**FINAL PROJECT**

**EXPLORATORY DATA ANALYSIS USING TABLEAU**

**BY**

**SADIA MAHBUB**

**INFO 5709**

**DATA VISUALIZATION AND COMMUNICATION**

**NAVYA REDDY VELVERTHI**

**SUMMER-2022**

**SECTION-003**

**INTRODUCTION:**

Health care costs in the United States have been rising steadily for decades and are projected to continue to rise. In 2019, the United States spent over $3.8 trillion on healthcare. A growing population, an aging population, disease prevalence or incidence, medical-care utilization, costs and intensity of services, and an increase in chronic illness are some of the factors that affect health care costs.

In 2020, U.S. spending is projected to grow by 9.7 percent in health care. It reached $4.1 trillion, or $12,530 per person of the country's gross domestic product, its health expenditure accounts for 19.7 percent.

Typically, customers shopping in the individual market choose from four levels of health coverage: bronze, silver, gold and platinum. Bronze plans have the lowest monthly premiums, but the highest deductibles and copays. Gold and platinum plans offer more financial protection if you get sick, but usually have higher monthly premiums. Silver plans are the most popular level of coverage because they fall between two extremes. This year, many people are eligible for tax credits that cover part of their premiums; The amount of tax credit is determined with the cost of the silver plan.

**Problem:**

In 2020, nearly 31 million Americans are uninsured. Due to lack of health care, an average of 18,314 Americans aged 25 to 64 die each year. Most people now choose not to buy health insurance due to high costs (Stasha, 2021).

Below the income, 200% of individuals at the federal poverty level (FPL) have the highest risk of becoming uninsured. Family insurance has increased Premiums for by nearly 55% over the past decade.

**Motivation:**

For most companies, using consumer data is critical. Customer characteristics can be important in making business decisions for insurance providers. As a result, being able to explore and extract value from such data is a valuable skill. In this project I try to evaluate and explore a dataset for medical costs in health insurance in the US to extract useful information and answer questions through inferences/results.

Some quarterly financial reports trading health insurance companies make publicly available monthly price data are used to create a preliminary picture of price, utilization, and enrollment growth, and recent significant changes in legislation that may take effect in one and two years. Inclusion when generating statistics. Fundamental changes in legislation after the release of the statistics are not necessarily included, however, and thus may contribute significantly to previously anticipated growth rates.

Now, machine learning, artificial intelligence, cloud computing, blockchain, and big data analytics have radically changed the global financial environment. As business models adapt and evolve to exploit the potential of these innovations, along with traditional emerging technologies, more and more companies are adopting digital transformation strategies, and Implementing barriers are slowly dissolving into different specialized domains across the industry. Insurance is no different. Data-driven decision-making is nothing new in the insurance industry.

This project focuses on a set of healthcare data from the United States Individuals with certain characteristics are included in the medical cost data. Data are accessible for four regions of the United States (Northeast, Southeast, Southwest, and Northwest). The original data set consists of 7 columns and 1339 rows of insured data, in which insurance charges are calculated based on the insured's age, SEX (gender), BMI, children(number), smoker, region, and charges. To allocate a unique identification number to each insured person, an additional 8th column 'Beneficiary ID' is added to the original data set. Age, gender, BMI, number of children, smoker, region, charge, and beneficiary ID are the eight attributes of the data set.

**RELATED WORK:**

I was looking for a dataset for health insurance studies on various websites, including Kaggle and Data.gov, before finding this dataset, which is ideal for my research. The dataset contains attributes that refer to various functions and I want to perform some operations as shown below.

There are hundreds of web pages and journal articles about health insurance on the Internet. There is a wealth of knowledge available on this topic. While browsing this site I found some papers related to health insurance. Some of the most important sources of information are Medicare.gov, the US Census Bureau. These websites have helped me to gather necessary information for my work. Different platforms, such as Kaggle and Healthdata.gov, provide different healthcare information and datasets.

