Stroke contributes to your insurance bills

Han Zhang

(hz5g21@soton.ac.uk/32672446)

I. Data Story Summary

Unlike other countries, which offer comprehensive national health insurance that provides basic medical coverage to everyone, their prices are fixed for everyone, while health insurance in the United States must be purchased through a government or private insurance company. This results in different health insurance bills for different people and leading opaque market-based pricing [1].

Annual medical insurance bills are growing with the time, and medical insurance billing rules are also evolving. How do insurance companies calculate the cost of your bill? What is the benchmark? When they are charging for insurance, customer attributes are critical in making business decisions. Now, it's time to demystify it. This data story is about the relationship between customer characteristics and their health insurance spending and insurance billing benchmarks.

From this data story, age, BMI, and stroke are the biggest factors affecting your health insurance bill. Stroke dominates your insurance bill. it is contributing to your health insurance bills. So reducing your stroke risk and keeping a balanced body is one way to reduce your insurance bills

II. Dataset Summary

The main dataset is from the kaggle website. Two secondary datasets are from CMS.gov and bea.gov.

A. *Healthcare*

This dataset contains 5110 pieces of data and 13 attributes [2]. Each attribute is associated with the target which is health insurance bill. Before data visualization, it was pre-processed. After reviewing the data, it was found that the target attribute had 201 missing values, so those missing values were deleted. There is only one record whose gender is other, which is considered an outlier, so this record is deleted. Before linear regression, categorical variables are processed through one-hot techniques [3] to visualize the relationship with insurance costs.

B. *Average health bills between 1991 and 2014*

This dataset contains per capita insurance bills for each state in the United States from 1991 to 2014.

C. *Average income in each state*

This dataset contains the per capita income of each state in the United States from 1991 to 2014.

III. Visualisations

A. *Map distribution of Average health insurance bills*

1) *Description*

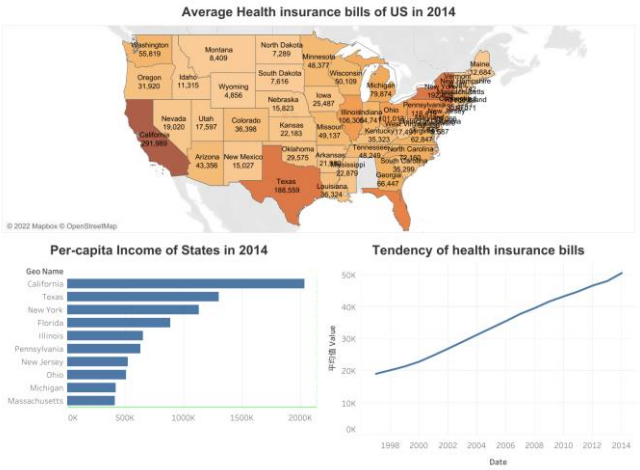

Fig 1: Interactive map

This visualization contains 3 small graphs. The largest of these, average health insurance bills of the US in 2014, shows the distribution of health insurance per capita in 2014. Those with deeper colours represent higher per capita health insurance costs. In the picture, it can be seen that California, Texas, New York, and Florida have the highest average health insurance bills. Second, per-capita income of states in 2014 shows the top 10 highest-income states in the United States. The states with the highest per capita insurance bills also have the highest per capita incomes. The tendency of health insurance bills shows trends in health insurance from 1991 to 2014. It can be seen that in these years, the per capita medical insurance cost has nearly doubled

2) *Justification*

From the map, it's nationally displayed the per capita medical insurance bills of each state in the United States. By designing into a map-like form in order to take advantage of the cognitive advantages of map maps, one is able to use the spatial relationship between map elements as a measure of similarity [4]. In this case, it appears to be by linking where people live to the cost of insurance per capita. Perceptions of people living in various states can easily observe their own insurance charges.

Using a horizontal bar chart to sort the per capita income of each state, the states with the highest income can clearly be ranked first. people can compare states with high per capita income to states with darker colours on the map. The comparison found that higher-income states had higher per capita health insurance bills.

The trend of per capita health insurance is represented by a line graph, showing that health insurance charges become higher and higher.

