Final Report

Name: J.M. Imtinan Uddin
ID: CBJ988
Course No: Math 5130

Title: Statistical Exploratory Analysis and Predictive Modeling for Healthcare Insurance Costs

Introduction

The healthcare sector is a cornerstone of societal well-being, serving as a critical indicator of a nation's development and quality of life. At the heart of this sector is the notion of accessibility and affordability of healthcare services, which are essential for ensuring the overall health and well-being of the population. In many countries, healthcare insurance is a fundamental mechanism that enables individuals to access necessary medical services without facing prohibitive costs. However, the landscape of healthcare insurance is complex and dynamic, influenced by a wide array of factors ranging from individual health conditions to macroeconomic trends. In recent years, the cost of healthcare has been a topic of increasing concern globally. With medical advancements and rising operational costs, the expenses associated with healthcare provision have escalated significantly. This increase directly impacts the cost of health insurance, making it a key area of concern for individuals, families, healthcare providers, and policymakers. The interplay between the costs of healthcare services and the premiums charged by insurance companies is intricate and multifaceted, necessitating a thorough and nuanced understanding. The rising healthcare costs have multifarious implications. For individuals and families, these costs can mean the difference between receiving necessary medical care and foregoing it due to financial constraints. For healthcare providers and insurers, they influence operational decisions, risk assessment, and policy development. At a broader level, for policymakers and government bodies, understanding these costs is crucial for designing effective healthcare systems that are both high-quality and financially sustainable.

Despite its importance, there is a knowledge gap in understanding the specific factors that drive healthcare insurance costs. Various elements, such as demographic shifts, changes in healthcare policy, advancements in medical technology, and evolving disease patterns, all contribute to this complexity. Additionally, individual factors like age, gender, lifestyle choices, and pre-existing health conditions play a significant role in determining insurance premiums. This project aims to address this gap by employing a data-driven approach to identify and analyze the key factors influencing healthcare insurance costs.

In sum, the objective of this project is not only to illuminate the factors affecting healthcare insurance costs but also to contribute to a more equitable and efficient healthcare system. By providing insights into the drivers of insurance costs, this study aims to aid stakeholders in making informed decisions and devising strategies that can optimize costs without compromising the quality of healthcare.

Data Collection and Preprocessing

In our study, we utilized a dataset named "insurance.csv" from kaggle which encompasses a range of variables potentially influencing healthcare insurance costs. The dataset, comprising 1338 observations, includes variables such as **age, sex, body mass index (BMI), number of children, smoking status, region, and insurance charges**. Initially, the dataset was loaded into R, and a preliminary inspection was conducted to understand its structure. This inspection revealed a mix of numerical and categorical variables.

To gain initial insights, we generated summary statistics for the dataset. This analysis provided key information on the distribution of each variable, including their central tendency, standard deviation, and dispersion. For instance, it highlighted the age range of the insured individuals and the average BMI. Detailed descriptive statistics were then calculated for numerical variables to further understand their characteristics. Additionally, we performed a missing value analysis, which is a crucial step in data preprocessing. Identifying and addressing missing data is essential to ensure the integrity and reliability of the analyses that follow. Overall, this preliminary phase laid a solid foundation for more advanced statistical analyses. It highlighted the importance of careful data preparation, including the transformation and cleaning of categorical variables and addressing any missing or anomalous data points.

	age	sex	bmi	children	smoker	region	charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622

```
## $age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   27.00   39.00   39.21   51.00   64.00
##
## $sex
## NULL
##
## $bmi
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.96   26.30   30.40   30.66   34.69   53.13
##
## $children
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.095   2.000   5.000
##
## $smoker
## NULL
##
## $region
## NULL
##
## $charges
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4740    9382   13270   16640   63770
```

```r
print(sd(data$charges))
```

```
## [1] 12110.01
```

```r
print(sd(data$bmi))
```

```
## [1] 6.098187
```

```r
print(sd(data$age))
```

```
## [1] 14.04996
```

Figure 1: 1st 5 rows of the dataset and descriptive summary

```r
# Assuming your data frame is named 'data'
data$sex <- ifelse(data$sex == "male", 1, 0)  # Assign 1 to male and 0 to female
data$smoker <- ifelse(data$smoker == "yes", 1, 0)  # Assign 1 to smoker and 0 to non-smoker
# Assuming your data frame is named 'data'
# Drop 'column_to_drop'
# One-hot encoding for 'region'
region_dummies <- model.matrix(~ region - 1, data=data)  # The '- 1' removes intercept column
colnames(region_dummies) <- gsub("region", "region", colnames(region_dummies))

# Combine with original data
data <- cbind(data, region_dummies)

data <- subset(data, select = -c(age_group, bmi_category,region))
```

```r
# Checking for missing values
missing_values_count <- sum(is.na(data))
print(missing_values_count)
```

```
## [1] 0
```

Figure 2: Processing categorical variables into numerical using one hot encoding

Exploratory Data Analysis(EDA)

The EDA phase began with the loading of the "insurance.csv" dataset into R, followed by an inspection of its structure. The dataset comprised 1338 observations across seven variables: age, sex, BMI, number of children, smoker status, region, and insurance charges. The variables were a mix of numerical (age, BMI, children, charges) and categorical (sex, smoker, region) types. A summary of the dataset revealed key descriptive statistics such as the minimum, median, mean, and maximum values for each variable. For example, the age of individuals in the dataset ranged from 18 to 64 years, and the mean BMI was approximately 30.66.

