# Medical Insurance Cost Analysis

Mia Saavedra, Tanya Arya, Amolika Kondapalli, Jisoo Park

School of Information, The University of Texas at Austin

I 310D: Introduction to Human-Centered Data Science

Professor Abhijit Mishra

May 6, 2024

## Introduction

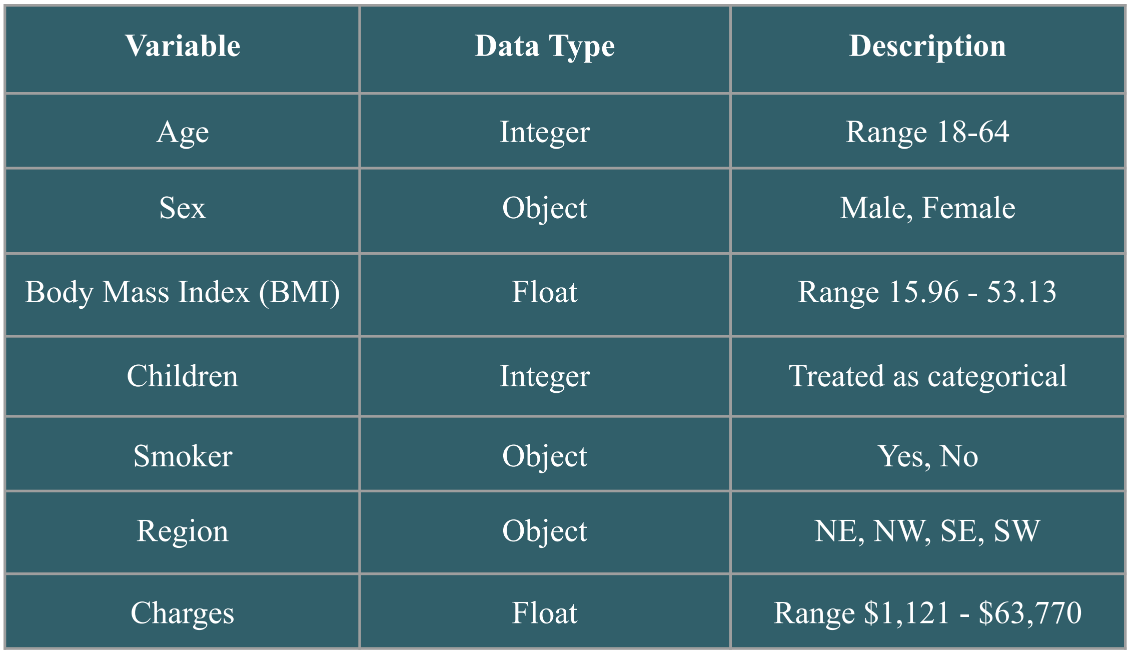
According to The Commonwealth Fund, a large population of insured working-age adults says it is very or somewhat difficult to afford health care in America (2023). Given the challenges many adults face in affording healthcare, understanding the factors influencing medical costs becomes crucial. By examining the variables impacting medical costs, including age, gender, BMI, smoking habits, number of children, and geographic region, we will determine which of these key factors has the most significant correlation with medical charges and if there is a difference between demographic groups.

Understanding the primary factor driving medical expenses can provide valuable insights for healthcare policy-makers, insurers, policymakers, researchers, and the general public, so resources can be more effectively distributed to address the most significant cost drivers in healthcare. First, healthcare providers can tailor services to better meet the needs of different patient demographics. Second, insurance companies can develop more accurate pricing models and risk assessments. Third, policymakers can better make decisions aimed at addressing disparities in healthcare access. Fourth, the findings of this project could contribute to researchers of healthcare economics and public health. Finally, increased transparency around the factors driving medical expenses can empower the general public to make more informed decisions about their insurance coverage.

A brief literature review was conducted to form a hypothesis. First, a report conducted by the National Bureau of Economic Research concludes that “medical spending more than doubles between ages 70 and 90” (Medical Spending of the Elderly, 2024). Second, the World Economic Forum found the “average working woman in the US spends 18% more on healthcare costs than a man” (Edmond, 2023). Third, the JME insurance agency reports that “families with children, especially families with multiple teenage children, the new rating factors have caused premiums to skyrocket" (Why Are Health Insurance Premiums for Children so High? – JME Insurance Agency, 2018). Fourth, a study published by BMJ journals found that "former smokers had the highest annual medical expenditures" (Swedler et al., 2019). Finally, a study conducted by the National Library of Medicine found that high medical charges are "especially high for people with severe obesity" (Ward et al., 2021). Therefore, the initial hypothesis suggests that increased insurance costs could be associated with older age, female gender, higher BMI (Body Mass Index), and having more children or being a smoker.

## Data Description

The dataset used for this project was "Medical Insurance Cost Prediction" from Kaggle.com (https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction). This data set is released under MIT License which grants the rights to copy, modify, publish, and distribute the Software. The copyright and permission notice was included in the GitHub and Jupyter Notebook as requested. The dataset comprises 2,700 rows and 7 columns for age, sex, body mass index (BMI), number of children, smoker status, region of the United States, and insurance charge value. The data type and description of variables are included below in Table 1. Before processing the data, pandas, numpy, script, researchpy, and matplotlib were imported to the Jupyter Notebook. The data was downloaded as a CSV file from Kaggle, uploaded to the notebook, and converted into a data frame using Python. The pre-processing included removing 1,435 duplicates and removing null attributes. 1337 non-null values were used for our data analysis. No transformation was performed on the values.



| Variable | Data Type | Description |
|---|---|---|
| Age | Integer | Range 18-64 |
| Sex | Object | Male, Female |
| Body Mass Index (BMI) | Float | Range 15.96 - 53.13 |
| Children | Integer | Treated as categorical |
| Smoker | Object | Yes, No |
| Region | Object | NE, NW, SE, SW |
| Charges | Float | Range $1,121 - $63,770 |

