REPORT

UNIVERSITY OF DERBY

STATISTICAL TECHINQUES

7MA505

STUDENT ID: 100757077

---

title: "Statistical Techniques Report"

author: '100757077'

date: "2024-12-10"

output: word\_document

---

Introduction

This report focuses on analyzing the key factors that influence health insurance charges based on statistical tests performed on the dataset. The data, sourced from Kaggle (https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset), includes 1,338 records with no missing or unclear values. It contains information about age, gender, BMI, number of children, smoking habits, and region, along with the insurance charges for each individual.

The report shows various statistical methods to investigate the relationships between these variables and medical costs. Special attention was given to smoking habits, age, and BMI, as they are expected to have the most noticeable impact on charges.

Through these tests, patterns and key relationships were identified, providing insights into how these factors interact and contribute to insurance costs. This report presents the results of these tests, supported by visualizations and detailed explanations, to highlight the practical implications of the findings.

```{r}

# Load necessary libraries

install.packages("dplyr")

install.packages("car")

install.packages("effsize")

install.packages("ggplot2")

install.packages("reshape2")

install.packages("pscl")

install.packages("pROC")

install.packages("dplyr")

library(pscl)

library(ggplot2)

library(car)

library(effsize)

library(reshape2)

library(pROC)

library(dplyr)

```

```{r}

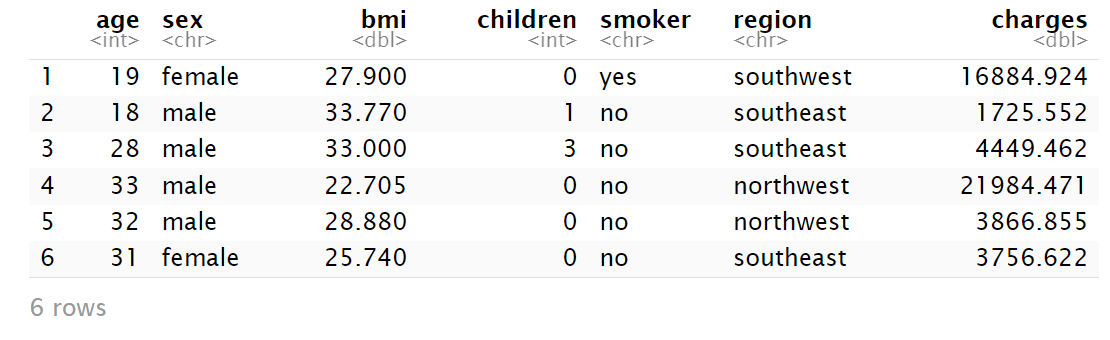
# Load the dataset

insurance <- read.csv("insurance.csv")

# Display the first few rows of the dataset

head(insurance)

```

 Figure 1: Dataset Display

The dataset captures demographic, lifestyle, and cost-related features for health insurance analysis.

The charges column is the primary target variable and its relationship with other features will make a difference.

Categories like smoker, sex and region are categorical variables, while age, bmi, children and charges are numerical.

```{r}

#Summary Statistics

# Summary statistics for the entire dataset

summary(insurance)

```

Analysis of Summary Output

The population is mostly middle-aged, with a median of about 39 years. The average age is also around 39, showing a fairly even distribution.

Most people are in the "overweight" category, with a median of 30.4 and an average of 30.66, which are quite similar.

Most people have one child, with a median of 1 and a little variation.

Smoker category is a yes/no category showing whether a person smokes or not.

The Sex variable shows if a person is male or female.

Region is a geographical variable, with categories for different regions: southwest, southeast, northwest, and northeast.

