Investigating Determinants of Variability in Insurance Costs Across the United States Healthcare Landscape

Dhanavikram Sekar, Hariharan Kumar, Indu Varshini Jayapal,
Naveen Vinayaga Murthy, Nidhi Choudhary

Abstract

This project investigates health insurance costs in the United States using an open-source GitHub dataset. Employing rigorous statistical methods, our study uncovers gender-based charge disparities, notable regional variations, and the influence of gender, smoking status, and geographic location on insurance costs. Utilizing univariate and multivariate analyses, along with hypothesis testing, our findings provide valuable insights into healthcare pricing disparities. This emphasizes the need for nuanced considerations in policy and practice, addressing the complexities revealed by our comprehensive analysis.

1 Introduction

As of 2022, the healthcare landscape in the United States is marked by a continuous increase in total national health expenditures, estimated to reach $4.5 trillion. This ongoing escalation underscores the enduring financial significance of the healthcare sector, carrying multifaceted implications for both individuals and the broader economy.

Insights from the Kaiser Family Foundation's 2022 Employer Health Benefits Survey shed light on updated health insurance cost trends. The average annual premium for employer-sponsored health insurance has witnessed an uptick, reaching $8,114 for single coverage and $22,221 for family coverage. These figures emphasize the persistent upward trajectory of healthcare expenses, highlighting the financial burdens faced by individuals and families. This further accentuates the need for a comprehensive understanding of insurance cost dynamics.

Despite continuous efforts to enhance healthcare affordability, individuals in the U.S. grapple with substantial out-of-pocket costs. The recent study from the Commonwealth Fund underscores this concern, revealing that, even with insurance coverage, deductibles and copayments significantly contribute to out-of-pocket expenses. These financial challenges can act as barriers to accessing necessary medical services, necessitating a closer examination of strategies to mitigate their impact on individuals' healthcare-seeking behavior.

A persistent challenge within the health insurance landscape is the variability in charges by insurance companies for similar services. This ongoing phenomenon complicates individuals' understanding of healthcare costs and reinforces the importance of advocating for transparency and fairness in pricing. As we navigate the complexities of rising healthcare expenditures, increasing insurance premiums, and the enduring issue of variability in charges, our research endeavors to uncover the factors contributing to the significant variation in insurance costs in the United States.

2 Data

Dataset Link

The dataset under consideration, derived from the specified GitHub repository, provides insights into insurance charges for individuals in the United States. The data consists of 1338 records and 7 variables, with no explicit time frame provided. Each record uniquely corresponds to an individual, and the dataset's structure incorporates both numerical and categorical variables, providing a comprehensive perspective on the factors influencing insurance charges.

The seven variables in the dataset are outlined as follows:

- Age: Age of individuals
- Sex: Gender of individuals
- BMI: Body Mass Index of individuals
- Children: Number of children or dependents covered by insurance
- Smoker: Whether an individual is a smoker or not
- Region: Geographical region of individuals in the United States
- Charges: Incurred insurance charges

The dataset includes both categorical and numerical features, with categorical features encompassing 'Sex,' 'Children,' 'Smoker,' and 'Region,' and numerical features comprising 'Age,' 'BMI,' and 'Charges.'

Below are the categorical features and the various categories:

- Sex: Male/Female
- Children: 0, 1, 2, 3, 4, 5
- Smoker: Yes/No
- Region: Northeast, Northwest, Southeast, Southwest

Below are the numerical features and their respective ranges:

- Age: 18 - 64
- BMI: 15.96 - 53.13
- Charges: $1,121 - $63,770

The 'Charges' column serves as our target variable, and any analysis conducted will be cross-tabulated with this variable to discern the factors influencing insurance charges. It is essential to acknowledge that bias might be present in the dataset, potentially influencing the outcomes of our analysis. For instance, if the dataset is predominantly composed of certain demographic groups or regions, it could introduce bias in our understanding of factors impacting insurance charges. Careful consideration and examination of potential biases will be crucial in drawing accurate conclusions.