**Table 1:** Variables for the dataset, their respective data type, and further description.

**Methodology**

The methods used to analyze the data include exploratory data analysis (EDA), statistical tests, and data visualization. For EDA, we produced histograms for numerical values (age, BMI, charges) along with descriptive statistics using the describe function in Python. For categorical variables, pie charts were produced to show the distribution of each group. The data visualization consisted of boxplots for the variables age, smoker status, and number of children against medical insurance charges. The statistical tests including correlation analysis, t-tests, ANOVA, and linear regression were completed using Python in the Jupyter Notebook. Correlation analyses were performed on the non-categorical values, age, and BMI, using pearsonr from the scipy.stats library. This method found the correlation coefficient, which allowed us to determine the strength and direction of the relationship between each variable and medical insurance charges. This value was squared to calculate R-squared which tells us how much of the variability in charges is attributed to each variable. However, the limitations of the correlation analysis were that it assumes a linear relationship and can't look at the effect of other variables other than the two being considered. A correlation matrix was made with Seaborn and Matplot libraries to create a simple visualization of the correlations between the numerical variables.

Two sample t-tests were conducted to determine if there was a significant difference between the means of the two groups. The variables used for this test were sex (male and female) and smoker (yes and no). The assumptions for this test include random samples, independent samples, and a normally distributed population or a large sample. Our sample is greater than 25, subjects can't be in both groups, and we can assume that data collection followed proper practices. The null hypotheses were that the mean charges of the male and female groups were the same ($\mu_{male} = \mu_{female}$) and that the mean charges of the smoker and non-smoker groups were the same ($\mu_{smoker} = \mu_{non\text{-}smoker}$). The alternate hypotheses were that the mean charges of the male and female groups differed ($\mu_{male} \neq \mu_{female}$) and that the mean charges of the smoker and non-smoker groups differed ($\mu_{smoker} \neq \mu_{non\text{-}smoker}$). For each test, we produced the t-statistic values and p-values. T-statistics measures the difference relative to the variation in the sample, but we only looked at the p-value to determine our conclusion. If the p-value was less than 0.05, there was a significant difference in the means of the groups meaning that we reject the null hypothesis. A limitation of the t-test is that it is sensitive to outliers which we did not remove from the data.

ANOVA  t-tests were conducted to determine if there was a significant difference between the means of any two groups. The variables used for this test were region (northeast, northwest, southeast, and southwest) and number of children (0, 1, 2, 3, 4, 5). The number of children variable was treated as a categorical variable because there is limited variety in values.

The assumptions for the ANOVA test include random samples, independent samples, similar variance, and a normally distributed population or a large sample. Our sample is greater than 25, subjects can't be in multiple groups, variances are relatively similar between groups, and we can assume that data collection followed proper practices. The null hypotheses were that the mean charges of the northeast, northwest, southeast, and southwest groups were the same ($\mu_{NE} = \mu_{NW} = \mu_{SE} = \mu_{SW}$) and that the mean charges between the number of children groups were the same ($\mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$). The alternate hypotheses were that the mean charges of the northeast, northwest, southeast, and southwest groups of at least two groups differed ($\mu_{NE} \neq \mu_{NW} \neq \mu_{SE} \neq \mu_{SW}$) and that the mean charges of at least two groups based on the number of children differed ($\mu_0 \neq \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$). For each test, we produced the f-statistic values and p-values. Similar to the t-statistic, the F-statistic measures the difference relative to the sample's variation, but we only looked at the p-value to determine our conclusion. If the p-value was less than 0.05, there was a significant difference in the means of at least two groups, meaning we reject the null hypothesis. However, post-hoc tests were not conducted, therefore, we don't know exactly which two groups had the most significant difference.

The linear regressions completed used the libraries Sklearn and Matplot. The variables used were age and BMI. This differs from correlation in which linear regression allows us to consider causality. The assumptions for linear regression models include random samples, a linear relationship, no influential outliers, and consistent spread. Our data may not be well suited for a linear regression model because there are outliers and weak linear relationships between variables. The linear regressions produce a coefficient for the slope of the line of best fit and its y-intercept. Theoretically, this equation can be used to predict charges based on the value of the variable in question. Our code produces the mean squared error which is the averaged squared difference between the predicted and actual values. A large mean squared error signifies that the values were widely dispersed and poorly predicted signifying a poor model. The opposite is true for a small mean square error. This model produces visualizations including the data points and line of best fit.

## Results

Before conducting statistical tests, we conducted EDA on the variables to better summarize the main features of the dataset. As seen in Figure 1, the age distribution is right skewed meaning our dataset included a generally younger population. Table 2 provides the descriptive statistics of the age column showing that the mean and median are about 39 years old. Figure 2 shows the normal distribution of BMI meaning that it was distributed on a bell curve. Table 3 provides further information about the BMI column and the mean and median are about 30. The distribution of charges is seen to be rightly skewed in Figure 3 with a higher frequency of charges under 10,000 and the mean charge being $13,279 as seen in Table 4. Table 4 also shows that the insurance charges had a large range of values ranging from the max charge being $63,770.43 and the minimum charge being $1,121.87. For the factors of smoking, sex, and region, we created pie charts to view their distributions. Both the sex distribution and region distribution are evenly split as seen in Figure 4 and Figure 5 respectively. The smoker distribution, however, in Figure 6 indicates that the dataset includes a considerable amount of non-smokers than smokers (around 60% more) which is important to note when testing the data further.

