

# Healthcare Insurance Analysis Process Book

The analysis of specific trends in the healthcare insurance industry in recent times

Presented To

Hamza Elhamdadi

University of Massachusetts Amherst

Prepared By

Gauri Arvind

Akshitaa Jay

Tiven Pangtono

May 12, 2025

# Table of Contents

---

|                                 |   |
|---------------------------------|---|
| Overview and Motivation .....   | 1 |
| Related Work .....              | 2 |
| Questions .....                 | 3 |
| Data .....                      | 4 |
| Exploratory Data Analysis ..... | 5 |
| Design Evolution .....          | 6 |
| Implementation .....            | 7 |
| Evaluation .....                | 8 |

# Overview and Motivation

---

Health Care refers to an organized provision of medical services to maintain and improve human health. However, in the United States, U.S. healthcare is more of a business, and patients are paying customers. A key component in this system is health insurance, which plays a crucial role in determining who gets access to which type of care and the costs involved.

One major concern for healthcare however is its rising costs. For example in 2025, the average monthly premium for the silver plan under the Affordable Care Act was approximately \$621 which was a 7% increase from last year. Despite being a mid-tier plan, the Silver plan is still quite expensive and can be financially burdensome for many even with deductibles, copayments, and even out-of-pocket.

In addition to costs, many Americans still face frustrating and sometimes devastating experiences of having their insurance claims denied. In fact, in 2023, insurers of marketplace plans denied about 20% of all in-network claims and even a higher percentage of out-of-network care. Shockingly, fewer than 1% of denials were ever appealed often due to the lack of awareness or how complex the appeals process is.

Overall, there are disparities when it comes to access, claim approvals, and cost-effectiveness that create concerns when it comes to the efficiency of the existing insurance model. Our goal is to help create an effective data-driven solution by identifying trends and helping to analyze these issues with factors such as income levels, unemployment status, and geographic location influence.

## Related Work

---

Although we unanimously agreed to do health insurance. We were somewhat influenced by the events that happened earlier that year involving Luigi Mangione. Luigi Mangione was infamous for the assassination of the United HealthCare CEO Brian Thompson in cold blood. His motivation is his deep-seated hatred of the health insurance industry and calling it “parasitic” and driven by corporate greed. Another factor that is implied to be part of this is the fact that he was suffering from spondylolisthesis, which occurs when a fracture causes a vertebra to slip out of alignment. Although there was a procedure that would correct this issue, it is speculated that Luigi had to fight time and again against his health insurance provider and they kept denying to cover his treatment.

The assassination of the Healthcare CEO also raised some questions about the U.S. Healthcare system. More specifically, how expensive it is as ¼ adults have skipped off on health care they needed due to the price tag. Which translates to about 100 million adults. Furthermore, investigations done by the U.S. Governmental Accountability Office in 2020 showed that the estimated prices for 20 selected brand prescriptions were more than 2-4 times higher than other prices in Australia, Canada, and France. In addition to the expensive prices, health insurance doesn’t even cover most of the medical costs and most people pay out of pocket even with deductibles. This is not only terrible for the person paying those costs, but even medical staff responsible for their care are being taken advantage of such as getting underpaid for their work. Furthermore, what we found disgusting was insurance subrogation where when insurance covers a cost for you, they make someone else responsible for paying that amount which is a euphemism for basically being a debt collector and harassing people. Some of the resources that inspired us include:

[https://abcnews.go.com/US/ceos-murder-sparks-widespread-criticism-us-health-care/story?id=116677583&utm\\_source=chatgpt.com](https://abcnews.go.com/US/ceos-murder-sparks-widespread-criticism-us-health-care/story?id=116677583&utm_source=chatgpt.com)

<https://www.cbsnews.com/news/luigi-mangione-healthcare-ceo-shooting-what-we-know/>

<https://www.youtube.com/watch?v=hJbzkwGkZ3E&t=1s>

# Questions

---

Overall, we have multiple questions that we aim to answer with our visualizations. Some of our initial questions are as follows:

- What is the best health insurance for patients based on patient statistics, healthcare needs, and insurance providers' costs and coverage?
- How many people in the US are insured compared to the number of people in the US who are uninsured?
- What are the average costs of health insurance in each state in the US?
- What are the percentages of all listed insurance plans that are used by state?
- What is the percentage of those who receive insurance from their employer?

As of the first milestone deadline, these questions have not changed. New questions will be added throughout the duration of the project.

For our final submission, we also wanted to answer the question of how much each state spends on healthcare expenditures, which is a modification of the third question. The second question is also modified to account for states specifically rather than the entire country. We did remove the first question due to an inability to find datasets related to insurance providers' cost and coverage and the broadness of the question.