There is a wide range in charges, with the average being much higher than the median, indicating that some people have very high costs.

```{r}

# Histogram of the 'charges' column

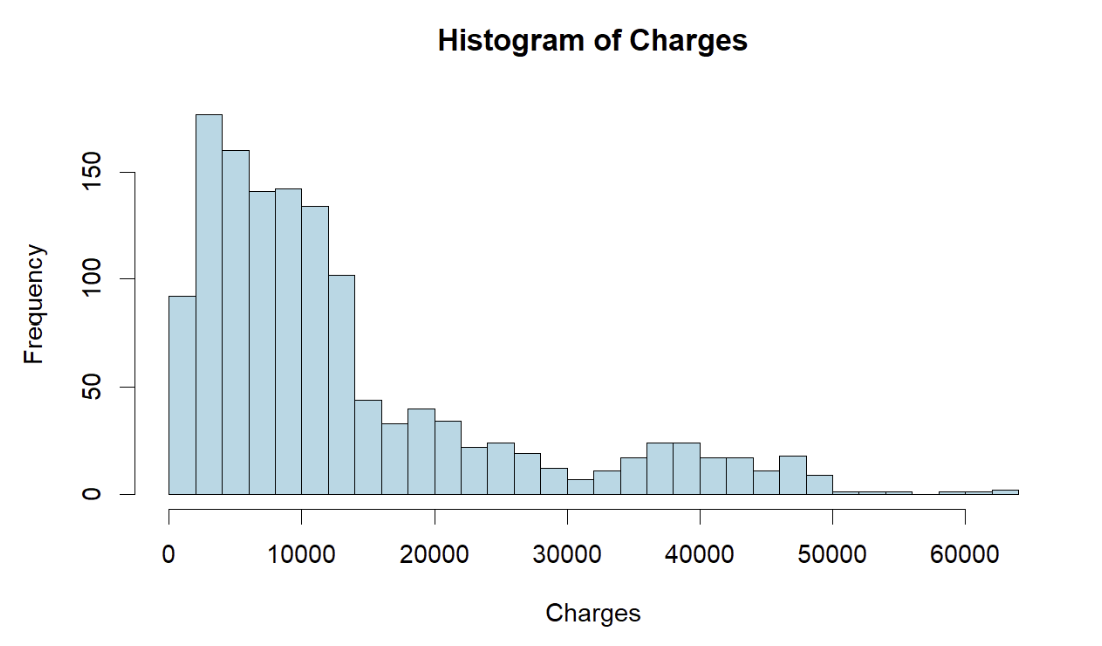
hist(insurance$charges, main = "Histogram of Charges", xlab = "Charges", col = "lightblue", breaks = 30)

# Q-Q plot for normality

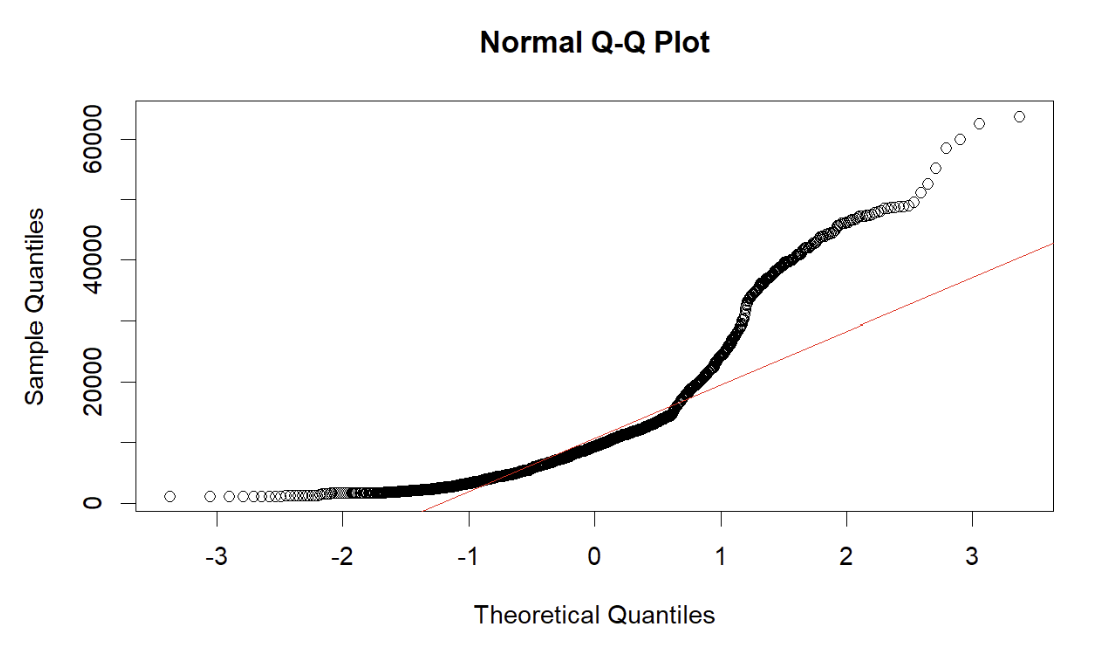
qqnorm(insurance$charges)

qqline(insurance$charges, col = "red")