The correlation coefficient between age and charges was 0.298 which indicates a moderately weak positive relationship between the variables. The R-square value was 0.888

meaning that 8.88% of the variability in charges is attributed to age. The correlation coefficient between BMI and charges was 0.198 indicating a weak positive relationship between the two variables. The R-squared value was 0.0392 which means that 3.92% of the variability in charges is attributed to BMI. The R-squared values for both of the variables are relatively low indicating that there are other factors involved in the variability of charges. The correlation matrix shown in Figure 7 shows the relationship between age, body mass index, and medical charges. A strong positive correlation would be close to 1.0, but the matrix results show that the relationships between these two variables are weak.

The t-test for sex resulted in a t-statistic of 2.124 and a p-value of 0.0338. Since the p-value was less than 0.05, we rejected the null hypothesis and concluded that there was a significant difference between the mean charges of the male and female groups. The boxplot for the two groups, Figure 8, shows that the two values have a similar median. However, the upper 50% quartiles of the male group had overall higher medical charges compared to the female group. The t-test for smokers resulted in a t-statistic of 46.645 and a p-value of 1.407e-282. Since the p-value was less than 0.05, we rejected the null hypothesis and concluded that there was a significant difference between the mean charges of the smoker and non-smoker groups. This is also reflected in Figure 9 as the overall median for the smoker group was higher in medical charges compared to the non-smoker group.

An ANOVA test between the region groups resulted in an F-statistic of 2.926 and a p-value of 0.03276. The ANOVA test between the groups for the number of children resulted in an F-statistic of 3.268 and a p-value of 0.0061. Both ANOVA tests resulted in p-values less than 0.05 meaning that we rejected the null hypotheses. This means that there is a significant difference in the mean charges between at least two region groups and a significant difference between at least two groups for the number of children. Based on Figure 10, the median charges of region groups are relatively similar, but the southeast group had overall higher values in the upper 50% quartiles. This could have increased the mean charges resulting in a difference between another region group. Figure 11 shows that the median charges between groups based on number of children are similar. The group with two children showed slightly higher charges compared to the other groups. However, since post-hoc tests were not conducted, we do not know exactly which groups have significantly different means.

We conducted two linear regression analyses for the non-categorical factors, Age and BMI. The age linear regression shown in Figure 12, resulted in a coefficient of 242.26, an intercept of 3532.09,  a mean squared error of 166,275,348.23, and an $R^2$ of 0.095. The predictive equation for charges was Charges = 242.26(Age) + 3532.09. The line of best fit does not suit the data well because the high values caused it to stray away from the linear trend below it. For the BMI linear regression shown in Figure 13, the result was a coefficient of 345.17, an intercept of 2488.57, a mean squared error of 174,251,720.52, and an $R^2$ of 0.0517. The predictive equation for charges was Charges = 345.17(BMI) + 2488.57. The linear regression visualization shows that the values have inconsistent spread around the line of best fit and don’t follow a linear trend. Both linear regressions produced a large mean squared error meaning that the predicted and actual values were widely dispersed and poorly predicted signifying a poor predictive model. Additionally, both models fail to meet the assumptions of a linear regression model.

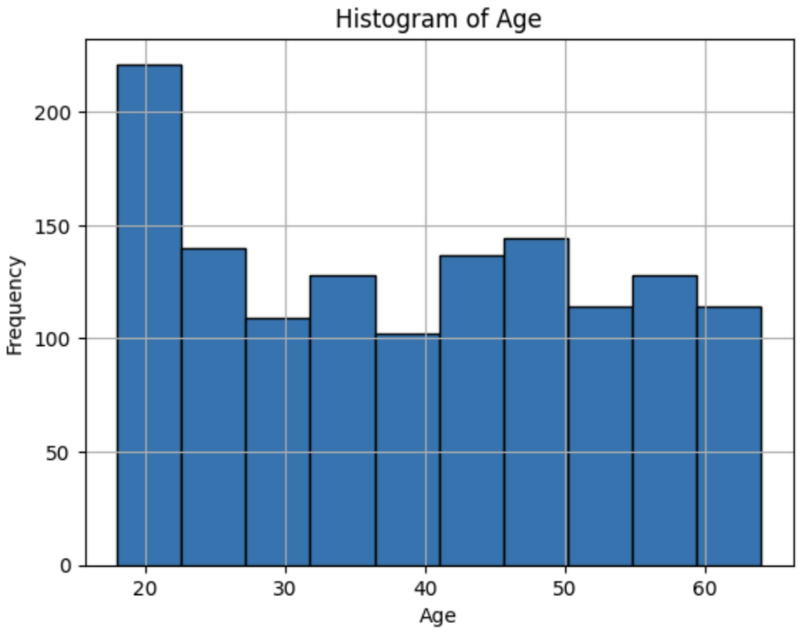


**Figure 1:** This distribution of age is slightly right-skewed.

| Descriptive Statistics: Age | |
|---|---|
| Count | 1337.00 |
| Mean | 39.22 |
| Standard Deviation | 14.04 |
| Min | 18.00 |
| Median | 39.00 |
| Max | 64.00 |

