Data visualization- Initial hand-in

Han Zhang hz5g21@soton.ac.uk

1 Story Background

Unlike other nations, which offer comprehensive universal medical insurance that provides basic medical protection to all people, medical insurance in the United States must be purchased through the government or private insurance companies. In the United States, health insurance refers to any program that assists in the payment of medical bills, whether through individually purchased insurance or social insurance. Medical expenditures for sickness are quite significant if medical insurance is not acquired. As a result, medical insurance is critical for every American.

2 Story outline

Since the birth of medical insurance in the United States, the charging standards have also been changing. How do insurance companies calculate the cost of your bills? and What is the benchmark? Have you considered it? Now is the moment to Uncover its mystery. So my data story is about the relationship between a client's characteristics and their medical insurance expenditures and the benchmark set for insurance bills. Many questions will be answered during the story. Do costs for persons who smoke, for example, differ greatly from those who don't? , Will persons with chronic illnesses and poor health face higher fees? or Will the rich be paid a higher fee?

My early opinion is that BMI and stroke have an impact on medical insurance costs since they can impair your health. BMI is a measure of physical health, and stroke, to some extent, indicates the likelihood of a client being admitted to the hospital. They all reflect the medical expenses of the insured. This can be proofed by Figure 1, the average health bills on these measurements. Medical insurance premiums are almost unaffected by gender (Male or Female) or residence type and the most crucial aspect is a stroke. Clients who have had a stroke, for example, are virtually charged twice as much as those who haven't. Furthermore, the majority of patients in Figure 2. c had a BMI between 15 to 40. Others out of this range charged more. People with those two characteristics will create more medical bills, resulting in a higher price. However, this cannot always be the case; they may occasionally fail to accurately represent insurance fees. Other circumstances, as illustrated in Figures 1. b and 1. c, might alter the fees. For example, Hypertension and heart disease will both be impacted the insurance fees.

3 Dataset Detail

Three datasets are collected to contribute to the data story. One of them is an hiscare.csv file collected from Kaggle. There are 5110 rows and 13 columns in all. The column names are ID, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, BMI, smoking_status, stroke, health_bills. It occupies 380KB. Some values are missing. Only 4909 records were found in the BMI and health bills columns, with 201 records missing. The age, avg_glucose_level, BMI, and health are numerical numbers. Others are categorical.

The other is a per_capita_bills.csv file that was obtained from cms.gov. It contains information on health insurance bills by the state of residence between 1991 and 2014. Although there are 30 tables, only one of them will be utilized in the data story. The file is 299KB in size and contains 62 rows of

US status information and 36 columns of per capita personal insurance bills by state, with no missing value.

The final file is a per_capit_income.csv file obtained from bea.gov. It contains data on the US Per Capita Personal Income and how it has changed between 1991 and 2014. There are 65 and 98 rows in all. The rows represent statistics about the United States, while the 98 rows represent per capita income, with no missing value.

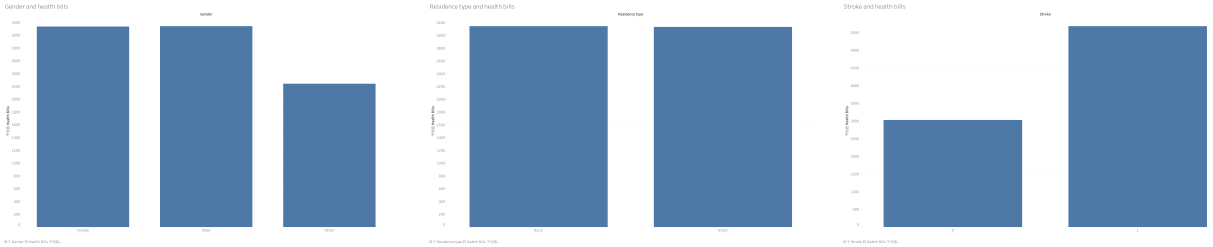
4 Visualisation Detail

For the Visualisation, firstly, I should deal with the missing value and abnormal value. For the 201 missing values, I decide to delete them and delete the abnormal value in gender (other).