# Data

---

The primary dataset we are utilizing is from Kaggle, which is an online data science and machine learning platform, which is accessible through this link:

<https://www.kaggle.com/datasets/hhs/health-insurance/data>. The dataset is based on health insurance coverage by state in the United States and originates from the US Department of Health and Human Services and the US Census Bureau. The data will be collected through a CSV file. There was an additional source of data considered but as it did not include all states, it was not used as a source of data for the first milestone.

From this dataset, the information we plan to derive from the data includes information on the state, the percentage of people who were uninsured in 2010 and 2015, the rate of change of those who were uninsured between 2010 and 2015. We may derive information on the number of people who are insured, the number of people who are insured by their company, health insurance coverage by the market, marketplace tax credits, average monthly tax credit, and state Medicaid expansion in 2016. For this dataset, the main processing that we are expecting to utilize is replacing missing values.

To clean the database, two main factors were cleaned, which included removing special characters in order to parse the data efficiently and handling empty values. In terms of the empty values, there were two scenarios that occurred. The first scenario was that the states of Connecticut and Maryland did not have a Medicaid Enrollment for 2013, and the Medicaid Enrollment Change was left empty. For this scenario, the value was replaced with -1 as a placeholder when parsing so that when it is added into the application, it can be handled more effectively. The second was that for State Medicaid Expansion, there was an empty value for the United States, which was replaced with FALSE as it would not have had an expansion regardless, and also serves as a placeholder.

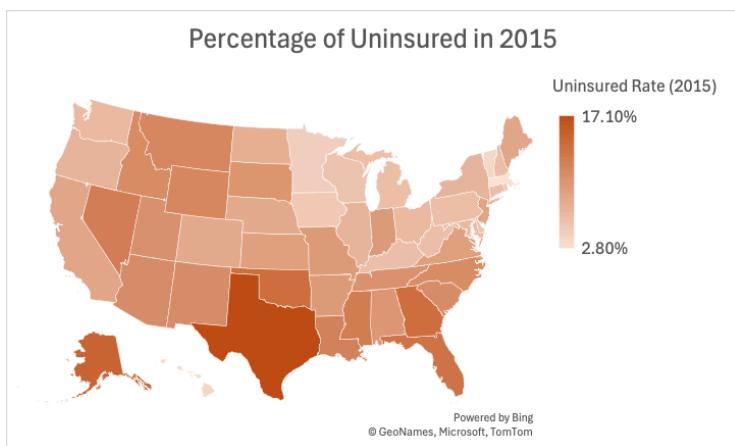
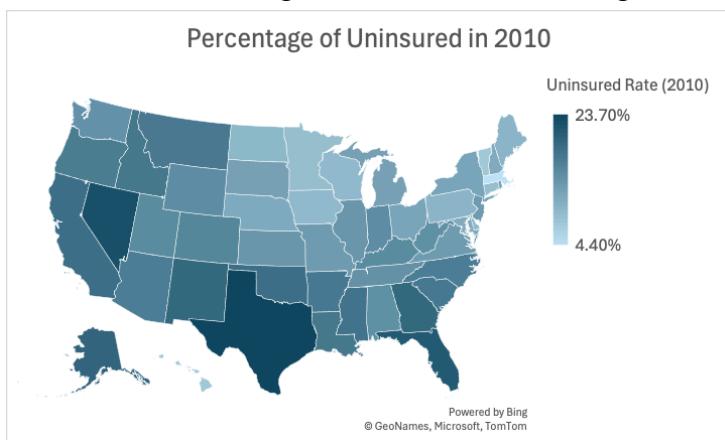
To answer our questions more thoroughly, we found an additional dataset that covers the health insurance coverage of the total population. This dataset includes records between 2008 and 2023. This is compiled by KFF based on the American Community Survey for the aforementioned years. We may derive information from all columns to answer our main questions. The link to this dataset is listed below: [Health Insurance Coverage of the Total Population, Multiple Sources of Coverage | KFF](#). This dataset was processed by removing special characters and parsing through each file with Python.

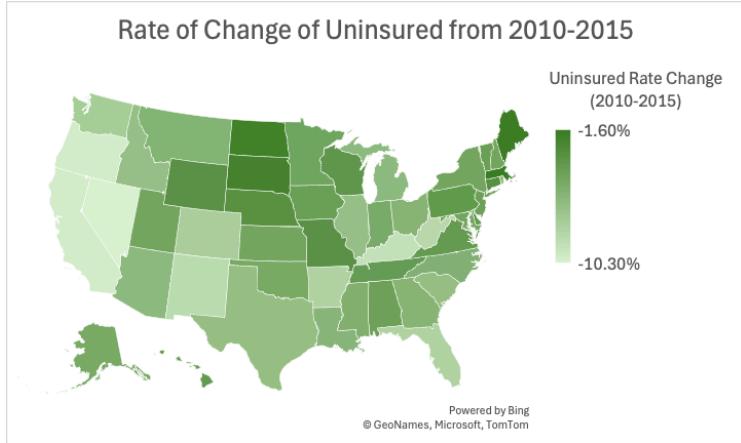
In addition, we have found another webpage that covers the total health expenditures of each state in a given year. This allows us to answer the question of what are the average costs of health insurance in each state across multiple years. This data source includes the location and

