the Outliers

October 2024: Team 3 Project 1 Submission and EDA

TEAM NAME: The Outliers

LINK TO GIT REPO: https://github.com/KathMac58/Team-3-The-Outliers

TEAM MEMBERS & ROLES:

NAME	EMAIL	GIT ID	ROLE
Fran Ruiz	franciscoruiz2025@u.northwestern.edu	fuijo	Git Collaborator Team Collaborator Presentation Development Answering Question # 3 + 5
Sam Sims	Sam.sims.13@gmail.com	SamSims	Git Collaborator Lead Developer Data Merging Answering Question # 1
Nurmaa Dashzeveg	nurmaa.dashzeveg@northwestern.edu	nkd2882	Git Collaborator Data Analyst Answering Question # 4
~~Valeria Figueroa~~	~~vfigueroa2828@gmail.com~~	~~vfig2828~~	~~Git Collaborator~~ ~~Data Management: Cleansing & Merging~~ ~~Answering Question # 5~~
Kathryn McAtee	Kathryn.mcatee@gmail.com	KathMac58	Git Collaborator – created Git Repo Project Manager/ Team Lead Final Report Aggregator Data Analyst Answering Question # 2

PROJECT TITLE: A glance into US Healthcare Expenditures & Health Outcomes

PROJECT BRIEF: US healthcare spending continues to increase year after year, and projections show no signs of slowing down. Is there a positive correlation between personal healthcare investment and mortality outcomes? Does this correlation change based on factors such as location, income level, receipt of government assistance, or the amount of money being allocated? Is the investment in healthcare worth it?

QUESTION TO BE ANSWERED; PROBLEM TO BE SOLVED: With the increasing investment individuals and families need to make for healthcare each year, regardless of whether they are on government or private insurance, there is a growing interest in understanding if the amount personally paid contributes to an individual's overall well-being.

The Outliers aims to provide peace of mind to consumers who have the same question in mind. For our initial project, we will specifically analyze CMS data/spending in comparison to Census and

the Outliers

Mortality data to determine if there is a correlation between healthcare spending, location, income, and mortality. However, we acknowledge that there are numerous other factors to consider to truly comprehend this complex issue.

TEAM ANALYSIS:

1. **Sam**: Demonstrate the healthcare expenditure over the year (2010-2017)
2. **Kathryn**: A look into government healthcare program personal spending.
3. **Fran**: The relation between total population and enrollment in Medicare between 2010 to 2017 in the US.
4. **Nurmaa**: The relation of high-income earners population and healthcare spend.
5. **Sam**: The statistics of mortality rates and healthcare spend.
6. So… Is the investment worth it?

HYPTHESIS: The Outliers believe that the average American is paying more for healthcare than they were in previous years, but the value of the investment is declining year over year. We believe the rate of increase does NOT correlate to mortality outcomes.

GENERAL ISSUES & CHALLENGES:

1. We had lots of great ideas at the beginning, but quickly realized the importance of finding a great idea that had an equally as great data set – it proved to be difficult, especially in the healthcare space.
2. We lost a team member half-way through, when we were already crunched from time and already recognized we 'bit off more than we could chew.'
3. Data
    a. Mapping data and fields across data sets from multiple sources proved to be difficult. There were a few data integrity deep dives after viewing our graphs, simply to make sure we were reporting accurate numbers and totals.
    b. Availability of complete data sets. We knew we were pressed for time at the get-go, so we quickly aligned on an idea and data set(s) so we could get moving. The problem was, the more we dug into the data, the more we had to adjust or pivot to make the data line up and the end report/graphs make sense.
    c. Merging challenges due to gaps in data, or fields do not like for:
        i. CMS data 10-year gap in PHI (Private Healthcare) data
        ii. CMS data gap in itemized costs by code (PHI only had 1 code)
        iii. County Health Rankings & Roadmaps Data had gaps in mortality data after cleansing which shifted our target analysis period from 10 years to 7 years.
    d. After data merges and joins, for Medicare we noticed there were more people enrolled year over year than the population of 65 and older in each state – either our population data was wrong or there is quite a lot of fraud happening in a few states!
    e. Made a few assumptions early on and misjudged how the numbers were represented where there was a data dictionary gap.
    f. Should have used Census library via Python instead of manipulating csv files in PowerBI

EDA

Datasets being used:

the Outliers

See .xls for full details: Data_Inventory.xlsx

1. Centers for Medicare and Medicaid Services Data: 12 data files
2. County Health Rankings & Roadmaps Data: 10 data files (combined to 1)
3. Census Data: 2 data files

Included in the .xls you will find the following for each file used:

- File Name
- Contents/ Brief Description
- Source of data
- Total # of Columns
- # of Columns Required
- Total # of Rows
- Total # Null
- Total # Duplicates
- Notes
- Data Required and files they are sourced from:
    - Item
    - Code
    - Region Name
    - State Name
    - Year (2010-2020)
    - Spend
    - Deaths
    - Population

the Outliers

Team 3: The Outliers
Datasets for Project 1

File Name	Contents	Source	Total # Col	# Col Req	# Rows	# Null	# Dups	NOTES	Data Required and files they're sourced from							
									Item	Code	Region_ Name	State_ Name	Year (2010-2020)	Spend	Deaths	Pop
US_AGGREGATE20.CSV	Total personal health care spending by state and by service, 1991-2020	SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	14	600	6	0	Each state has 10 different rows, numbered with a Code 1 through 10, representing each of the code keys annotated in the data dict tab ;the values in each year are different for each code - they are not duplicates and need to be included	X	X	X	X	X	X		
US_POPULATION20.CSV	US Population by State, 1991-2020	NOTE: Population obtained from the U.S. Bureau of the Census, released December, 2020. SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	14	60	6	0	Code 11 only: Population_Enrollment	X	X	X	X	X			X
US_PER_CAPITA20.CSV	Per capita personal health care spending by state and by service, in dollars, 1991-2020 (aggregate spending divided by population)	NOTE: Population obtained from the U.S. Bureau of the Census, released December, 2020. SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	14	600	6	0	Each state has 10 different rows, numbered with a code 1 through 10, representing each of the code keys annotated in the data dict tab ;the values in each year are different for each code - they are not duplicates and need to be included 6 rows that are null in 1 column because there is no associated state	X	X	X	X	X			
MEDICAID_AGGREGATE20.CSV	Total Medicaid personal health care spending by state and by service, 1991-2020	SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	15	600	6	0	Each state has 10 different rows, numbered with a code 1 through 10, representing each of the code keys annotated in the data dict tab ;the values in each year are different for each code - they are not duplicates and need to be included 6 rows that are null in 1 column because there is no associated state	X	X	X	X	X	X		
MEDICAID_ENROLLMENT20.CSV	Medicaid enrollment by state, 1991-2020	NOTE: Enrollees measured in fiscal person years, adjusted to a calendar year basis. SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	15	60	6	0	Code 11 only: Population_Enrollment 6 rows that are null in 1 column because there is no associated state	X	X	X	X	X			X

the Outliers

File Name	Contents	Source	Total # Col	# Col Req	# Rows	# Null	# Dups	NOTES	Data Required and files they're sourced from							
									Item	Code	Region_ Name	State_ Name	Year (2010-2020)	Spend	Deaths	Pop
MEDICAID_PER_ENROLLEE20.CSV	Per enrollee Medicaid personal health care spending by state and by service, 1991-2020 (aggregate spending divided by enrollment)	SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	15	600	6	0	Each state has 10 different rows, numbered with a code 1 through 10, representing each of the code keys annotated in the data dict tab ;the values in each year are different for each code - they are not duplicates and need to be included 6 rows that are null in 1 column because there is no associated state	X	X	X	X	X			
PHI_AGGREGATE20.CSV	Total private health insurance personal health care spending by state and by service, 2001-2020	SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	27	15	60	6	0	Code 1 only: Personal Health Care Y2001-2020 only; does not start from 1991 like the other files, which is why the total Col # is 10 lower than the others. 6 rows that are null in 1 column because there is no associated state	X	X	X	X	X	X		
PHI_ENROLLMENT20.CSV	Private Health Insurance Population by State, 2001-2020	SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	27	15	60	6	0	Code 11 only: Population_Enrollment Y2001-2020 only; does not start from 1991 like the other files, which is why the total Col # is 10 lower than the others.	X	X	X	X	X			X
PHI_PER_ENROLLEE.CSV	Per enrollee private health insurance personal health care spending by state and by service, in dollars, 2001-2020 (aggregate spending divided by enrollment)	SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	27	15	60	6	0	Code 1 only: Personal Health Care Y2001-2020 only; does not start from 1991 like the other files, which is why the total Col # is 10 lower than the others.	X	X	X	X	X			
MEDICARE_AGGREGATE20.CSV	Total Medicare personal health care spending by state and by service, 1991-2020	SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	15	600	6	0	Each state has 10 different rows, numbered with a code 1 through 10, representing each of the code keys annotated in the data dict tab ;the values in each year are different for each code - they are not duplicates and need to be included	X	X	X	X	X	X		