```

 Figure 2: Histogram of Charges

The histogram displays a right-skewed distribution, where the majority of the charges are clustered around lower values. This indicates the presence of outliers or extreme values, likely affecting the mean.

 Figure 3: Normal Q-Q Plot

The Q-Q plot shows that the data deviates from a normal distribution. Most of the points follow a curved pattern, especially in the ends, suggesting the data is not normally distributed.

```{r}

#Descriptive statistics grouped by gender

insurance %>%

group\_by(sex) %>%

summarise(

Mean\_Charges = mean(charges, na.rm = TRUE),

Median\_Charges = median(charges, na.rm = TRUE),

Mean\_BMI = mean(bmi, na.rm = TRUE),

Median\_BMI = median(bmi, na.rm = TRUE),

Count = n()

)

# Descriptive statistics grouped by smokers and non-smokers

insurance %>%

group\_by(smoker) %>%

summarise(

Mean\_Charges = mean(charges, na.rm = TRUE),

Median\_Charges = median(charges, na.rm = TRUE),

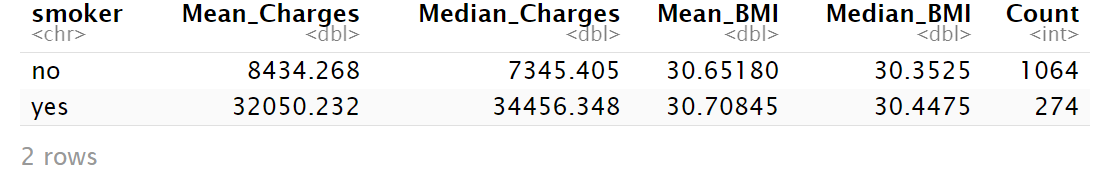
Mean\_BMI = mean(bmi, na.rm = TRUE),

Median\_BMI = median(bmi, na.rm = TRUE),

Count = n()

)

```

 Figure 4: Datasets Statistics Group by Smoking

Statistics Grouped by Gender

There are very few differences between genders when it comes to insurance charges. The data shows that gender doesn’t have a significant impact on the cost of insurance.