the total health spending based on a specific year, which ranges from 1991 to 2020. The data itself is sourced from the Centers for Medicare and Medicaid Services (CMS) Office of the Actuary, which creates the Health Expenditures by State of Residence and Health Expenditures by State of Provider. The site that sources the data is KFF, and the link to the dataset is listed here: [Health Care Expenditures by State of Residence \(in millions\) | KFF](#). This dataset was also processed by removing special characters and parsing through each file with Python.

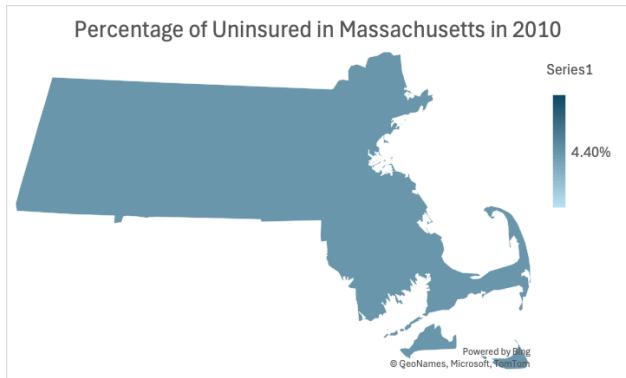
# Exploratory Data Analysis

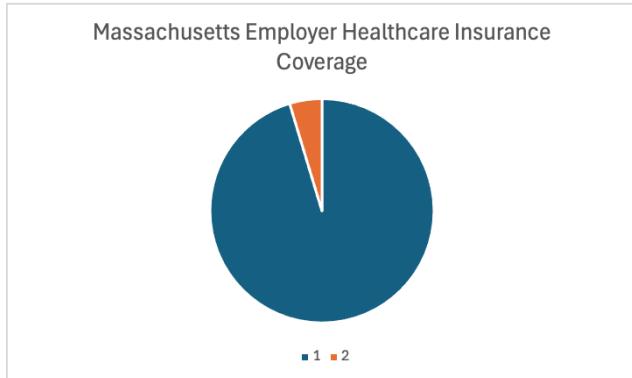
To initially view the data, we used several different visualizations based on our final draft of our prototype design, which is currently based on our primary dataset. These visualizations were created in Excel. The first type of visualizations used were map visualizations of the percentage of those who were uninsured in 2010, 2015, and between 2010 and 2015. These visualizations show that some regions of the United States have varying levels of those who are uninsured that stay relatively similar over time, and the percentage of those who are uninsured falls over five years. In terms of the rate of change of the uninsured, it also seems to vary by region in the United States, with most West Coast states seeing the greatest change and with Midwest and Northeast states seeing the least amount of change. These three visualizations are shown below:





The second set of visualizations is based on the focused portion that the user selects. This visualization is based on Massachusetts for the time being. The first set is a percentage of uninsured in Massachusetts in 2010, which displays the percentage for the entire state rather than by smaller increments such as county or town. In addition, the color hue is not currently relative to other states, which creates difficulty in discerning its meaning in Excel. The second visualization is a pie chart of the types of health insurance coverage in Massachusetts, where the dark blue coverage is employer healthcare coverage in 2015 and the second is marketplace healthcare coverage in 2016. This shows that overall, most of Massachusetts has healthcare insurance under employers, and has very little marketplace healthcare insurance coverage. There is one additional visualization for creating a table of key information for each state's health insurance, but this was not feasible to do in Excel. The main two visualizations are shown below:



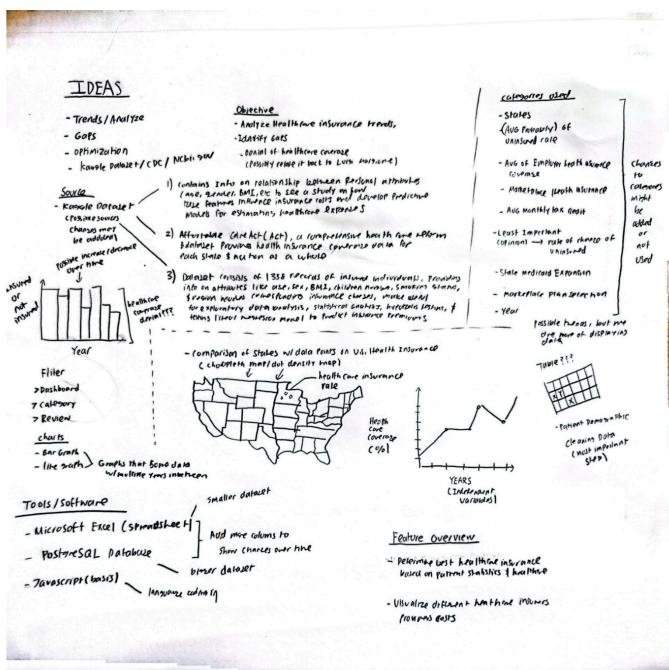


As of Milestone 1, these designs are the main sources of inspiration, except for the map describing the Rate of Change of Uninsured from 2010-2015. These designs will be used to inform the decisions made to create the overall visualizations during the project. The only change would be to not include the colormap series in the Percentage of Uninsured by state, as the entire state is described by one specific color and was not a part of the original design of the focus portion of the visualization.

# Design Evolution

Our initial visualizations for our prototype designs are as follows, based on the Five-Design Sheet Methodology. As of the first milestone, these were the visualizations considered.

**Initial Prototype Design:** This design is just brainstorming ideas that we are going to use. This includes the objective which is to analyze healthcare insurance trends, identify gaps, and possibly the denial of healthcare coverage(optional objective). This included types of graphs such as a choropleth map/dot density map to show the percentages across the healthcare insurance rates, line and bar graphs to show healthcare coverage and insured/not insured people over the years, and tables with patient demographics (optional idea). Furthermore, this includes a feature overview that would determine the best healthcare insurance based on patient statistics and healthcare and visualize different healthcare insurers and providers' costs.



CS CamScanner

**Alternative Prototype Design 1:** This design includes a geographical map of the United States and the percentage of those who are uninsured by state. When one state is chosen, additional information about insurance rates and types of insurance are displayed in a separate portion in

the form of the state itself and a pie chart. The geographic map shows the percentages across states to provide a visual and immediate understanding of uninsured rates across the United States. The selection of an individual state pinpoints the uninsured rate to a specific state, and the pie chart shows the distribution of health insurance plans for each state.

LAYOUT

metadatas

TII, Healthcare Insurance Analysis Initial Design 1  
Group members: Gauri Aravind, Architaka Jais, Tien  
Target audience: Government  
Task: Complete the second sheet on the first initial design  
sheet # 2

Operations

- The user can select a year to display the percentage of those who are uninsured in each state.
- The data is pre-loaded and starts with the year 2010.
- Selecting the state will reveal more health insurance information about the state.
- The graph is shaded in a continuous colormap so that the user is able to view the difference in uninsured rate by state.
- The user can only select one year at a time and cannot view data from different years.
- In the focused portion, the user can also view changes per year.

FOCUS

Example: The user selects Colorado. A panel below the US map displays the following information:

CO

Year  
2010  
2015  
2020

Employer Health Coverage  
 Marketplace Health Coverage

Discussion

- Some advantages include the continuous colormap to view information in a fluid manner, the geographical map to provide context, and a separate window to view more information about each state.
- Some disadvantages include the limited amount of information that can be displayed in a pie chart and the lack of a hover option to view information by state.
- The dashboard is suitable for our analysis of health insurance.
- It is somewhat possible to implement and somewhat scalable for more data but is limited to the US.
- The overall design is clear in purpose.

Alternative Prototype Design 2: Similar to the previous design, a geographic map of the United States is shown with the percentage of those who are uninsured in the United States. Unlike the previous map, when a state is selected or hovered on, it shows the state itself as well as a table of health insurance information for each state. The inclusion of the table alongside the geographic map allows for all the data to be shown, allowing the user to understand the distribution of health insurance plans for each state.

Metodala

Title: Healthcare Insurance Analysis Initial Design 2  
 Group Members: Gavin Arvind, Architaa Josy,  
 and Nivin Rengaraju  
 ToDos: Complete the third sheet of the Five  
 Design sheet methodology  
 Sheet #: 3

Operations

- The user can select a year to display the percentage of those who are uninsured in each state.
- The data is pre-loaded and starts in 2010.
- Scrolling or hovering over the bars will reveal more health insurance information about the state.
- The additional information is displayed in the bottom half of the screen.
- The user can only select one year at a time and in the focused portion, the user can select different years without it having an effect on the above map.

Discussion

- Some advantages include that all information can be displayed on the visualization, that the user can hover over and select the state, and that the color map shows a continuous change in the data.
- Some disadvantages are that the table is not as abstract, multiple years cannot be viewed at a time, and the years are not distinguishable by color.
- The dashboard is suitable for our analysis.
- It is somewhat feasible, more so than the previous visualization due to the table and it is somewhat scalable.
- The overall design is clear in purpose.

| Health Insurance Coverage    | Value |
|------------------------------|-------|
| Employer Health Insurance    | ~     |
| Marketplace Health Insurance | ~     |
| Marketplace Tax Credits      | ~     |
| Average Monthly Tax Credit   | ~     |

### Alternative Prototype Design 3:

As with the previous two designs shown above, we have incorporated a map design that allows the user to select which state they would like to see a data visualization for insured vs uninsured percentage demographics along with the change from year to year in a line graph, with a dotted line representing uninsured patient percentages vs the solid line representing the insured patient percentage.