the Outliers

File Name	Contents	Source	Total # Col	# Col Req	# Rows	# Null	# Dups	NOTES	Data Required and files they're sourced from							
									Item	Code	Region_ Name	State_ Name	Year (2010-2020)	Spend	Deaths	Pop
MEDICARE_ENROLLMENT20.CSV	Medicare enrollment by state, 1991-2020	NOTE: Enrollees reflect the number of persons enrolled in the Hospital and/or Supplementary Medical Insurance programs as of July 1, 1991-2020. SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	15	60	6	0	Code 11 only: Population_Enrollment	X	X	X	X	X			X
MEDICARE_PER_ENROLLEE20.CSV	Per enrollee Medicare personal health care spending by state and by service, 1991-2020 (aggregate spending divided by enrollment)	SOURCE: Centers for Medicare & Medicaid Services, Office of the Actuary, National Health Statistics Group.	37	15	600	6	0	Each state has 10 different rows, numbered with a code 1 through 10, representing each of the code keys annotated in the data dict tab ;the values in each year are different for each code - they are not duplicates and need to be included	X	X	X	X	X			
2010 - 2017 Death Rate	2010-2020 National Mortality Data by State	SOURCE: County Health	3	3	408	0	0	Data set was to include 2010 - 2020, but due to years 2018 & 2019 missing # of deaths we decided to consolidate it down to 2010 - 2017. There was a total of 8 files that were pulled. Each sheet had between 135 to 150 rows and 3141 rows of data. The sheet were put together to create one sheet with three columns and 408 rows.				X	X		X	
Income	2010-2020 US Income by State	SOURCE: Census	58	5	420	0	0	There was a total of 8 files that were pulled. Each sheet had between 135 to 150 rows and 3141 rows of data. The sheet were put together to create one sheet with t44 columns and 420 rows.				X	X			
age & sex	2010-2020 US Age and Sex by State	SOURCE: Census	58	5	420	0	0	There was a total of 8 files that were pulled. Each sheet had between 455 columns and 52 rows of data. The sheet were put together to create one sheet with 44 columns and 420 rows.				X	X			X

the Outliers

Data Dictionary

Description
Numerical code assigned to each Item
Identifies health spending level of aggregation/payer/service/good, enrollment, or population; appropriate units/scale (e.g. millions of dollars)
Level of aggregation by geography
Numerical Code Assigned to each Region, for sorting purposes
Bureau of Economic Analysis Region Name
U.S. State Name
Spending, population, or enrollment level for 1991
Spending, population, or enrollment level for 1992
Spending, population, or enrollment level for 1993
Spending, population, or enrollment level for 1994
Spending, population, or enrollment level for 1995
Spending, population, or enrollment level for 1996
Spending, population, or enrollment level for 1997
Spending, population, or enrollment level for 1998
Spending, population, or enrollment level for 1999
Spending, population, or enrollment level for 2000
Spending, population, or enrollment level for 2001
Spending, population, or enrollment level for 2002
Spending, population, or enrollment level for 2003
Spending, population, or enrollment level for 2004
Spending, population, or enrollment level for 2005
Spending, population, or enrollment level for 2006
Spending, population, or enrollment level for 2007
Spending, population, or enrollment level for 2008
Spending, population, or enrollment level for 2009
Spending, population, or enrollment level for 2010
Spending, population, or enrollment level for 2011
Spending, population, or enrollment level for 2012
Spending, population, or enrollment level for 2013
Spending, population, or enrollment level for 2014
Spending, population, or enrollment level for 2015
Spending, population, or enrollment level for 2016
Spending, population, or enrollment level for 2017
Spending, population, or enrollment level for 2018
Spending, population, or enrollment level for 2019
Spending, population, or enrollment level for 2020
Average annual growth rate for spending, population, or enrollment, 1991-2020

Code Key for CMS data

CODE	DESCRIPTION
1	Personal Health Care
2	Hospital Care
3	Physician & Clinical Services
4	Other Professional Services
5	Dental Services
6	Home Health Care
7	Prescription Drugs and Other Non-durable Medical Products
8	Durable Medical Products
9	Nursing Home Care
10	Other Health, Residential, and Personal Care
11	Population or Enrollment

Key takeaways from EDA:

- Overall: After organizing and cleansing the data selected, our team was excited to see that these datasets were clean and complete over the course of multiple years. There were a few findings our team made note of that drove changes to our scope, to ensure we are selecting the right scope of analysis to answer our questions and prove our hypothesis above.

- For the CMS Data:
  - The CMS files need to be merged down to 4 files.
  - For the 3 files that have 27 cols instead of 37 cols (Private Healthcare: PHI data), there are 10 years missing – spans from 2001-2020, whereas the others have data starting from 1991; need to select years to analyze appropriately.
  - For the files that have 600 rows instead of 60, it is because each state has an itemized list of healthcare spend or size (10 per state); "code" will be important to include in the dataset, where we initially thought, we would not need to.
  - Each file has 6 cells that are 'Null' in the state column; this is because those rows/totals are rolled up to the region level, so that data is not associated with a particular state.

- For the County Health Rankings & Roadmaps data:
  - The County Health Rankings & Roadmaps files need to be merged down to 1 file.
  - Our team was initially going to analyze the years 2010-2020 across datasets, but after cleansing this dataset, noticed there were two years of mortality data missing (2018-2019). Because of this, we changed our target analysis years from 2010-2020, to 2010-2017 across CMS, CHR, and Census.

LESSONS LEARNED:

Project Learnings

- Data: Without proper data cleaning, the analysis could have led to inaccurate insights, such as misrepresenting population or Medicare enrollment trends due to faulty or missing data.
- Data: We saw the benefit of quickly selecting an idea with a 'good enough' data set, despite a few challenges along the way. Staying organized and having effective communication as a team contributes just as much to the success of the project as the code itself. Understanding the data and understanding the questions to be answered is the most important.
- Data: It all starts with the data. We quickly realized the importance of finding a good, solid data set would make all our lives easier. The quality of the data supersedes the quality of the idea itself.
  - A lot of what we found was aggregated or cut down to the bare minimum, so it was difficult to get substantial evidence to prove our hypothesis.
  - Checking data quality and the resources of data are essential to be considered before starting analysis.
  - Explore data and create a detailed strategy for the analysis. Data is broad enough to create numerous outputs, so focusing on the specific questions will save time and reproducibility.
- Data: It was challenging to find complete datasets and obtain them without having to pay for a subscription. It is important to fully understand the data before using it, as different datasets may be formatted differently. We encountered issues with incomplete and differently formatted data,

which affected our ability to analyze the information effectively. For example, we had to adjust our project timeline due to a 2-year gap in mortality data, and we also faced challenges in merging data from various sources due to differences in the orientation and format of the datasets. In addition, we found it difficult to interpret certain health and wellness data, so we ended up using only the number of deaths per state. These challenges could have been mitigated by using standardized tools and ensuring better communication with the team member responsible for pulling the data.

- Scope: The most important lesson we learned from this project was about managing the scope. Our biggest challenge was not knowing how large the project should be as our first one and coordinating everyone's schedules with the limited time we had to complete it. We kept adding more datasets to gather enough information, then we discussed the available resources and questioned the project deadlines. We should have been more selective in choosing our topic, rather than letting our curiosity and excitement lead us astray without a clear goal and plan. This all boils down to research. We should have taken more time to determine our goal instead of rushing due to time constraints. For our first project, instead of seeking something new and exciting to research, we should have chosen something simpler to get a feel for working as a team and understanding the necessary scope and project timeline. We took on too much of a challenge for our first project; the need to remember less is more.
- Time Management: Make sure to commit and pull more often instead of working days on code before pushing it and to make sure that the code that you are working on is up to date, so you are not trying to solve issues that someone has already solved in their code. There were multiple times when we were working with coding that we realized we were trying to do something that someone else has already done which made us less efficient as a team. This also goes hand in hand with meeting more regularly, which we found to be difficult with everyone having conflicting schedules. We should have tried hard to have daily check-in meetings to see where everyone was and if anyone was struggling with an issue. We also should have had a more concise idea of how we were going to get to our answers and what we needed to do to answer each of our questions.
- Be friends with GOOGLE

Team Learnings

- With such a brief period to deliver, we quickly realized we needed to lean on each other's strongest skills to deliver.

