Project Proposal

Investigation of Insurance Charges

Adrien Sesco

## Introduction, Context, and Goals

The United States has a for-profit medical system which leads to high prices for treatment and insurance alike. This project aims to analyze a small amount of insurance data with a limited number of variables and observations in order to gain a basic understanding of what contributes most to insurance prices. The variables in this data are age, sex, BMI (body mass index, a health metric), number of children, smoker (true/false), region, and charge in USD. There are 1,338 observations for these columns, so anything more than a basic analysis would be impractical or impossible with this data. Any attempt to apply an analysis of this data to real-life scenarios would also be impractical or impossible. It will only work for a basic analysis as conducted in this project.

## Methodology

This project aims to analyze the data through correlation values and visualizations. These methods were chosen because they allow for easy communication of results and are not too complicated. Visualizations were also chosen because of the number of different categorical variables present in the data. For example, the data from different regions look largely the same with about the same correlation values. Visualizations make the differences between data in each region easier to spot and understand, which saves time and avoids confusion about the significance of these differences.

## Limitations

The only limitation of this data is its size. It is nowhere near as comprehensive as the data sets used by real insurance companies to determine charges. This leads to unexplainable trends in the data and means that any prediction would be unsuitable for real world applications. Such limited data cannot be used for anything other than a basic analysis like this one.

However, the methodology used in this analysis would be usable for real-world applications if extensive data was available. It is because of this that the only limitation of this project is the data set used with it.

## Conclusion and Findings

In conclusion, the project found that BMI and charge are positively correlated. This is due to the fact that BMI decreases life expectancy and increases the likelihood that the insurance company will have to pay out. This means that they must charge more or risk losing profit to provide care to their customers. Likewise, smoking also had a big impact on the charge as

smoking also decreases life expectancy. Sex contributed slightly and varied differently in each region. Region also showed slightly different trends, but the data is not able to explain why as it is limited compared to what is used to calculate the charge. None of the other variables contributed to the charge. There are clearly other variables that would contribute to the charge in real-life applications, but they are not included in this data set which is its main limitation.

## References

https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset/code