Do smoking status, age, gender and BMI affect health insurance and by how much?

Warren Wang, ID: 30260960

Divya Seth, ID: 30070431

Gulshan Laskar, ID: 30256534

Chandanarchutha Namburu, ID: 30263600

**Purpose:**

The project aims to estimate the difference in cost of personal insurance for individuals based on various factors such as age, gender, BMI, and smoking habits.

**Population:** The population comprises of 662 females and 676 males in the age range of 18-64 from different regions of Canada.

**Variables of interest: there are 1338 observations and 7 different variables.**

**Categorical variables**: Sex, smoking

**Numerical variables**: Age, BMI and Charges (insurance cost)

**Statistical Inferences performed**: Hypothesis testing, bootstrap confidence interval.

**Parameters to be assessed:**

**Effect of Age, Smoking, Gender and BMI on insurance cost:**

1. Insurance cost in smokers and non-smokers
2. Insurance cost according to Age
3. Insurance cost according to gender
4. Insurance cost according to BMI level

# **Data Collection Methods:**

In our analysis, we sought to explore the relationships between a range of factors and medical charges. To achieve this, we required a dataset suitable for performing hypothesis testing, calculating confidence intervals, and conducting linear regression analyses. After extensive research, we selected a dataset from Kaggle, a renowned platform for data sharing and collaboration.

## **Dataset Overview:**

The dataset we utilized is the Insurance Charges dataset, which contains 1,339 rows and 7 columns. The columns and their respective descriptions are as follows:

**Age**: The age of the primary beneficiary.

**Sex**: The gender of the insurance contractor categorized as either female or male.

**BMI (Body Mass Index)**: An objective measure of body weight relative to height, where healthy BMI typically ranges from 18.5 to 24.9.

**Children**: The number of children covered by health insurance or the number of dependents.

**Smoker**: Indicates whether the individual is a smoker or not.

**Region**: The beneficiary's residential area in the U.S., categorized into northeast, southeast, southwest, and northwest.

**Charges**: The individual medical costs billed by health insurance.

This dataset provides a comprehensive view of the factors influencing medical charges, making it an ideal dataset for our analysis.

**Data Cleaning:**

During the initial stages of working with the dataset, we took several steps to ensure its readiness for our analyses. The dataset, sourced from Kaggle, was well-organized and contained the necessary variables for our study, which included columns like age, sex, BMI, and medical charges. Despite the overall cleanliness of the data, we conducted preliminary checks to identify and address any duplicates or missing values. Outliers were also reviewed to ensure they wouldn't distort our statistical tests. By cleaning the data, we were able to maintain its integrity and ensure reliable results for the analysis.

A screenshot of a computer

Description automatically generated

**Issues Encountered:**

While the dataset itself was structured and accessible, our main challenge lay in verifying its suitability for the specific statistical tests we intended to perform. We needed to ensure that the numerical data aligned with the requirements of hypothesis testing, confidence intervals, and regression models. This meant paying close attention to the distribution of variables and any anomalies that might impact on the validity of our analyses. Additionally, we had to ensure that treating insurance charges as a dependent variable was appropriate, given the variety of factors contributing to healthcare costs.