Statistics Grouped by Smoking

Smoking has a clear and strong connection to higher insurance charges. Smokers pay noticeably more than non-smokers. Interestingly, BMI is quite similar between smokers and non-smokers, which suggests that the higher costs for smokers are more related to smoking-related health risks rather than differences in BMI.

```{r}

#Correlation Analysis

# Calculate correlations between numeric variables

cor\_matrix <- cor(insurance %>% select(age, bmi, children, charges))

# Display correlation matrix

print(cor\_matrix)

# Heatmap of correlations

heatmap\_data <- melt(cor\_matrix)

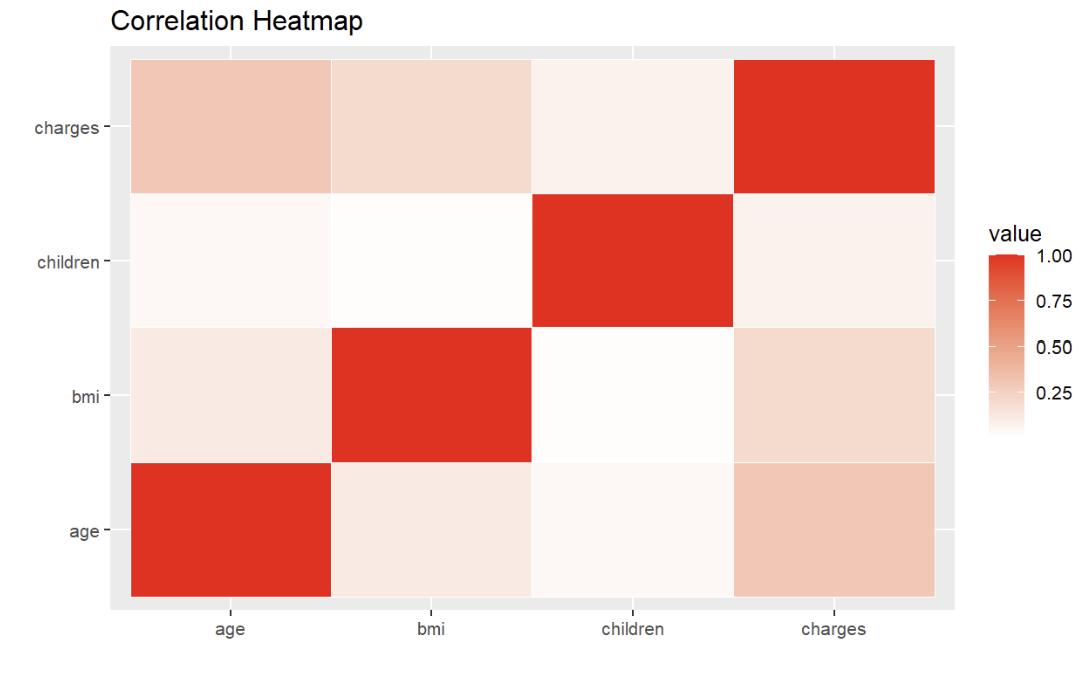
ggplot(heatmap\_data, aes(x = Var1, y = Var2, fill = value)) +

geom\_tile(color = "white") +

scale\_fill\_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +

labs(title = "Correlation Heatmap", x = "", y = "")

```

 Figure 5: Correlation Heatmap

Correlations with charges

Among all variables, age has the strongest correlation with insurance charges, although the relationship is still moderate. BMI and the number of children show weaker correlations with charges, suggesting that these factors may not be the main contributors to insurance costs when compared to stronger factors like smoking.

