Technical Report

Group 5

December 10, 2024

Introduction

For the final assessment, Group 5 examined mortality rates from the *NCHS Leading Causes of Death United States* dataset. This data was then analyzed with data based on cigarette smoking, obesity, Medicaid, and Medicare on Tableau.

Summary

By connecting the mortality rate dataset to other datasets, we learn key information about mortality. We discover that smoking correlates with lung disease and cancer, obesity correlates with diabetes, heart disease, and stroke. In addition we learn that Medicaid and Medicare spending is best spent on the more preventable death causes.

Dataset Deep Dive

For each of the four datasets, we deep dive

Smoking Process

The dataset used to examine cigarette smoking rates is the *Behavioral Risk Factor Data: Tobacco Use (2011 to present)* dataset from *data.gov*. The only necessary information in the dataset under the *MeasureDesc* column is the **SmokingStatus** value. All the non SmokingStatus values under *MeasureDesc* column are removed to avoid mass filtering in the dataset. For the *Gender*, *Race*, *Age*, and *Education* column, all the rows without **Overall**, **All Races**, **All Ages**, and **All Grades** values for each respective column were filtered out to further remove unnecessary data. This successfully removed all of the non **Cigarette Usage** under the *TopicDesc* column. In addition, the decluttering removed all of the year ranges (for example, “2018-2019”) from the dataset. At this point, the cigarette smoking data was ready to be analyzed.

The first thing to do was connect the mortality dataset with the tobacco dataset. The two datasets were joined by both the year values (named **year1** and **YEAR**) and state values (named **State** and **LocationDesc**).

The first utilized visualization for smoking data is the scatter plot utililyzing the **Age-adjusted Death Rate** on the x axis and **Data Value** on the y axis. The “Data Value” after the data cleaning represents exclusively the percentage of people who smoke in the state, I renamed the y value to **Smoking %** Smoking Process

The dataset used to examine cigarette smoking rates is the *Behavioral Risk Factor Data: Tobacco Use (2011 to present)* dataset from *data.gov*. The only necessary information in

the dataset under the *MeasureDesc* column is the **SmokingStatus** value. All the non SmokingStatus values under *MeasureDesc* column are removed to avoid mass filtering in the dataset. For the *Gender*, *Race*, *Age*, and *Education* column, all the rows without **Overall**, **All Races**, **All Ages**, and **All Grades** values for each respective column were filtered out to further remove unnecessary data. This successfully removed all of the non **Cigarette Usage** under the *TopicDesc* column. In addition, the decluttering removed all of the year ranges (for example, "2018-2019") from the dataset. At this point, the cigarette smoking data was ready to be analyzed.

The first thing to do was connect the mortality dataset with the tobacco dataset. The two datasets were joined by both the year values (named **year1** and **YEAR**) and state values (named **State** and **LocationDesc**).

The first utilized visualization for smoking data is the scatter plot utilizing the **Age-adjusted Death Rate** on the x axis and MEDIAN(**Data Value**) on the y axis. The "Data Value" after the data cleaning represents exclusively the percentage of people who smoke in the state. I then added a regression line in order to see the correlation visualized. Red was the color set on all scatterplots as a team rule for more focus attention purposes. Next, I implemented tabpy by calculating the r value (otherwise know as correlation). This visualization serves to determine which causes of death are more correlated with smoking rates.

The next visualization is a line chart which visualizes the smoking rate for all 50 states plus Washington DC. This visualization serves to show how smoking rates have been gradually decreasing across years. The x axis is **Year** while the y axis is Median(Data Value), which is truly the smoking rate. **State** was set as a detail value to separate each state into its own line.

On the dashboard titled *Smoking Dashboard*, I placed the line chart on top and the scatter plot below it. I added text which summarizes the results on the top left of the dashboard. Below the text is a filter labeled "Filter states by smoking rate", which utilizes a parameter named *Filter Small Values*. This parameter ranges from 1 to 51 and as the value increases, more states appear on the line graph. This allows the user to selectively view only the top median smoking rate value states. For the scatterplot, I displayed the tabpy calculated r value to the top left of the visualization. Below is a filter selecter which allows the user to select the age adjusted death rate for each cause of death. Below the cause of death filter is a year slider, allowing the user to select a year to see the correlation of.