There are a lot of Health Insurance Data visualization Works available on the internet. Kathryn Keisler-Starkey and Lisa N. in the US Census on Health Insurance in the United States. A job posted by Bunch: 2019 which includes the percentage of insured, uninsured, type of health insurance coverage, etc. Edward Berchik posted a work on the uninsured in the United States, including demographics of uninsured individuals by age, sex, race, region, education, occupation, income, etc. A work posted by Douglas Conway on health insurance for young adults includes details such as which states have the most uninsured adults (US Census Bureau, 2020).

**DATASET INFORMATION:**

It is a visual method of evaluating data sets and summarizing key features. EDA is mostly used to look at data that can provide knowledge beyond that available through hypothesis testing or structured modeling. Exploratory data analysis begins with basic questions. Exploratory data analysis helps in analyzing complex data, as well as summarizing data and drawing conclusions or conclusions from it.

In this project, I used exploratory data analysis to generate various hypotheses and questions, as well as address them by analyzing and visualizing the data we encountered during the phase. The dataset has 1339 rows and 7 columns. We note that BMI, age, and charge are numerical attributes; and gender, region smokers, and children are unique characteristics. We notice that the 'child' column is being treated as an integer datatype, even though it contains categorical data. There is no missing data in the data frame. All the figures seem reasonable, and the data is valid.

The following graphic illustrates the current state of the insurance industry's digital transformation, as well as the introduction of analytics, elements of the high-performing digital insurance industry.

**Tableau**: The data set for this project is visualized and analyzed using Tableau desktop software. It is the most widely used data visualization method for analyzing complex data in a short amount of time (less than a minute). Basically, Tableau is a business intelligence application that allows users to visualize and analyze data. With the Tableau platform, data from software like SQL, Big Data, Hadoop and cloud-based MS-Excel can be analyzed and visualized in the form of maps, graphs and reports. A filter choice is available for filtering features based on user requirements. Color is used to separate or visualize data so that it is easily understood at a glance. Tableau's advantages include fast visualization, ease of use, and high performance. Each graph can be given a title. Dashboards are used to analyze multiple sheets at once. Graphs can be created using charts such as scatterplots, pie diagrams, bar graphs, maps, and line graphs.

Visualization graphs used in this project include pie chart, bar graph, scatter plots, box plot, heat maps, circle view, packed bubbles

**Dataset source-link:** <https://www.kaggle.com/teertha/ushealthinsurancedataset>

**Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| **VARIABLES** | **DATATYPES** | **DESCRIPTION** |
| Age | Number | Age of primary beneficiary |
| Sex | String | Insurance contractor gender (female / male) |
| BMI | Number | Body mass index provides a measure of body weight, whether weight is relatively high or low relative to height, an objective index of body weight (kg/m^2) using a height-to-weight ratio, ideally 18.5 to 24.9 |
| Children | Number | Number of children covered by health insurance / Number of dependents of each primary beneficiary. |
| Smoker | Yes/No | Smoker / Non – smoker |
| Region | String | The beneficiary's residential area in the US (northeast, southeast, southwest, northwest). |
| Charges | Number | Individual medical costs billed by health insurance. |
| Beneficiary\_ID | Number | Unique identification number assigned to each beneficiary. |

**Methods**: The dataset is secured by downloading it from Kaggle.com website

link: <https://www.kaggle.com/teertha/ushealthinsurancedataset>

It is saved as a .csv file which is to be viewed in Tableau Desktop. The visualization is made by following statements

Following Visualizations are done by following:

1. **Analysis-**

* Analysis of Region.
* Analysis of Age
* Analysis of BMI
* Analysis of Smoking Habits
* Analysis of Children
* Analysis of Gender

1. **HYPOTHESES/RESULTS:**

* Are the charges dependent on age in both genders (male and female)?
* Does the charge for smokers vary much from non-smokers? If yes does the charge depend on the age of the smoker?
* Is smoking habit related to region? If yes, does the average charge vary significantly in each region?
* Does the charges distribution for no. of children vary by region?
* Does gender have any effect on BMI? If not, how does BMI for Female change based on numbers? Children?
* Does the charge vary based on BMI for both Genders? Does the charge raise with increase in BMI?