One interesting observation from the correlation heatmap is the relationship between age and BMI. While their correlation isn’t particularly strong, they tend to work together in influencing charges. This means that as people get older, their BMI might increase, and the combined effect of age and BMI contributes to higher insurance costs.

```{r}

# Scatterplots for relationships

ggplot(insurance, aes(x = children, y = charges, color = smoker)) +

geom\_point(alpha = 0.5) +

geom\_smooth(method = "loess", se = FALSE) +

labs(title = "Scatterplot: Children vs Charges by Smoking Status", x = "Children", y = "Charges") +

theme\_minimal()

ggplot(insurance, aes(x = region, y = charges, fill = region)) +

geom\_boxplot(alpha = 0.7) +

facet\_grid(smoker ~ sex) +

labs(

title = "Combined Effect of Region, Smoking, and Gender on Charges",

x = "Region",

y = "Charges",

fill = "Region"

) +

theme\_minimal() +

theme(strip.text = element\_text(size = 10, face = "bold"))

ggplot(insurance, aes(x = age, y = charges, color = bmi)) +

geom\_point(alpha = 0.6) +

scale\_color\_gradient(low = "blue", high = "red") +

facet\_wrap(~ smoker) +

labs(title = "Effect of Smoking on Age and BMI vs Charges Relationship",

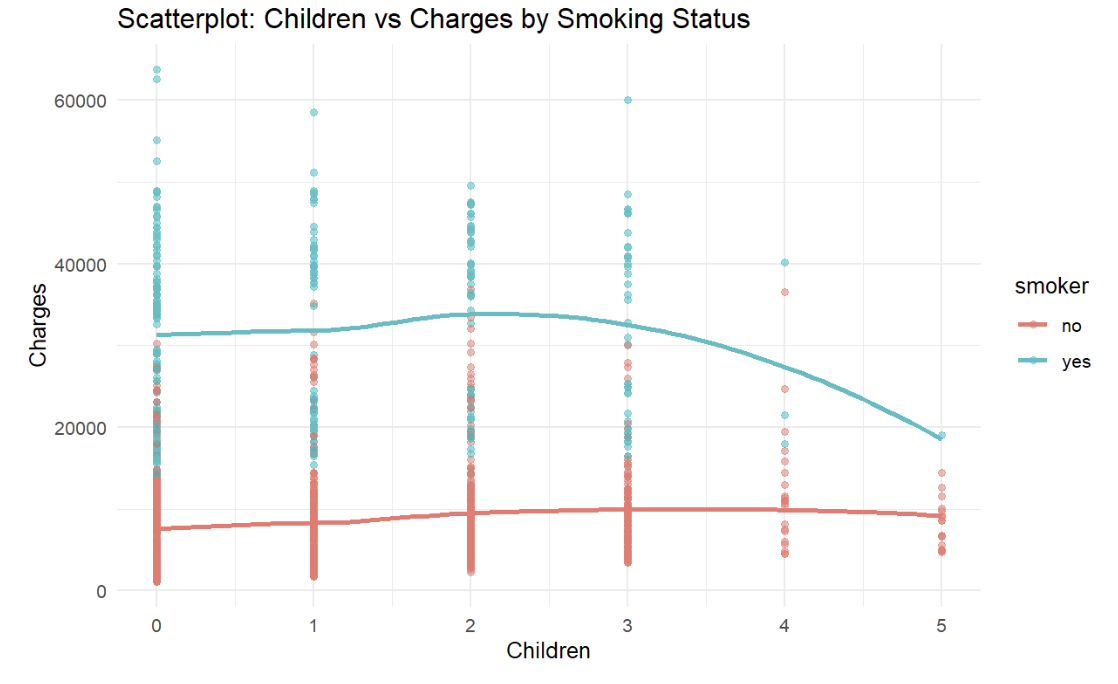
x = "Age",

y = "Charges",

color = "BMI") +

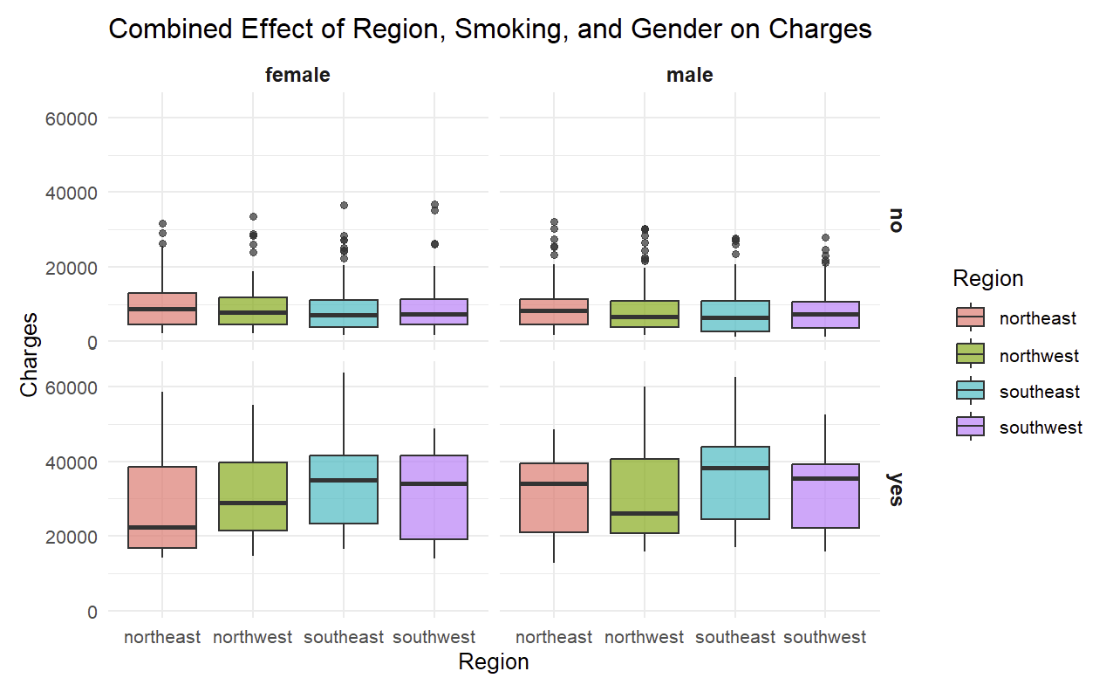
theme\_minimal()