3) *Narrative Design Patterns*

The Narrative design pattern of this visualization uses compare and Users-find-themselves. First, through the use of maps, the subconscious allows US residents to find out where they live and then find their average per capita insurance. Secondly, through the comparison of interactive

map and histogram, it revealed the potential relationship between per capita income and per capita health insurance. The audience has a guess that per capita income is positively correlated with per capita health insurance.

4) Strengths and Weaknesses

The advantage is that it is easy to correlate familiar geographic locations with per capita health insurance costs, so people can find the per capita health insurance costs for their location. The disadvantage is that the contrast is not obvious. Only the previous high-income states can be sure, and no reliable conclusions can be drawn.

5) Improvements

Try to link per capita insurance spending with per capita income, that is, the proportion of health insurance spending in income. This behaviour can make people more readable. Besides, it can show the different situations of the map by changing the time.

B. Categorical Feature

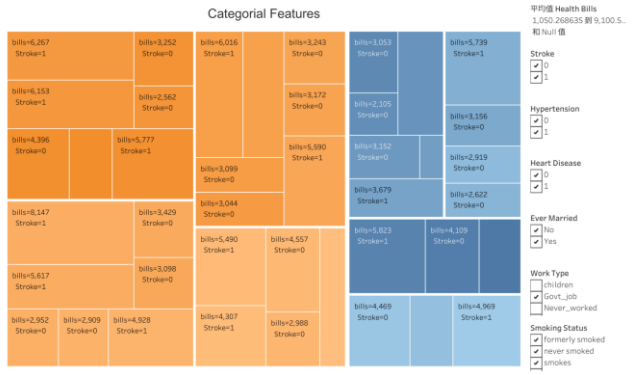

Fig 2: Treemap

1) Description

This treemap shows the impact of different customer attributes on their own health insurance costs. The area of the squares represents the average health insurance bill. The larger the area, the higher the cost.

2) Justification

Treemap is used to effectively represents the difference between the whole and the parts. Since the data has 8 categorical attributes, it is difficult to express the impact of 8 categorical features on health insurance charges through other visualizations from an overall perspective. How to judge the impact of a feature on self-insurance charges, that is the impact of having this feature and not having this feature on insurance bills. This makes the entire format hierarchical and can clearly express the impact of different features on the target [5].

3) Narrative Design Patterns

The narrative design pattern used is compare. The treemap is used to compare individual-to-individual differences. Allows individuals to explore differences of different attributes. And the influence of 8 features on the target value is fully represented.

4) Strengths and Weaknesses

The benefit is to create an overall review of the impact of different categorical variables. Insured people can check different characteristics according to their own situation and check the per capita insurance cost. Compare the impact of customer characteristics on the health bills. The disadvantage is that the information is too complicated and the gap is not obvious. Too many features are in the treemap which divides treemap into many small squares, and finding information in these squares is a troublesome thing [5].

5) Improvements

One Improvement is to represent visualizations in the form of a sunburst or tree. Taking the average insurance expenditure as the root node, then each layer is a feature, and different characteristics will have different influences on the average insurance expenditure insurance. Sunburst and tree can more effectively represent the influence of features on health bills.

C. Correlation coefficient of categorical features

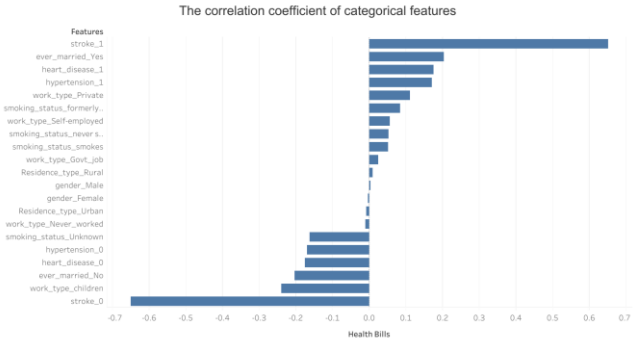

Fig 3: Bar chart of features

1) Description

Figure 3 shows the correlation coefficient of categorical attributes. Since categorical variables cannot calculate the correlation coefficient, one-hot technology is used to convert categorical variables into numerical variables. The results show that stroke is the one that most affects health insurance bills.

2) Justification

The histogram can show the impact of categorical variables on the health bills, and the negative and positive signs can show the positive and negative effects of different values on the target. For example, stroke_0 has the effect of reducing health insurance bills, and stroke_1 has the effect of increasing insurance bills.