**Analysis of Region:**

Here, the four regions are equally represented in each dataset.

Chart, pie chart

Description automatically generated

**Analysis of Age:**

Here, I have details about age. We have health insurance Vs median age. Age start is 18 ages to 64 ages.

Bar chart

Description automatically generated with low confidence

**Analysis of BMI:**

Here, the highest BMI male is 53.13 and female is 48.07.

**Chart, bar chart

Description automatically generated**

Here, both Gender has the same BMI. That is, gender has no effect on BMI. The median body mass index is 30 The BMI distribution of the insured is approximately normal, with a mean of 30.66 and a median of 30.4. There are a total of nine outliers in the BMI distribution, all of which are on the high side. The highest BMI ever recorded is 53.13.

**Chart, box and whisker chart

Description automatically generated**

**Analysis of Smoking Habits:**

Smokers are overrepresented in the data. Smoking rates are higher among men than among women. Of the 1338 insured, 274 (20.5 per cent) were smokers, the rest non-smokers. Among the 274 smokers, the proportion of males was higher (159) than females (115).

Chart

Description automatically generated

**Analysis of Children:**

**Chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Analysis of Gender:** The dataset is almost evenly split between males and females, with 676 males (50.5 percent) and 662 females (49.5 percent).

**Graphical user interface, chart, pie chart

Description automatically generated**

**HYPOTHESES/RESULTS:**

As exploratory data analysis helps in summarizing the data, it helped me to answer the following questions or hypotheses.

1. **Are the charges dependent on age in both genders (male and female)?**

**Graphical user interface, chart, scatter chart

Description automatically generated**

Here, I used scatterplots to illustrate and evaluate the above hypotheses. A scatter plot usually shows relationships between variables, with variable values ​​represented by points. The value of each data point is determined by placing the point on the vertical and horizontal axes. Automatic select I have Gantt Bar.

The scatterplot above is created using Tableau. Colors are used to distinguish males and females. The X-axis represents the age groups of people. Age groups are created using age bins in the table. Each bin represents a 5-year period. Y-axis represents charge. From the above scatterplot it is clear that older people are charged more than younger people in both genders. However, higher males have higher charges. So, from the above graph it can be concluded that the charges are age dependent, with increasing age the charges increase.

**b**) **Does the charge for smokers vary much from non-smokers? If yes does the charge depend on the age of the smoker?**

**Chart, scatter chart

Description automatically generated**

Here, the above hypotheses were visualized and evaluated using scatter plots. A scatter plot illustrates the relationship between variables by using points to represent variable values. The location of the points on the vertical and horizontal axes determines the importance of each data point. Automatic select I have Shape chart.

Tableau was used to create the scatterplot above. Smokers and non-smokers are separated by color. Age groups of people are represented on the X-axis. Tableau's age bins are used to create age groups. Each bin represents a five-year period. Charges are represented on the Y-axis. Smokers, regardless of age, pay higher charges than non-smokers, as seen in the scatterplot above. On the other hand, older age smokers claimed more money than younger age smokers. Consequently, it can be assumed that charges for smokers are significantly different from non-smokers and that charges for smokers are independent of age.

**c) Is smoking habit related to region? If yes, does the average charge vary significantly in each region?**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

Here, the above graphs were created using the visualization tool 'Tableau' to illustrate and evaluate the above hypotheses. The bar graph was chosen for this visualization because it is commonly used to compare things between different groups or to track changes over time. From the bar graph above, we can see that smokers are evenly distributed across the four regions. That means the smoking habits of people in different regions are similar. Also, we see that the average charge is similar for each region. Thus, we can conclude that the region has the least correlation with the charge and smoking habit.