```

 Figure 6: Children vs Charges by Smoking Status

1. Children vs. Charges by Smoking Status

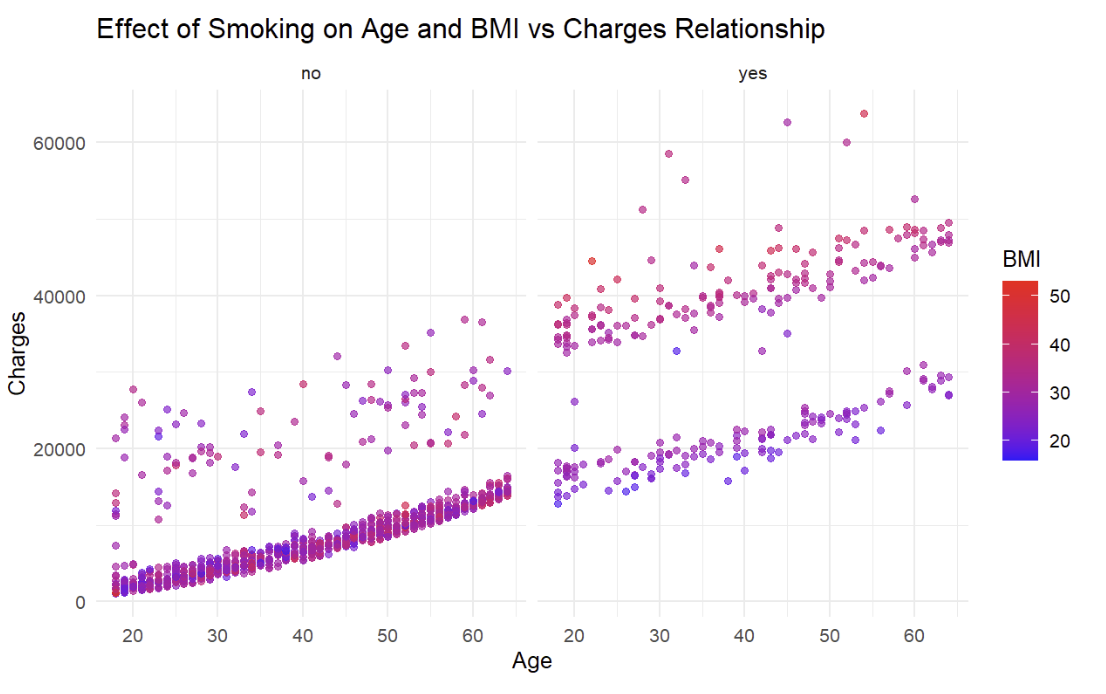
For smokers, the number of children doesn’t seem to affect charges much, but their overall charges are much higher than those of non-smokers. For non-smokers, having more children makes little difference in their charges. Overall, smoking has a much bigger impact on costs than the number of children.

 Figure 7: Effect of Region, Smoking, and Gender on Charges

2. Combined Effect of Region, Smoking, and Gender on Charges

The Southeast region has the highest charges for both smokers and non-smokers, while the Northwest and Southwest have lower costs. Male smokers in the Southeast face the highest charges, and female non-smokers in the Northwest have the lowest.

Smoking is the main factor driving up costs, but regional differences also play a role. These differences might be due to variations in healthcare costs, lifestyle, or health risks. Gender also has a small effect, male smokers tend to pay slightly more than female smokers, possibly due to gender specific health issues or habits.