WISH LIST:

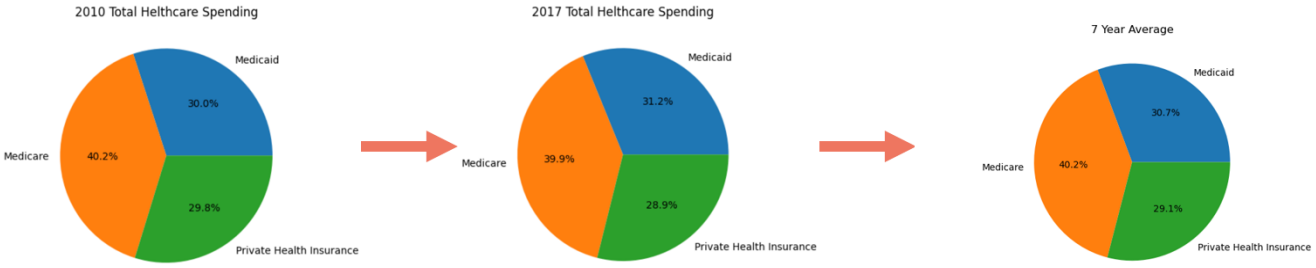
- Look to compare additional data other than just mortality rate, to understand if the investment is worth it.
- Access to quality healthcare data – too much aggregated data sets with gaping holes.
- Ability to overlay Medicare and Medicaid to see, of the total populations enrolled, who is enrolled in both. This would provide a more accurate depiction of total spend per person/population.

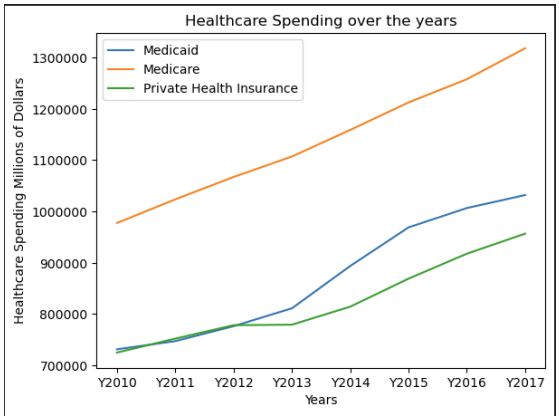
ANALYSIS & CONCLUSIONS:

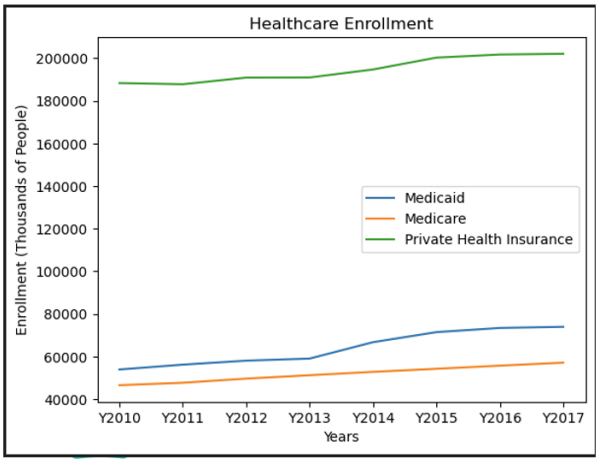
[Numbered Objectives align with the 'Areas to Analyze' included in Presentation]

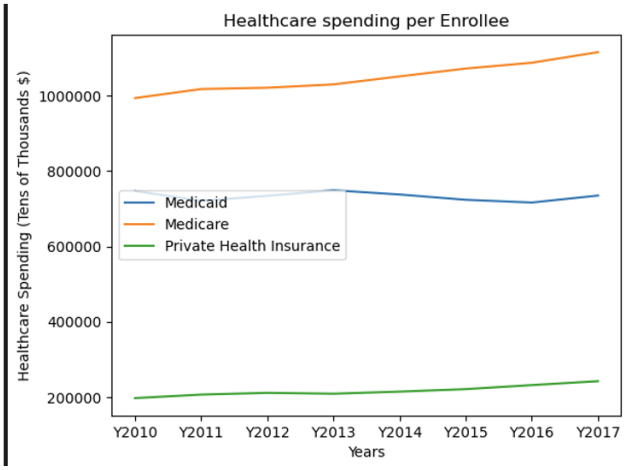
Objective 1: A look into Healthcare Expenditure Spending over time

There is an increase of spending over time, however, the amount of spending per person has not increased as much, meaning the increase is more likely related to population than the price of the insurance. The ratio of the 3 types of spending has also stayed close to the same over the 7-year period. Meaning that each of the different types of spending are increasing at a similar rate.





