the Outliers

October, 2024: Team 3 Project Brief 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
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 Data Analyst Answering Question # 2

Project Title: US Healthcare Expenditures: Is personal healthcare investment making a positive impact on health outcomes?

Project Brief: US Healthcare spend continues to increase year over year[1]. Is the investment worth it? And does that answer change based on where you live, how much you make, if you receive government assistance, or where all of that spend is being allocated? The Outliers plan to figure that out.

Questions to be answered:

1. **Sam: Has personal healthcare spend truly increased over the years?** How much did US healthcare spending increase year over year, by state, from 2010-2020?
2. **Kathryn: Do government programs decrease total personal healthcare spend?** Do those using government programs actually spend less on healthcare than those that have private insurance? How much of the total US healthcare spend does Medicaid/Medicare account for vs. Private Insurance?
3. **Do states with older populations have a higher level of Medicare spend?**
4. **Do wealthier populations (by state) invest more for health insurance?** *(Include correlation graph of Medicare, Medicaid, and PI compared to overall spend)*
5. **Is there correlation between state mortality rates and healthcare spending?**
6. **So… Is the investment worth it? [Written Analysis]**

[1] https://www.healthsystemtracker.org/chart-collection/u-s-spending-healthcare-changed-time/
https://www.statista.com/statistics/184955/us-national-health-expenditures-per-capita-since-1960/

Hypothesis: The Outliers believe that for 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 match the positive correlation to mortality outcomes.

EDA

Datasets being used[2]:

See .xls for additional details: Data_Inventory.xlsx

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

Key takeaways from EDA:

- Overall: After organizing and cleansing the data selected, our team was excited to see that these datasets, for the most part, were fairly clean and complete over the course of multiple years. There were a few findings our team made note of, that ultimately drove changes to our scope, to ensure we're 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 col instead of 37 col (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 wouldn't 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 years 2010-2020 across datasets, but after cleansing this particular 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.

[2] CMS.gov Health Expenditures by State of Residence, 1991-2020: https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/state-residence
https://www.countyhealthrankings.org/health-data/methodology-and-sources/data-documentation/national-data-documentation-2010-2022