3 Analyses

3.1 Data Pre-processing

Before commencing the analysis, it is crucial to verify the cleanliness of our data. The dataset underwent thorough examination for the presence of null values using standard methods, such as the ‘isnull()‘ or ‘info()‘ functions in Python with Pandas. Fortunately, our dataset exhibited an absence of null values, eliminating the need for preprocessing in terms of null value removal. Duplicate rows were identified using the ‘duplicated()‘ function in Python with Pandas, uncovering one duplicate record. This specific record was removed, preserving the first occurrence. As a result, the dataset now consists of 1337 records and 7 variables. The potential for bias will be acknowledged and addressed in subsequent analyses to ensure the robustness and fairness of our findings.

3.2 Univariate Analyses

Univariate Analysis is a statistical method employed to scrutinize each variable within a dataset independently. The objective is to gain insights into the distribution, central tendency, variability, and inherent patterns of individual variables. This analysis involves creating visualizations tailored to the nature of each variable. Categorical variables are represented through bar plots, providing a visual summary of the frequency distribution of distinct categories. On the other hand, numerical variables are depicted using histograms, allowing for a graphical representation of the data’s distribution and concentration around specific values. Through Univariate Analysis, one can unveil the unique characteristics and trends associated with each variable, laying the groundwork for a more comprehensive understanding of the dataset.

3.3 Multivariate Analyses

Multivariate Analysis explores complex relationships among multiple variables simultaneously, providing a holistic understanding of their interplay and collective impact. In the context of our study, we employ this analytical approach to unravel intricate patterns in insurance charges. By delving into the interrelationships between gender, geographic region, and health indicators, we aim to uncover nuanced insights that go beyond the scope of individual variables. The following analyses navigate the multifaceted landscape of insurance charges, shedding light on factors that collectively contribute to the variability observed in our dataset.

Which gender tends to have higher insurance charges?

This analysis focuses on exploring the intricate relationship between gender and insurance charges. The central objective is to identify patterns and trends in the distribution of insurance charges in the context of gender. By visualizing the distribution using box plots and five-number summary, this analysis unveils statistical parameters such as minimum value, maximum value, the inter-quartile range and potential outliers of the insurance charges for both genders. This investigation allows the insurance companies to re-formulate their strategic policies and pricing models based on gender, as it ensures a data-driven, fair pricing structure that will align with risk assessments.

Does region play a role in this relationship?

To perform a more nuanced analysis of the relationship between gender and insurance charges, the region variable is also brought into the context. This introduces a spatial dimension to the analysis. To consider all four variables, a violin plot has been used. This plot allows discerning different variations in the distribution of charges for males and females in all four of the regions - Northeast, Northwest, Southeast and Southwest. The outcomes of this analysis can guide an insurance company to refine its pricing strategies at a regional level.

Are unhealthy people at a risk of getting higher medical insurance charges?

To answer this question, BMI and Charges variables are taken into consideration. To visually analyze this relationship, a scatter plot has been constructed, to illuminate the potential patterns or trends that may exist. To complement the graphical exploration, a numerical correlation test has also been performed to bolster the results of the visual analysis. To fit the smokers into this picture, the observations/the people who smoke have been displayed differently with a different color in the same scatter plot.

3.4 Hypotheses Testing

Below are the four hypotheses that have been formulated based on the sample dataset. As the exact statistics of the entire population is unknown, the categorical variables are used and compared to the distribution of insurance charges across each group within the feature. For hypothesis testing, a null and an alternative hypothesis is defined, and a statistical test is used to determine whether the null hypothesis can be accepted or rejected.

Hypothesis 1

- Null Hypothesis : H_0 - Charges of both men and women are equal
- Alternate Hypothesis : H_1 - Charges of men is greater than women

Hypothesis 2

- Null Hypothesis: H_0 - Charges for both smokers and non-smokers are equal
- Alternate Hypothesis : H_1 - Charges for both smokers and non-smokers are not equal

Hypothesis 3

- Null Hypothesis : H_0 - Charges for all regions are equal
- Alternate Hypothesis : H_1 - Charges for all regions are not equal

Hypothesis 4

- Null Hypothesis: H_0 - Proportion of smokers in male and female are same
- Alternate Hypothesis : H_1 - Proportion of smokers in male and female are not same

Below are the different tests chosen and assumptions made to evaluate our hypotheses.