**Table 2:** Descriptive statistics for age column.

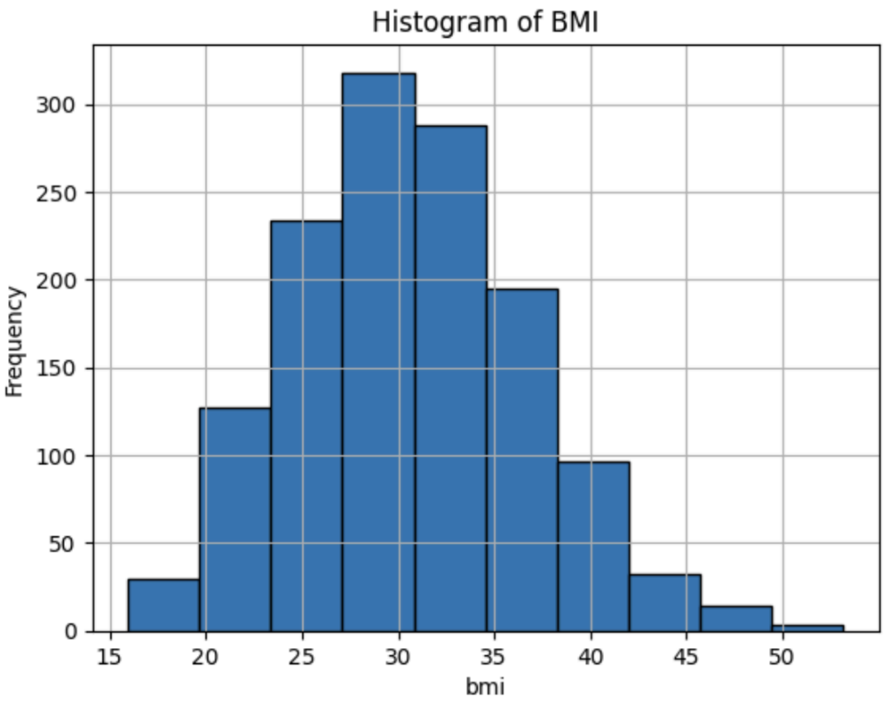


**Figure 2:** The distribution of BMI is normally distributed.

| Descriptive Statistics: BMI | |
|---|---|
| Count | 1337.00 |
| Mean | 30.66 |
| Standard Deviation | 6.10 |
| Min | 15.96 |
| Median | 30.40 |
| Max | 53.13 |

**Table 3:** Descriptive statistics for BMI column.

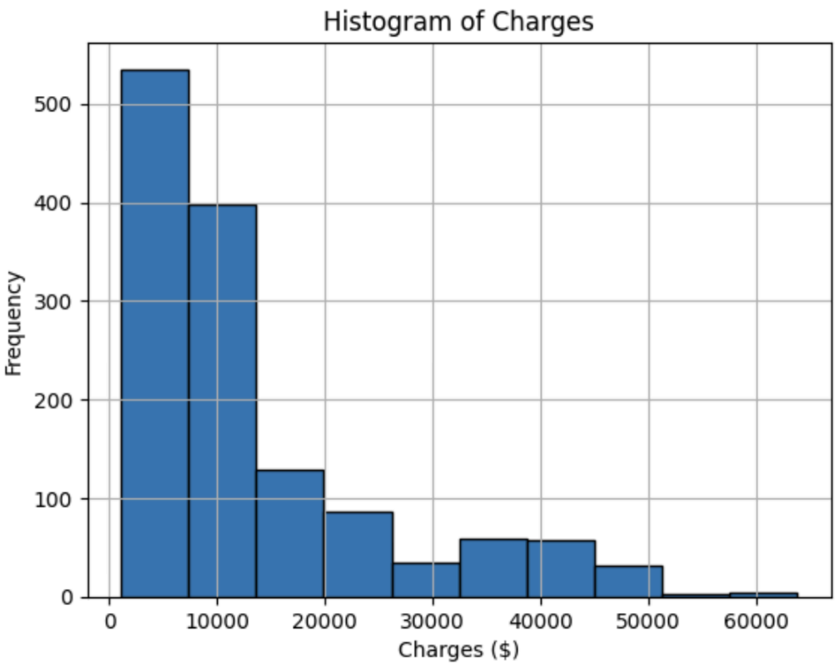


**Figure 3:** The distribution of charges is right-skewed.

| Descriptive Statistics: Charges | |
|---|---|
| Count | 1337.00 |
| Mean | 13,279.12 |
| Standard Deviation | 12,110.36 |
| Min | 1,121.87 |
| Median | 9,386.16 |
| Max | 63,770.43 |

