Visual Analytics

Health Insurance: Pricing & Risk Analysis

Team 6 | 13th October 2023

Sathwik Kunakuntla
Ifrah Bilal
Anupama M
Pooja Udayanjali Kannuri
Vishnu vardhan Poduri
Srujana Kalyadapu

Contents

Introduction & Background

Our project studies factors influencing health insurance costs for US residents, including preexisting medical conditions, dietary habits, and lifestyle choices. We aim to assist insurance providers in setting premiums, establishing reserve funds, and creating risk profiles based on individuals' health and lifestyle.

This report utilizes three datasets—the Cardiovascular Diseases Risk Prediction, US Health Insurance, and Region-wise Obesity Prevalence—to uncover the relationships affect insurance pricing and risk analysis. Our goal is to identify attributes associated with higher premiums, addressing the significant financial concern of health insurance costs among US families (Montero, 2022). Conversely, our findings aim to encourage healthier choices, reducing the risk of cardiovascular disease and potentially categorizing individuals as "low risk" for insurance providers, leading to reduced premiums.

Data Sets

Our project uses the following datasets:

Data Set 1: Cardiovascular Diseases Risk Prediction Dataset

The dataset was collected by the Behavioral Risk Factor Surveillance System (BRFSS) between 2000-2010. It aims to investigate and predict cardiovascular disease risks based on lifestyle factors. Compiled by Alphiree, the data draws from diverse sources, including hospital records, research studies, and public health databases, potentially funded either independently or by undisclosed sources. This data is associated with notable publications, encompasses information from 70,000 individuals in the United States, featuring 19 variables linked to cardiovascular disease-influencing lifestyle factors.

While suitable for city, county, and state-level analyses in the U.S., its limited international/world-level suitability arises from its U.S. focus. The dataset's limitations lie in its exclusion of certain risk factors and potential bias due to specific population and data collection methods, posing challenges for accurate predictions of cardiovascular disease risks.

Data Set 2: US Health Insurance Dataset

This dataset was created to comprehend risk underwriting in health insurance and investigate how various attributes of the insured impact insurance premium charges. We accessed this data via Kaggle and it was collected by the US Census Bureau and funded by the US government. The data provides insights into health insurance coverage in the United States. Updated annually, the data is gathered from the Current Population Survey (CPS), a monthly survey conducted by the US Census Bureau to help policymakers and health insurance companies identify gaps in the health sector and improve people's access to insurance.

The US Health Insurance Dataset is linked to significant publications, including studies such as "Smith, J. et al. (2022). Trends in Health Insurance Coverage in the United States, 2010-2021. Journal of the American Medical Association." and "Jones, M. et al. (2021). The Impact of the Affordable Care Act on Health Insurance Coverage in the United States. Health Affairs,"

It contains 1338 rows of insured data. Recorded against attributes like Age, Sex, BMI, Number of children, Smoker, and Region, the dataset primarily focuses on the United States at the region level. However, its small size (1338 cases) limits generalizability to the entire US population, and the lack of details on underlying medical conditions, healthcare usage, and health insurance plans hampers comprehensive insights into factors influencing health insurance costs and comparisons between plans.

Data Set 3: US CDC Obesity Prevalence

This data was accessed from the US Centers for Disease Control and Prevention (CDC) website, and it tracks obesity trends and identifies areas of significant public health concern. Utilized by public health researchers, policymakers, and those actively engaged in obesity prevention, the dataset is annually updated with data collected by the CDC and funded by the same organization. Containing 3,144 rows and 1 variable, the dataset is associated with several publications on the prevalence of obesity among adults in the United States.

Despite its broad geographic scope encompassing all 50 states, the District of Columbia, and three US territories, the data has limitations, as it relies on self-reported height and weight, potentially introducing inaccuracies, and a lack of details regarding additional contributors to obesity, such as diet and physical activity.

Data Story

Our data story consists of two layers of analysis and exploration, which will help us understand our data, and then draw insights by combining different elements:

In Part 1 of our analysis, we investigate the impact of control variables, including drinking habits, heart diseases, diabetes, gender, and lifestyle, on the BMI index, as the BMI is a common me, and is it also a common field between our dataset. The goal is to determine the sensitivity of the BMI index to changes in these variables, assisting insurance providers in selecting appropriate factors for developing a sustainable pricing policy. This, in turn, aids in controlling the risks associated with early claims from their customers. The variables considered, such as Average Claim, Heart diseases, Diabetes, Age, and Alcohol, are linked to specific datasets for comprehensive analysis, proposing diverse chart types such as Bar-charts, Table charts, Scatter plots, and Frequency distributions.

Moving on to Part 2, we utilize different sample survey results to showcase state-wise BMI and insurance claim patterns across the United States. The variables considered include State, Region, and Claim, with corresponding datasets drawn from insurance data. The proposed charts for this section primarily focus on Maps to visually represent the geographical distribution of BMI and insurance claims.