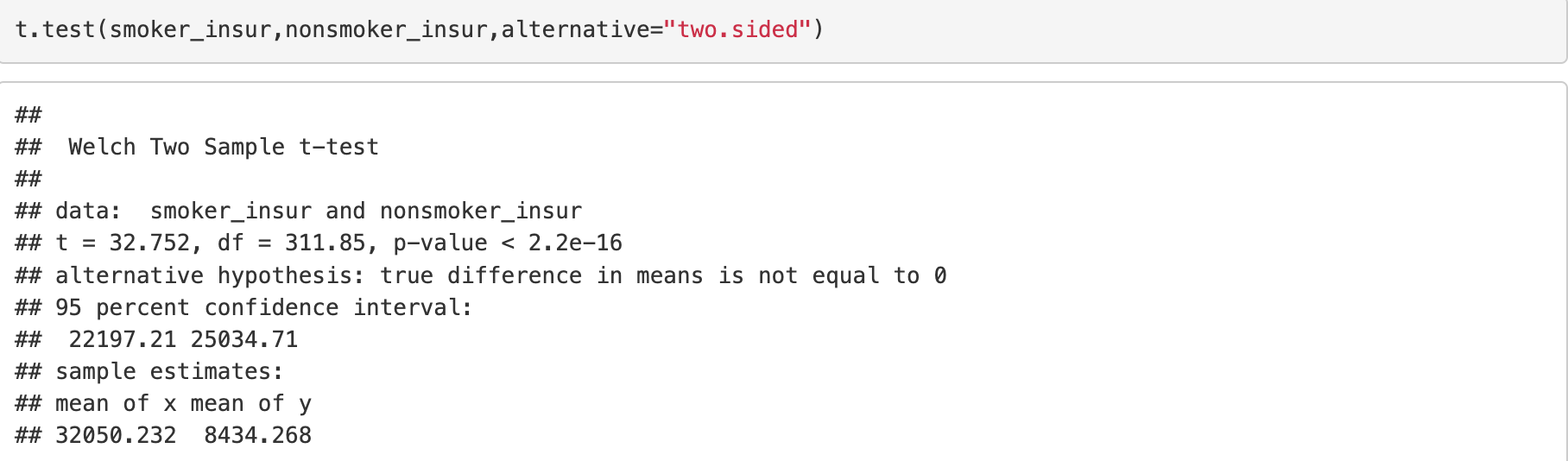
**Statistical Analysis:**

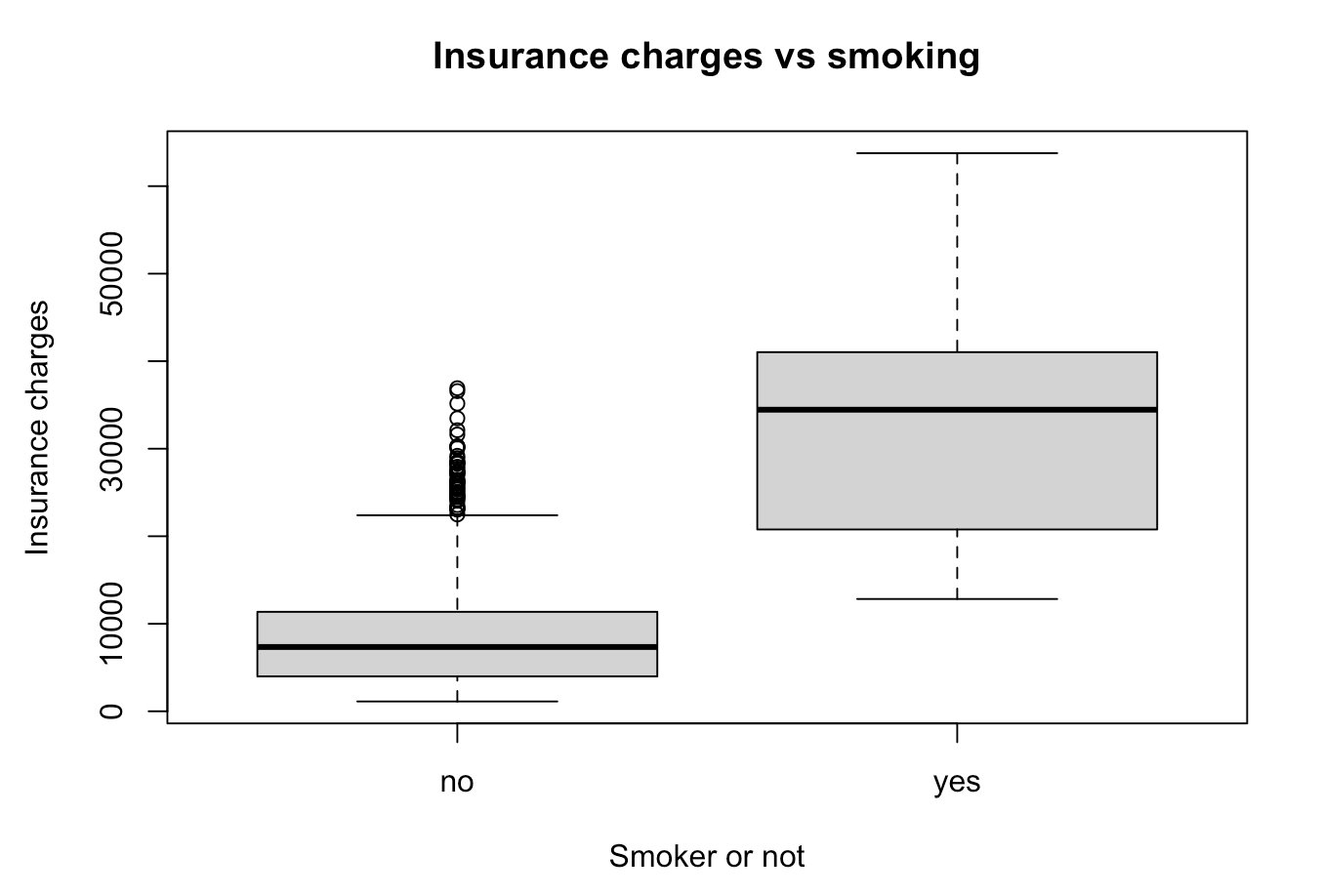
**Insurance cost vs. Smoking Status**

**Hypothesis Testing 1**: Does smoking affect health insurance charge?

**Null hypothesi**s: There is no difference between smokers and nonsmokers' insurance amount.

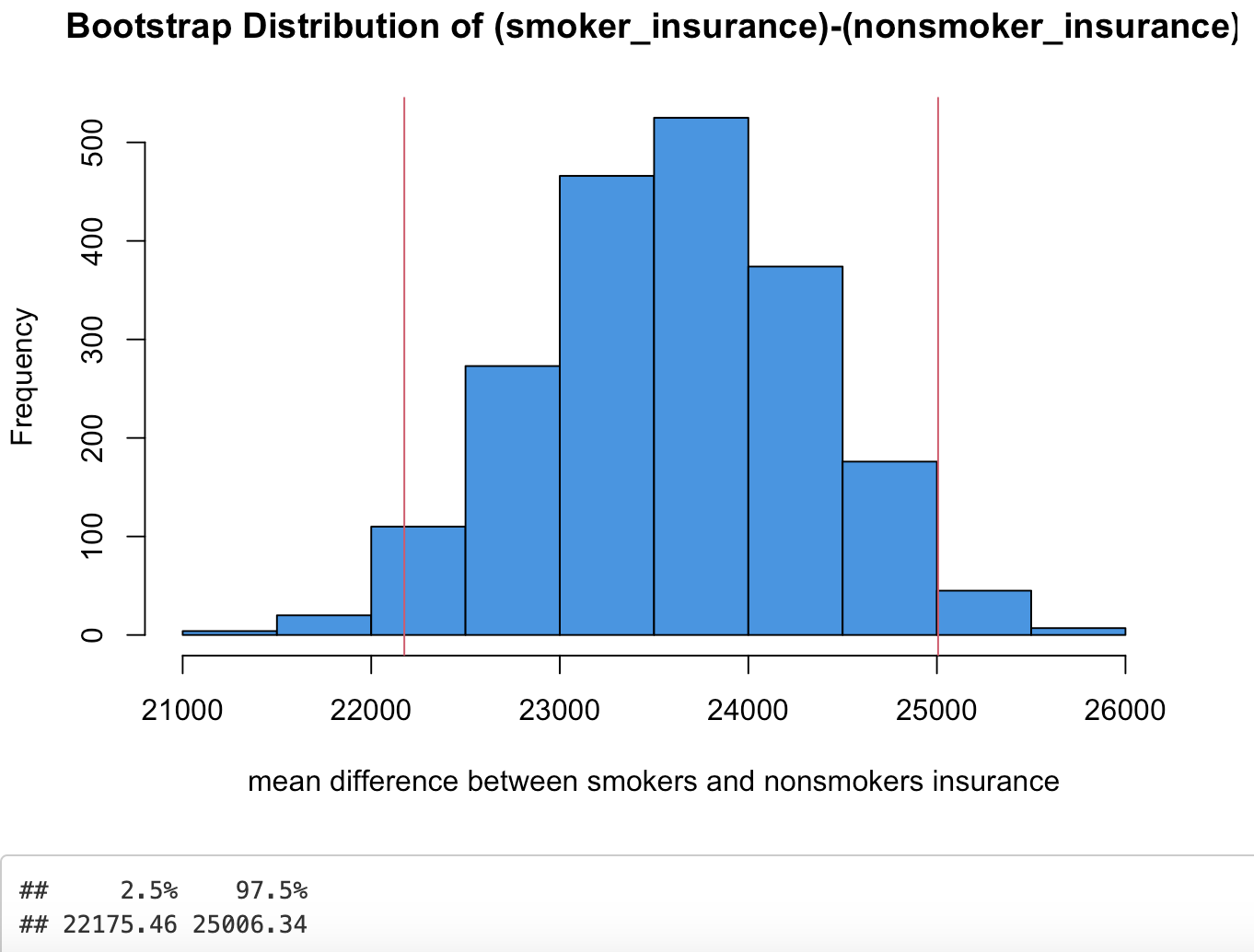
**Alternative hypothesis**: Smokers pay more insurance than nonsmokers.



The p-value here is 2.2e^-16 which is a lot less than the significance level which is 0.05. Therefore, we reject the null hypothesis and conclude that smokers pay more than nonsmokers.

We made a boxplot to confirm the result of hypothesis testing. As we can see, the boxplot of smokers is a lot higher than that of the