**Table 4:** Descriptive statistics for charges column.

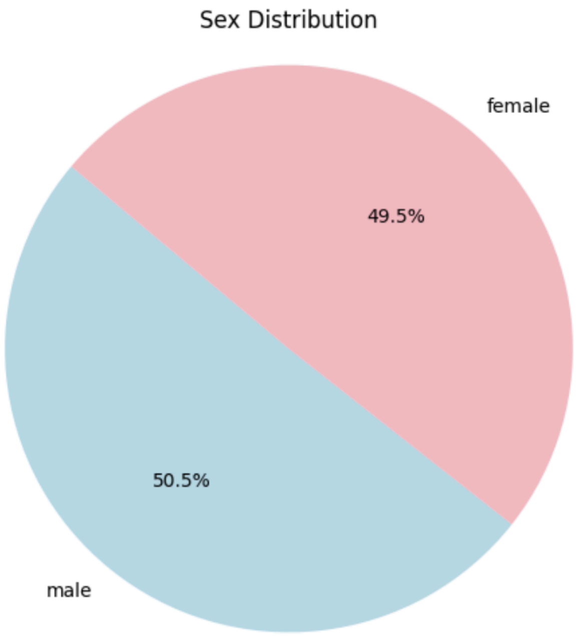


**Figure 4:** Sex Distribution: There is an almost even split between the male and female groups.

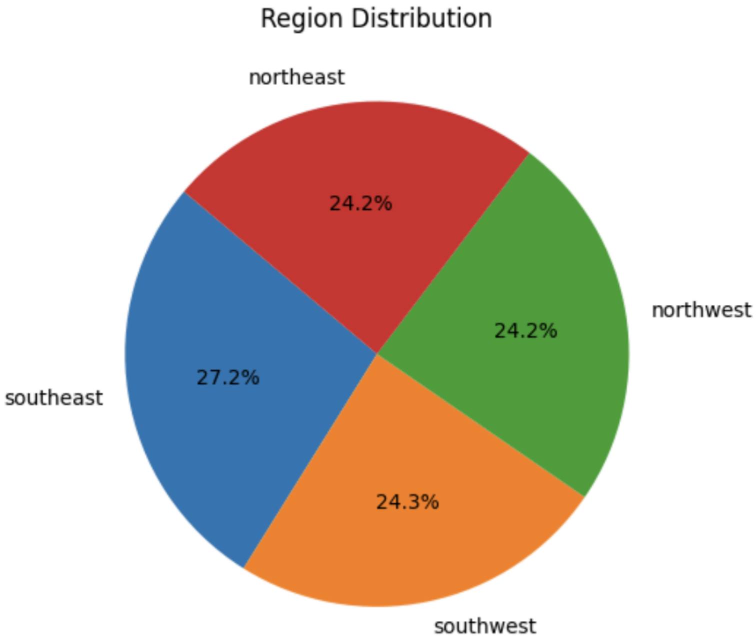


**Figure 5:** Region Distribution: The 4 regions are almost evenly distributed.

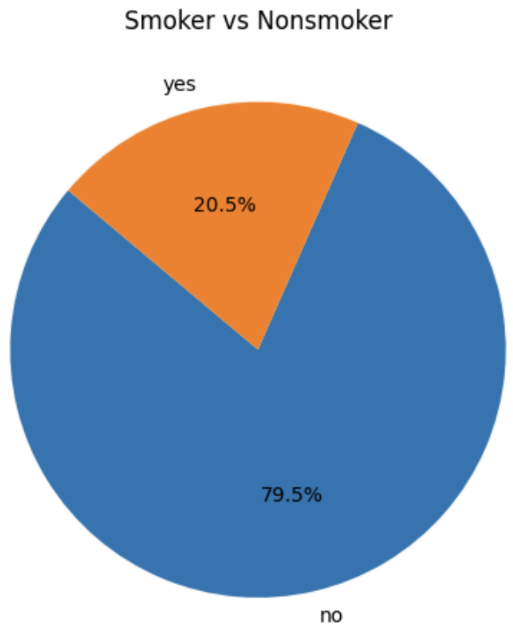


**Figure 6:** Smoker Distribution: The dataset includes more non-smokers than smokers.

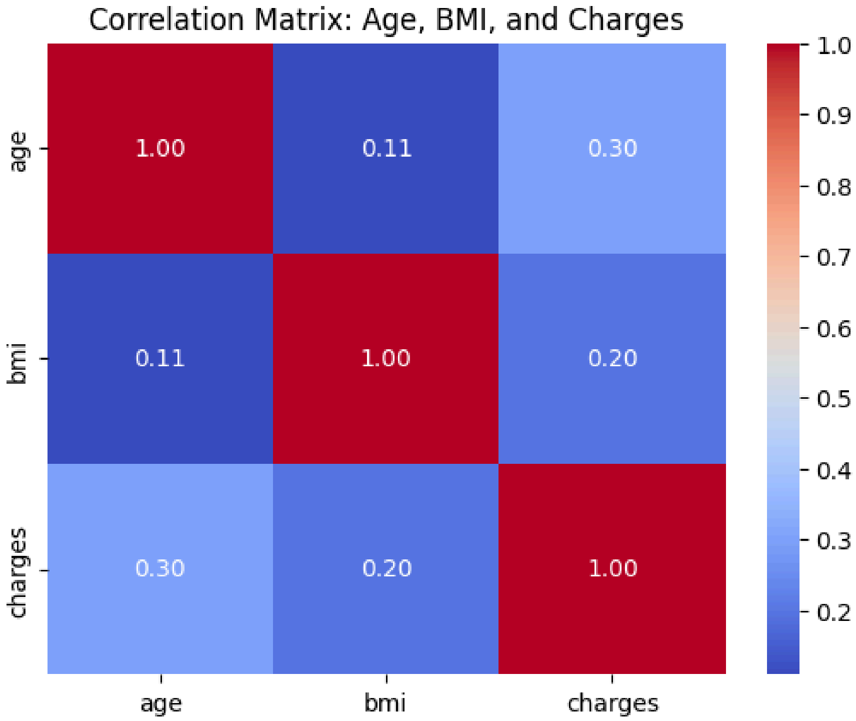


**Figure 7:** Correlation matrix between age, BMI, and charges.

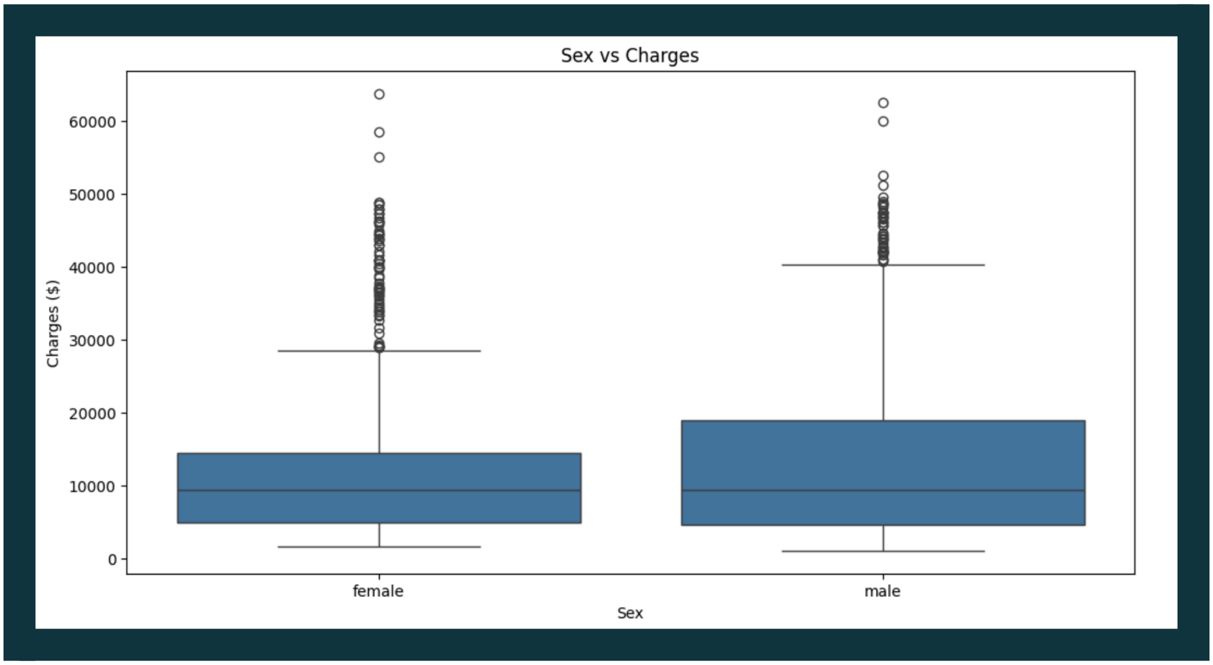


**Figure 8:** While the median is around the same for both female and male groups, the upper 50% of the male group had higher charges than the female group.