Obesity Process

The raw obesity data is from *Behavioral Risk Factor Data: Overweight and Obesity (2011 to present)* dataset from *data.gov*. This data contained many surveys from across state and years to be used in calculating obesity rates. In order to use this data properly, it was processed outside of Tableau in Python to calculate obesity rates by state and year. This cleaned data was then loaded into Tableau and merged on state and year.

The dashboard *Obesity Dashboard (Aidan)* contains five visualizations and two filters. The filters are for year and cause of death. The Cause Name filter is a select-one filter contining all causes of death. The Year filter is a slider containing all years of obesity data. The top graph is from *Obesity Rate vs Death Rate* and is a graph of correlation between the obesity rate and death rate, which is filterable by year and cause. The bottom map is from *Obesity Rate Map* and is a map of obesity rate by state, filtered by year. The other three are text from sheets *Obesity Rate, Death Rate,* and *Correlation r-value*. The Obesity Rate box displays the national obesity rate for the filtered year. The Death Rate box contains the national death rate for the filtered cause within the filtered year. The Correlation box shows the correlation r-value between the obesity rate and the filtered cause of death.

Medicaid Process

Data Wrangling and Cleaning

The Medicaid dataset is made of three different CSV files from Centers for Medicare and Medicaid Services. These datasets are Medicaid Spending per Enrollee, Total Medicaid Spending, and Total Medicaid Enrollment. These together have total Medicaid enrollment, total Medicare spending, and Medicare spending per enrollee data from 1999 to 2017. This data also has groups for the type of Medicare under the label ‘Code’, the state name, and region name.

To connect these three datasets to the source Leading Cause of Deaths dataset, the datasets were reshaped in Python to a long format. A long format makes ‘Year’ one column, allowing connection to the source dataset by state name and year. After linking datasets through Tableau, there were a few null values ignored because that data was missing state name since it was aggregated by the whole United States.

Graphs

"Total Deaths by Stroke" is a line graph between total deaths in America from 1999 to 2017. This was created by placing 'total deaths (sum) filtered by stroke' on the y – axis and 'year' on the x – axis. The purpose of this graph is to show the change in stroke deaths over time.

"Death Rate vs Medicaid Spending for Short-Term and Long-Term Care" is a scatter plot of 'Avg. Age-adjusted Death Rate' on y-axis and 'Avg. Spending per Enrollee' on x – axis. This was then filtered by 'state name' and then grouped by a calculated field called 'group codes' that groups spending by the type of spending. The purpose of this graph is to show that Short-Term care is more effective at decreasing deaths.

"Death Rate vs Medicaid Spending for Stroke, Cancer, and Heart Disease" is a scatter plot between 'Avg. Spending per Enrollee' and 'Avg. Age-adjusted Death Rate' filtered by stroke, cancer, and heart disease. The purpose of this graph is to show that chronic illnesses like cancer and heart disease are less influenced by Medicaid spending than acute illnesses. This because chronic illnesses can be influenced by many other factors such as environment, genetics, and behaviors.

Medicare Process

Data Prepping

The data which Medicare was used from was also combined with 3 different CSV files from the Medicare Services. After having found the datasets, we had to aggregate the 3 files specifically so we could see the correct data instead of a numerous number of random fields. After having found out what Tableau's easy aggregation method was, first put the 3 into there and decided to find relationships between the three.

The 3 datasets that were linked to our primary dataset (NCHS Leading Cause of Death) were all shaped out of the Medicare data source, including Medicare enrollment for enrollees, total Medicare spending, along with Medicare spending per enrollee. All of this data goes back from 1999 to 2017. Along with the Medicare data, a few variables from the primary source stood out, such as region name, state name, mortality rate, and year. Noticed that different null values were apparent and decided to remove them to make the graphs look cleaner without unreliable data points.

After spending some time looking at the data, I found out that we could use the 'year' column and the 'state name' column to cross reference all the data. This correctly aggregated the 3 into 1 master dataset, which made it very easy to graph visuals in Tableau. Making it easy to compare the year, state name, region name, and code of region from the primary dataset to the other 3 connecting data.