nonsmokers. Even the highest insurance amount a nonsmoker pays is just a little higher than the median of smokers.

**How much more insurance do smokers pay?** We performed a bootstrap sample difference confidence interval between the smokers and nonsmokers.



As can be seen from the plotted confidence interval, smokers can pay as much as $25000 dollars more in health insurance than nonsmokers.

In conclusion: smokers pay more health insurance than nonsmokers with a 95% bootstrap confidence interval of ($22175.46, $25006.34)

**Hypothesis Testing 2**: (Between Age Groups and charges)

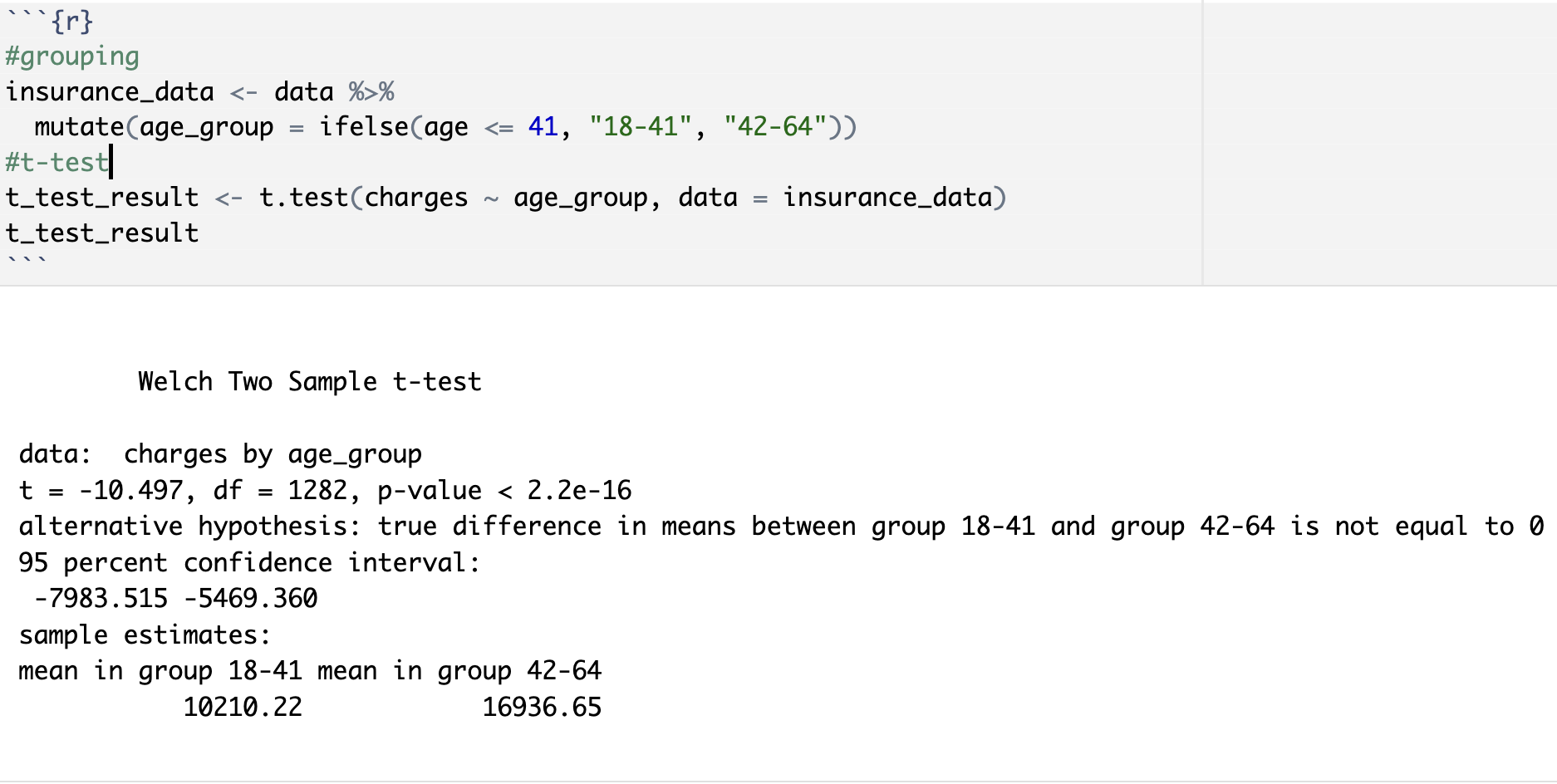
We want to test whether there is a significant difference in the average medical charges between two age groups:

**Group 1**: Ages 18 to 41

**Group 2:** Ages 42 to 64

**Null hypothesis (H0):** There is no difference in the average medical charges between the two age groups.

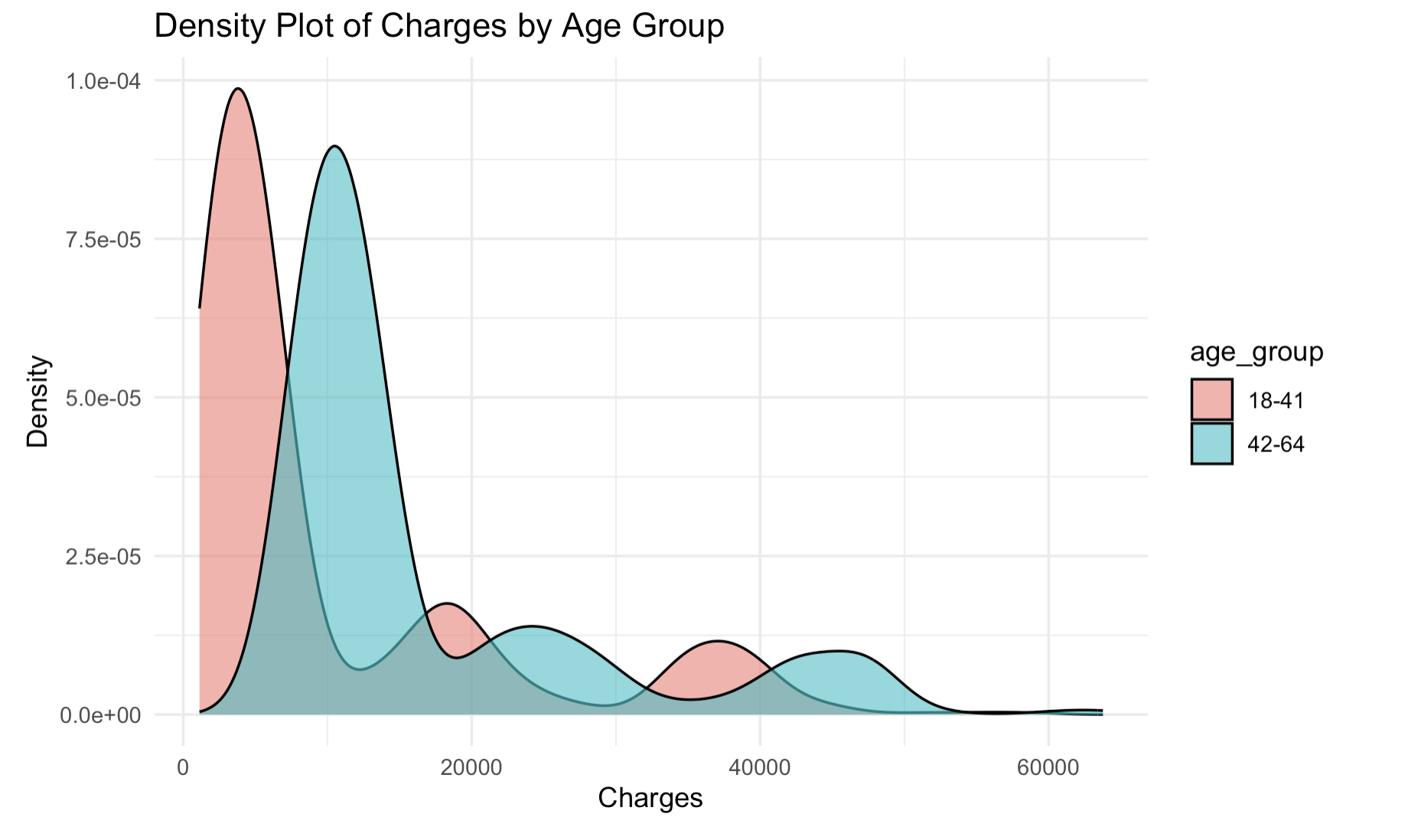
**Alternative hypothesis (Ha):** There is a significant difference in the average medical charges between the two age groups.