3.4.1 One Tailed (Right Tailed) Two Sample T-test

One Tailed Two Sample T-test is used to determine if there is a significant difference in statistical parameters between two independent groups. There are two types of One Tailed Two Sample T-test. Right tailed test has a rejection region on the right tail and the Left tailed test has a rejection region on the left tail. When the alternative hypothesis involves comparison of statistical parameters like mean/proportion using greater than sign ($\mu_a > \mu_b$), right tailed Two Sample T-test is employed. When the less than sign is used ($\mu_a < \mu_b$), left tailed Two Sample T-test is used. But before proceeding with the T-test, the underlying assumptions of the T-test must be satisfied.

- Normality of the sample
- Independence of data points

Normality of the sample: The entire sample of the data must be approximately normal for satisfying the normality of the sample. Since the data used in this case is positively skewed, transformations like box-cox transformation can be utilized for converting the data into normal data. But, checking for normality is always not necessary, as the real world data is never going to be normal. Instead, if the sample size is large, the sample mean aligns with the principles of central limit theorem i.e. the sample means are approximately normally distributed. From Fig. 2, it can be seen that in the insurance charges column, as the sample size increases, the distribution of sample means approaches normality. So the assumption of normality is satisfied.

Independence of data points: The random sample is independently and identically distributed. There is no repetition of data points across different groups . This satisfies the independence of data points assumption.

In this case the charges of men and women are compared. So a t-test is appropriate for testing whether charges of men and women are equal or not. The null hypothesis of t-test is assumed as the charges of men and women are same and the alternative hypothesis is assumed as charges of men greater than the women. Once t-test is performed, t-statistic is computed which is the measure that compares the sample mean with the population mean, considering the standard deviation and sample size. Corresponding p-value is computed which is compared with the critical value. If p-value (probability value) is less than the critical value we can reject the null hypothesis and accept the alternative hypothesis.

3.4.2 Two Tailed Two Sample T-test

A two tailed two sample test is used to compare the means of continuous variable between two independent groups and check if they are significantly different in both the directions. The only difference between the one tailed and two tailed tests is that, in one tailed t-test the rejection region is either right side (if $\mu_a > \mu_b$) or left side (if $\mu_a < \mu_b$), but in two tailed t-test the rejection region is on both sides ($|\mu_a| > |\mu_b|$). The assumptions of this test are similar to the above test.

In this case charges between smokers and non-smokers are compared. This test is appropriate for checking how significantly insurance charges differ between smokers and non-smokers. The null hypothesis assumes that the charges of smokers and non-smokers are equal and the alternative hypothesis is assumed that charges are not the same. As our hypothesis states that charges are not equal for smokers and non-smokers, this test is ideal as the rejection region lies in both directions.

Once t-test is performed, t-statistic is computed which is the measure that compares the sample mean with the population mean, considering the standard deviation and sample size.

Corresponding p-value is computed which is compared with the critical value. If p-value (probability value) is less than the critical value we can reject the null hypothesis and accept the alternative hypothesis.

3.4.3 ANOVA test

Analysis of Variance or ANOVA is the test performed when the number of samples/groups to be compared is greater than two. The t-test or z-test cannot be used for performing hypothesis testing for more than two samples/groups. Since the comparison of charges across the regions is tested and the number of groups is four: North East, North West, South East, South West, it is better to compare the statistical parameters with the ANOVA test. ANOVA test analyzes the difference of means among the different groups in the sample. It is used to check whether the means are equal or different for various groups. But before proceeding with ANOVA test, the underlying assumptions of ANOVA test must be satisfied. The three major underlying assumptions of ANOVA test are:

- Normality of the sample
- Homogeneity of Variances
- Independence of data points

Normality of the sample: The number of data points in each group is high. With the reference to the Central limit theorem, the means of each group will be normally distributed. So, the normality condition is satisfied.