Visualizations

Once having the dataset ready to analyze, my first step was to zoom in on the data and decided to look at some sort of growth rate throughout the years, to see if there's a trend in the enrollees of Medicare over time. First had to create a calculated field to adjust for the growth rate using the data given. Growth rate over time, basically I wanted to see a trend line about the amount of spending per enrollee by # of enrollees every year.

After having created my calculated field, I wanted to see the trend over time for Medicare Enrollees. The results were what I expected, no necessary pattern between the years and growth rate. Many spikes and drops are printed due to policy changes or environmental changes. What stood out was the data point set in 2006, it was incredibly high with a rate of .14, meaning this was the most money spent per enrollee using Medicare ever. After having done research, I noticed why there were several drops and spikes in the line graph, due to the Medicare Plan D: introduction of prescription drug coverage. Clearly after having such a year in 2006, 2007 leveled off even though spending was still expensive believe it or not. But overall, the trend I saw from 1999-2017 spikes up, then decreases, spikes up in 2006, then drops down all the way to 2012, and has slowly been on the rise ever since. My reasoning for the dramatic drop was due to the year-by-year comparison of growth. Once you have such a killer year like 2006, it's going to be hard to top that the very next year.

After having seen the growth rate over time, I wanted to next investigate which leading causes held most of the deaths, specifically over regions of the US. One of our datasets had provided different regions throughout the U.S. to split up; Far West, Rocky Mountains, Southwest, the Plains, Great Lakes, Mideast, Southeast and then New England. I set up the graph to be a stacked bar chart to cross reference cause of death by color in each region. After having finally corrected the graph, had many takeaways I thought were interesting. One being that I noticed more along the lines of the West Coast was that they had more fatalities and deaths from heart disease, along with the East Coast having more of a cancerous environment. Some explanations I have from this analysis are that due to environmental changes or access to healthcare, these might determine why more heart diseases occur in the West and cancer is caused in the East. No accurate determining factor has been stood out yet.

After having made my first 2 visualizations, I wanted to backtrack a moment and look at another point of view. I decided to make a map over the U.S. and wanted to track how much spending does each state spend per enrollee, if there was a geographical difference. After having made my map with the average spending per enrollee designated to color and state name on the tooltip. I could see every state and how much they spend per user, all ranging from about $1,000 – $2,000. The results I got back weren’t as surprising, more

populated areas such as California, Texas, Florida, and New York all had higher spendings than any other state. My first thought was average spending per enrollee has a very strong correlation with highly populated areas. My vision for the map was to try and see if, did spending per enrollee help with causes of death, for both West coast and East coast, or if it was just a coincidence.

After finding out the top 3 causes of death in the US, I wanted to solely look at those 3 alone; heart disease, cancer and stroke. Just to get a clearer look at the leading causes, I felt it was important to add this visual because it helps show relationship between cause of death and location. As we can see heart disease beats everybody else on the West Coast, but more cancer spikes towards the East coast. Then of course you have stroke, which was neutral throughout the U.S. One thing that caught my attention, was that why is it that heart disease and cancer both have huge effects on life in both East and West coast, when looking at my map, those are the same areas that have the most spending per enrollee. I thought that was sort of messed up. The people who need Medicare the most are getting the money for it; they just have that little correlation between cause of death. Which makes more sense, because the older everyone gets, the less likely we are going to be healthy, more vulnerable to sickness and diseases, etc.

Key Conclusion and Takeaways

After having settled down with my visuals and created my dashboard, I have come to a few conclusions. The reasoning for more heart diseases in the West and cancer in the East could be due to environmental changes, the East has more polluted areas dealing with vehicles, along with the West might have more populated obesity? Either or, none of these have a high correlation to Medicare.

After relooking over every takeaway I drew from this topic, Medicare has little to no correlation to cause of death in the U.S. Although with little correlation to death, Medicare does have a high correlation to populated areas. All this said, doesn't mean Medicare is useless for elderly people, it still tremendously helps with health. I personally believe it has something to do with preventive and unprevented diseases. Such as cancer and the flu, you can't prevent cancer like you could with the yearly flu. Same for Medicare, it doesn't help as much with heart disease or strokes rather than it does with less severe causes of death.