In analyzing the medical charges between two age groups, we found compelling evidence indicating a significant difference in mean charges incurred by individuals. The analysis produced the following key results:

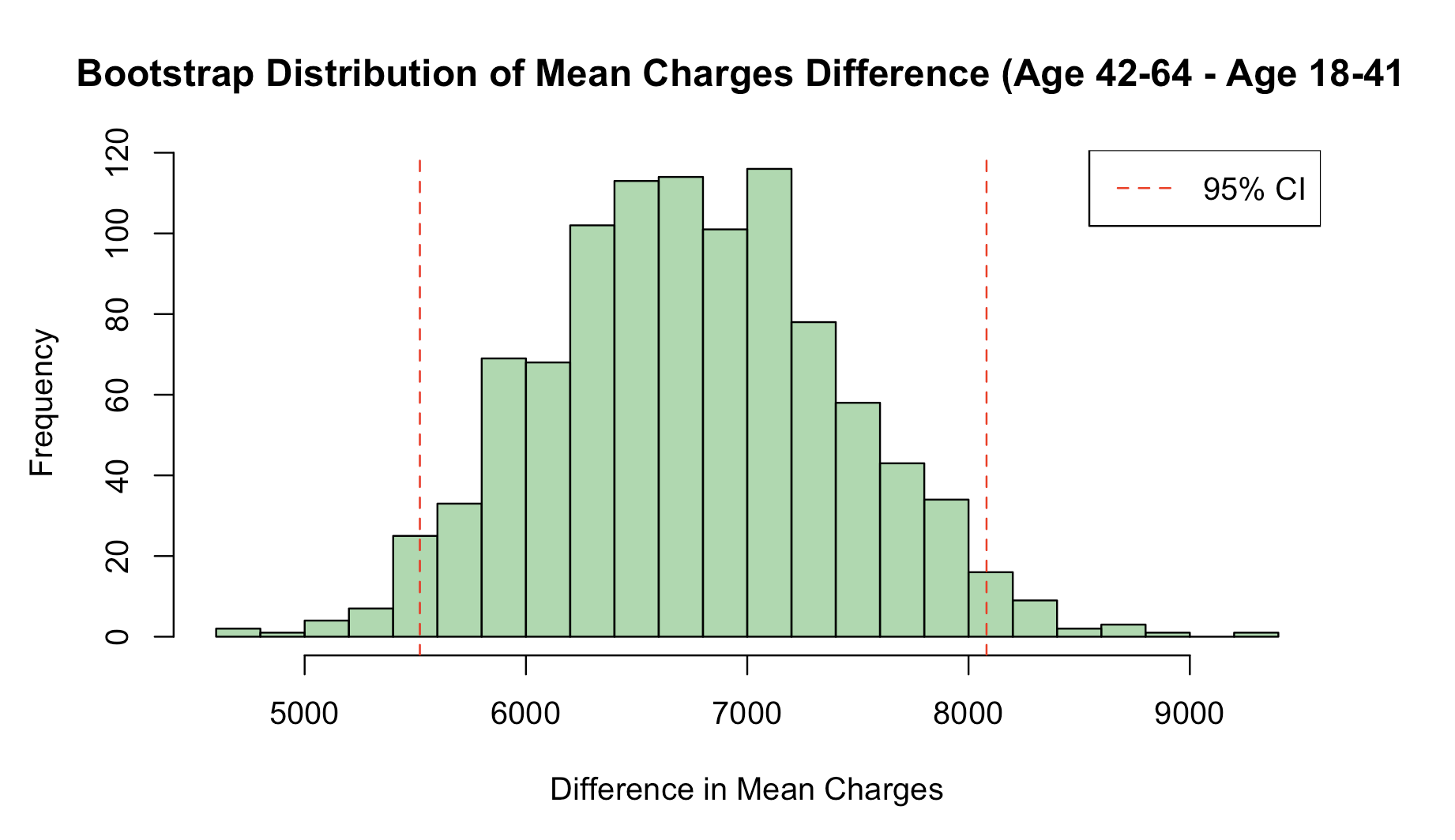
**p-value:** < 2.2e-16: Since the p-value is much lower than the significance level of 0.05, we reject the null hypothesis. This strongly supports the conclusion that there is a significant difference in medical charges between the two groups.

The findings reveal that individuals in the 42-64 age group incur significantly higher medical charges compared to those in the 18-41 age group. This discrepancy may be attributed to a variety of factors, including increased medical needs and more frequent healthcare utilization as individuals age.

Furthermore, the density plot of charges by age group visually supports the findings of the t-test, providing additional justification for the observed differences.

The age group (42-64) exhibits a wider distribution of medical charges, indicating greater variability in expenses. In contrast, the G1 age group (18-41) shows a higher density of individuals with lower medical charges.

Additionally, when we applied bootstrapping to calculate the confidence interval, we obtained consistent results.