 Figure 8: Effect of Smoking on Age and BMI vs. Charges

3. Effect of Smoking on Age and BMI vs. Charges

Smokers have much higher charges than non-smokers, no matter their age or BMI. Among smokers, charges go up with both age and BMI, showing how these factors combine to increase costs. For non-smokers, charges do rise with age, but BMI doesn’t have the same impact as it does for smokers. This suggests that smoking makes BMI a more important factor in costs.

To conclude, smoking, along with region, BMI, and gender, causes big differences in insurance charges. The Southeast stands out as the region with the highest costs, especially for male smokers.

Additionally, high BMI leads to even higher charges for smokers, showing how smoking and BMI together can strongly increase costs.

```{r}

#Non-Parametric Test: Mann-Whitney U Test

# Mann-Whitney U Test (non-parametric)

mann\_whitney <- wilcox.test(charges ~ smoker, data = insurance)

# Display the result

print(mann\_whitney)

```

Test Statistic (W):

The Wilcoxon test statistic is W = 7403, which represents the sum of ranks for one of the groups. While the exact value itself doesn’t provide much insight, it is used to calculate the p-value for the test.

P-Value:

The p-value is extremely small. This provides strong evidence to reject the null hypothesis, meaning the distributions of charges for smokers and non-smokers are not the same.

Alternative Hypothesis:

The alternative hypothesis states that the median charges for smokers and non-smokers are not equal. The test results support this hypothesis, confirming that there is a significant difference between the two groups.

Conclusion:

These results align with earlier findings, such as the boxplot analysis. Smokers not only have higher average charges but also a significantly different overall distribution compared to non-smokers.

```{r}

# Kruskal-Wallis Test

# Testing whether charges differ by region

# Perform the Kruskal-Wallis test

kruskal\_test <- kruskal.test(charges ~ region, data = insurance)

# Print the result

kruskal\_test

```

The Kruskal-Wallis test showed no significant differences in median charges between regions (p > 0.05). This means that, based on this dataset, location does not seem to have a big impact on medical charges. Instead, factors like smoking and BMI appear to play a much larger role in the variation of costs. While it might be surprising since healthcare access and costs can vary by region, this result shows that personal health and lifestyle have a stronger influence than where someone lives. Future research could look into socioeconomic or environmental factors within regions for a more detailed understanding.

```{r}

# Chi-Square Test for Independence

# Testing the relationship between smoking status and region

# Create a contingency table

table\_smoker\_region <- table(insurance$smoker, insurance$region)

# Perform the Chi-Square test

chi\_square\_test <- chisq.test(table\_smoker\_region)

# Print the result

chi\_square\_test

```

The p-value of 0.06172 suggests that we fail to reject the null hypothesis. This means there isn’t enough evidence to conclude a statistically significant association between the variables: smoker and region.

```{r}

#Variance Equality: Levene's Test

# Levene's test for equal variances

levene\_test <- leveneTest(charges ~ smoker, data = insurance)

print(levene\_test)

```

The F-statistic is 332.61, which is very large, indicating a significant difference in the variances between the two groups. The p-value is < 2.2e-16, which is extremely small. This suggests that we reject the null hypothesis of equal variances.

Levene's test assumes that the variances of the two groups are equal. Given the very low p-value, we reject this null hypothesis and conclude that the variances in charges for smokers and non-smokers are significantly different.

```{r}

#Effect Size: Cohen's d

# Calculate Cohen's d for charges between smokers and non-smokers

effect\_size <- cohen.d(charges ~ smoker, data = insurance)

print(effect\_size)

```

The negative sign in Cohen's d indicates that the Southwest region has lower charges compared to the Southeast region. The confidence interval supports this by confirming that the effect size is large and consistently negative, meaning the Southeast region consistently has higher charges than the Southwest region.

```{r}

#Regression Analysis

#Basic Linear Regression

# Simple linear regression: Predicting charges

model <- lm(charges ~ age + bmi + smoker, data = insurance)

summary(model)

#With Interaction Terms

# Linear regression with interaction: BMI and Smoking

interaction\_model <- lm(charges ~ bmi \* smoker, data = insurance)

summary(interaction\_model)

```

Model 1 shows that age, BMI, and smoking status are all significant predictors of insurance charges, with smoking having the largest impact on charges.