To delve deeper into the data, we conducted descriptive statistical analysis for numerical variables. It was observed that the age of insured individuals spanned from 18 to 64, with a

median of 39. Similarly, the BMI ranged from 15.96 to 53.13, indicating a diverse dataset in terms of body weight. We also verified the dataset for missing values and found none, ensuring the integrity of our subsequent analyses. The EDA was significantly enhanced by visualizations created using ggplot2 in R. We generated histograms for 'age', 'BMI', 'children', and 'charges', offering a visual representation of their distributions. Bar plots for the 'smoker' variable and the count of males and females provided insights into the categorical data distributions.

Scatter plots were used to explore relationships between age, BMI, and insurance charges, while box plots compared charges across smoker status, number of children, sex, and region. These visualizations were pivotal in identifying potential patterns and anomalies in the data.

For instance, the scatter plot of age versus charges showed how insurance charges varied with age, and the box plot of smoker status against charges highlighted the impact of smoking on insurance costs. In summary, the EDA provided a comprehensive understanding of the dataset, laying a solid foundation for more advanced statistical analysis and modeling. The combination of descriptive statistics and visualizations facilitated a deeper exploration of the data, revealing critical insights that would inform the subsequent stages of our research.

Figure 3 presents a series of six plots that provide a statistical summary of a healthcare insurance dataset. The first plot is a histogram detailing the age distribution of individuals, indicating a diverse age range without significant outliers. Next, a histogram of Body Mass Index (BMI) values shows a **bell-shaped curve, typical of a normal distribution**, suggesting that most individuals' BMIs are concentrated around the median. The third histogram illustrates the number of children among the insured individuals, with the majority having none or one child, and fewer having larger families. A bar plot comparing smokers to non-smokers reveals a greater prevalence of non-smokers within the dataset. The fifth plot, a histogram of insurance charges, is **right-skewed**, showing that lower charges are more common among the population, while high charges are less frequent. Finally, a bar plot displaying the count of males and females shows a nearly balanced distribution between the two genders, with a marginal predominance of males. Collectively, these plots encapsulate key demographic and insurance-related characteristics of the dataset, which could be critical for subsequent analytical and predictive modeling efforts.