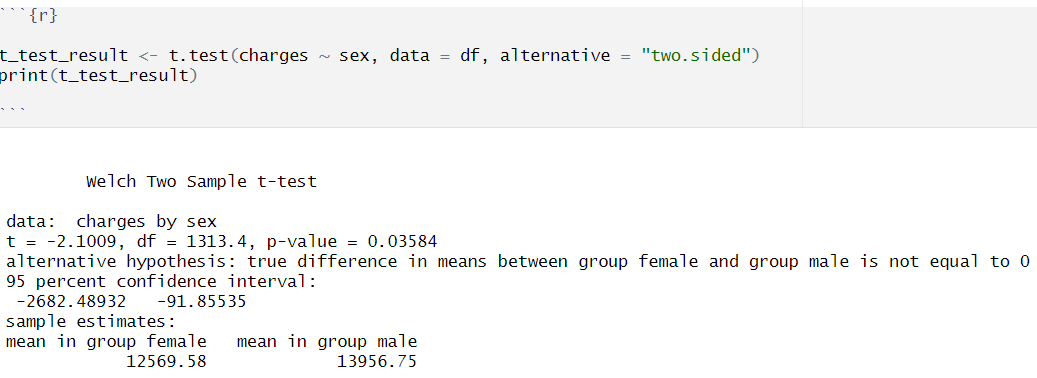
This indicates that individuals in the 42-64 age group tend to incur average charges that are between $5,507 and $8,055 higher than those in the 18-41 age group. These bootstrap confidence interval results further substantiate the findings of the t-test. So,

**Hypothesis testing 3**: (Sex vs charges)

The Impact of Gender on Insurance Costs: Do Males pay Higher Charges?

**Null hypothesis**: There is no significant difference in the mean insurance charges between males and females.

**Alternate Hypothesis**: There is a significant difference in the mean charges between males and females.

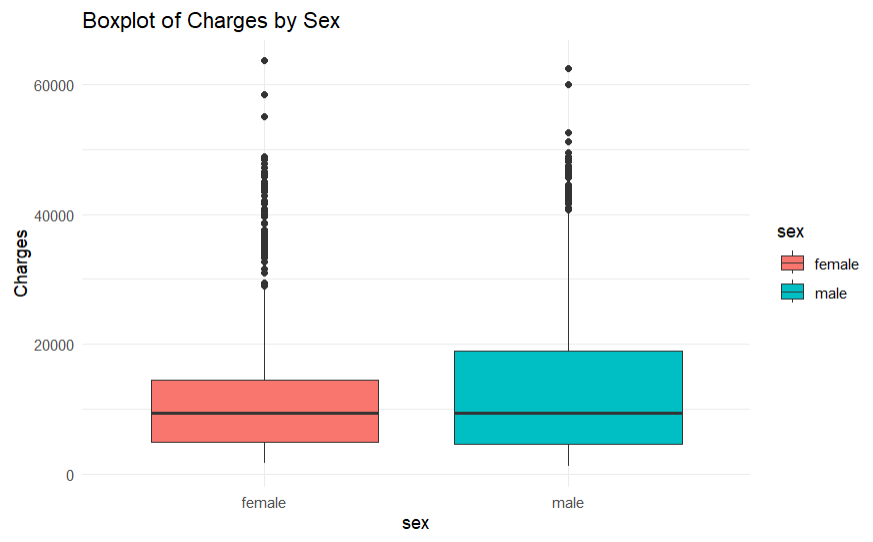


The test shows that the average insurance charges for males ($13956.75) are significantly higher than those for females ($12569.58), indicating that males pay more on average for insurance.

The p-value from t-test is less than 0.05, so we reject the null hypothesis. This indicates that there is a statistically significant difference in mean insurance charges between males and females.

It is important to note that p-value is close to the threshold, so the difference may be significant, but it may not represent an extremely strong effect. the difference, while present is not overwhelmingly large, and any small variations in data could potentially affect the outcome.to strongly reject the null hypothesis, a larger sample size would be beneficial.

However, when bootstrapping is applied, we generate a much larger number of samples through re-sampling. If the result from bootstrapping is consistent with the t-test, then we can confidently reject our null hypothesis.



The variability suggests that the charges for males are more spread out, indicating a wider range of insurance costs. while the means differ significantly, Interestingly, median insurance charges are the same for both groups

**Outliers:** These outliers represent natural variability in the sample. These may result from high risk factors such as severe health conditions or more advanced insurance coverage. While they affect the mean charges, they are kept in the data set as they represent real world scenarios.

A graph of a number of differences

Description automatically generated

The confidence interval for the difference is shown by the red dashed lines, ranging from 99.59 to 2625.96. This interval does not include zero, providing strong evidence that the average charges for males are higher than for females, which supports our results from our hypothesis testing on sex vs charges.

The consistency between the t-test and the bootstrap results allows us to confidently conclude that there is a meaningful difference in mean insurance charges between the two groups.

**Hypothesis testing 4:**

**Problem:**

To study the effect of BMI levels on the cost of personal insurance**.**

**Introduction:**

According to Life Insurance Insights (2024), life insurance premiums are decided in part based on your BMI (height and weight ratio), which is considered a health risk indicator in underwriting. Every insurance company has a range to classify a normal healthy weight, which is used to decide on the rate of personal insurance and the coverage (Mylona, et.al, 2020), but on an average a higher BMI may mean higher payment for insurance cost or in some cases insurance might also be declined (Sun Life, n.d). The data on personal insurance cost was thus assessed to determine how much on average an individual is expected to pay more if they have BMI, not in the normal range as per WHO BMI cutoffs (WHO, 2021).