**FOCUS:**

A line graph titled "Healthcare Reform Timeline" with the y-axis labeled "Healthcare Reform Percentage" and the x-axis labeled "years". The graph shows a dashed line starting at 0% in 2013, rising to about 40% by 2014, dipping slightly, then rising again to approximately 60% by 2016.

**KEY:**

- state insured
- state uninsured

lines will be color assigned to a state, like Colorado is green

**metadara**

**Title:** Healthcare Insurance Analysis

**Author:** Savitri Arvind, Prashita Ray, Riven Pongrone

**Task:** Complete fourth sheet of design sheet methodology sheet. 4

**Operations:**

- user can select a state on the map to see the line graph for insured vs uninsured patient percentages by state
- user can select one state at a time

**DISCUSSION:**

- advantage: prevents overcrowding of multiple lines representing all states all in one graph, as it breakdown data per state and allows for a clean/straight to the point visualization.
- disadvantage: is not being able to do a comparison among states data, as each state is individually represented on a different line graph. To fix this issue, we can develop a feature that allows overlap between two line graphs each representing a different state to do a comparison between the data.

**LAYOUT:**

A hand-drawn map of the United States with a box labeled "my state" pointing to the state of Maryland. Below the map, the text "Insurance coverage by state" is written.

### Finalized Design:

The final design is more refined with the layout from the United States Map and the line graph. There will possibly be a menu option that allows you to switch between the two graphs and a feature that will be added to make the tables more abstract and allow comparison by creating an overlap between two line graphs each representing a different state. For the Focus, I added the pie chart on design one for a more appealing visualization however it might be excluded due to lack of information unless a key is in place. I've also included the table and year legend as the alternative option. If all goes well we can implement all of them together.

**LAYOUT**

**Metadata**

**Title:** Healthcare Insurance Analysis

**Group Members:** Nitin Pangarkar, Gauri Arvind, Akashna Jay

**Task:** Final Design for Five sheet Methodology

**Operations, Line Graph**

- User can select a state on the map to see the line graph for insured vs uninsured percentage by state
- User can detect one state at a time

**Operation (United States Graph)**

- User can select & display the percentage of those who are uninsured in each state
- The data is pre-loaded and starts in 2010
- Additional information is displayed in bottom half of the screen

- User can select any year in time, if in the forward format, the user can select different years without having to go back

Forward portion, the user can view changes per year

- Graph shaded in a continuous colormap so user can view difference in uninsured rate by state

**DISCUSSION**

- The Dashboard is suitable for analysis
- Overall design is clear in purpose
- Somewhat coarse, due to somewhat heavy scaling for more data, but it is limited to United States

**Advantages**

- Some advantages include continuous colormap to view information in a fluid manner, the geographic map to provide context, & a separate window
- Some advantages are the overlapping of multiple lines as it is continuous, data per state follows for a crown/straight line point visualization

**Disadvantages**

- Some disadvantages are that the bubble is not as abstract, multiple years cannot be viewed at a time, and the years are not distinguishable by color
- Can't do comparison among state data as each state is individually represented on a different line graph
- Fix to this would be develop a feature that allows overlap of two line graphs such representing a different state to do a comparison between the data

CS CamScanner

The current design for the first milestone, which is based on the finalized prototype design, follows the key design principles of the course. The first set of principles it follows are the effectiveness and expressiveness principles. In terms of the effectiveness principle, the utilization of spatial regions, color hues, and saturation to show the rate of uninsured by state as well as the utilization of color hue, area or length, and/or aligned position in the remaining graphs shows that the most effective channel has been used for visualizing the data in each graph. In addition, utilizing a continuous colormap for the map allows for minute differences in uninsured rates to be displayed effectively throughout the map visualization. In terms of the line chart, using different one-dimensional markings allows the line to be easily distinguishable by the user. As of the first milestone, we did not deviate from the project proposal.