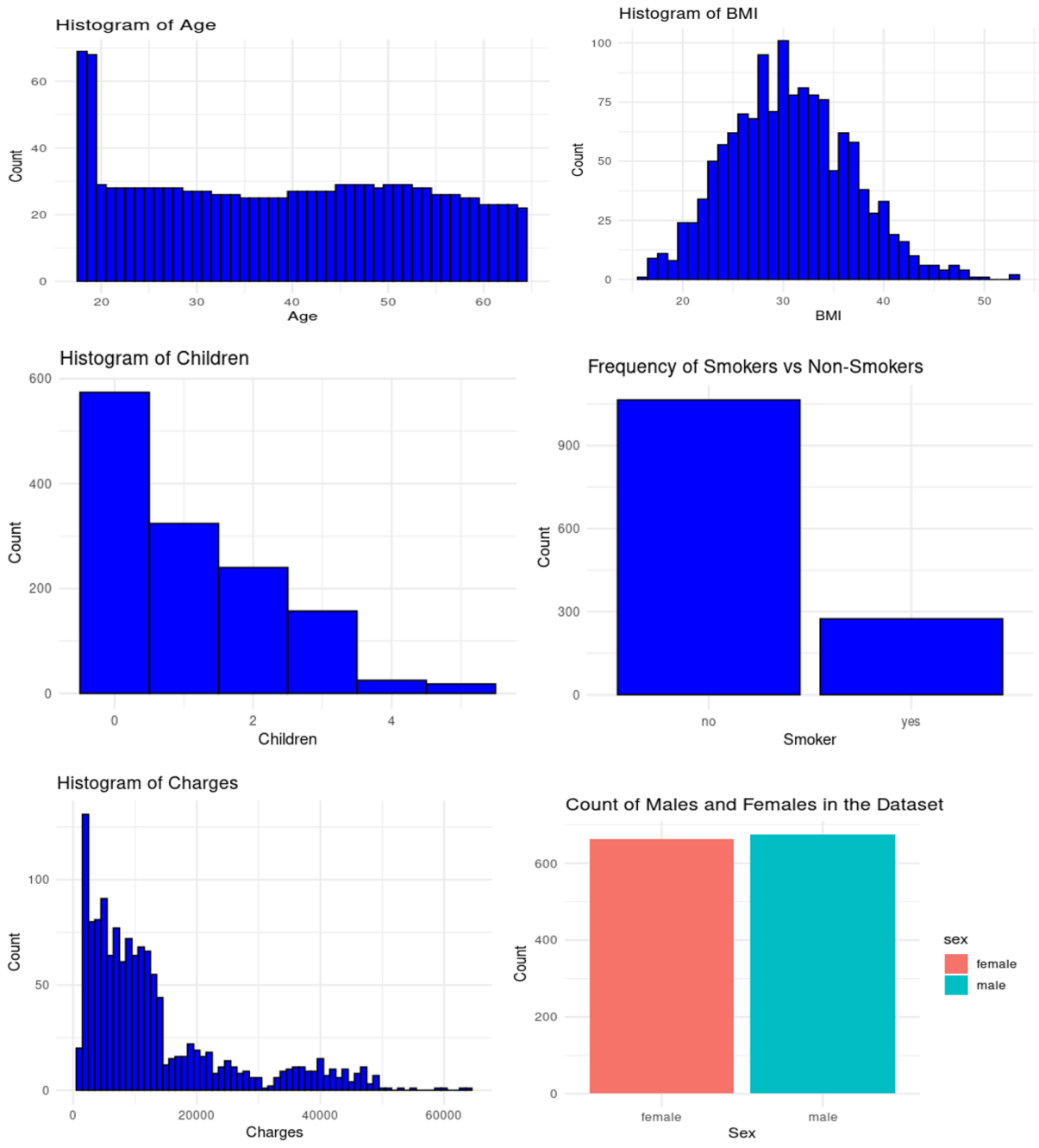

Figure 3: Histogram of variables

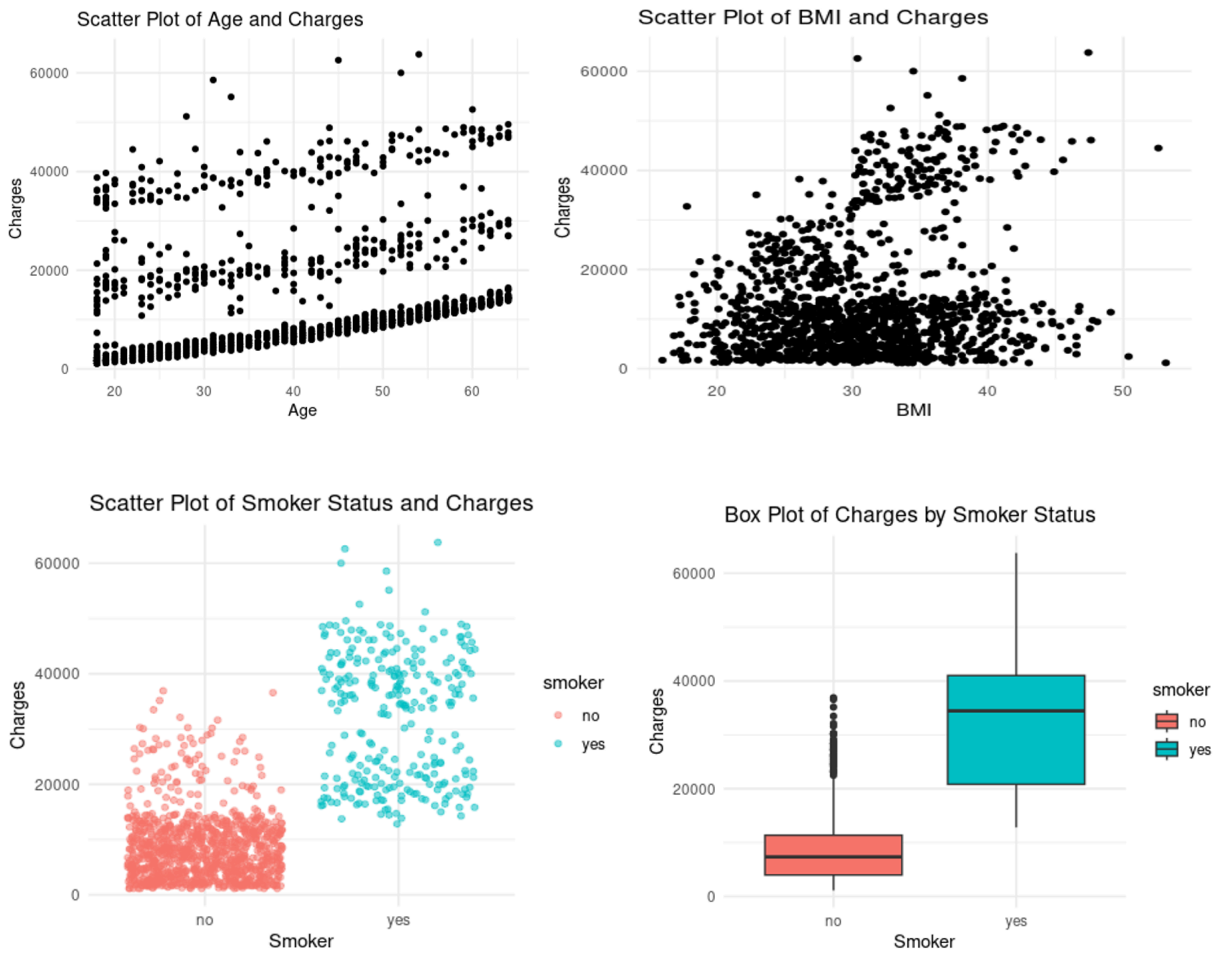

Figure 4: Scatter plot of variable vs Charges

The provided Figure 4 contains four plots that explore the relationships between insurance charges and other variables within a healthcare dataset. The first plot is a scatter plot of age against charges, showing a spread of charges at each age with a slight increase in variability as age increases. There doesn't appear to be a clear trend or pattern, indicating that while age might affect charges, it is not the sole determinant. The second plot is a scatter plot of Body Mass Index (BMI) against charges, which reveals a more discernible pattern. As BMI increases, there appears to be a trend of increasing charges, particularly noticeable for higher BMI values. This suggests a possible correlation where individuals with higher BMI may incur higher insurance charges. The third plot is a scatter plot of smoker status against charges, color-coded to distinguish smokers from non-smokers. This plot shows a clear distinction between the two groups: non-smokers tend to have lower insurance charges, while smokers have a wider range and generally higher charges, indicating that smoking status is a significant factor in insurance charges. The fourth plot is a box plot of charges by smoker status, providing a different view of the same relationship shown in the scatter plot. It confirms that smokers tend to have higher median charges than non-smokers, with the interquartile range for smokers being substantially higher as well. This box plot also shows outliers in insurance charges, particularly within the

smoker category. These plots collectively highlight the importance of BMI and smoker status as factors that can influence the cost of healthcare insurance charges. The visualizations underscore the higher charges associated with smokers and suggest that higher BMI is linked with increased charges, all valuable insights for insurance cost modeling and risk assessment.