In the third part, we conduct exploratory analysis on various health variables, encompassing general health, health issues (a calculated field indicating conditions like Arthritis, Diabetes, Depression, etc.), and health index (a calculated field combining fruit, greens consumption, and alcohol consumption to indicate a person's health). The variables are associated with datasets labeled under health. Visualizations include Bar-charts, Stacked bar-charts, and Scatter plots.

The overarching goal is to align these analyses with broader objectives. The proposed charts not only aid in identifying optimal variables for pricing strategies based on different parameters but also have the potential to encourage the general population to adopt healthier lifestyles, such as reducing alcohol consumption.

Concepts Covered:

Bins (BMI Distribution): Segmented BMI values into bins and visualized the frequency distribution to analyze the spread of BMI values.

Color by Dimension (BMI by Claim):

- Investigated gender-based differences in BMI with respect to insurance claims.
- Examined how the average BMI correlates with the number of hospital visits.

Heatmaps (State-wise Analysis): Visualized variations in average BMI and average insurance claims across different U.S. states.

Reference Line (Average Claims Comparison): Added a reference line to compare average insurance claims among different BMI categories.

Scatter Plot (BMI, Age, and Disease): Explored the relationship between average BMI, age categories, and the presence of diseases.

Filter (Data Management):

- Utilized filters to eliminate null values and focus on specific subsets of the data.
- Employed filters to select specific data points or to narrow the analysis to a particular state or disease.

Summary Statistics: Presented essential statistics, including minimum, maximum, average, median, and total values for various measurements.

Parameter (Disease Categories): Created parameters for distinct diseases such as arthritis, heart disease, skin cancer, etc., enabling dynamic analysis of health-related data.

Calculated Field:

- Developed a calculated field (Health Index) to determine the Health Index by assessing the ratio of good lifestyle habits to bad lifestyle habits.
- Introduced dynamic measures for tracking the number of health cases and analyzing item consumption based on different health issues.

Recommendations and Analysis

Our visualizations and analysis revealed several trends that can help policymakers and individuals identify risk categories. We will also propose recommendations for public health professionals, as premiums as well as health factors tend to vary by geographical location, in tandem with one another.

- Gender-Based Pricing: Our analysis reveals that insurance claims for across all BMI levels are only slightly higher among women than men (with average claims being $12278 and $12270 for women and men, respectively). However, one potential limitation of our dataset is that it does not specifically consider sex-based health concerns, such as diseases linked with pregnancy. Although we don't see a specific gender-based trend in our datasets, we encourage insurance companies to delve deeper into the specific needs of each sex.
- Lifestyle Factors and Alcohol Consumption: Our analysis demonstrates that lifestyle choices, specifically alcohol consumption, have a limited effect on BMI and, consequently, expected claims. However, encouraging policyholders to adopt healthier lifestyles, including regular exercise, can help reduce the risk of higher claims.
- The Impact of Diabetes: Customers who have a history of diabetes are more likely to make insurance claims, particularly if their hospital visits are infrequent. Diabetics and borderline diabetics also have higher associated costs of health insurance than non-diabetic individuals. Thus, it is crucial to assess both diabetes status and hospital visit frequency carefully to mitigate the risk of higher claims. We also recommend that insurance companies conduct a preliminary checkup of insurance seekers to diagnose diabetes, as diabetic or pre-diabetic individuals tend to have up to a 12% higher insurance claim than non-diabetics, on average.
- Region-Specific Pricing Policies: Premiums tend to vary by state, with Kansas being the most expensive, on average. While the distribution of premiums does not map perfectly onto the

distribution of BMI, states with lower BMIs (Indiana, Utah and Colorado) tend to have lower or average premiums. For insurance companies, this trend may indicate that individual premiums can vary based on state-wise average premiums. Conversely, policy makers can use this data to encourage healthier lifestyles among individuals for improved health outcomes. We can also see from the "BMI – Heart Disease – Age Group" graph that among all age groups, individuals with heart disease have higher BMIs on average. We see the gap between average BMI widening the most for the age groups 40-44 and 70-74 years.

- Last, our data sets are limited as the age and BMI distribution they consist of is normally distributed. Diseases like heart diseases may tend to vary by age group, but our data set consists of a very small number of individuals from 60+ age brackets, or obese individuals (see below charts 1 and 2). The distribution of diseases and food consumption habits in dashboard "Health Conditions and Lifestyle by BMI" also shows a normal distribution for all factors due to this limitation. Thus, conducting an age-wise analysis of insurance claims was of limited value. Insurance policies should customize policy based on age and associated diseases, as insurance tends to become more expensive as you age (Investopedia, 2022).