Appendix

Name of Field	Code	Where it is used	Purpose of the field
R Value	SCRIPT_REAL("import numpy as np return np.corrcoef(_arg1,_arg2)[0,1]",MEDIAN([Age-adjusted Death Rate]),MEDIAN([Data Value]))	Age Adjusted Death Rate vs Smoking Rates	To calculate the r value (correlation)

Name of Field	Parameter	Where it is used	Purpose of the field
Filter Small Values	Values 1 to 51	Smoking Rates 2011-2017	To keep the top median states by smoking percentage exclusively (controllable with a slider still)

Name of Field	Formula	Where it is used	Purpose of the field
Valid Obesity / Death Rate Data	NOT ISNULL([Obesity Rate]) AND NOT ISNULL([Age-adjusted Death Rate])	Correlation r-value	To be able to calculate correlation between obesity rate and death rate, since obesity rate has less data which would show up as null

Name of Field	Formula	Where it is used	Purpose of the field
Years of Obesity Data	IF [Year] IN (2011,2012,2013,2014) THEN STR([Year]) ELSE NULL END	Obesity rate map	To only display years with obesity data on a filter slider instead of all years

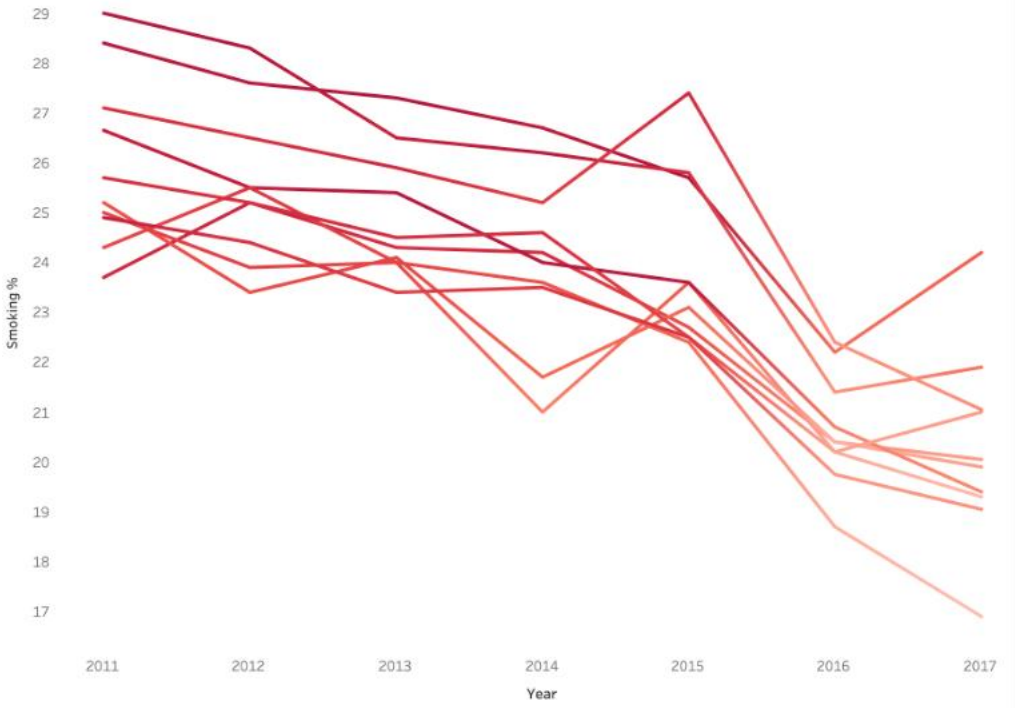
Name of Field	Formula	Where it is used	Purpose of the field

Growth Rate	(SUM([Medicare spending PER enrollee]) - LOOKUP(SUM([Medicare spending PER enrollee]), -1)) / LOOKUP(SUM([Medicare spending PER enrollee]), -1)	Growth Rate of Medicare Enrollees - line chart	To find a trend of spending per enrollee by # of enrollees / year
Name of Field	**Formula**	**Where it is used**	**Purpose of the field**
Calculated correlation	SCRIPT_REAL(' from scipy import stats result = stats.linregress(_arg1,_arg2) return result.rvalue ', AVG([Obesity Rate]),AVG([Age-adjusted Death Rate]))	Correlation r-value	To calculate the correlation r-value between obesity rate and death rate