Homogeneity of the Variances: ANOVA test also assumes that the variance of each group must be equal to each other. Since the data in our case is skewed, Levene's test of variances is used for comparing the variance of each sample. Levene's test functions well with skewed data and is robust.

Levene' Test for equal variance: Levene's test takes different samples as input and analyzes the samples and gives the result whether the samples have equal variances or not. It initially assumes the null hypothesis that all the samples have equal variances. If the statistical value (p value) is lesser than the critical value then the sample does not have equal variances. If the value is greater than the critical value then the samples have equal variances.

Exception: Even if the levene's test for equal variance fails, there is an exception for homogeneity of the variances. If the sample size of each group is large and equal, it can be assumed that the homogeneity of the variances condition is satisfied.

Independence of data points: The random sample is independently and identically distributed. There is no repetition of data points across different groups . This satisfies the independence of data points assumption.

Once all the underlying assumptions are met, the one way ANOVA test is conducted. The reason behind choosing the one way ANOVA test is that there is only one factor/independent variable (charges) which is being considered. The null hypothesis of ANOVA test is assumed as insurance charges across all the regions are equal and the alternative hypothesis is that the insurance charges across all the regions are not equal. After the test is performed, based on the ratio of variances between the groups and variance within the groups (f - value). Following that corresponding p-value is calculated and compared with the critical value to check whether the null hypothesis can be rejected or cannot be rejected. The results of the test are further discussed in the results section.

3.4.4 Chi-Square Test

Chi-Square test is the test performed when two categorical variables are compared. Chi-square test helps in analyzing whether there is any association between categorical columns. The data must be converted into contingency table format for carrying out Chi-Square test. But before proceeding with the Chi-Square test the underlying assumptions of the test must be satisfied.

- **Independence & Mutual Exclusivity of data points:** The random sample is independently and identically distributed. There is no repetition of data points across different groups . This satisfies the independence & mutual exclusivity of data points assumption.
- **Expected Frequencies:** The frequency distribution of each cell in the contingency table must not be small. Contingency tables between gender and smoker variables are computed and frequencies are not small satisfying this second condition.

Once all the assumptions of the chi-square test are met, the chi-square test can be carried out. The two most commonly used Chi-Square tests are: Pearson Test, Likelihood Ratio Test

In this case, as the sample size is large, Pearson Chi-Square test is used which performs much better than the latter. The initial null hypothesis of the chi square test assumes that the proportion of smokers in both male and females are the same and an alternate hypothesis is that the proportion of smokers in both male and females are not the same. After the test is performed, p value is calculated and compared with the critical value to check whether the null hypothesis can be rejected or cannot be rejected. The results of the test are further discussed in the results section.

4 Results

4.1 Univariate Analyses

4.1.1 Bar Plots

All of the bar plots display the count of different categories of the categorical variable.

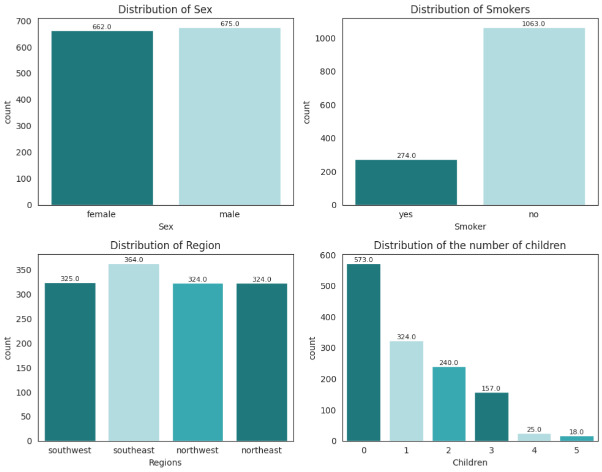

Figure 1: Bar Plots for Categorical Variables

In Figure 1, the first graph is generated to assess the count of male and female patients. As depicted in the plot, the counts are approximately equal. The second plot illustrates the count of smokers and non-smokers, revealing a substantial difference. The size of non-smokers is approximately four times that of smokers. The third plot exhibits the count of patients in different regions. The patient numbers in the Southwest, Northwest, and Northeast regions are equal, with slightly more patients in the Southeast. The last graph, pertaining to the 'children' column, illustrates the number of children per patient. The plot indicates that a majority of patients have no children, and there are very few patients with 4 or 5 children, a common trend in the current generation.

