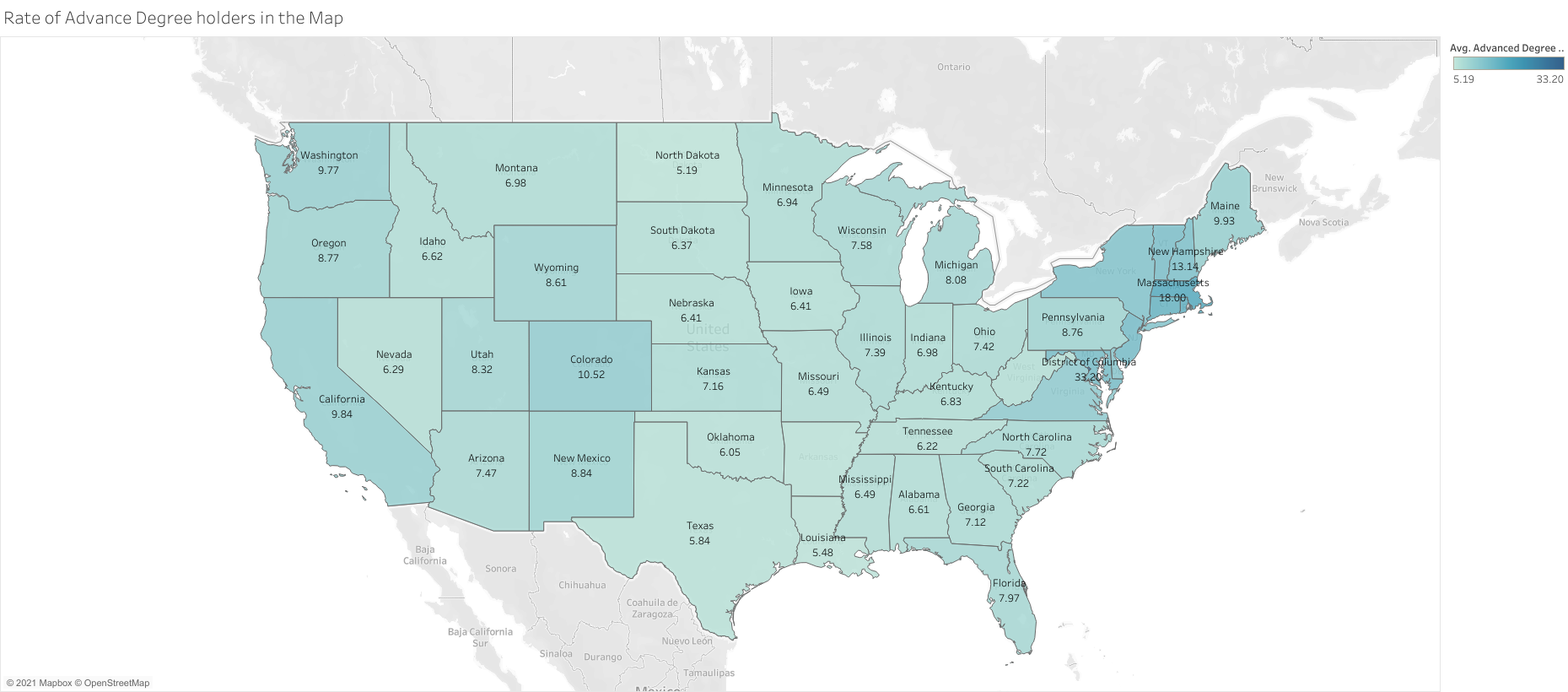
The above dashboard has a number of visualizations that will advise the user on where they would like to move. This specific example shows details for the State of Pennsylvania, which can be set to any other state in the country. The stacked columns above show the average monthly costs of living in each county of Pennsylvania based on how the county voted. The costs include average rent, monthly average costs for food and utilities. The values show that the counties where the cost of living is higher are generally more right leaning and voted for the Republican party this past election. We also have a bar graph where the columns show the spread of the different races of people that live in Pennsylvania. The map shows the spread of each county by the political affiliation as well, along with the average cost of a meal in each county.



The bubble graph shows which States have the best health access quality. The larger the bubble, the better their overall healthcare system quality. Looking at the graph above, it shows Texas has the best healthcare facilities off of all the other states in the country.



The above visualization shows the rate of college graduates by state. The larger the block and darker the shade of red, the more the percentage of the population have a bachelor’s degree. The value underneath the name of the state shows the rate of college graduates.



The heatmap shows what the rate of the population by each state has an advanced degree. The darker the shade of blue, the higher the rate of advanced degree holders. This would also give our user a general idea of states they might consider moving to, especially considering if the value of education in their surroundings is important to the user.

## Conclusion

Overall we found that the ETL and data cleaning stages were the hardest. As we looked at data from various sources, aligning the data format and various codes and conversions used all over the various agencies that compiled the data was the most time consuming and frustrating stage of our project.

In addition, we actually found that there was a large quantity of data online for the types of data we were looking for. We found some datasets that looked promising were actually incomplete or were of low quality. Thus, it took more effort for us to sift through a long list of data we downloaded from various sources.

As we spent a good chunk of the semester on the ETL and cleaning process, the stage of data visualization turned out to be the easiest and most straightforward. This tells us that, in BI it is crucial to spend as much effort as possible in cleaning and standardizing data before jumping to the visualization stage.

We learned quite a bit throughout this process, the most important of which was to start the ETL process earlier. In addition, we learned that it is important to carefully examine datasets before spending time and effort in trying to integrate them into the ETL and warehousing process. A good data source “filtering” strategy or protocol would have helped us at the start. For instance, we could have only looked at government and high reputation websites as they tend to have the most complete and highest quality data.

Our data itself has taught us a great deal. For instance, now we know that Texas looks like a great place to live with good healthcare and cheap cost of living. That we found various facts that surprised and informed us, tells us that the project was a success and would prove very useful for all remote workers who are considering a move.

1. Kilbride, J. (2021). IBM Study: COVID-19 Is Significantly Altering U.S. Consumer Behavior and Plans Post-Crisis. IBM. Retrieved December 5, 2021, from https://newsroom.ibm.com/2020-05-01-IBM-Study-COVID-19-Is-Significantly-Altering-U-S-Consumer-Behavior-and-Plans-Post-Crisis [↑](#footnote-ref-0)
2. (n.d.). Economist report: Remote Workers on the move. Upwork. Retrieved December 5, 2021, from https://www.upwork.com/press/releases/economist-report-remote-workers-on-the-move. [↑](#footnote-ref-1)
3. <https://www.countyhealthrankings.org/reports/county-health-rankings-reports> [↑](#footnote-ref-2)
4. https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model [↑](#footnote-ref-3)
5. https://www.census.gov/topics/education/educational-attainment.html [↑](#footnote-ref-4)
6. https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/ [↑](#footnote-ref-5)
7. Gundersen, C., Strayer, M., Dewey, A., Hake, M., & Engelhard, E. (2021). Map the Meal Gap 2021: An Analysis of County and Congressional District Food Insecurity and County Food Cost in the United States in 2019. Feeding America. [↑](#footnote-ref-6)
8. https://www.huduser.gov/portal/datasets/50per.html [↑](#footnote-ref-7)
9. https://www.move.org/utility-bills-101/#Average\_Utility\_Costs\_by\_State [↑](#footnote-ref-8)
10. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ [↑](#footnote-ref-9)