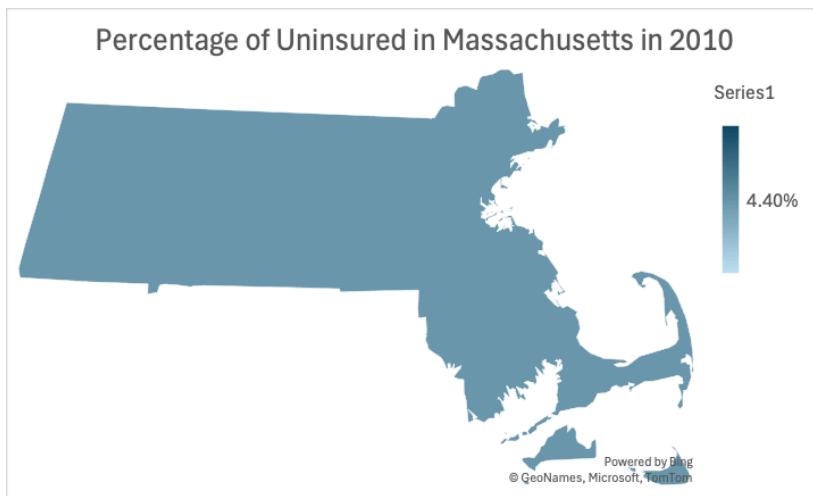
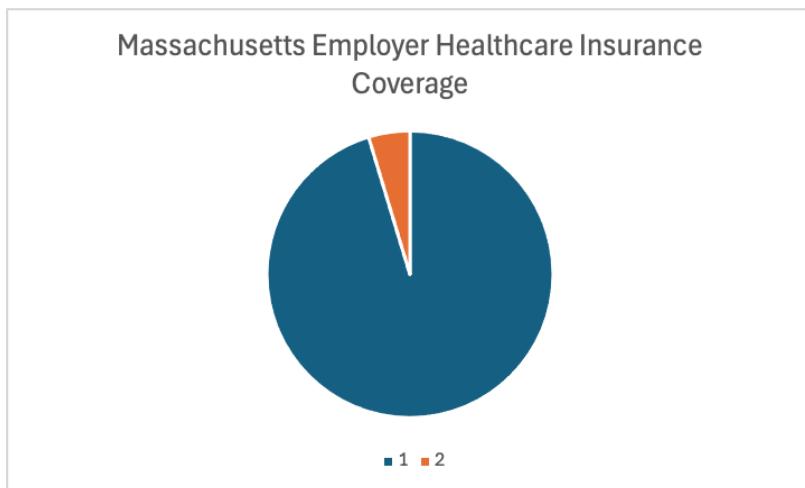
After the first milestone, there were a few changes made to the design for the focus portion of the application based on availability of designs given the data. While developing visualizations, it was noted that the visualization of the state as well as the table would not be possible with either the data in the dataset itself or the functionalities available in d3. With this in mind, we sought to create focus visualizations that were within the scope of the data and d3, which were changed to create a bar chart displaying the change in Medicaid enrollment for each state between 2013 and 2016. This was done to highlight key portions of the data effectively and ensure that the visualization itself is true to the intent established previously. However, due to the sizes of the data values visualized, the natural log was used to show the general change between the years instead of the precise value change. In addition, when designing the pie chart, there was a discrepancy with the data used, specifically that it belonged to different years. To change this, two key points from the same year were used, which focused on Marketplace Health Insurance Coverage and Marketplace Tax Credits. Both of these follow the Expressiveness Principle as well as the Effectiveness Principle by using the most effective channels for the most important data. Specifically, the pie chart uses aligned position, area, and color, while the bar chart uses aligned position, length, and color.

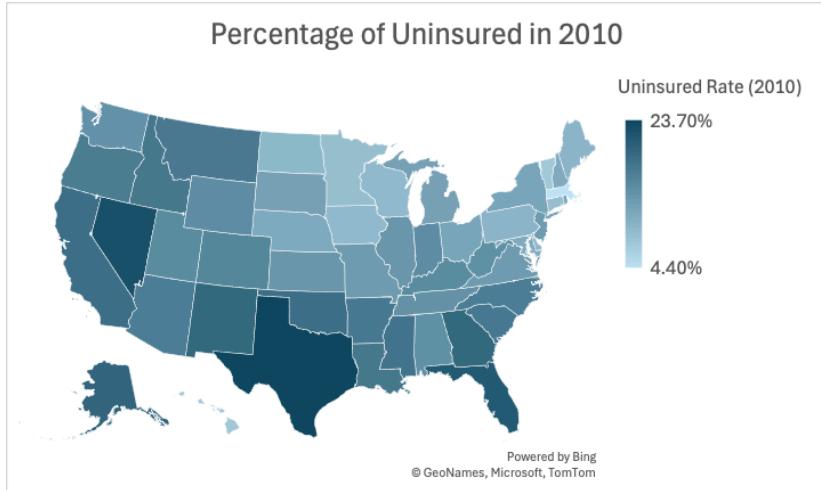
Afterwards, there were additional alterations made to the visualizations in the focus section. The pie chart was modified to represent the percentage of individuals that were insured versus the percentage of individuals that were not insured. In addition, a stacked bar chart was added to show the amount of different types of insurance coverage by year for the given state. The stacked bar chart displays this information by year from 2008 to 2023. This was done to show the distribution of insurance types by year and to show changes in insurance enrollment for each category over time. The categories were displayed with a categorical colormap to display the distinct categories of insurance types effectively. In terms of the Effectiveness Principle, the data follows the identity principles in that it uses the most effective channels for the most important aspects and follows with the less important aspects, specifically with spatial region and color hue. The visualization also follows the Expressiveness Principle as it does not include chart junk.