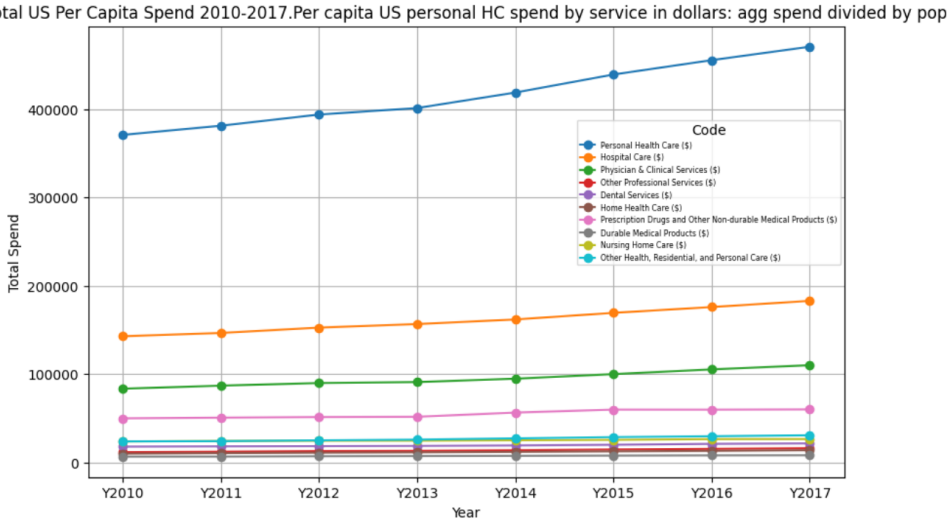


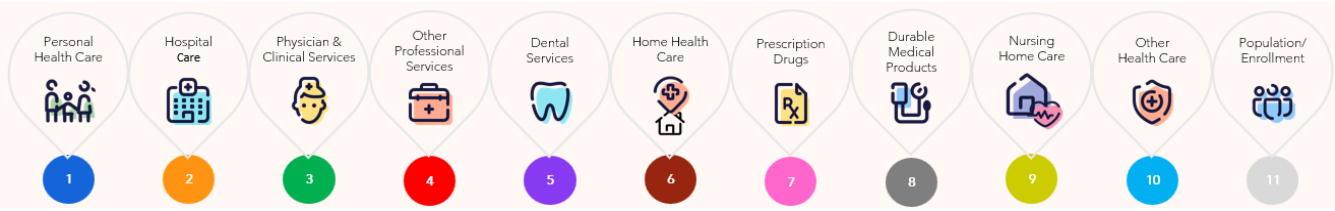


the Outliers

Objective 2: A look into Personal Healthcare SpendingTrends

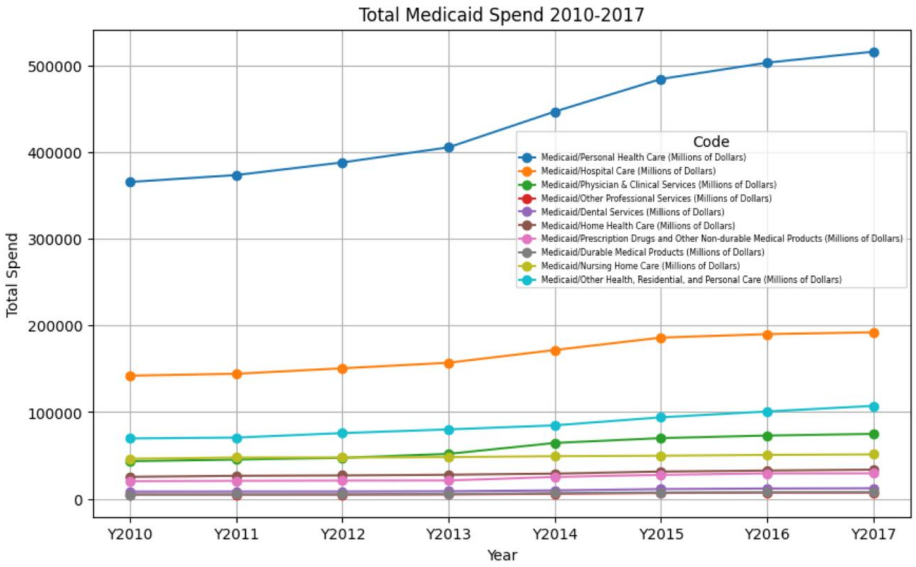
Healthcare is getting more expensive, and personal health spending is increasing year over year – no surprise there. But rather than just looking at total spending, we want to see where the spending was increasing. This segment looked at government program spending to understand if it was increasing year over year on a variety of levels. We originally were also going to include Private Healthcare Spend, but the data only included enrollments, so it was out of scope for this particular analysis.

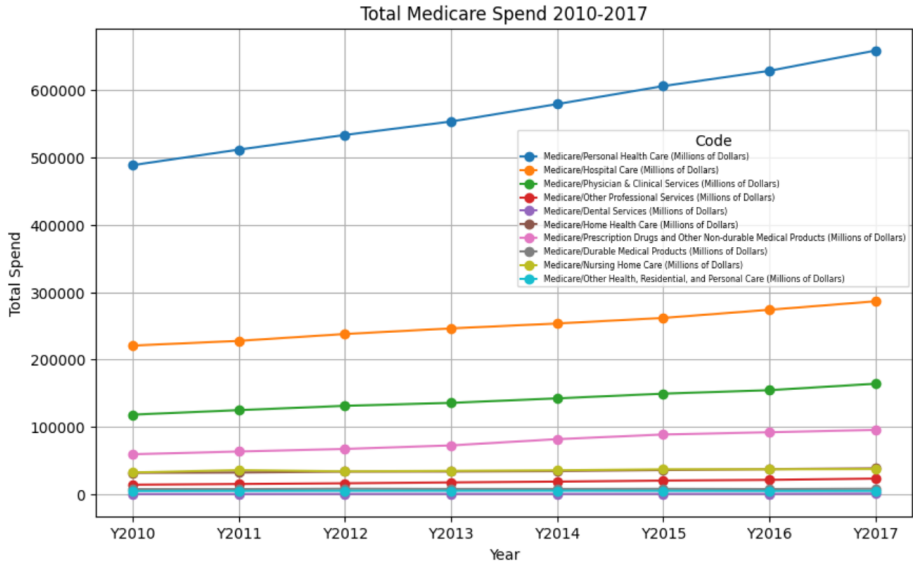




We first looked at the total spend per capita by service and year, as indicated by the graph above. Certain data sets were broken down by code/service (Medicare, Medicaid), so we were able to understand if spending was increasing, where it was increasing, and by how much. The year over year increase is consistent across programs, in some states more than others, with the top 3 areas of spending being Personal Health Care, Hospital Care, and Physician & Clinical Services.

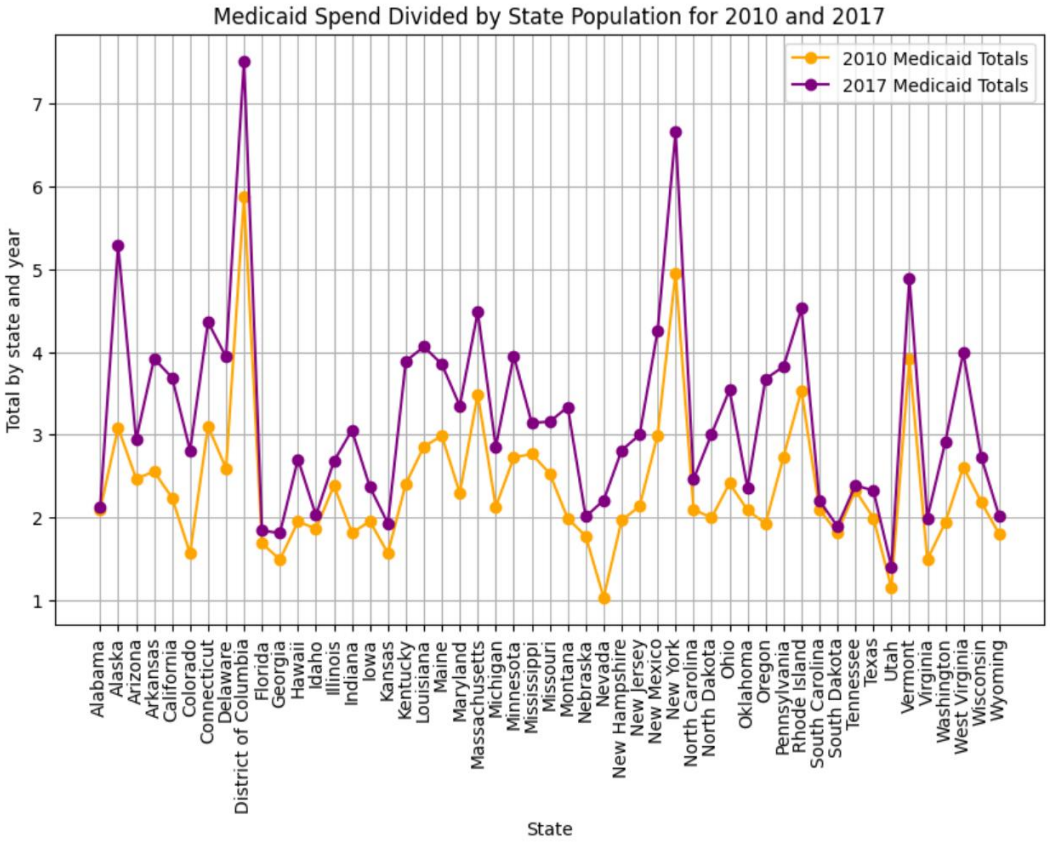
We next looked at the spending by service at the government program level - Medicare and Medicaid, per the graphs below. With clear data showing overall personal health spend was increasing, we wanted to see if we saw the same trend in government program spending, knowing it is heavily subsidized. Not only do you see the same upward trend, you can see sharper increases on Medicaid starting around 2014 (a bit shocking since those are typically the group that require extra financial support) vs. those on Medicare (those 65 & older) – Medicare is also increasing but at a steady pace year over year.

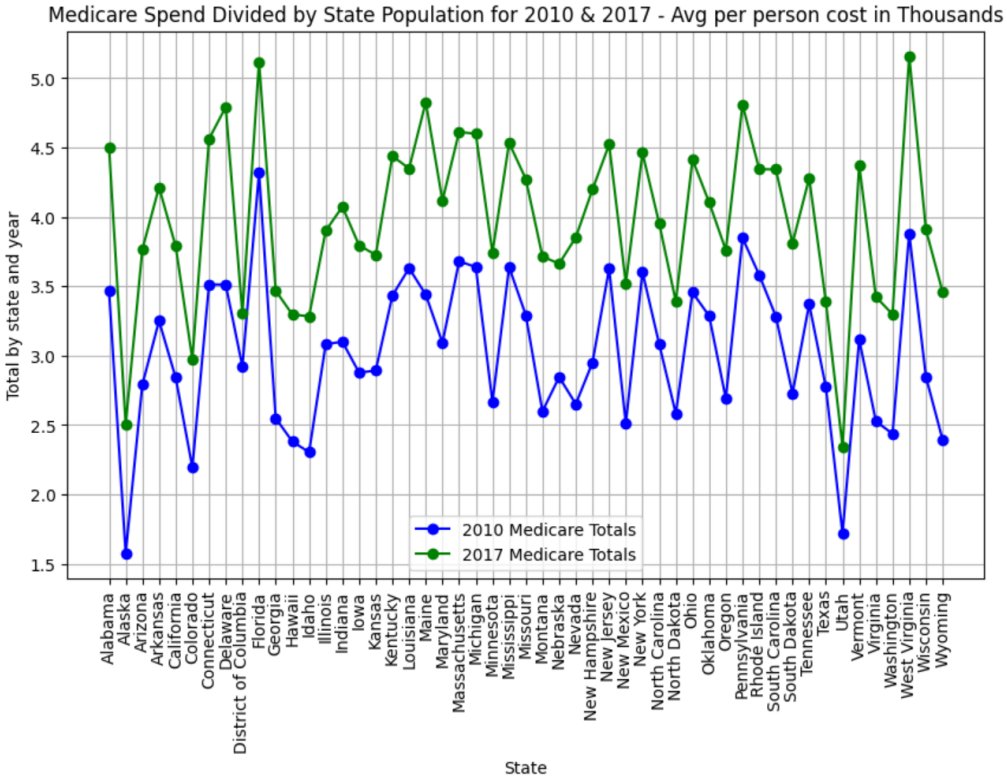




Finally, diving one level deeper, we looked at the spend by state for government programs, you will notice both saw increases from 2010-2017, some states more than others. We overlaid Census population data with CMS spend per state so we could ensure there was an accurate depiction of spend, knowing state populations vary greatly. Key call out is that Medicaid is seeing a higher spend than Medicare, though Medicare is seeing sharper increases year over year.