4.1.2 Histograms

Histograms are plotted to check the distributions of the numerical variables.

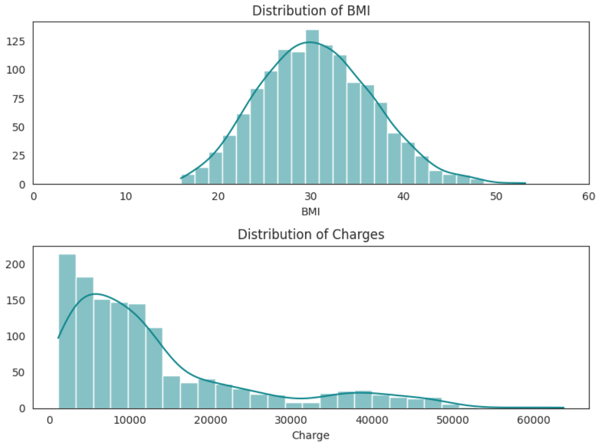

Figure 2: Histograms for Numerical Variables

In Figure 2, the first graph is created to analyze the distribution of the BMI variable. The distribution appears to be normal, with most BMIs centered around the mean of 30. The histogram of insurance charges, however, indicates a non-normal distribution. It is skewed to the right, revealing that the majority of patients incur insurance charges up to fifteen thousand dollars.

4.2 Multivariate Analyses

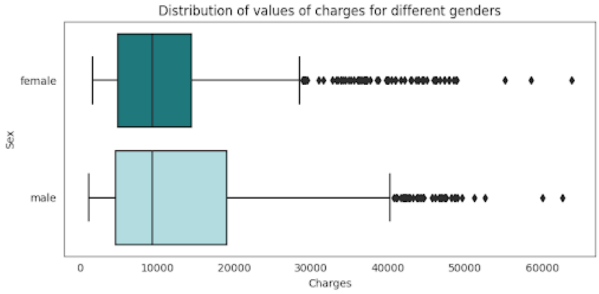

Figure 3: Which gender tends to have higher insurance charges?

From the boxplot 3, it is noticeable that the distribution of charges for males tend to be wider compared to that of females, although the median values are fairly similar. The 75th percentile of charges is higher for males by an approximate amount of about $5000 and there is an even higher difference between the maximum values of both distributions. There is also a considerable number of outliers in both the distributions. From these findings, we can conclude that females tend to have lower charges than males. This could be due to various factors like income, driving habits, or types of insurance coverage. More data is needed to explore the underlying reasons.

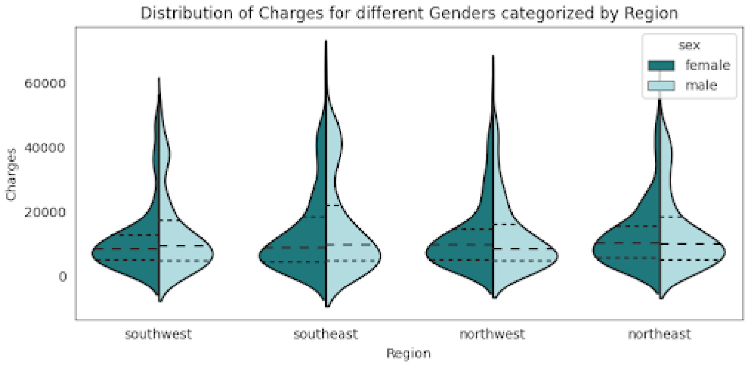

Figure 4: Does region play a role in this relationship?