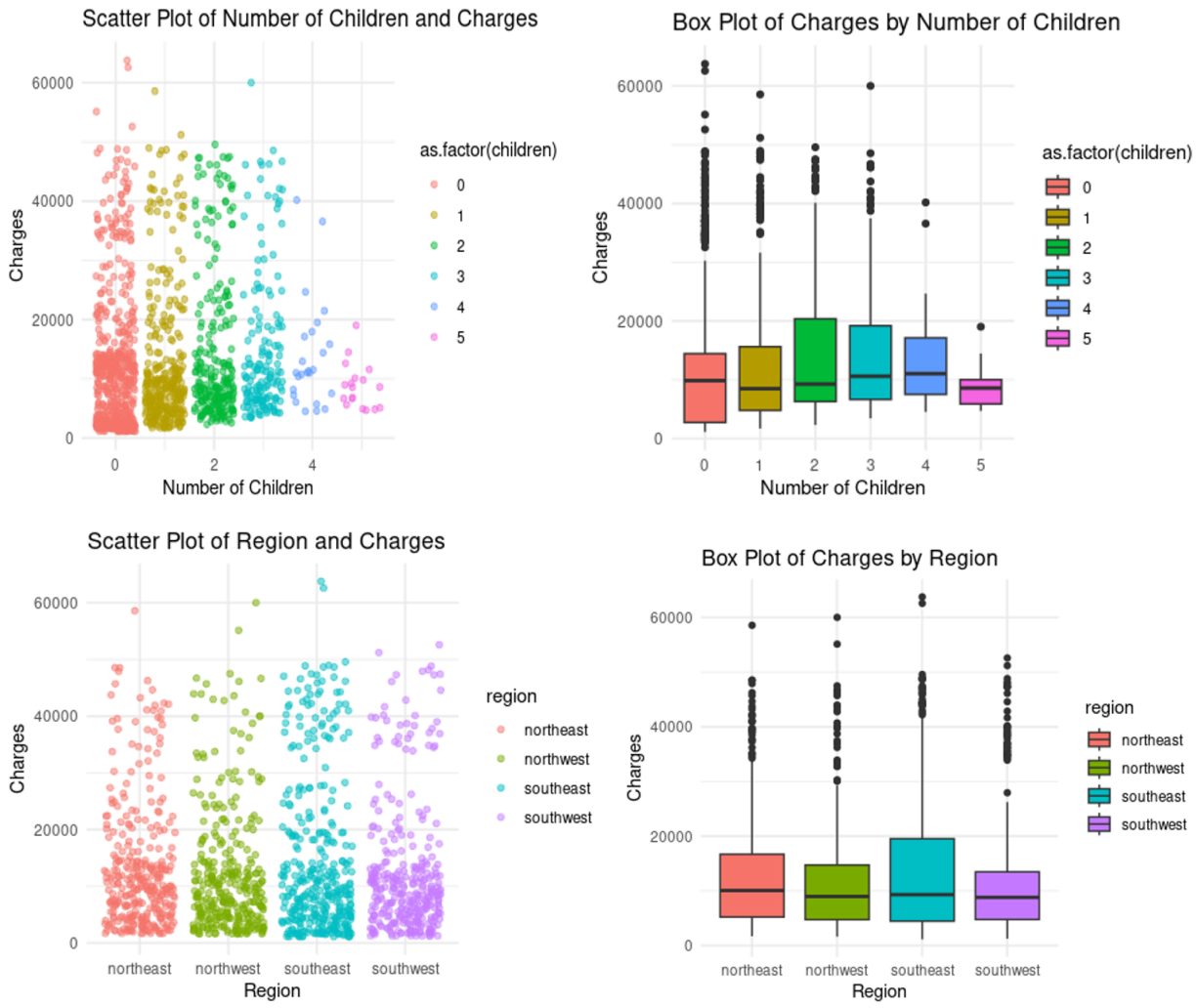

Figure 5: Scatter plot of variable vs Charges

Figure 5 comprises four plots that examine the relationships between insurance charges and two categorical variables: the number of children and the region of the policyholders. The first plot is a scatter plot showing the number of children on the x-axis and insurance charges on the y-axis. The data points are color-coded for the number of children ranging from 0 to 5. The plot

indicates that there is no clear increase or decrease in charges based on the number of children, but there is a noticeable concentration of data points at the lower end of the charge scale regardless of the number of children. The second plot is a box plot of insurance charges categorized by the number of children. Each category from 0 to 5 children is represented by a different color. The median charges do not show a consistent trend with the increasing number of children, and there are outliers across all categories, suggesting that having more children does not consistently relate to higher or lower charges. The third plot is a scatter plot of the region versus charges, with each region denoted by a different color. The regions included are northeast, northwest, southeast, and southwest. The distribution of charges seems to be fairly uniform across all regions without a discernible pattern, indicating that the region may not be a strong predictor of insurance charges. The fourth plot is a box plot that represents the insurance charges by region. Similar to the scatter plot, the box plot shows that median charges across different regions do not vary dramatically. However, there are outliers present in each region, particularly in the southeast, suggesting some regional variations in the higher end of the charges. These plots together suggest that while the number of children and region may have some influence on insurance charges, they are not the primary drivers of the cost. The data indicates that other factors at play may have a more significant impact on insurance charges.

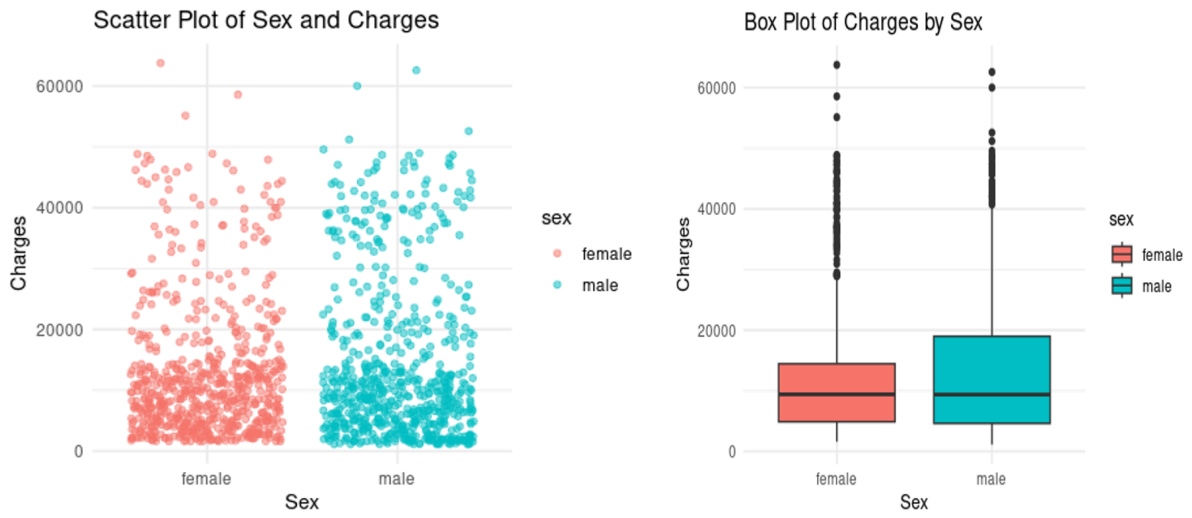

Figure 6: Scatter plot of variable vs Charges

Figure 6 presents features two plots that analyze the relationship between sex and insurance charges.