3) Narrative Design Patterns

This narrative design pattern is concretise. The impact on health insurance is specified in different bars. And classify the impact as positive or negative.

4) Strengths and Weaknesses

The advantage is that it can well show the influence of the feature on the target value. However, the audience may be confused. For example, stroke is a categorical variable, but there are 2 values in the graph, stroke_0, and stroke_1.

5) Improvements

Redundant values for categorical variables can be removed. Keep the original categorical variables and rank them by their influence on health insurance bills. This is more intuitive and the audience is not confused

D. Numerical Features

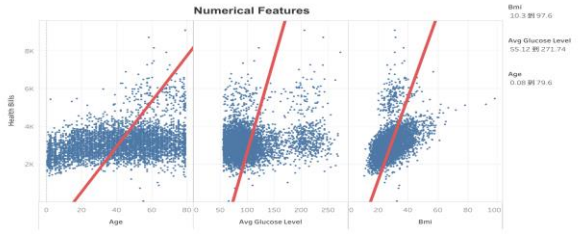

Fig 4: Numerical features

1) Description

The graph shows the impact of three numerical variables on health bills. As shown, all three variables are positively correlated with health bills, with BMI being the most strongly associated with health bills.

2) Justification

The relationship between numerical variables and health bills is represented overall by using trend line. It can be seen that they are positively correlated with health bills [6]. it can be seen their distribution at the same time.

3) Narrative Design Patterns

This narrative design pattern uses users-find-themselves. Select its own features by dragging the filter to the right. Then look at which of these three characteristics has the greatest impact on you. Different characteristics may lead to different results.

4) Strengths and Weaknesses

The Strength is that it can display the impact of the three numerical attributes of the overall population on health bills, and the filters can be a drag to display the impact of some people. However, the Weakness is that the information is too complicated, and it is difficult for people without statistical knowledge to understand

5) Improvements

An improvement is to label their slopes and correlation coefficients. This makes the image more readable.

E. Slection features

Fig 5: suitable choice of features

1) Description

The Error shows the effect of the number of features on the linear regression predictions. Drag the alpha to see that the error keeps rising. At alpha = 40, the slope becomes larger. so alpha = 40 is a suitable choice. The Train Model of alpha =40 shows the optimal number of features and a comparison between the predicted and true values.

2) Justification

By dragging the alpha, you can see that the error is rising. At alpha = 40, the slope becomes larger. so alpha = 40 is a suitable choice. By comparing the actual value and the predicted value, it can be found whether the prediction result is good or bad.

3) Narrative Design Patterns

The narrative Design Patterns use users-find-themselves. Find the appropriate alpha to perform linear regression by dragging alpha.

4) Strengths and Weaknesses

It filters out the features that have the most impact on health bills. Disadvantages are more difficult to understand for those without a machine learning background.

5) Improvements

The Train Model of alpha = 40 image can be transformed into a moving image, and the real values generated by different alphas can be compared with the prediction accuracy. Drag the alpha and you can see that the results of the Train model are changing at the time. It can more intuitively show the impact of features on health bills.

IV. Conclusion

This data story explains some of the factors in the health insurance bill and explains to the audience that, of these factors, stroke is the most influential factor. Data Stories leverages the strengths and weaknesses of charts and the narrative patterns learned in data visualization courses to design visualizations. Most visualizations are done with the tableau software. In this class, I learned that it is important to know what you know, and it is also important to communicate what you know through visualization and try to get others to understand your point of view.

References

[1] "Health care prices in the United States - Wikipedia", En.wikipedia.org, 2022.

[2] "health insurance bills", Kaggle.com, 2022.

[3] J. Brownlee, "Why One-Hot Encode Data in Machine Learning?", Machine Learning Mastery, 2022. [Online]. Available: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/.

[4] M. Hogräfer, M. Heitzler and H. Schulz, "The State of the Art in Map-Like Visualization", Computer Graphics Forum, vol. 39, no. 3, pp. 647-674, 2020. Available: 10.1111/cgf.14031.

[5] B. Johnson, "TreeViz", *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*, 1992. Available: 10.1145/142750.142833.

[6] "Visualizing Data with Pairs Plots in Python", *Medium*, 2022. [Online]. Available: https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166.