One thing we wanted to investigate that proved to be challenging due to gaps in data, was to overlay personal spend for Medicare and Medicaid – if you're on Medicare, there's a chance you could be on Medicaid, but if you're on Medicaid, it doesn't necessarily mean you're on Medicare. That perspective would further refine our perspective regarding personal healthcare investment for Medicare and Medicaid, which could help us more precisely answer the question if their investment was working for them.





Objective 3: A look into Enrollment - The relation between total population and enrollment in Medicare between 2010 to 2017 in the US.

the Outliers

The purpose of this report is to analyze the trends in total population and Medicare enrollment across U.S. states from 2010 to 2017. Utilizing data from the U.S. Census Bureau and Medicare enrollment statistics, this report seeks to understand how population changes correlate with Medicare enrollment eligibility and actual enrollment figures.

	Year	State	Total Population	Age Categories 65 Years And Over
0	2010	Alabama	4712651.0	636208.0
1	2010	Alaska	691189.0	50457.0
2	2010	Arizona	6246816.0	830827.0
3	2010	Arkansas	2872684.0	405048.0
4	2010	California	36637290.0	4066739.0
...	...	...	...	...

Table 1: Data set 1 Census data

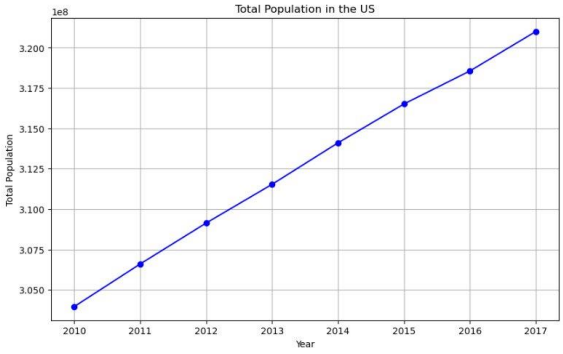
	State	Medicare 2010	Medicare 2011	Medicare 2012	Medicare 2013	Medicare 2014	Medicare 2015	Medicare 2016	Medicare 2017	Code	Item
0	Alabama	845	864	896	922	947	968	988	1006	11	Medicare Enrollee Population (Thousands)
1	Alaska	66	69	73	76	80	84	88	92	11	Medicare Enrollee Population (Thousands)
2	Arizona	930	962	1009	1051	1096	1140	1180	1224	11	Medicare Enrollee Population (Thousands)
3	Arkansas	531	542	559	572	584	594	607	617	11	Medicare Enrollee Population (Thousands)
4	California	4757	4901	5111	5294	5476	5645	5814	5966	11	Medicare Enrollee Population (Thousands)
5	Colorado	625	650	688	722	755	786	816	840	11	Medicare Enrollee Population (Thousands)
6	Connecticut	568	577	595	608	619	630	642	654	11	Medicare Enrollee Population (Thousands)
7	Delaware	149	154	161	168	175	181	187	193	11	Medicare Enrollee Population (Thousands)
8	District of Columbia	78	80	82	84	87	88	90	91	11	Medicare Enrollee Population (Thousands)
9	Florida	3375	3472	3621	3757	3900	4040	4164	4289	11	Medicare Enrollee Population (Thousands)

Table 2: Data set 2 Medicare data

Some of the main finding from these data sets are:

1. Total population in the US from 2010 to 2017: The purpose of this analysis is to investigate the total population trends in the United States over several years. Using census data, the population totals were grouped by year to identify changes and patterns in population growth. The results were visualized using a line chart, offering a clear representation of how the population has evolved over time. The dataset used for this analysis contains population data from the U.S. Census, covering multiple years. The steps taken for the analysis are as follows:

- Data Grouping: The data was grouped by year, and the total population for each year was calculated using the groupby() function.
- Visualization: A line plot was created to display the total population for each year, with markers representing each year on the line.
- Findings: The plot clearly shows a steady increase in the U.S. population over the years. Each point on the graph represents the total population for a specific year. The population growth is gradual, with no sharp spikes or declines over the given time, indicating consistent population growth. The total population in the United States has been steadily growing over the analyzed period. This is consistent with historical population trends, driven by factors such as natural population growth (births exceeding deaths) and immigration. Understanding population trends is crucial for various policy decisions, including resource allocation, infrastructure development, and social services planning.



Top Ten States by Medicare Enrollment in 2010

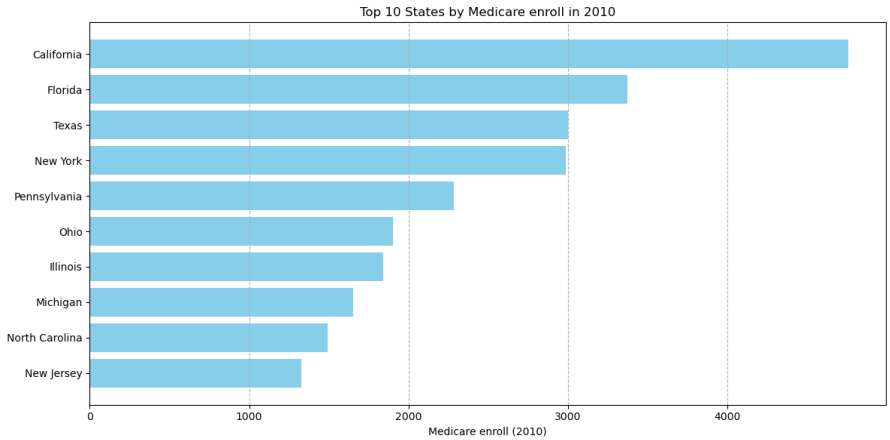
This report analyzes the top ten states with the highest Medicare enrollment in the United States for the year 2010. Medicare is a federal health insurance program that primarily serves individuals aged 65 and older, as well as certain younger people with disabilities. Understanding the distribution of Medicare enrollees by state provides insight into healthcare demands and resource allocation at both state and federal levels. The analysis was performed using the new_merged_data dataset, which contains both population and Medicare data across different years. The following steps were taken:

- Data Filtering: The dataset was filtered to include only records for the year 2010.
- Top 10 States Selection: The nlargest() function was used to identify the top 10 states with the highest Medicare enrollment in 2010.
- Visualization: A horizontal bar chart was created to visually represent the top 10 states by Medicare enrollment for that year.

	State	Medicare 2010
4	California	4757
9	Florida	3375
43	Texas	3001
32	New York	2988
38	Pennsylvania	2283
35	Ohio	1901
13	Illinois	1839
22	Michigan	1651
33	North Carolina	1490
30	New Jersey	1327

Table 3: Top ten states by Medicare enrollment in 2010

- Key observations: The state with the highest Medicare enrollment in 2010 was California, with an enrollment figure of 4,747,000 people. The states at the top of the list, including Florida and Texas, have notably larger populations, which contributes to their higher Medicare enrollments. The chart shows a significant drop between the top few states and those ranked lower in the top 10, reflecting the varying levels of population across different states. The states with the highest Medicare enrollments tend to be those with larger populations and higher proportions of elderly residents. For example, states like Florida, Texas, and California often appear at the top of such lists due to their large elderly populations. These states may face higher healthcare demands, which could impact the allocation of Medicare funding and healthcare resources.