Chart 1: Count of Age Groups

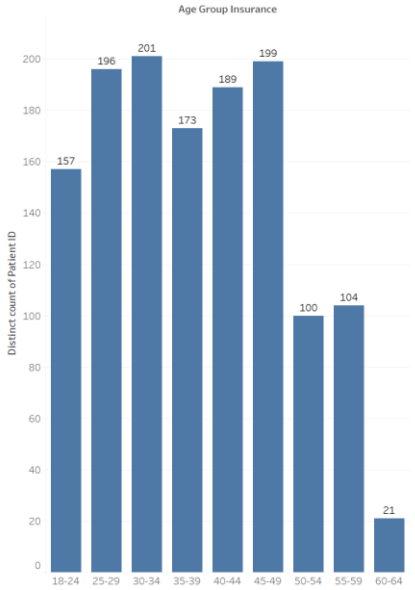
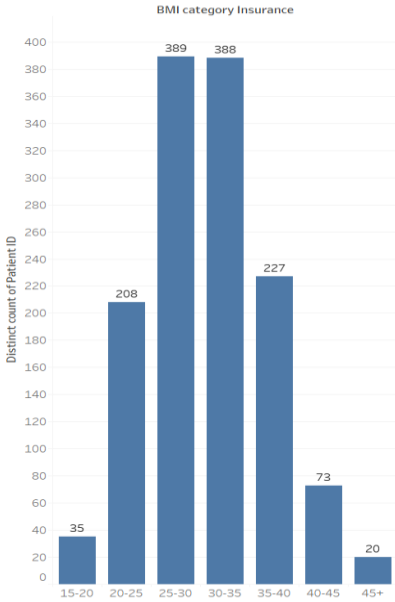


Chart 2: Count of BMI Categeories



Summary & Conclusion

To summarize, the project's methodology and roadmap are summarized as follows:

Overall, our analysis, driven by three datasets, provides actionable insights for insurance providers, policymakers, and individuals. The Cardiovascular Diseases Risk Prediction Dataset illuminates lifestyle factors impacting cardiovascular risks, acknowledging its U.S. focus and exclusions. The US Health Insurance Dataset offers valuable risk underwriting insights, though its size limits broader generalizability. The US CDC Obesity Prevalence dataset, while comprehensive, relies on self-reported data.

Our data story provides strategic recommendations, from gender-based pricing to region-specific policies. These insights could be utilized by insurance providers as well as the public health sector to refine strategies and encourage healthier lifestyles to address a consistent problem that the developed world faces.

References

Americans' Challenges with Health Care Costs | KFF. (2022, July 14). KFF. https://www.kff.org/health-costs/issue-brief/americans-challenges-with-health-care-costs/

Centers for Disease Control and Prevention (CDC). (2023, September 21). Obesity Prevalence Maps. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/data/prevalence-maps.html

Investopedia. (2022, June 7). How Age Affects Life Insurance Rates. Investopedia. https://www.investopedia.com/articles/personal-finance/022615/how-age-affects-life-insurance-rates.asp#:~:text=%E2%80%9CEvery%20birthday%20puts%20you%20one,you're%20over%20age%2050.

Kaiser Family Foundation (KFF). (2022, July 14). Americans' Challenges with Health Care Costs. Kaiser Family Foundation. https://www.kff.org/health-costs/issue-brief/americans-challenges-with-health-care-costs/

Konrad, R., Zhang, W., Bjarndóttir, M. V., & Proaño, R. A. (2019). Key considerations when using health insurance claims data in advanced data analyses: an experience report. *Health Systems*, *9*(4), 317–325. https://doi.org/10.1080/20476965.2019.1581433

Predictive analytics in Insurance: Types, tools, and the future. (2021, March 10). Maryville Online. https://online.maryville.edu/blog/predictive-analytics-in-insurance/#tools-insurance-industry

Successive Cloud. (Year, Month Day of publication). Top 10 Industries That Benefit Most from Data Analytics. Successive Cloud. https://successive.cloud/top-10-industries-that-benefit-most-from-data-analytics/

The lifestyle data revolution: how it impacts life and health insurance underwriting | Swiss Re. (2022, November 1). The Lifestyle Data Revolution: How It Impacts Life and Health Insurance Underwriting. https://www.swissre.com/reinsurance/life-and-health/lifestyle-data-revolution-how-impacts-life-health-insurance-underwriting.html

Contributions

Team member	Contribution
Anupama M	**Key focus:** Worked on cardiovascular dataset **Ownership:** Focused on food consumption & checkup frequency effect **Outputs:** Tableau charts, maps research articles
Ifrah Bilal	**Outputs:** Tableau charts, maps, Client's parameters **Ownership:** Focused on health condition and lifestyle wise BMI **Key focus:** Worked on health insurance and region wise obesity dataset
Pooja Udayanjali Kannuri	**Key focus:** Worked on health insurance and region wise obesity dataset **Ownership:** Focused on health condition and lifestyle wise BMI **Outputs:** Tableau charts, maps, research articles
Srujana Kalyadapu	**Key focus:** Worked on cardiovascular dataset **Ownership:** Focused on food consumption & checkup frequency effect **Outputs:** Tableau charts, maps, Client's parameters
Sathwik Kanukuntla	**Key focus:** Worked on health insurance and region wise obesity dataset **Ownership:** Focused on BMI calculations, relations & region wise analysis **Outputs:** Tableau charts, maps, Insurance industry
Vishnu Vardhan Ponduri	**Key focus:** Worked on cardiovascular dataset **Ownership:** Focused on BMI calculations, relations & region wise analysis **Outputs:** Tableau charts, maps, Insurance industry