The left plot is a scatter plot with sex on the x-axis and insurance charges on the y-axis. The data points are color-coded, with one color representing female and another representing male. This plot indicates a wide range of charges for both sexes, with a considerable overlap between males and females. It does not show any clear pattern or significant differences between the charges for the two sexes. On the right is a box plot of charges by sex. The plot divides charges

into female and male categories, displaying the median, quartiles, and outliers for each. The median insurance charges for both sexes are represented by a line within the boxes. While both boxes have outliers on the higher charges end, the median and interquartile ranges appear to be quite similar for both sexes. This suggests that there are no significant differences in the central tendency of charges between females and males, though variability and extreme values exist in the data for both groups. Together, these plots suggest that while individual insurance charges vary, there is no obvious distinction between the sexes in terms of overall insurance charges. This information could be critical in understanding and ensuring gender equity in insurance pricing.

Hypothesis Testing

For column which has more than two groups ANOVA hypothesis test was used.
In our hypothesis testing alpha(significance level) was 0.05. A P-value less than alpha denotes a highly significant variable that rejects the null hypothesis where the null hypothesis assumes the variable is not significant.

Age

```
#Hypothesis
# Load necessary libraries
library(ggplot2)

# Categorizing age
data$age_group <- cut(data$age, breaks =
c(0, 30, 60, Inf), labels = c("Young",
"Middle-Aged", "Old"))

# ANOVA for age groups
anova_age <- aov(charges ~ age_group,
data = data)
print(summary(anova_age))
##               Df    Sum Sq   Mean Sq F
value Pr(>F)
## age_group      2 1.346e+10 6.729e+09
49.19 <2e-16 ***
## Residuals   1335 1.826e+11 1.368e+08
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
ggplot(data, aes(x = age_group, y =
charges)) +
    geom_boxplot() +
    ggtitle("Charges by Age Group") +
    xlab("Age Group") +
    ylab("Charges")
```

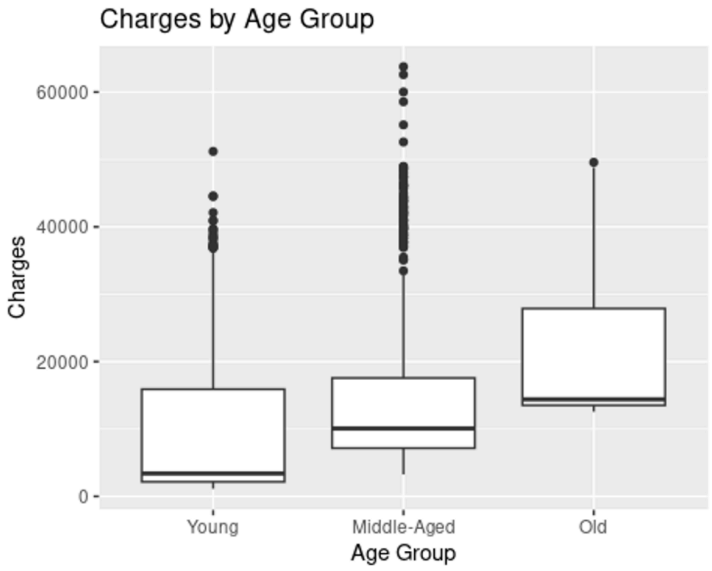

Figure 7: Hypothesis testing of Age

The R script conducts an ANOVA(more than two groups) to assess if there are significant differences in charges between three age groups: Young (0-30 years), Middle-Aged (31-60

years), and Old (61+ years). The ANOVA results indicate a significant effect of age group on charges ($p < 2e\text{-}16$), with an F value of 49.19, suggesting strong differences in charges between the groups. A boxplot is also created to visualize these differences in charges across the age categories. The significant p-value leads to the rejection of the null hypothesis, affirming that age has a statistically significant impact on charge.

SEX

```
# T-Test for sex
t_test_sex <- t.test(charges ~ sex, data
= data)
print(t_test_sex)
##
##  Welch Two Sample t-test
##
## data:  charges by sex
## t = -2.1009, df = 1313.4, p-value =
0.03584
## alternative hypothesis: true
difference in means between group female
and group male is not equal to 0
## 95 percent confidence interval:
##  -2682.48932   -91.85535
## sample estimates:
## mean in group female   mean in group
male
##             12569.58
13956.75
ggplot(data, aes(x = as.factor(sex), y =
charges)) +
    geom_boxplot() +
    ggtitle("Charges by Sex") +
    xlab("Sex") +
    ylab("Charges")
```