Lastly, we added an interactive line chart that displays the costs of healthcare for each state over the course of the years 2008 to 2020. This graph is interactive in that the state's line itself can be selected to be included from the chart, selected to be removed from the chart, and have the color displayed at the bottom of the chart in a legend. The individual state lines in the chart are also displayed using a categorical colormap. In terms of the Effectiveness Principle, it follows it effectively for the magnitude channels of aligned position, length, and color. It also follows the Expressiveness Principle due to a lack of chart junk.

# Implementation

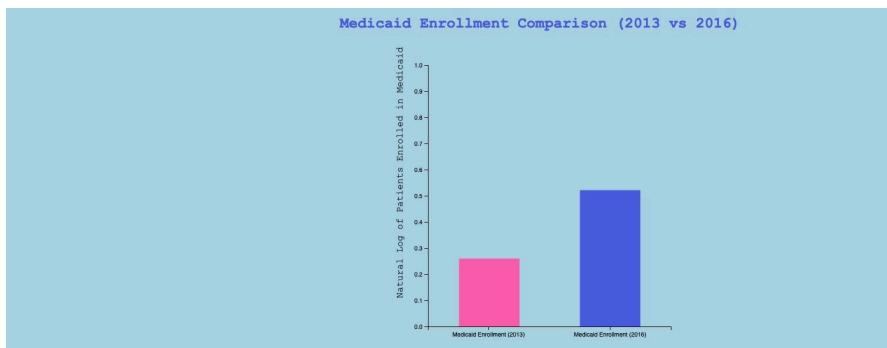
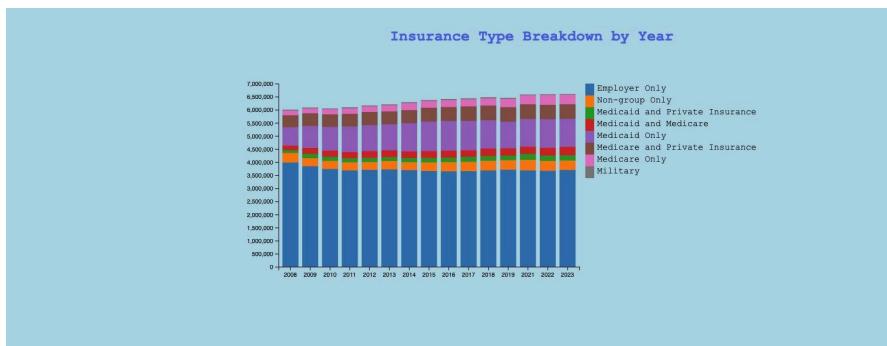
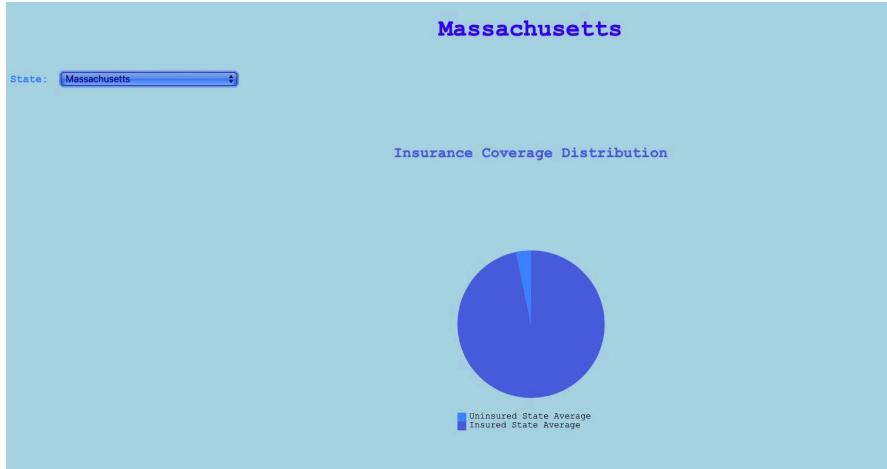
We currently don't have any visualizations running on our website as of now. We have pictures of example visualizations we plan on implementing. We aim to develop a visualization of the U.S map as well as a drop down menu or an interactive map where the user is able to either click on or select the state they are from and see the visualizations for that state. Below we have images of what a visualization for Massachusetts would possibly look like a pie chart as well as a map of the state. As well as the percentage of uninsured overall U.S map by year here below we have an example of 2010 uninsured U.S map.