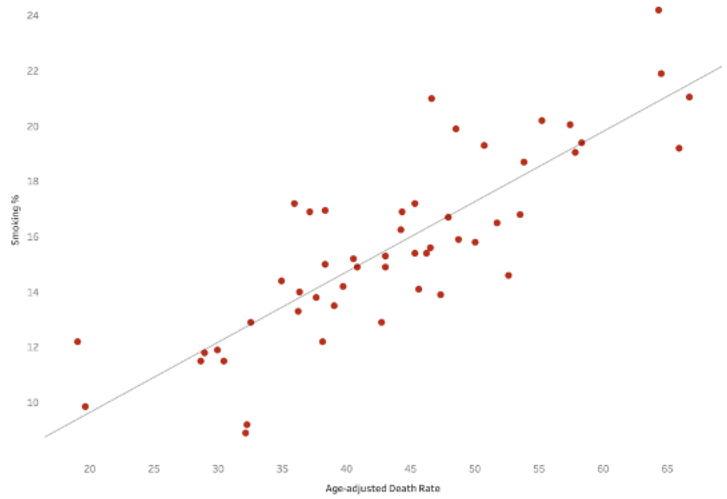
Name of Field	**Formula**	**Where it is used**	**Purpose of the field**
Put something here	SUM([Profit])/SUM([Sales])	Sales by County	To understand how much of a profit each of the items sold is turning relative to the sales of the product.

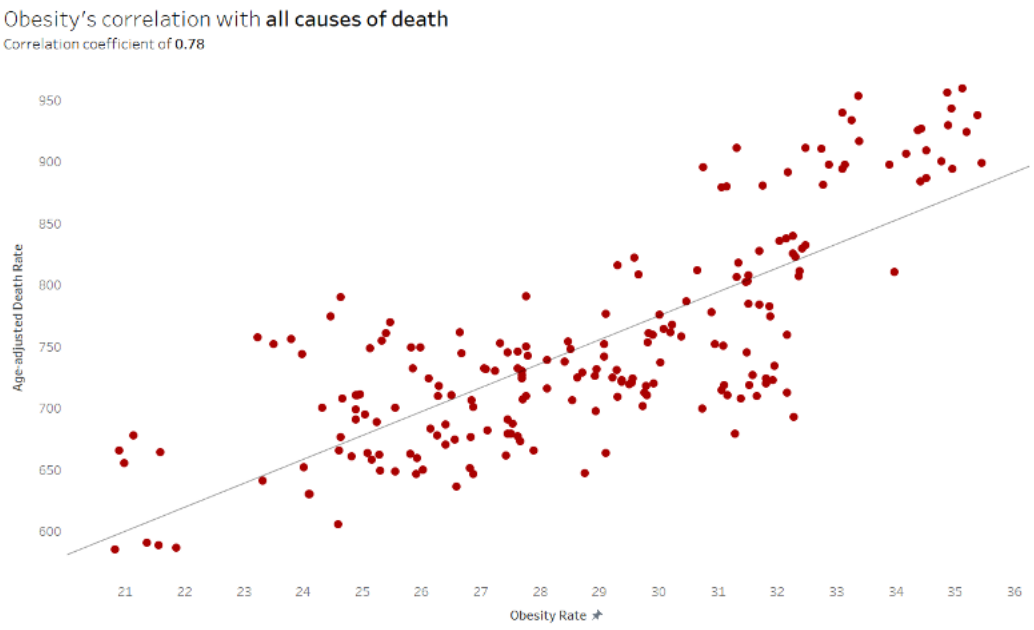
Name of Field	**Formula**	**Where it is used**	**Purpose of the field**
group codes	IF [code] IN (1, 2, 3) THEN "Short-Term Care" ELSEIF [Code] IN (4, 5, 6, 7) THEN "Long-Term Care" ELSE "Other" END	Code	To group types of spending (‘Code’) by either Long-Term or Short_Term care