The last ten states by Medicare Enrollment in 2010

This report presents an analysis of the 10 states with the lowest Medicare enrollment in the United States for the year 2010. Medicare enrollment is an important metric, as it reflects the number of elderly or eligible individuals receiving healthcare coverage under the federal Medicare program. By identifying the states with the lowest enrollment, we can explore potential reasons for such low figures and the possible healthcare needs of these states.

The analysis was performed on the new_merged_data dataset, which contains population and Medicare data across various years. The specific focus is on the year 2010. The following steps outline the methodology used:

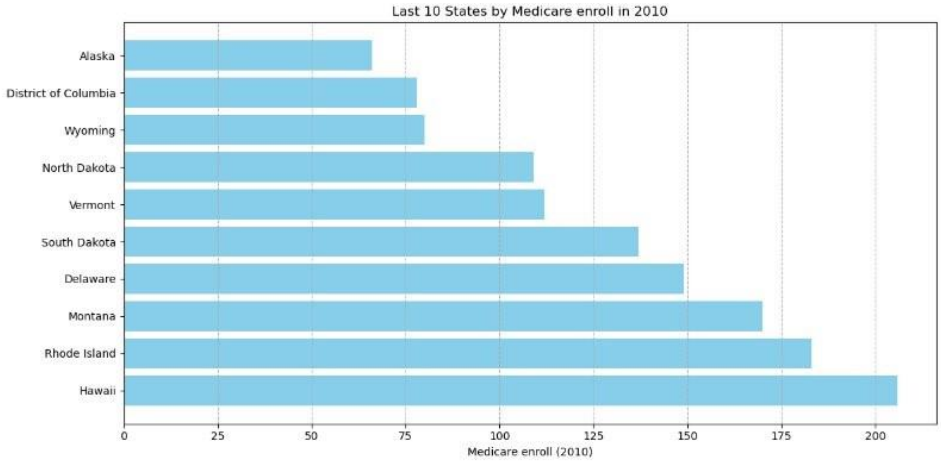
- Data Filtering: The dataset was filtered to include only the records for the year 2010.
- Lowest 10 States: The nsmallest() function was used to identify the 10 states with the lowest Medicare enrollments in 2010.
- Visualization: A horizontal bar chart was created to display the bottom 10 states by Medicare enrollment for 2010, with the states ranked from lowest to highest.

	State	Medicare 2010
1	Alaska	66
8	District of Columbia	78
50	Wyoming	80
34	North Dakota	109
45	Vermont	112
41	South Dakota	137
7	Delaware	149
26	Montana	170
39	Rhode Island	183
11	Hawaii	206

Table 5: The least ten states in Medicare enrollment in 2010

The analysis highlights the 10 states with the lowest Medicare enrollment in 2010, revealing significant variation in enrollment across different regions of the United States. Alaska and District of Columbia had the lowest enrollment numbers, while the gap between the lowest and the 10th state on the list is still substantial. Understanding the factors behind these low enrollment figures could help policymakers and healthcare providers address potential disparities in access to healthcare resources in these states. Further research could explore how Medicare enrollment has changed in these states over time and how it compares to healthcare coverage trends at the national level.

the Outliers



Comparison of Medicare Enrollment in 2010 and 2017 with changes over the years:

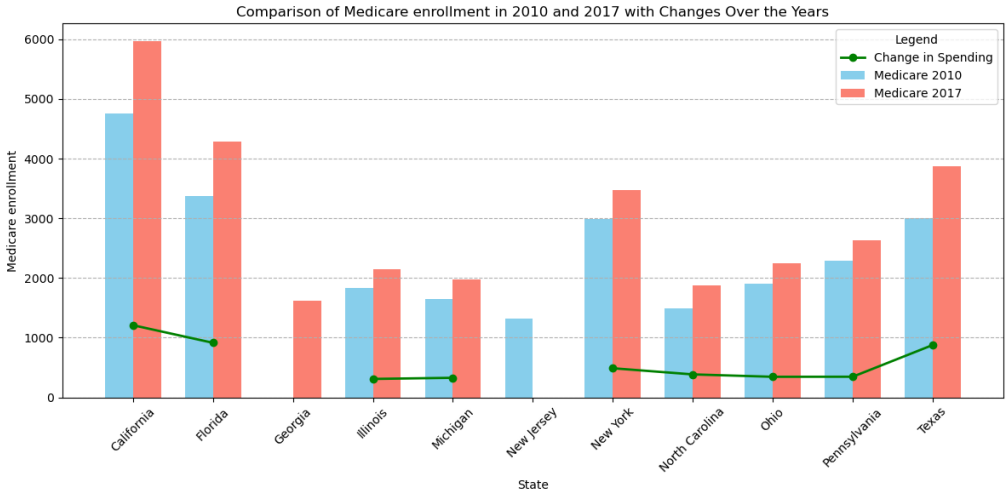
This report presents a comparative analysis of the top 10 states with the highest Medicare enrollments for the years 2010 and 2017. Medicare, a critical federal healthcare program, is designed to provide health insurance to people aged 65 and older, as well as certain younger individuals with disabilities. The analysis highlights changes in Medicare enrollments over time and identifies trends in spending across states.

The analysis was conducted using a merged dataset containing population and Medicare data for both 2010 and 2017. The following steps outline the methodology:

- Data Filtering: The dataset was filtered to include records from 2010 and 2017.
- Top 10 States Selection: For both years, the top 10 states with the highest Medicare enrollments were identified using the nlargest() function.
- Data Merging: The two DataFrames containing the top 10 states for 2010 and 2017 were merged, allowing for a side-by-side comparison of Medicare enrollments in each state.
- Visualization: A grouped bar chart was created to display Medicare enrollment in both years, along with a line plot to represent the change in Medicare enrollment from 2010 to 2017. The green line in the chart above represents the change in Medicare enrollments between 2010 and 2017 for each state. This change provides valuable insights into the states where Medicare enrollment has increased or decreased the most over the seven-year period.
- Significant Growth: Some states, such as California and Florida, experienced significant growth in Medicare enrollment, reflecting either an aging population or improvements in Medicare access.
- Stable or Declining Enrollment: States like Ohio had stable enrollments, with only minor changes between the two years.
- Geographical Patterns: There may be regional differences contributing to the trends, with some regions showing higher increases in Medicare enrollments due to demographic shifts or state-specific policies related to healthcare.

The comparison of Medicare enrollments in 2010 and 2017 reveals significant variation across states in both absolute enrollment numbers and the rate of change over time. The analysis highlights states with the highest Medicare enrollments, such as Texas, and those with significant increases, such as California.

These trends may reflect demographic changes, such as an aging population in some states, or differences in state healthcare policies. Policymakers can use this information to better understand healthcare needs and resource allocation across states.



Medicare Enrollment in the US per State

This report provides a comparative analysis of Medicare enrollment by state for the years 2010 and 2017. Medicare, a vital healthcare program in the United States, supports millions of individuals over the age of 65 as well as certain younger individuals with disabilities. This analysis aims to illustrate how Medicare enrollments varied across states in these two years, providing insights into trends and potential regional differences.

The analysis was conducted using a merged dataset containing Medicare enrollment data from both 2010 and 2017 for each state. The following steps outline the methodology used:

- Data Filtering: The dataset was filtered to select data from the specified years, 2010 and 2017, allowing for a direct comparison of Medicare enrollments per state.
- Line Plot Visualization: Line plots were created to visualize Medicare enrollments per state in both 2010 and 2017. Each state's enrollment is represented on the x-axis, while the y-axis displays the number of enrollees in thousands for both years.

The line plot comparison reveals distinct trends:

- Rapid Growth in Enrollment: States like California, Florida, and Texas saw rapid growth in enrollments, which may reflect the states' growing elderly populations or healthcare expansion efforts.
- Moderate or Minimal Changes: In contrast, smaller states such as Rhode Island and Montana experienced smaller or more gradual changes in Medicare enrollments.