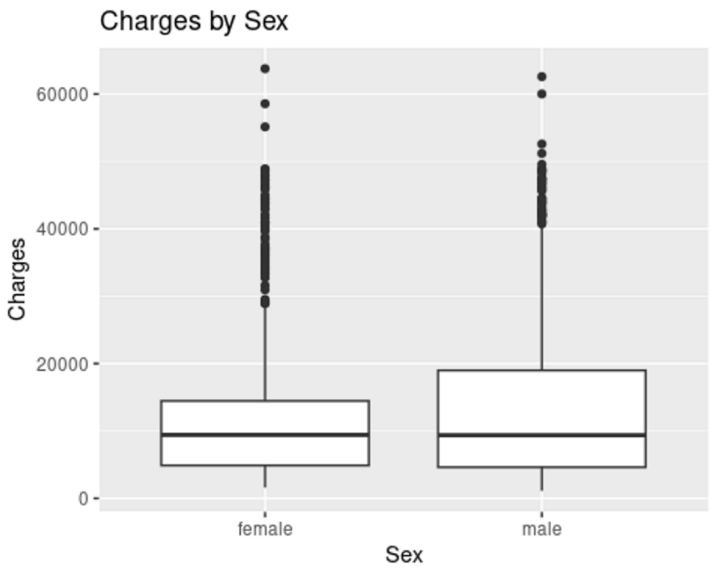

Figure 8: Hypothesis testing of Sex

The Welch Two Sample t-test output from the R script indicates a significant difference in charges between females and males, with a p-value of 0.03584 suggesting the difference is statistically meaningful. The test yields a t-value of -2.1009 and a 95% confidence interval for the mean difference in charges ranging from -2682.49 to -91.85. This means that males are likely charged more than females, with female charges averaging $12,569.58 and male charges averaging $13,956.75. A corresponding boxplot visualization would show this disparity in charges between the sexes.

BMI

```r
# Categorizing BMI
data$bmi_category <- cut(data$bmi,
breaks = c(0, 18.5, 25, 30, Inf),
labels = c("Underweight", "Normal",
"Overweight", "Obese"))

# ANOVA for BMI categories
anova_bmi <- aov(charges ~
bmi_category, data = data)
print(summary(anova_bmi))
##                 Df    Sum Sq   Mean
Sq F value Pr(>F)
## bmi_category     3 7.956e+09
2.652e+09    18.8  6e-12 ***
## Residuals    1334 1.881e+11
1.410e+08
## ---
## Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
ggplot(data, aes(x = bmi_category, y
= charges)) +
    geom_boxplot() +
    ggtitle("Charges by BMI
Category") +
    xlab("BMI Category") +
    ylab("Charges")
```

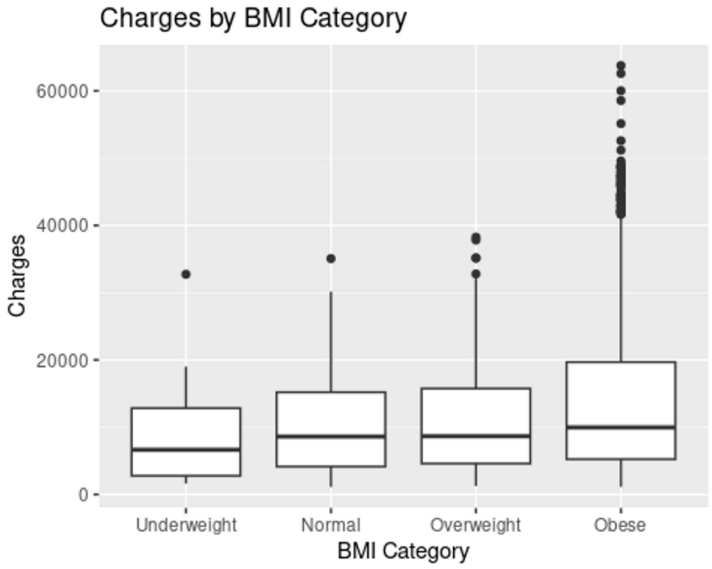

Figure 9: Hypothesis testing of BMI

The provided Figure 9 output details the results from an ANOVA test that evaluates the impact of BMI categories on medical charges. The BMI is categorized into four groups: Underweight, Normal, Overweight, and Obese, based on established BMI ranges. The ANOVA results show a significant difference in charges across these BMI categories with an F value of 18.8 and a highly significant p-value (6e-12), suggesting that the BMI category is a strong predictor of the variability in charges. A boxplot is suggested to visualize this relationship, displaying the spread and central tendencies of charges within each BMI category, which would further illustrate the disparities in charges related to different BMI classifications.

Number of Children

```
# ANOVA for number of children
anova_children <- aov(charges ~
as.factor(children), data = data)
print(summary(anova_children))
##                      Df     Sum
Sq    Mean Sq F value  Pr(>F)
## as.factor(children)     5
2.397e+09 479383343   3.297 0.00579
**
## Residuals            1332
1.937e+11 145403382
## ---
## Signif. codes:  0 '***' 0.001
'**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
ggplot(data, aes(x =
as.factor(children), y = charges))
+
    geom_boxplot() +
    ggtitle("Charges by Number of
Children") +
    xlab("Number of Children") +
    ylab("Charges")
```

