An Exploration of Healthcare Data

Does Income, Race and/or Age Affect Health Insurance?

Kileigh Boggess

A picture containing person, indoor

Description automatically generated

# Overview

* The esteemed health insurance industry is no stranger to using data to drive decision making. Even so, in today’s swiftly transformation into the digital age, insurance is still struggling to adapt and benefit from these new technological advances compared to other industries.
* On the contrary, these developing technologies such as AI and Block Chain, have brought such a drastic change to the health insurance world, and Data Analytics sits at the fore front of this transformation.
* We found this dataset to be helpful in exploring and understand how various attributes of those insured can affect their insurance premium. This dataset contained 1338 rows of data that included these as attributes: Age, Sex, Number of Children, BMI, Region, and Smoker.
* Just looking at the data, we noticed that very few people had more than two children (75% had two or less children). We also saw that the amount of insurance claimed was highly skewed – most people required basic Medicare and only a handful suffered from diseases that would cost them more to get rid of.

## Exploratory Data Analysis

We started off by cleaning the data as well as organizing it. We did this by looking for unique values, converting the strings into numbers, and creating a dictionary that would display each datatype and the columns that contained those datatypes.

* We witnessed that the ‘age(int64), ‘bmi’(float64) and ‘charges’(float64) are numerical characteristics of the data. We also noticed that ‘sex’, ‘region’, and ‘children’ are categorical characteristics of the data. And even though the ‘children’ column is treated as an integer datatype, it also contained categorical information – we thought this was important to include here.
* We saw that there was no missing data from the dataframe – which as a data scientist, is always a nice treat!

## Correlation

From there, we began to look to see if there was any known correlation between the different attributes within this dataset.

Based on this heatmap, we were able to conclude that smoking is intensely correlated (in the positive direction) with that of premium charges, but also displayed a weak correlation (in the positive direction) between age and BMI with premium charges.

Graphical user interface, chart, treemap chart

Description automatically generated

Then we wanted to look at the different characteristics within the premium charges in our dataset.

Based on our observation of this heatmap, we concluded that again, the smoking habit was strongly correlated with premium charges (0.79) but was extremely weakly positively correlated with premium charges (0.068, 0.057), and region had the absolute weakest positive correlation with the premium charges (0.0036).

Graphical user interface, application

Description automatically generated

For fun, we thought it would be interesting to explore the relationship between BMI, Premium Charges, and Smoking Status. You can see that correlation below.

Chart, scatter chart

Description automatically generated