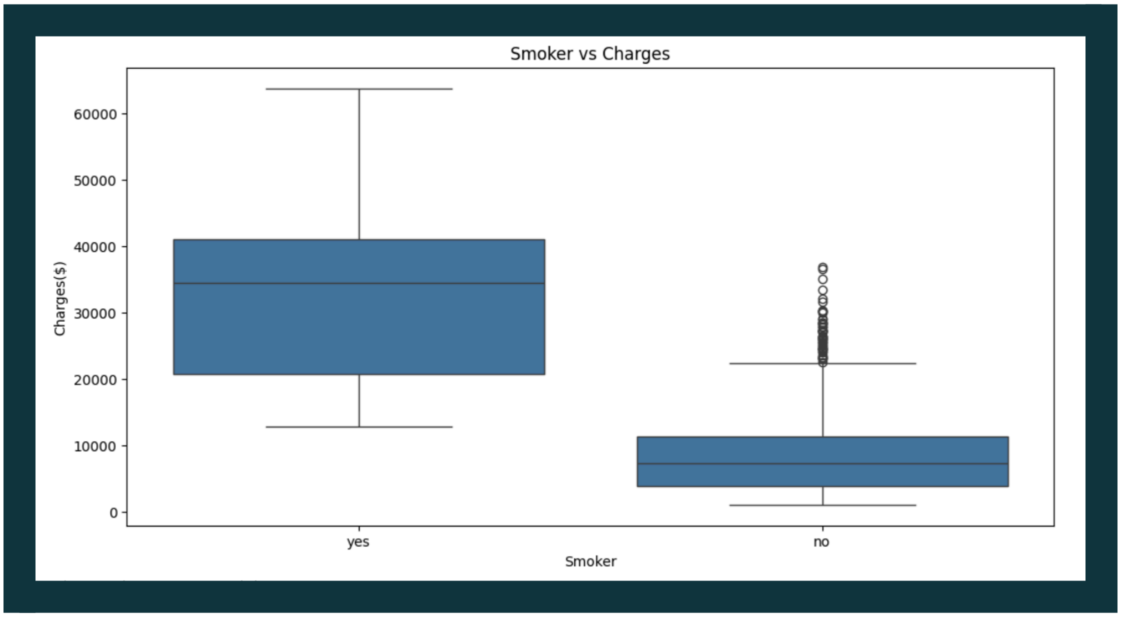


**Figure 9:** The smoker group had evidently higher medical insurance charges compared to non-smokers.

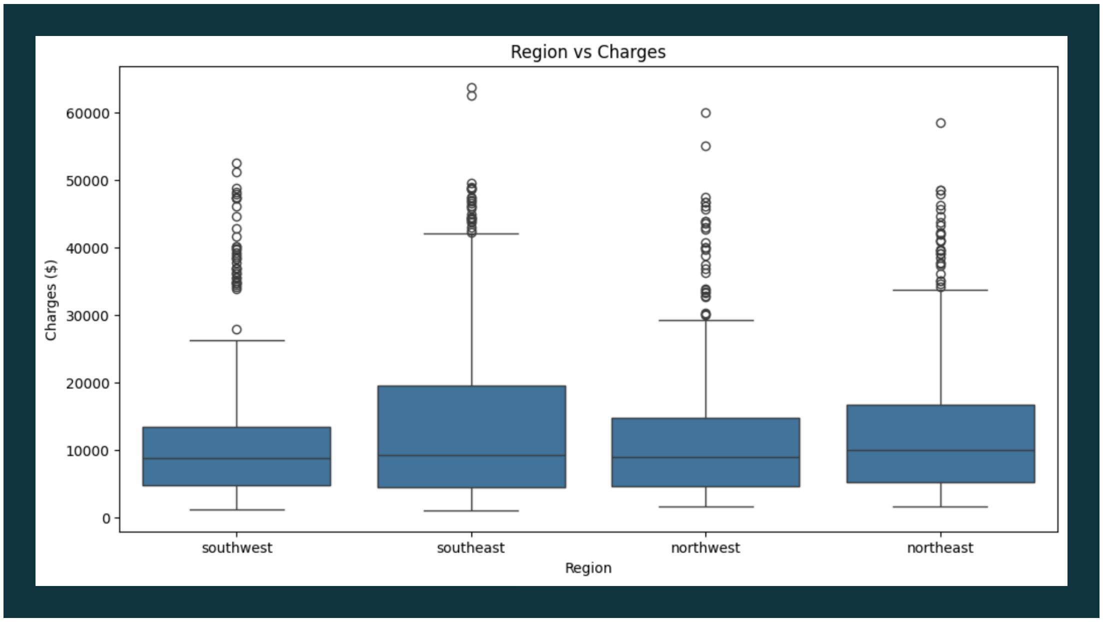


**Figure 10:** The region groups had similar medians, but the southeast group generally had higher charges.

