Healthcare dataset

Business Scenario

Problem statement:

A significant public health concern is the rising cost of healthcare. Therefore, it's crucial to be able to predict future costs and gain a solid understanding of their causes. The insurance industry must also take this analysis seriously. Healthcare insurance providers may use this analysis to make various strategic and tactical decisions.

Objective:

This project aims to predict patients' healthcare costs and identify factors contributing to this prediction. It will also be useful to learn the interdependencies of different factors and comprehend the significance of various tools at various stages of the healthcare cost prediction process.

SQL

Week 01

In explorative data analysis, a few steps needed to be followed here which consisted of combining multiple datasets into a simple dataset. Further, we needed to explore more scenarios and complete the below steps.

1. Collate the files so that all the information is in one place
2. Check for missing values in the dataset
3. Find the percentage of rows that have trivial value (for example, ?), and delete such rows if they do not contain significant information
4. Use the necessary transformation methods to deal with the nominal and ordinal categorical variables in the dataset
5. The dataset has State ID, which has around 16 states. All states are not represented in equal proportions in the data. Creating dummy variables for all regions may also result in too many insignificant predictors. Nevertheless, only R1011, R1012, and R1013 are worth investigating further. Design a suitable strategy to create dummy variables with these restraints.

6. The variable NumberOfMajorSurgeries also appears to have string values. Apply a suitable method to clean up this variable.
7. Age appears to be a significant factor in this analysis. Calculate the patients' ages based on their dates of birth.
8. The gender of the patient may be an important factor in determining the cost of hospitalization.
9. The salutations in a beneficiary's name can be used to determine their gender. Make a new field for the beneficiary's gender.
10. You should also visualize the distribution of costs using a histogram, box and whisker plot, and swarm plot.
11. State how the distribution is different across genders and tiers of hospitals
12. Create a radar chart to showcase the median hospitalization cost for each tier of hospitals
13. Create a frequency table and a stacked bar chart to visualize the count of people in the different tiers of cities and hospitals

Week 02

1. To gain a comprehensive understanding of the factors influencing hospitalization costs

   a. Merge the two tables by first identifying the columns in the data tables that will help you
   in merging
   b. In both tables, add a Primary Key constraint for these columns

2. Retrieve information about people who are diabetic and have heart problems with their average age, the average number of dependent children, average BMI, and the average hospitalization costs
3. Find the average hospitalization cost for each hospital tier and each city level.
4. Determine the number of people who have had major surgery with a history of cancer
5. Determine the number of tier-1 hospitals in each state

Python

Week 01 & Week 02

All the above steps were implemented in Python as well.

Tableau

Dashboards created with several parameters trying to make sense of the dataset and visualize it, like the below charts:

1. Cancer history of smokers
2. BMI vs. charges
3. Hospitalizations per months
4. Monthly charges
5. Number of children vs Hospital tier
6. Number of surgeries vs charges
7. Gender wise smokers
8. Surgery charges based on cancer history

Combined these charts into 3 dashboards,