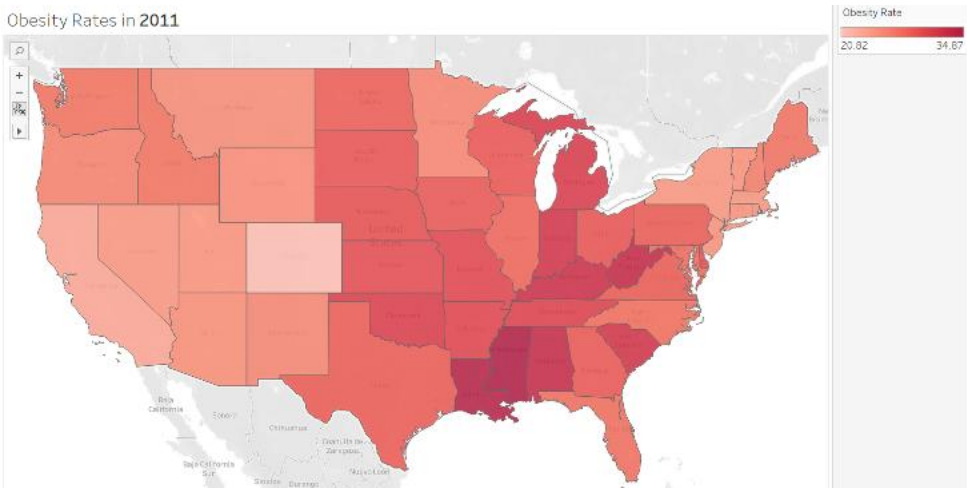
Smoking rates are **decreasing** yet many states still have high smoking rates

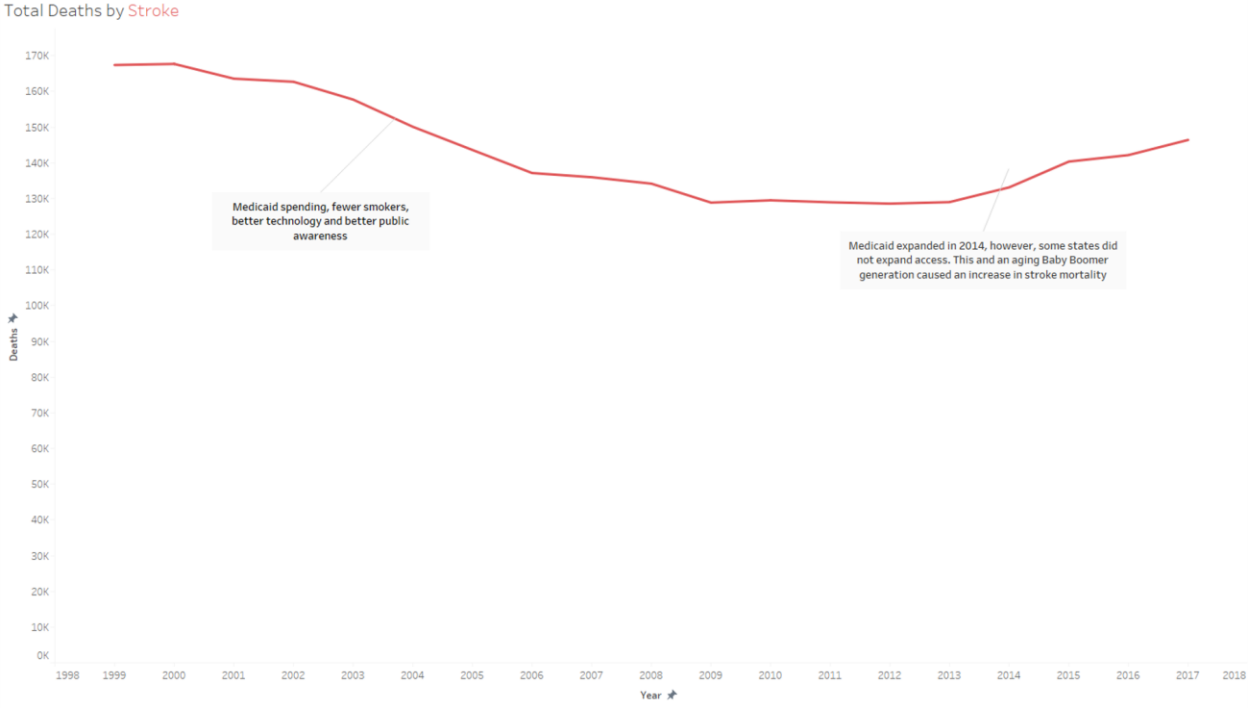


Smoking is correlated the most with **lung disease** (CLRD)