Regional Differences

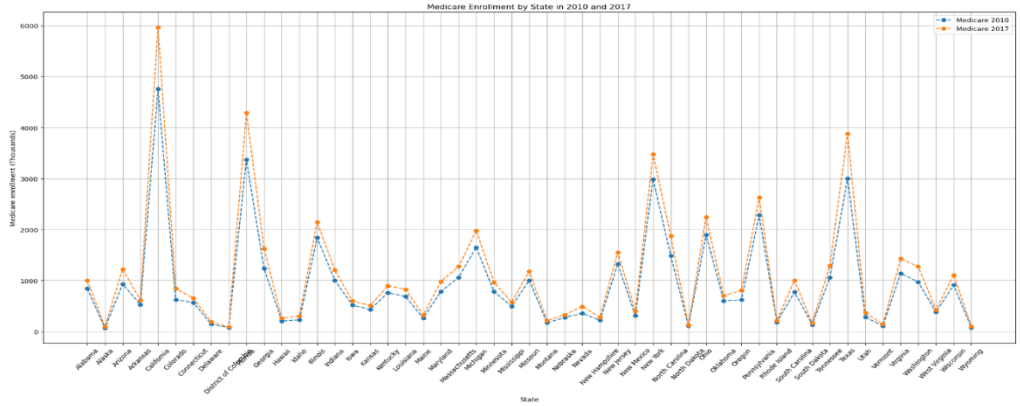
The differences in Medicare enrollment growth across states could be attributed to:

- Population Aging: States with higher percentages of aging populations such as Florida naturally see more significant increases in enrollments.
- State Healthcare Policies: Local policies and Medicare outreach programs may also influence the degree of growth in Medicare enrollments.

the Outliers

The analysis of Medicare enrollment data for 2010 and 2017 reveals substantial variability across states. Larger, more populous states, especially those with growing elderly populations, saw the greatest increases in enrollments, while smaller states experienced more stable trends.

These insights highlight the importance of understanding state-specific demographic shifts and healthcare policies when assessing future Medicare demands.



Top ten States by Medicare Enrollment in 2010, 2014 and 2017

The purpose of this analysis is to compare the Medicare enrollment trends across the top 10 states in the years 2010, 2014, and 2016. By visualizing the Medicare enrollment for these years, we aim to identify patterns and changes in the enrollment numbers over time.

The dataset used contains Medicare enrollment data per state for the years 2010, 2014, and 2016. The code filters the data for each of these years and selects the top 10 states with the highest Medicare enrollment in each year.

Key steps in the analysis:

- Filtering Data by Year: The dataset is filtered to extract data for the specific years (2010, 2014, and 2016).
- Identifying Top 10 States: The nlargest() function is used to select the top 10 states by Medicare enrollment for each year.
- Plotting the Data: A line chart is used to visualize Medicare enrollment for the top 10 states across the selected years. For each year, a line is plotted for each state showing its enrollment numbers.

Key Findings:

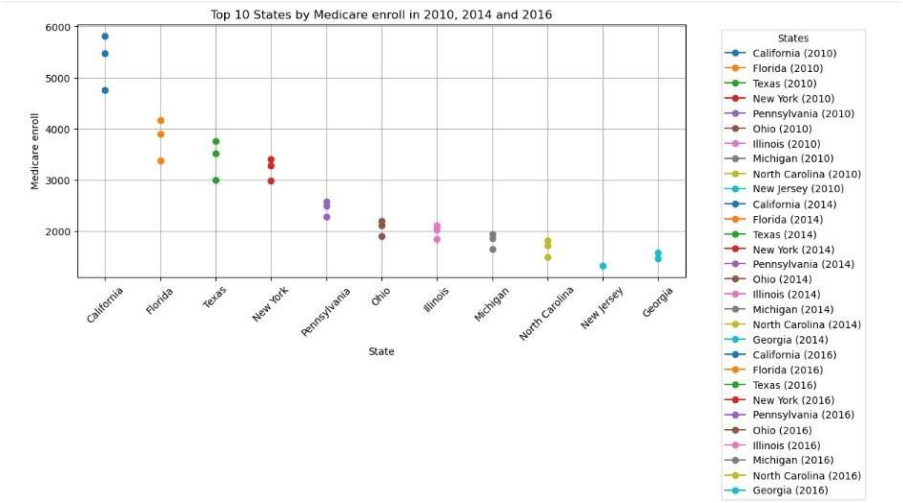
- State-Level Medicare Enrollment: The line plot for each year shows the top 10 states with the highest Medicare enrollments. Enrollment patterns show variation from state to state, with some states consistently ranking high across multiple years.
- Trend Over the Years: By plotting the data for three different years (2010, 2014, and 2016), the chart illustrates how Medicare enrollments have changed within the top 10 states over time. Some states show steady growth, while others fluctuate in enrollment numbers across the three years.

Interpretation

- Consistent States: States that consistently appear in the top 10 for all three years have large populations of eligible individuals for Medicare.

- Year-on-Year Changes: The chart shows states with increasing or decreasing Medicare enrollments, which may indicate demographic shifts, changes in state healthcare policies, or other factors.

The visual comparison of Medicare enrollments in 2010, 2014, and 2016 highlights key states with high Medicare enrollments and reveals trends over time. This information can be useful for healthcare policymakers to assess the demand for Medicare services and plan resource allocation accordingly.



Objective 4: A look into Income - The relationship between high earners and healthcare spending.

Income varies based on job types and perhaps legal wills or inheritance. Income can indeed differ widely depending on the type of job, the industry, level of experience, location, and many other factors. Health expenses can vary depending on the healthcare provider, location, type of care, and the healthcare system in each state in United States. In the current project, we aim to see if health care expenses correlate with income level of people.

To answer our question, we analyzed US_POPULATION20.CSV for population by states throughout years 2010-2017, US_PER_CAPITA20.CSV for spendings by states throughout years and Income.csv for household income by states collected from United States Census Bureau (refer to attached data_inventory.xlsx).

Strategy of the analysis

First, all data were cleaned using .dropna() (Fig.1). After cleaning, data showed equal number of contents when checking with .count, and reduced number of columns and rows when use .shape. Later, useful columns for further analysis were selected and saved in new dataframes. In US_POPULATION.csv, population were multiplied by thousand to create total population and the years in Year-columns were shifted to the rows by using looping and .concat(). Income per person in different states was calculated by dividing Income.csv by number of total populations. Next, using looping and .concat(), columns_Year is shifted to the rows and using .pivot_table() contents in the Items column were listed as columns to analyze US_Per_Capita20.csv to create same arrangement of the table as

in Income.csv. After that, the values of the spendings in different health care were combined to estimate total expenses of health care. Finally, all data were merged by utilizing .merge() for further analysis.

Figure 1. Strategy of data analysis

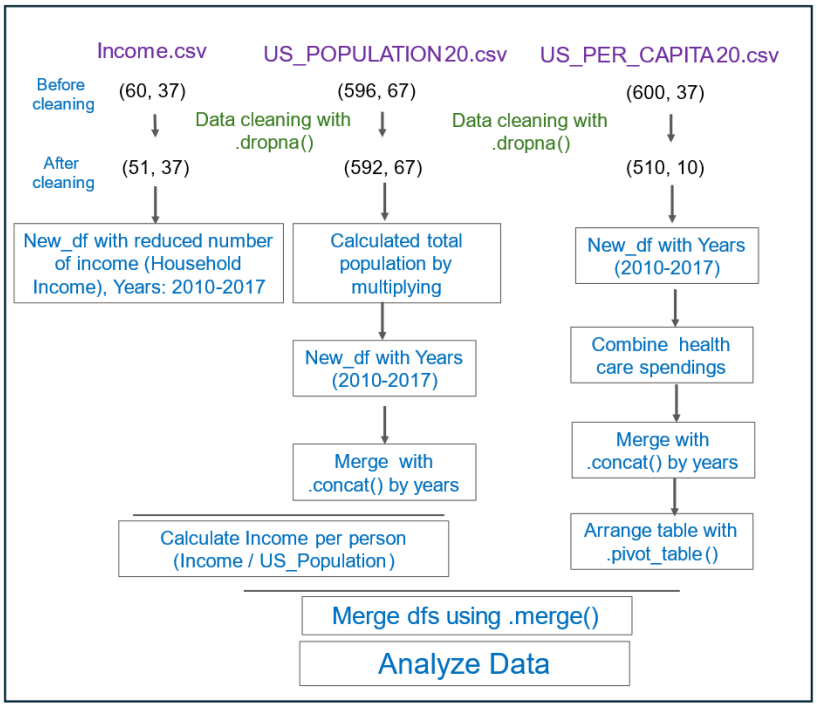


Figure 1. Three different csv files (in purple) used for the analysis. First data cleaned with .dropna() (in green). After cleaning, data were processed as indicated in steps in square box (in blue).