**Hypothesis:**

**Null Hypothesis**: The BMI has no effect on the personal insurance cost

**Alternate Hypothesis**: The BMI effects personal insurance cost

**Variables of interest include:**

1. Personal insurance cost
2. BMI: Body mass index (BMI) is a health indicator of nutritional status and defines the health risk situation. It is calculated as a ratio of body weight(kg) and height (in meter square). According to the World Health Organization, Body mass index cut off classification is as follows:

|  |  |
| --- | --- |
| **BMI** | **Nutritional status** |
| Below 18.5 | Underweight |
| 18.5–24.9 | Normal weight |
| 25.0–29.9 | Pre-obesity |
| 30.0–34.9 | Obesity class I |
| 35.0–39.9 | Obesity class II |
| Above 40 | Obesity class III |

(**Source**: World Health Organization - A Healthy Lifestyle: WHO Recommendations)

**Method**:

The normal range of BMI for a healthy adult within the normal weight range is between 18.5 to 24.9. For ease of analysis, in this current data set, we classified all others (except normal weight) as at risk, as we believe they are at higher risk of developing health issues as compared to the individuals in the normal weight BMI range. Thus, we divided the entire data into 2 subsets.

1. **Subset 1**: BMI normal: In the range (18.5-24.9)

2. **Subset 2**: BMI at risk: above and below the normal range.

Accordingly, we filtered the data set into individuals who have a normal BMI and at- risk BMI. A total of 222 individuals were included in the BMI normal subset and 1116 were included in the BMI at-risk subset.

**Statistical analysis:**

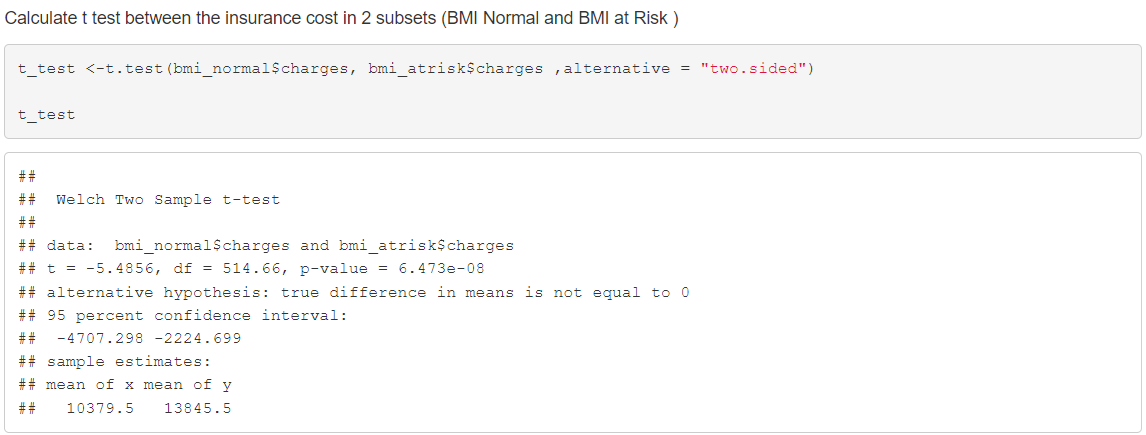
The variable charges were assessed between the two BMI subsets. The summary statistics for both the subsets are given below:

**Table2. Summary statistics for 2 subsets:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **n** | **Min** | **Max** | **Mean** | **SD** | **Missing value** |
| **Subset 1** | 222 | 1121.874 | 35069.37 | 10379.5 | 7500.713 | 0 |
| **Subset 2** | 1116 | 1131.507 | 63770.43 | 13845.5 | 12755.59 | 0 |
|  |  |  |  |  |  |  |

1. **2-sample t test:**

A 2-sample t test was conducted to calculate the difference in mean values in insurance cost between the two subsets.



A graph of a number of different colored squares

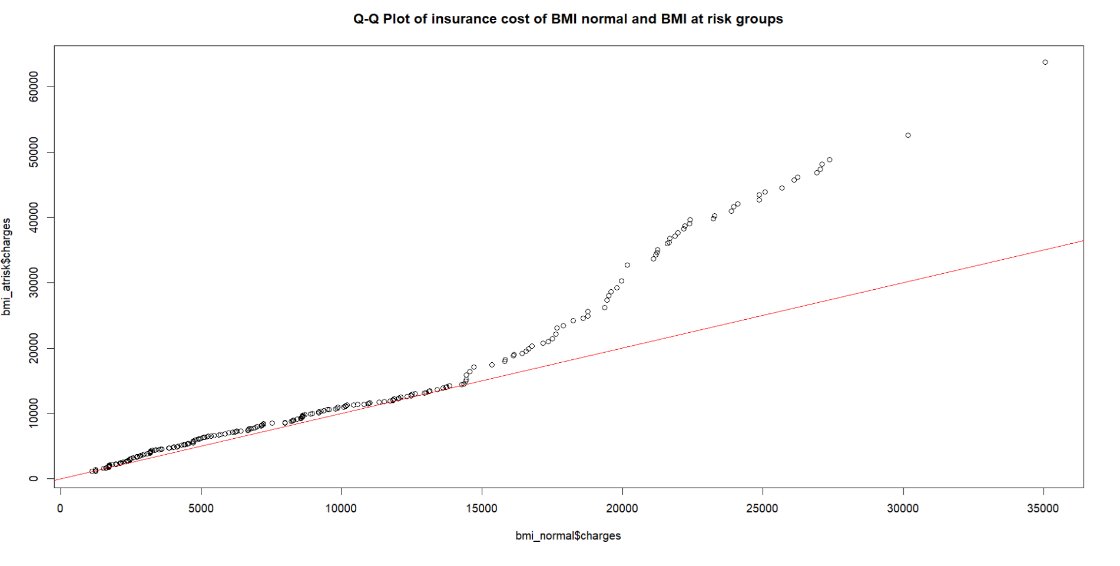
Description automatically generated

Similar is described in the histograms plotted above which shows a significant difference in cost of personal insurance with a high cost for individuals with BMI at risk (higher than the normal range).