The violin plot 4 further complements the outcomes of the previous analysis. The distribution of charges is similar to the box plot, with both genders having right-skewed distributions with longer tails towards higher charges. In all four of the regions in the United States, the median values tend to be similar for both genders and in all regions, males get higher insurance charges than females. The violin plot also shows that there is a wider range of charges for males, as the violin for males is wider than the violin for females. This finding from a business perspective can save a lot of work, as we can confirm that charges are distributed similarly in all regions of the United States. Thus the insurance corporation does not have to formulate different plans for different regions.

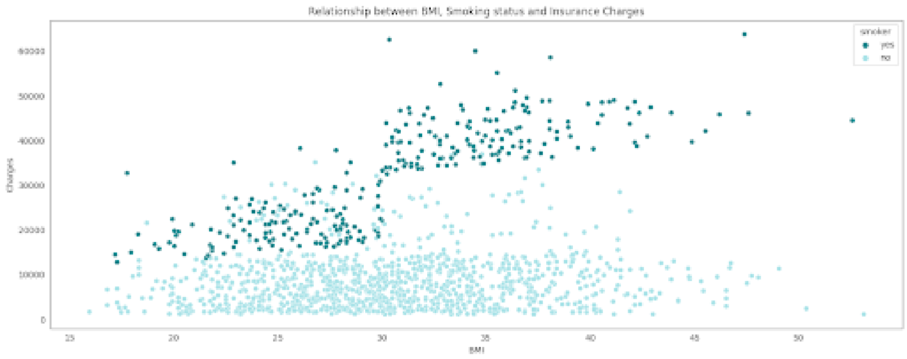

Figure 5: Are unhealthy people at a risk of getting higher medical insurance charges?

Based on the scatter plot 5 , it is visible that there are not any noticeable relationship between BMI values and the charges of a person. The correlation value between these two random variables was 0.19840083 which suggests that there is no strong positive/negative correlation between these two variables, supplementing the visual findings of the scatter plot.

The plot suggests that smokers tend to have higher charges than non-smokers, even at similar BMI levels. This is evident by the general presence of smoker data points above the non-smoker data points. However, there is still some overlap between the two groups, indicating that BMI isn't the sole factor influencing the difference in charges.

4.3 Hypotheses Testing

In this case the critical value is defined as 0.05 or confidence interval is set as 95%. This critical value is based on the level of significance and defines the threshold which can be used to reject the null hypothesis. Then the p-value (probability value) obtained from our hypothesis testing is compared to the critical value. If p-value is less than 0.05 (critical value), it can be concluded that there is less than 5% chance that the null hypothesis is true, and hence it can be rejected. The choice of critical value is subjective and depends on the problem statement and context. An in-depth domain knowledge is required for fixing the correct critical value.

4.3.1 Hypothesis 1 Test Results

Test Name	Critical Value	t value	p value
One Tailed Two Sample Test	0.05	2.1275	0.0169

As the p Value is less than the critical value, the null hypothesis can be rejected and the alternative hypothesis which states that the insurance charges for men are greater than those for women can be accepted.

4.3.2 Hypothesis 2 Test Results

Test Name	Critical Value	t value	p value
Two Tailed Two Sample T-test	0.05	46.6447	1.4067e-282

As the p Value is less than the critical value, the null hypothesis can be rejected and the alternative hypothesis which states that the insurance charges for both smokers and non-smokers are not equal can be accepted.

4.3.3 Hypothesis 3 Test Results

Levene's Test Results:

Test Name	Critical Value	s value	p value
Levene's Test of Equal Variance	0.05	5.5535	0.00086

As the p value is less than the critical value, the samples do not have equal variances.

Number of samples of various regions:

Region	Number of Samples
North east	324
North west	324
South east	364
South west	325

Even though the samples failed the homogeneity of variances for levene's test, since the number of samples are high and are almost equal, it can be assumed that the homogeneity of variances is satisfied.