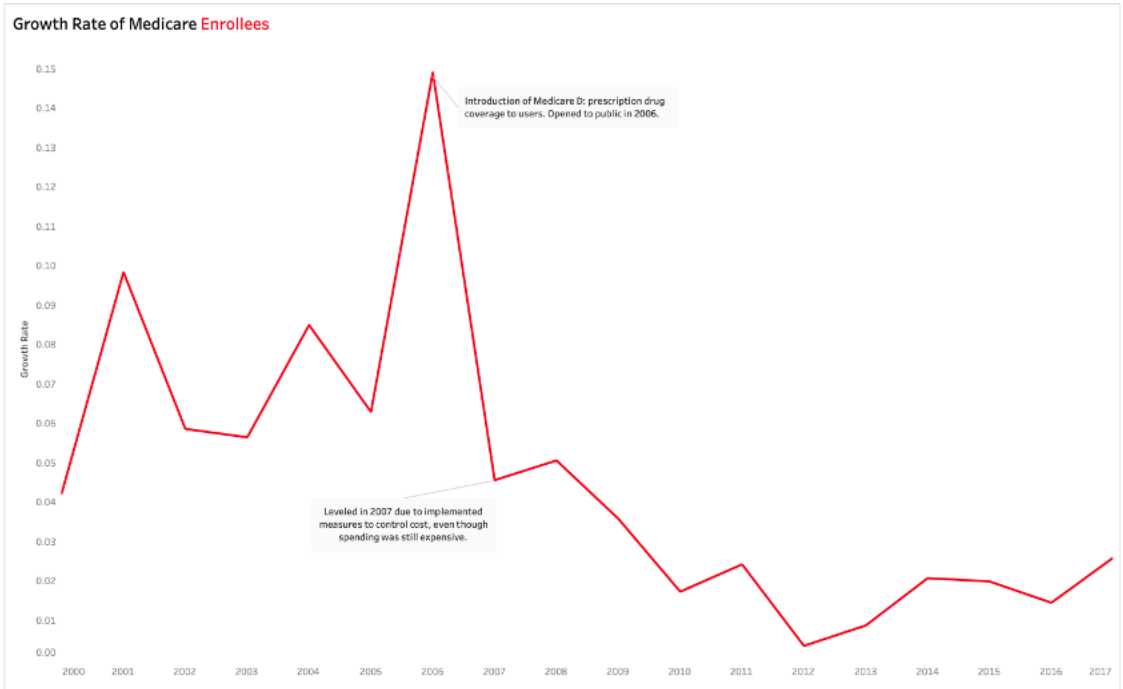



Obesity's correlation with all causes of death
Correlation coefficient of 0.78
Age-adjusted Death Rate
950
900
850
800
750
700
650
600
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
Obesity Rate


Obesity Rates in 2011
Obesity Rate
20.82
34.67

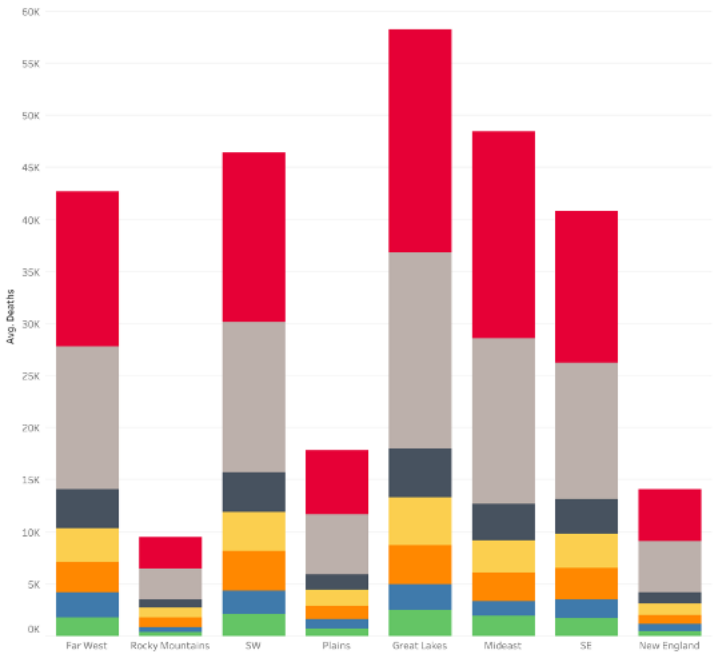

Total Deaths by Stroke
170K
160K
150K
140K
130K
120K
110K
100K
90K
80K
70K
60K
50K
40K
30K
20K
10K
0K
Deaths
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year
Medicaid spending, fewer smokers, better technology and better public awareness
Medicaid expanded in 2014, however, some states did not expand access. This and an aging Baby Boomer generation caused an increase in stroke mortality