Output of the analysis

Comparing income levels of 2010 to 2017 of States with their health care spendings show mild positive correlation (R=0.39, P=1.188e-16) indicating wealthy population spends more in health care. Income levels of 51 states (includes District of Columbia) of US show small differences; the highest income was in District of Columbia, Maine, and Vermont and the states with lowest income were Utah, Hawai and California suggesting average earning of States in US is balanced. States like California where most of the richest people live has big diversity in terms of income level. (Fig. 3).

Figure 2. Correlation between Income and health care spendings

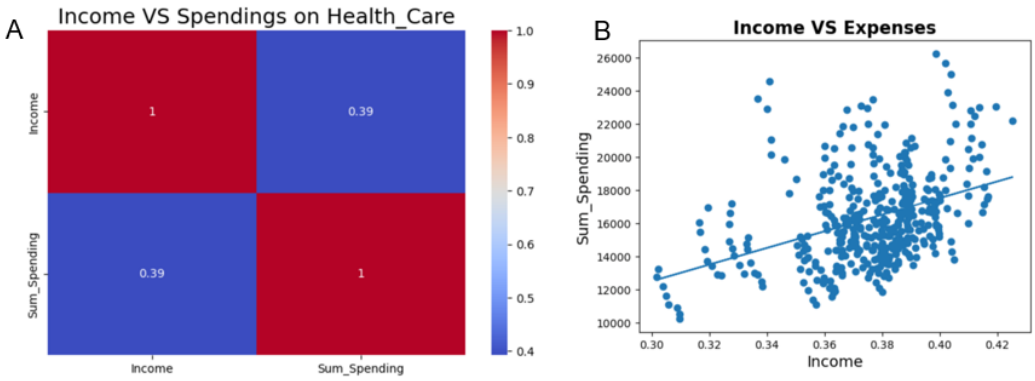


Figure 2. Total mount of income estimates from US Census Bureau (Income) were compared with total health care expenses and shown in heat map (A) and scatter plot (B, R=0.3921, p=1.88e-16

Figure 3.

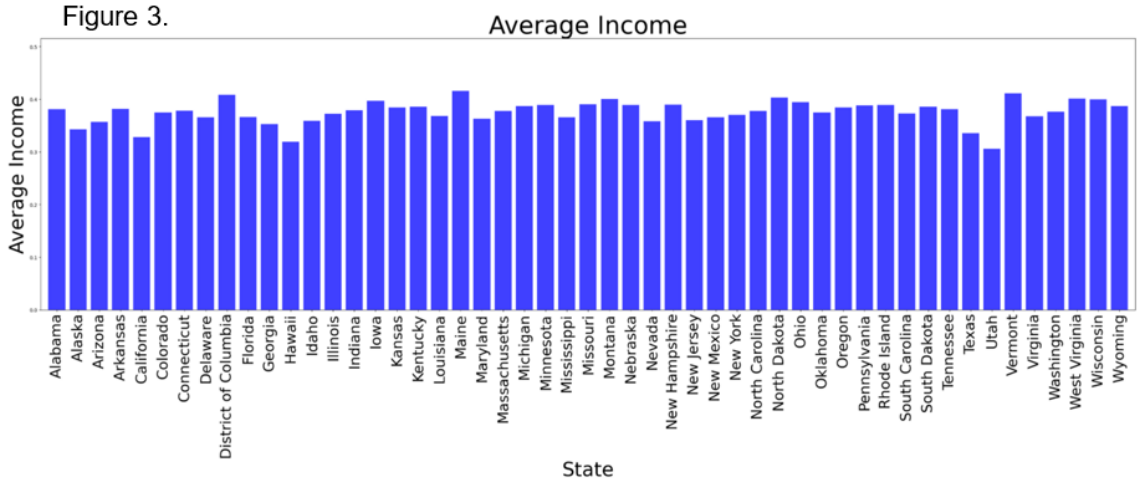


Figure 3. Average Income were calculated with the mean of values in years (2010-2017).

Alternative potential strategy: In the current analysis, I combined all spendings together and compared with income level. It might be better to categorize spendings into health serious illness-related and healthcare services for pleasure will estimate better correlation with income.

Lesson learned:

1. Understanding the data and understanding the questions to be answered is the most important.
2. Exploring in data deeply and make a detailed strategy for the analysis and collect the information (codes) that is useful for this analysis to answer the specific questions are also required. Data are broad enough to create numerous outputs, focusing on the specific question will save time and reproducibility.
3. Checking data quality and the resources of data are essential to be considered before starting analysis.
4. Be friends with GOOGLE

What should be done: (1) In the current analysis, I have combined different expenses of health care to compare Income level and mild correlation of expenses with Income. If expenses were categorized,

it might lead to a better outcome. (2) The income values of per person was less than one which was less than expenses on health care. The original value of Income might be in thousands; however it was not indicated in the resource. It traced back in the future.

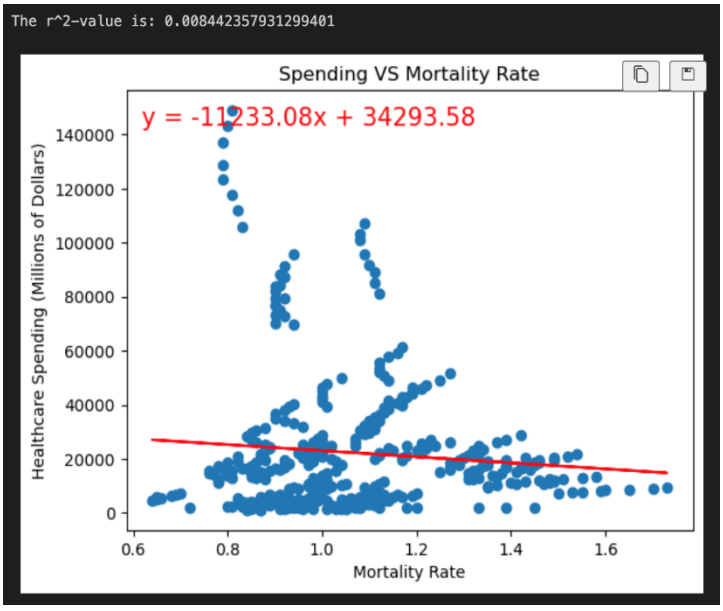
Conclusion

Analyses of income levels versus healthcare expenses provide a mild positive correlation, indicating that people spend, and purchase healthcare products based on their income level.

Discussion

(1) People with higher incomes tend to invest more in preventive care and healthier lifestyle choices (e.g., gym memberships, organic foods, wellness programs), which can result in higher overall healthcare spending. These expenditures reflect the ability to prioritize long-term health.
(2) Although higher-income people can afford routine healthcare and prevention, lower-income people tend to spend a higher percentage of their income on healthcare in emergency situations. But the overall spending remains lower due to limited access to services.
(3) Higher-income groups tend to purchase more expensive, high-quality healthcare products—such as prescription drugs, supplements, medical devices, or advanced treatments—which increase total healthcare spending.
(4) Wealthier individuals are more likely to have **comprehensive health insurance** that covers a wider range of medical services, often allowing them to seek specialized care or advanced treatments, which they can afford either through insurance premiums or out-of-pocket spending.
(5) The results of Average Income analysis of States showed California and Utah as a lowest income states, however, these states are known to have high income earnings indicating the variation of the income is high in those states.

Objective 5: Mortality rate versus government spending



For this comparison we took the total deaths in each year from 2010 to 2017 and divided it by the population to account for some states having much larger populations that others to get the

percentage of people who died in a state each year. After that we looked at the total spending of each of those states also divided by the population of those states to get the per capital spending of each state for each year. Then we made a scatter plot of the 2 values, Mortality Rate and Per capita spent on insurance to see if there was a correlation. We found an Rsquared value of .26 or that there was about a 26% relation between the money spent on insurance and the death rate of the state. This shows a weak correlation between the 2.

Conclusion Related to our Hypothesis:

The data shows that increased healthcare investment does not lead to better mortality outcomes – in fact if anything it has a slight positive correlation meaning that in places with higher healthcare spending there is also a higher mortality rate so, in the context of this analysis, our team was accurate in our prediction that higher health spend does not correlate to a better mortality rate.

Though the Outliers Hypothesis turned out to be true, we *do* believe the investment is worth it, though not in the sole context of mortality rate. There are many other factors that are part of a health society not just solely mortality rate and given enough time to find and understand other factors that lead to a healthier society we would have attempted to see if there is a correlation between those and healthcare spending. It is also important to note that this spending is for healthcare services and not health care research.