ANOVA Test Results:

Test Name	Critical Value	f value	p value
One way ANOVA test	0.05	2.9696	0.03089

As the p Value is less than the critical value, the null hypothesis can be rejected and the alternative hypothesis which states that the insurance charges across all regions are not equal can be accepted.

4.3.4 Hypothesis 4 Test Results

Contingency Table:

Smoker	No	Yes
Sex		
Male	516	159
Female	547	115

This table depicts the contingency table between the two category columns Smoker and Sex(Gender).

Chi-square Test Results:

Test Name	Critical Value	Chi-square coefficient value	p value
Chi-Square Test	0.05	7.3929	0.0065

As the p Value is less than the critical value, the null hypothesis can be rejected and the alternative hypothesis which states that the proportion of smokers in male and female are not the same can be accepted.

5 Conclusion

In conclusion, a comprehensive analysis was undertaken, encompassing both univariate and multivariate perspectives, to shed light on crucial factors influencing insurance charges in the United States healthcare landscape. Intriguing patterns within the dataset were revealed through univariate analyses. Particularly noteworthy is the highlighting of a significant disparity in the count of smokers versus non-smokers, emphasizing the necessity for a nuanced understanding of these lifestyle choices.

The findings were deepened by multivariate analyses conducted through boxplots and violin plots. Wider charge distributions were exhibited by males compared to females, suggesting a potential gender-based influence on insurance costs. Regional disparities were further accentuated by the violin plot, consistently portraying higher charges for males than females across all regions.

Conclusions were bolstered by hypothesis testing, providing statistical evidence for assertions. The acceptance of alternative hypotheses in various scenarios affirmed that gender, smoking status, and regional differences significantly impacted insurance charges. Notably, higher charges tended to be incurred by males than females, and the uniform distribution of charges across regions facilitated streamlined insurance planning.

Despite visual indications of a potential influence of BMI and smoking status on charges, it was suggested by hypothesis testing that these factors alone may not be sufficient to explain the observed differences. Further investigation and data collection were deemed necessary to explore the intricate relationships between these variables and insurance charges.

In summary, the complexity of factors contributing to insurance charges was underscored by the findings. Gender, smoking habits, and geographical location emerged as influential elements, warranting careful consideration in policy formulation and healthcare planning. The nuances revealed through the analysis offered valuable insights for stakeholders in the healthcare industry, guiding more informed decision-making and resource allocation.

In addition to the findings, acknowledgment and addressing of certain challenges that could influence the interpretation of results were deemed essential. The first notable challenge was related to the assumption of ”Independent and Identically Distributed” (i.i.d.) datapoints within the insurance dataset. The validity of this assumption was considered critical for the robustness of statistical analyses, and any deviation from it could introduce bias or affect the generalizability of conclusions. Careful consideration and potential adjustments were deemed necessary to mitigate the impact of any non-i.i.d. characteristics.

Furthermore, recognition was given to the fact that the dataset used in the analysis represented a small sample rather than the entire population. While hypothesis testing results provided valuable insights into this sample, caution was advised in generalizing these findings to the broader population. The inherent limitations of a small sample size were acknowledged to potentially affect the reliability and applicability of conclusions to the larger population. Future research was suggested to aim for obtaining more extensive and representative datasets to enhance the external validity of results.

6 References

1. Data Source: Insurance Dataset
2. Hypothesis Testing of Health Insurance Data on Kaggle
3. Insurance EDA and Hypothesis Testing on Kaggle
4. Insurance Claims EDA and Hypothesis Testing on Kaggle
5. scipy.stats.ttest_ind Documentation
6. ANOVA Tutorial
7. scipy.stats.chisquare Documentation
8. Chi-Squared Test for Machine Learning
9. Central Limit Theorem and Hypothesis Testing
10. Are My Data Normal?
11. ANOVA Violations of Assumptions
12. Kaiser Family Foundation 2022 Employer Health Benefits Survey
13. scipy.stats.levene Documentation
14. State of U.S. Health Insurance 2022 Biennial Survey
15. National Health Expenditures 2022 Highlights