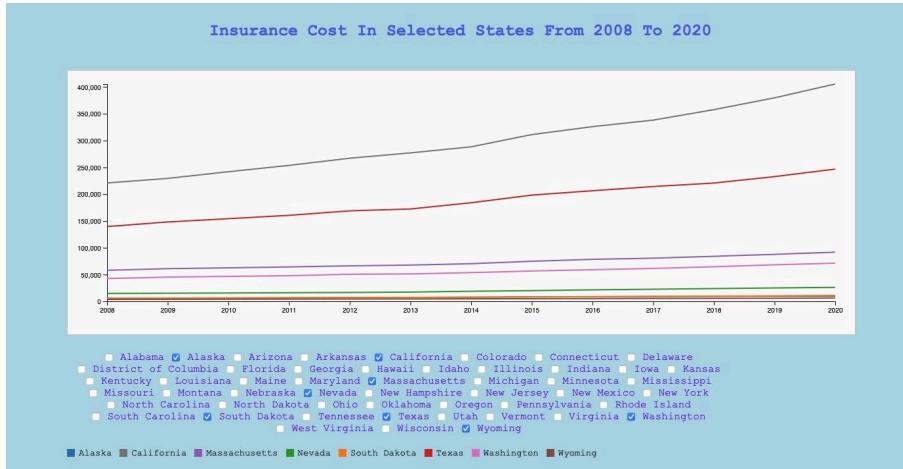


In terms of the focus section, the implementation for this section is as follows. For the pie chart, the aim was to display the difference between Marketplace Health Insurance Coverage and Marketplace Tax Credits in 2016. This highlights the percentage of tax credits that can cover health insurance, and displays it for each state. As a result, users can visualize how much their tax credits cover their insurance.

The focus section has been modified to include a new version of the pie chart as well as a stacked bar chart. The pie chart now refers to the Uninsured State Average and the Insured State Average from 2008 to 2023, and is distinguished by color. The stacked bar chart covers the Insurance Type Breakdown by Year for the years 2008 to 2023. This specifically covers information based on Employer Only, Non-group Only, Medicaid and Private Insurance, Medicaid and Medicare, Medicaid Only, Medicare and Private Insurance, Medicare Only, and Military. This allows users to view the distribution of healthcare insurance types by year as well as the number of enrolled individuals for each category. The images below describes the three visualizations of the focus charts and their ability to change based on the choice of state in the dropdown.



The main section, which is under the Main tab of our webpage, includes our main visualization of our line chart as mentioned in the Design Evolution. This specifically covers costs for the states selected from 2008 to 2020. The chart allows users to select the states they would like to see and deselect the states that they would prefer not to view. The image below describes these elements.



## Evaluation

---

We learned about handling the NA values in our data which could lead to inconsistencies and performing data cleaning on our insurance data set and decided on a larger dataset as the data set we initially selected did not answer many of our research questions as there was minimal data. We now have working visualizations including a bar chart and a pie chart for each state. The pie chart contains the number of insured vs uninsured people in avg from 2008 to 2023. The bar chart contains the breakdown of insured patients and how they are insured: Military, Employer, Private Insurance and much more by year in a stacked bar chart, with the total number of patients among all insured in the y axis.

From the pie chart, we learned the distribution of uninsured individuals versus insured individuals on average for each state. From the stacked bar chart, we learned what types of insurance are the most popular, the change in enrollment of different insurance types each year, and the overall change in the number of people enrolled in insurance. For the bar chart, we learned the distribution of Medicaid enrollment between 2013 and 2016 as well as the change in enrollment for Medicaid. From the line chart, we learned how state costs for insurance changes each year and how some states face less changes and less overall costs compared to other states.

Our visualizations helped to answer some of our research questions. In terms of the fourth question, the stacked bar chart shows the general range of types of insurance chosen, which gives the user a general sense of what types are more common for a given state for each year. It also helps to answer the final question. While the stacked bar chart does not show a percentage, it shows relatively that Employer insurance is the most common for the states. The pie chart answers the second question. Similarly, while the pie chart does not show percentages, the overall distribution of uninsured versus insured across a given state is present. The line chart answers our third question to display how the costs of insurance for a state changes over time, and gives a sense of the average cost over the course of the given timeframe.

Our visualizations do have a few major limitations. The first main limitation is the lack of interactivity for the focus charts specifically, which had complexities given the way the tabular design was made. The second limitation was interactivity with respect to the main chart, which also created complexities due to the selection nature for each state and the display of a single state in the focus section. The third and final limitation was the ability to scale the y axis based on state for the line chart, which is a little complex due to the fact that the default state of the visualization is empty.