Death Rate vs Medicaid Spending for Short-Term and Long-Term Care
Avg. Age-adjusted Death Rate
60
55
50
45
40
35
30
25
20
15
10
5
0
0K
1K
2K
3K
4K
5K
6K
7K
8K
9K
10K
11K
Avg. Spending per Enrollee

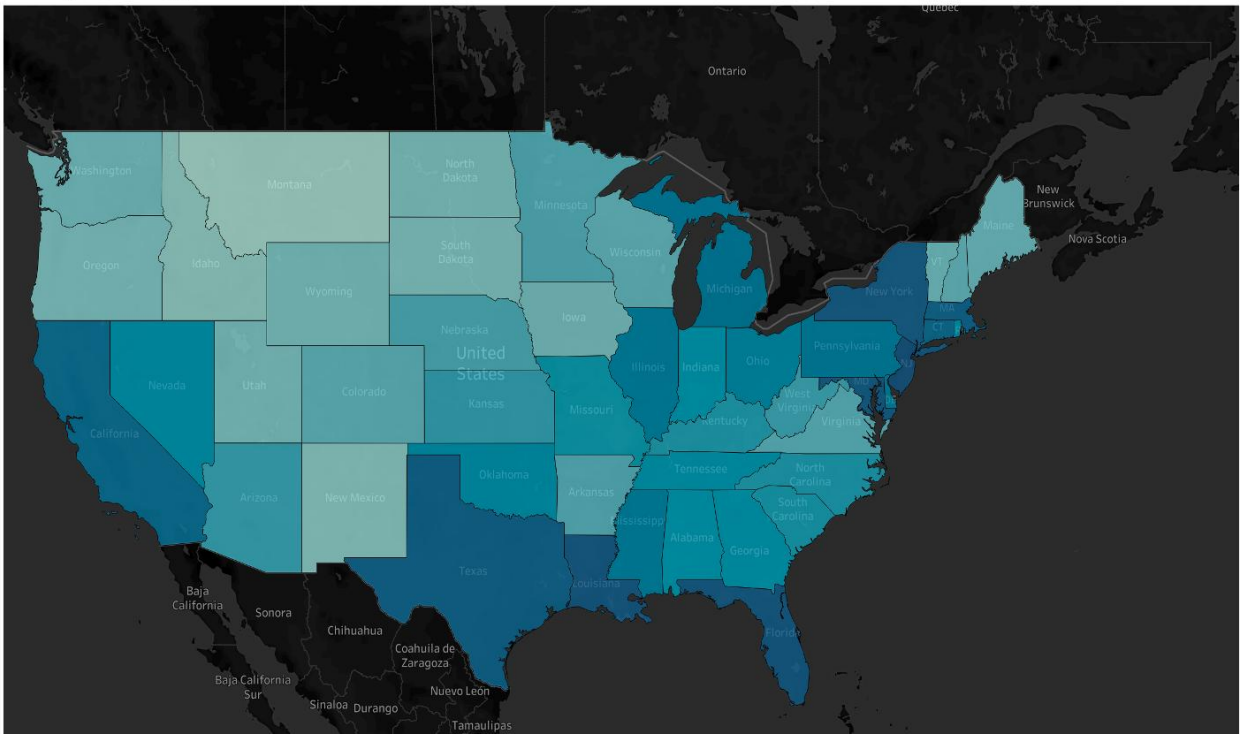

Growth Rate of Medicare Enrollees
Growth Rate
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Introduction of Medicare D: prescription drug coverage to users. Opened to public in 2006.
Leveled in 2007 due to implemented measures to control cost, even though spending was still expensive.

Death Rate vs Medicaid Spending for Stroke, Cancer, and Heart Disease

Average Cause of Death by Region



Average Spending per Person by State



Stroke, Cancer, **and Heart Disease** **in the US**