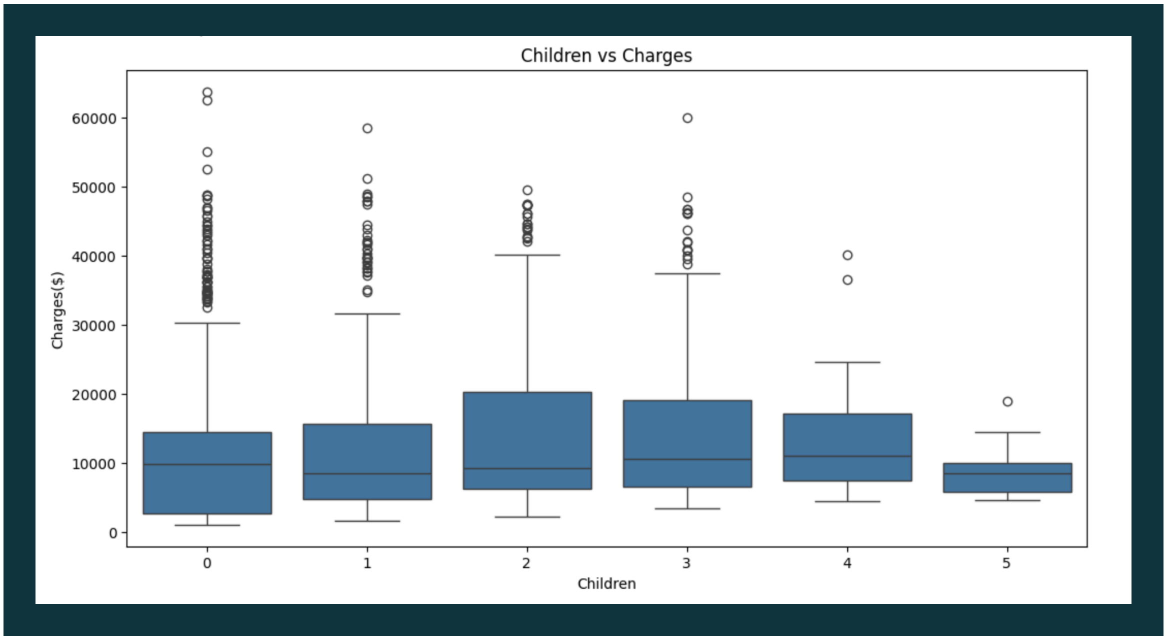


**Figure 11:** There is a slight difference between the number of children and medical charges. The median is relatively the same for all the groups.

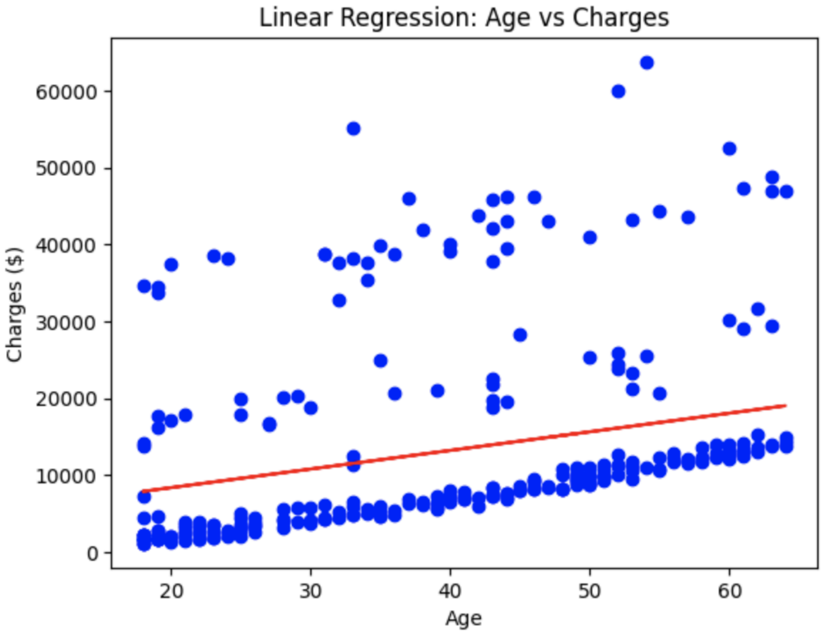
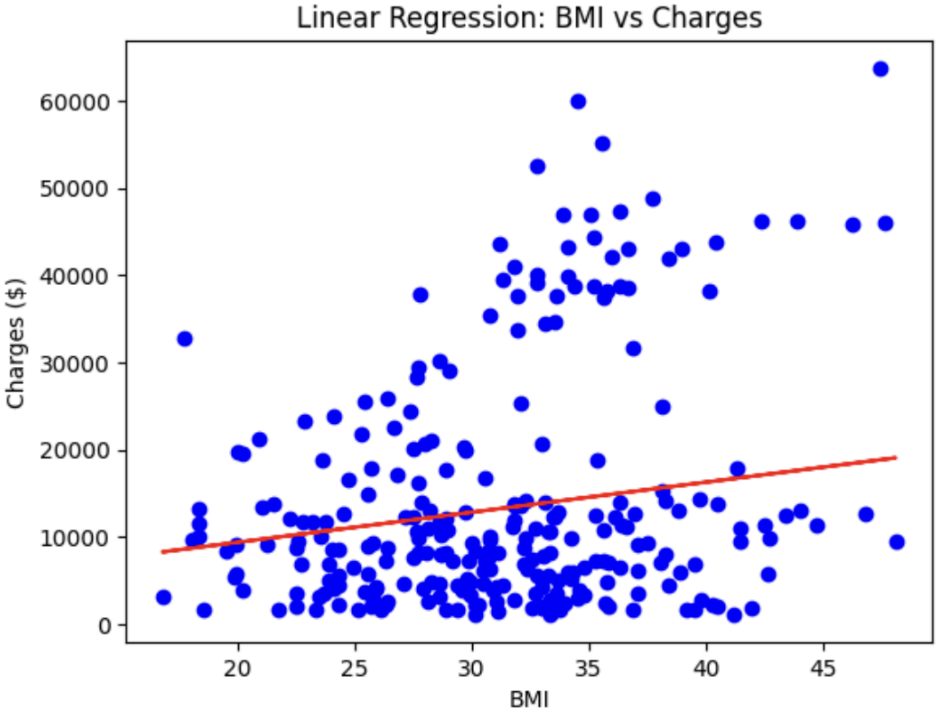