Then I would formally utilize a histogram graphic for each numerical value and focus on the highest bar, which symbolizes its cluster, to analyze how the benchmark of medical insurance bills works. The bar charts are then applied to category statistics, allowing distribution patterns to be discovered shown in Figure 1. Then, in order to determine the obesity rate, we separated the BMI values into four categories: Underweight is defined as a BMI which less than 18.5 and BMI located in 18.5 and 30 indicate the normal weight, whereas BMI bigger than 30 indicates obesity. As a result, we can make a pie chart. Then we create a dispersion between BMI and healthcare expenses shown in Figure 2. c. Next, using a heatmap shown in Figure3, determine the relationship between numerical numbers. In this way, we can know which numerical numbers affect insurance bills. Finally, for stroke and smoking kinds, we may design group bars that can have an overall view of those two categories. Finally, Using a box plot to show the relationship between work type and health bills, so that, we can know the dependence between work type and health bills.

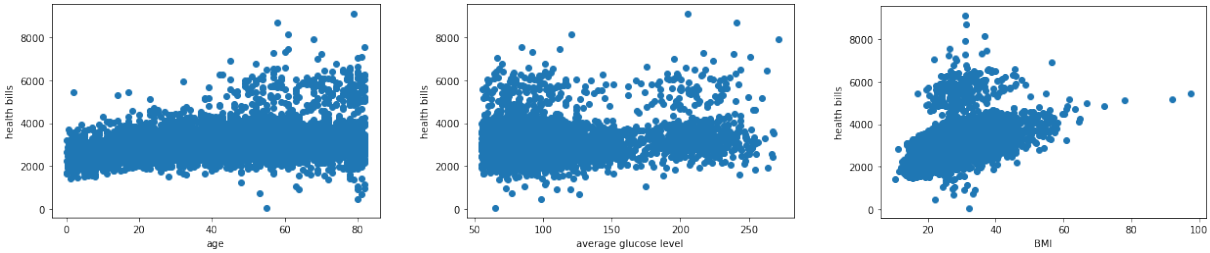
As for analyzing the per capita bills.csv and per capita income.csv, I would like to generate a map that compares the health bills and their income. So that you may perceive the disparity between the rich and the poor. In addition, the bottom of the map will be placed so that you can observe how health bills have changed over time and compare them to income growth as Figure 4 shows. This will reveal how quickly healthcare costs have risen.

A Sketches of graph



(a) Gender (b) Residence (c) Stroke

Figure 1: Categorical numbers



(a) age (b) average glucose level (c) BMI

Figure 2: numerical numbers

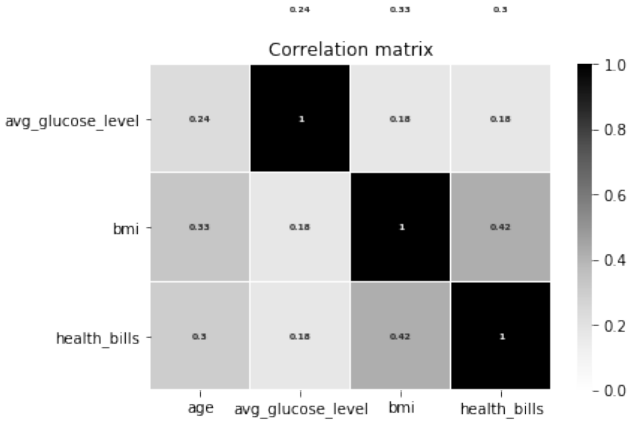


Figure 3: Heatmap

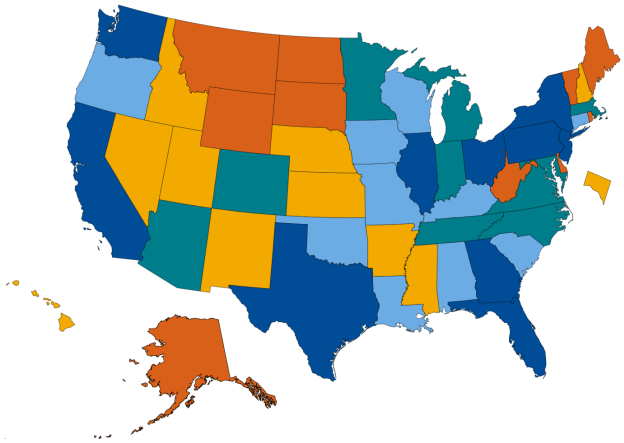


Figure 4: interactive map