1. **Bootstrap distribution and confidence interval**

When data is non normal or has outliers, creating a bootstrap distribution is useful, as it does not assume as specific distribution and provides more accurate reflection of variability.

Accordingly, we checked the normality of the subsets (Insurance cost of Normal and at-risk group) using QQ Plot. As given below:



The plot shows non normal values and outlier values as most of the points does not coincide with the qqline.

Also, A larger sample size derived from resampling through bootstrap method gives a more robust and accurate analysis. Hence, we did a bootstrap distribution of the current data subsets and found the Confidence interval

A graph showing a number of bmi

Description automatically generated

**Confidence interval at 95 % confidence level:**

At 95 % confidence level, our estimate of the mean insurance cost is between 2161.963 - 4639.211. The confidence interval is positive, which indicates that the insurance cost lies in the confidence interval.

**Result:**

The test result indicate a significant difference between the insurance charges in both subsets ( BMI Normal and BMI At Risk), the mean of personal insurance cost for individuals having normal BMI levels (18.5 -24.9 , as per WHO classification), is significantly lower (10379.5 ) than the mean cost for the people who have normal BMI , stating that they might be at risk for developing health problems (13845.5).

**t- statistics:**

The higher absolute value of t statistics indicates great difference in variability of the two groups while the negative sign indicates that the mean values in subset with BMI levels in the normal range is lower than the mean values of insurance cost in the subset with BMI in at risk group.

**p-value:**

Also, smaller p value than the .05 significance level, the difference in means of insurance cost is significant and is unlikely due to random chance.

**Confidence interval:**

The confidence interval for the difference in means is between -4707.298 and -2224.669. This range does not include 0, supporting the fact that there is a significant difference between the means. Also, the negative bound at each end signifies that the difference in insurance cost for BMI in normal group is consistently lower than the other group.

We can predict with 95 % confidence that the difference in insurance cost in the two subsets is significant and as the p value is less than .05, we reject the null hypothesis and accept that alternate hypothesis.

**Conclusions:**

* After the analysis, smokers pay more health insurance than non-smokers as the p-value from the two-sample t-test is less than the significance level 0.05. The t-test confidence interval is ($22197, $25034). We performed a bootstrap sample mean difference with a 95% confidence interval of ($22175.46, $25006.34) which quantified the difference.
* Based on the results, our analysis reveals that individuals aged 42-64 face significantly higher average medical charges compared to those aged 18-41. Both the t-test and bootstrapping methods consistently demonstrated this difference, with the t-test estimating a mean increase of $5,469 to $7,983 and the bootstrap confidence interval closely aligning at $5,507 to $8,055. These findings strongly suggest that healthcare costs increase with age, likely due to greater medical needs in the older age group. The consistency between both statistical approaches strengthens the validity of this conclusion and emphasizes the budgetary impact of aging on healthcare expenditures.
* Based on the analysis, there is a significant difference in average insurance charges between males and females. The t-test shows that males, on average, pay more ($13,956.75) compared to females ($12,569.58), with a p-value of 0.03584, allowing us to reject the null hypothesis. The bootstrap method confirms this result, with a 95% confidence interval for the difference in mean charges ranging from $99.60 to $2,625.96, further indicating that males tend to have higher charges.
* With the t test the confidence interval derived -(2224.846 - -4707.298) is consistent with the confidence interval derived from bootstrap distribution (-2161.963 - 4639.211). The cost of insurance is different for both groups. Thus, we can say that the individual with a BMI in at risk range (above and below) the BMI normal cutoffs) might have to pay approximately 2200 - 4600 $ more than the person with BMI in the normal range

**References:**

1. Sun Life Staff. (2023, January 30). *Life Insurance for Smokers*. Life Insurance For Smokers | Sun Life Canada. <https://www.sunlife.ca/en/insurance/life/how-does-smoking-affect-your-life-insurance/>
2. Aguiar-Díaz, I., & Ruiz-Mallorquí, M. V. (2023, October). Financial Knowledge and Private Health Insurance: Does Age Matter? *Healthcare (Basel)*, 11(20), 2738. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10606321/>
3. Walker, E. (2023, September 5). How Age Impacts Your Health Insurance Costs. *People Keep*. <https://www.peoplekeep.com/blog/how-age-impacts-your-health-insurance-costs>
4. Life Insurance Insights. (2024, July 10). Impact of height and weight on life insurance in Canada: How is height and weight used in life insurance? What is BMI and why is it important in life insurance? [Blog post]. Life Insurance Insights. <https://www.lifeinsuranceinsights.com/impact-of-height-and-weight>
5. World Health Organization. (2021, May 6). A healthy lifestyle - WHO recommendations. World Health Organization. <https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations>
6. Mylona, E. K., Benitez, G., Shehadeh, F., Fleury, E., Mylonakis, S. C., Kalligeros, M., & Mylonakis, E. (2020). The association of obesity with health insurance coverage and demographic characteristics: A statewide cross-sectional study. \*Medicine, 99\*(27), e21016. <https://doi.org/10.1097/MD.0000000000021016>4. Sun Life. (n.d.). Life insurance adult build tables. Sun Life. Retrieved October 14, 2024, from <https://suncentral.sunlife.ca/en/client-service/new-business-and-underwriting/applying-and-underwriting/life-insurance-underwriting/life-insurance-adult-build-tables/>