**Figure 12:** Linear regression model for age and charge



s.

**Figure 13:** Linear regression for BMI and charges.

## Conclusion

### Key Findings

When considering correlation, the most significant data resulted in age and body mass index having the strongest correlation compared to the rest of the factors. These factors still were seen to have a moderately weak correlation to the charges, but we can conclude that they had the greatest effect in comparison to the other factors. Additionally, both the male group and smoker group had overall higher medical charges as seen in the box plot figures,  which we initially hypothesized. Based on the ANOVA tests, we can reject the null hypothesis for both region and number of children stating that there is a significant difference in charges among different regions and based on the number of children. However, with these results, we still acknowledge that there were low correlations between our variables and some of our statistical tests failed the assumptions. Therefore, our conclusions should be acknowledged with caution.

### Implications

There are five main stakeholders involved in the conclusions of this research project. First, healthcare providers can use this data to better tailor services to meet the needs of different patient demographics. Second, insurance companies can use this model to develop more accurate risk assessments. Third, policymakers need medical insurance charge data to make informed decisions to address disparities in healthcare access. Fourth, researchers can use this data to contribute to healthcare economics and public health. Finally, the public can benefit from the findings as transparency about medical charge patterns to make better lifestyle and insurance choices.

## Limitations

One limitation of this project is the time frame. If given more time we could do more analysis on demographic factors or perform further statistical analysis on the trends. We could have also had the ability to add more factors to test against insurance charges to have a more comprehensive report. It would also be helpful to find correlations between the variables themselves. Another limitation is that the data is centered around American citizens, so adding international datasets will make it more comprehensive. We were unable to find information as to how and when this data was collected, therefore, the analysis was completed on the assumption that the data was representative and properly collected.

## Future Work

For a more comprehensive analysis, it would be helpful to analyze more factors such as health conditions, disabilities, and diets. These additional factors play a large role in people's health, thus being relatable to their medical charges. It would also be beneficial to do more research into the type of medical charges people face and even different insurance companies. By considering the medical insurance charges of international subjects, we could identify how charges between Americans and non-Americans differ. Differences between subjects with similar demographic characteristics could suggest governmental and economic factors that influence the insurance industry.

## Appendix: (Q and A)

1. How can this research be more human-centered?

Ensuring that data is human-centered includes making sure that the data is ethically sourced and contains no biases. Therefore, sourcing more datasets that include holistic demographics from diverse populations can help reduce bias and become more applicable to larger groups.

## Code

Below is the link to our GitHub repository which includes the Jupyter Notebook, python code, presentation slides, and CSV file for the data set.

https://github.com/tanyaarya10/1310DFinal.git

## References

Comparing several means (one-way ANOVA) — Learning Statistics with Python. (2022). Github.io. https://ethanweed.github.io/pythonbook/05.03-anova.html

GeeksforGeeks. (2018, June 10). EDA Exploratory Data Analysis in Python. GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/

Matplotlib Pie Charts. (2024). W3schools.com. https://www.w3schools.com/python/matplotlib_pie_charts.asp

Medical spending of the elderly. NBER. (2024). https://www.nber.org/bah/2015no2/medical-spending-elderly#:~:text=Medical%20spending%20by%20the%20elderly,percent%20of%20all%20medical%20spending.

M Rahul Vyas. (2024). Medical Insurance Cost Prediction. Kaggle.com. https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction

Renesh Bedre. (2018, October 22). How to Perform ANOVA in Python. RS Blog. https://www.reneshbedre.com/blog/anova.html

Sara R. Collins, Shreya Roy, and Relebohile Masitha, Paying for It: How Health Care Costs and Medical Debt Are Making Americans Sicker and Poorer — Findings from the Commonwealth Fund 2023 Health Care Affordability Survey (Commonwealth Fund, Oct. 2023). https://doi.org/10.26099/bf08-3735

Swedler, D. I., Miller, T. R., Ali, B., Waeher, G., & Bernstein, S. L. (2019). National medical expenditures by smoking status in American adults: an application of Manning's two-stage model to nationally representative data. BMJ open, 9(7), e026592. https://doi.org/10.1136/bmjopen-2018-026592

T-test with Python. (2015). Pythonfordatascience.org. https://www.pythonfordatascience.org/independent-samples-t-test-python/

US women are paying billions more for healthcare than men every year. (2023, October 22). World Economic Forum. https://www.weforum.org/agenda/2023/10/healthcare-equality-united-states-gender-gap/#:~:text=Working%20women%20in%20the%20States,to%20new%20research%20by%20Deloitte.

Ward, Z. J., Bleich, S. N., Long, M. W., & Gortmaker, S. L. (2021). Association of body mass index with health care expenditures in the United States by age and sex. PloS One, 16(3), e0247307–e0247307. https://doi.org/10.1371/journal.pone.0247307

Why are health insurance premiums for children so high? – JME Insurance Agency. (2018, January 16). https://jmeinsurance.com/why-are-health-insurance-premiums-for-children-so-high/