Model 2 highlights an important interaction: BMI has a stronger effect on charges for smokers, adding an extra $1,389.76 per unit of BMI.

Conclusion:

While BMI and age are significant predictors of charges on their own, the interaction between smoking and BMI suggests that higher BMI results in a more significant increase in charges for smokers compared to non-smokers. This emphasizes the importance of considering how multiple factors combine to affect charges.

```{r}

#Confidence Intervals

# Confidence interval for the mean charges of smokers

smoker\_ci <- t.test(charges ~ smoker, data = insurance)$conf.int

print(smoker\_ci)

```

The confidence interval shows that it is 95% confident that the true difference in mean charges between smokers and non-smokers falls between $25,034.71 and $22,197.21. The negative values indicate that, on average, smokers incur higher medical charges than non-smokers, which is consistent with the previous statistical tests.

```{r}

# Example: Compare charges to a hypothetical median value

median\_charges <- 10000 # hypothetical value

wilcox.test(insurance$charges, mu = median\_charges,

alternative = "greater", conf.int = TRUE)

wilcox.test(charges ~ smoker, data = insurance, conf.int = TRUE)

```

1. Wilcoxon Signed Rank Test (One-Sample Test)

The p-value is less than the significance level, which means we reject the null hypothesis. The data provides strong evidence that the median charges in the dataset are significantly greater than 10,000.

2. Wilcoxon Rank Sum Test (Mann-Whitney U Test)

The p-value is extremely small, indicating a highly significant difference in charges between smokers and non-smokers. The negative difference confirms that non-smokers incur significantly lower charges compared to smokers. Additionally, the confidence interval does not include zero, further supporting the conclusion of a significant difference between the two groups.

```{r}

# Convert smoker to a binary factor

insurance$smoker <- factor(insurance$smoker, levels = c("no", "yes"))

# Inspect dataset structure

str(insurance)

# Logistic regression to predict smoker status

logit\_model <- glm(smoker ~ age + bmi + charges,

data = insurance, family = binomial)

# Summarize the model

summary(logit\_model)

# Odds Ratios for Interpretability

exp(coef(logit\_model))

# Calculate pseudo-R2

library(pscl)

pR2(logit\_model)

```

The logistic regression model predicts smoking status based on age, BMI, and medical charges.

Intercept

This represents the log-odds of being a smoker when all predictors are zero, which indicates a high likelihood of smoking at baseline.

Age

Each year increase in age reduces the likelihood of smoking, showing a strong negative relationship (p = 3.02e-14).

BMI

A higher BMI is associated with a lower likelihood of smoking (p = 6.60e-16).

Charges

Higher medical charges slightly increase the likelihood of smoking (p < 2e-16).

Model Fit

The McFadden R2 is 0.77, indicating a strong fit, explaining 77% of the variability in smoking behavior.

To conclude, the model suggests that age and BMI are negatively associated with smoking, while medical charges have a small positive association with the likelihood of being a smoker.

Conclusion

The findings confirm that smoking is the most significant factor influencing health insurance charges, with smokers incurring much higher costs than non-smokers. Other variables like age and BMI also contribute significantly, with higher values generally leading to increased costs. Differences between regions, particularly between the Southeast and Southwest, were statistically significant. The analysis also highlighted how BMI impacts charges more heavily for smokers compared to non-smokers.

The statistical techniques used, such as the Wilcoxon Rank Sum Test and Kruskal-Wallis test, were particularly useful for dealing with imbalanced data. These tests helped to reveal significant differences between groups, even when the data didn’t follow normal distributions.

Future Improvements

There are several opportunities for further research:

- Adding more variables, such as data on exercise, diet, or medical history, could provide a fuller picture of the factors affecting charges.

- Exploring how charges change over time, if temporal data becomes available, could reveal trends and long-term effects.

- Conducting separate analyses for each region could help better understand regional differences and allow for more tailored predictions.

- Using clustering methods to group individuals with similar characteristics could lead to personalized recommendations or more precise insights.