**d) Does the charges distribution for no. of children vary by region?**

**Chart, bar chart

Description automatically generated**

Here, to illustrate and test the above hypotheses, the above graph was created using the visualization tool Tableau. To compare the children's charges in the four regions of data collection, the above bar graph was created using the analysis tool 'Tableau'. I chose a bar graph for analysis because it is commonly used to compare items between different categories. Colors are used to distinguish between numbers of children. From the above graph the charges for children in each region are very similar. This means that charges for children are not affected by the region of the country.

**e) Does gender have any effect on BMI? If not, how does BMI for Female change based on numbers? Children?**

**Chart, bar chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

Here, I explain these two pictures. To illustrate and test the above hypotheses, the above graph was created using the visualization tool Tableau. To compare women's BMI based on numbers. Children. I chose a bar graph for analysis because it is commonly used to compare items between different categories. Colors are used to distinguish between numbers of children. From the graph, it can be observed that the BMI of both sexes is the same. This means that gender has no effect on BMI. The average BMI is 30. Also, the BMI distribution in women is the same as for all children. So, it can be concluded that gender or not/. Children do not affect women's BMI.

**f) Does the charge vary based on BMI for both Genders? Does the charge raise with increase in BMI?**

**Chart, scatter chart

Description automatically generated**

Here, Scatterplots were used to illustrate and test the above hypotheses made with the tableau. Color differentiates males and females. On the X-axis, the median BMI is represented. On the Y-axis, the charges are reflected. Higher charge values ​​are associated with higher BMI, as seen in the graph above. Heteroskedasticity is observed in the data. Additionally, based on BMI, charges are almost identical for males and females.

**Results:** In this health insurance dashboard, we have listed the highest BMI for men and women. They have four regions, northeast, northwest, southeast, and southwest, where I found women, men, and children of various ages. Smokers are overrepresented in the data. Smoking rates are higher among men than among women. Of the 1338 insured, 274 (20.5 per cent) were smokers, the rest non-smokers. Among the 274 smokers, the proportion of males was higher (159) than females (115). The minimum charge is 1121, and the maximum charge is 63770. The charge distribution has 139 external values, all on the higher side. 1338 data points. Maximum amount paid is 63770.

**Discussions:** Explaining the above visualization helps insurance providers to create very useful healthcare plans or benefit designs. Age, sex, BMI, child, smoker and nonsmoker, region, charge, and beneficiary-ID, and average beneficiary in the dataset was 39 years. In addition, various assumptions help us visualize clear graphs of each factor. Many people are counted as non-smokers in the data. In fact, 85 percent of those covered have no more than three children. Body mass index of the region's population was not affected by gender. In terms of charges, the charges have steadily increased as the beneficiary ages. On the other hand, charges for smokers are not affected by their age. Based on BMI, the charges are the same for both sexes. There is no correlation between the number of children a woman has and her BMI. Charges are affected by the person's age and BMI. On the other hand, gender, region, and children are associated with lower charges.

**Future work:** In the new study, I need to work on some more datasets that I have created some visualizations and create additional charts. There is a lot of data available on health insurance helpful for future work. On the other hand, information is less available. Thus, more data set development and visualization are required. This work can be extended to uncover more common health problems for which a study of health-care rates is conducted, as well as how much they increase over time and demand by analyzing data from other states and/or countries. This information will be used to help develop and implement workplace wellness policies. The results of the study will be used to identify areas that could benefit from intervention. An examination of actual healthcare costs will alert company officials to the consequences of various diseases and conditions that add significantly to total healthcare costs. thought of working with Python it might have some extra features to grow the project to a certain extent. Finally, the current dataset can be used and manipulated in Python.

**References:**

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical#:~:text=U.S.%20health%20care%20spending%20grew,For%20additional%20information%2C%20see%20below>.

<https://www.healthaffairs.org/doi/abs/10.1377/hlthaff.2022.00113?journalCode=hlthaff>

<https://www.investopedia.com/insurance/why-do-healthcare-costs-keep-rising/>

<https://policyadvice.net/insurance/insights/how-many-uninsured-americans/>

<https://www.tableau.com/about/blog/2017/11/data-visualizations-help-consumers-get-better-deal-health-insurance-2018s-open>