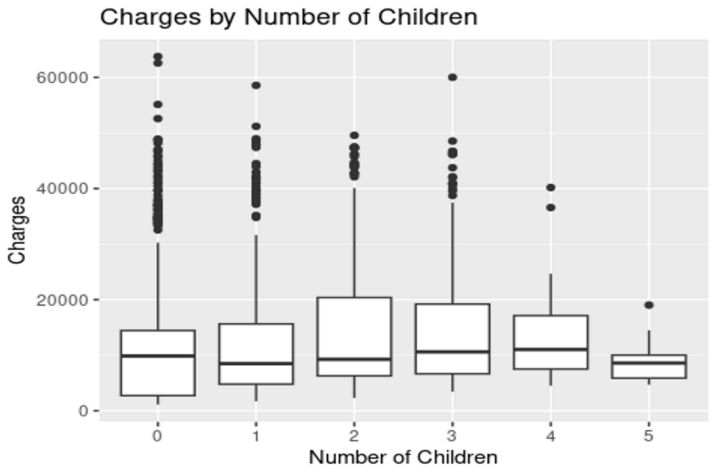

Figure 10: Hypothesis testing of Number of Children

The above R script conducts an ANOVA to investigate if the number of children a patient has is associated with differences in medical charges. The number of children variable is treated as a factor in the ANOVA, suggesting it's a categorical variable in this analysis. The ANOVA results reveal a significant effect with an F value of 3.297 and a p-value of 0.00579, indicating that there are statistically significant differences in charges among the different numbers of children categories. The boxplot visualization is intended to display these charge differences across the categories, providing a visual summary of the data distribution and central tendency for charges by the number of children.

Smoker Status

```
# T-Test for smoker status
t_test_smoker <- t.test(charges ~ smoker,
data = data)
print(t_test_smoker)
##
##  Welch Two Sample t-test
##
## data:  charges by smoker
## t = -32.752, df = 311.85, p-value <
2.2e-16
## alternative hypothesis: true
difference in means between group no and
group yes is not equal to 0
## 95 percent confidence interval:
##  -25034.71 -22197.21
## sample estimates:
##  mean in group no mean in group yes
##          8434.268         32050.232
ggplot(data, aes(x = as.factor(smoker), y
= charges)) +
    geom_boxplot() +
    ggtitle("Charges by Smoker Status") +
    xlab("Smoker Status") +
    ylab("Charges")
```

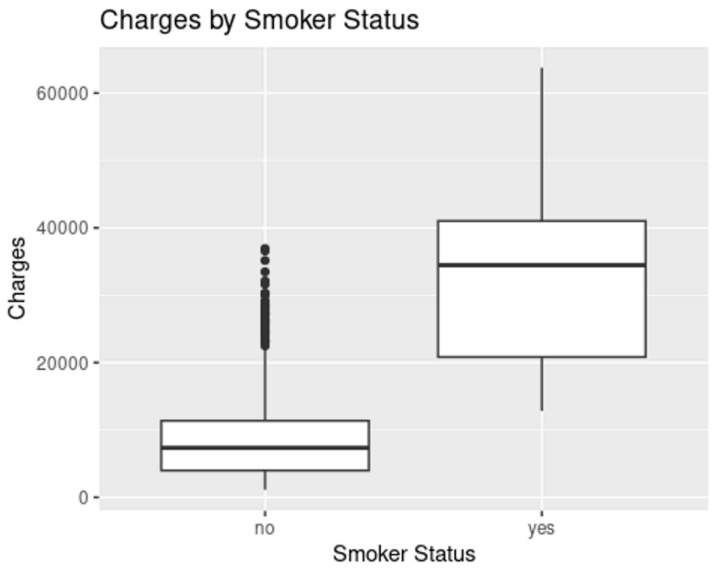

Figure 11: Hypothesis testing of Smoker Status

The above R script output provides the results of a Welch Two Sample t-test comparing medical charges between smokers and non-smokers. The t-test yields a very high t-value of -32.752 with degrees of freedom approximately 311.85, reflecting a significant difference in charges between the two groups. The extremely small p-value (< 2.2e-16) indicates this difference is highly statistically significant. The confidence interval for the difference in means between non-smokers and smokers ranges from -25034.71 to -22197.21, suggesting smokers have considerably higher charges, with mean charges for non-smokers at $8434.268 and for smokers at $32050.232. A boxplot is also prepared to visually represent the disparity in charges based on smoker status, with labels distinguishing smoker from non-smoker groups.

Region

```
# Assuming 'region' is already a
factor in your data
# ANOVA for region
anova_region <- aov(charges ~
region, data = data)
print(summary(anova_region))
##                Df    Sum Sq
Mean Sq F value Pr(>F)
## region          3 1.301e+09
433586560    2.97 0.0309 *
## Residuals    1334 1.948e+11
146007093
## ---
## Signif. codes:  0 '***' 0.001
'**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
ggplot(data, aes(x = region, y =
charges)) +
    geom_boxplot() +
    ggtitle("Charges by Region") +
    xlab("Region") +
    ylab("Charges")
```

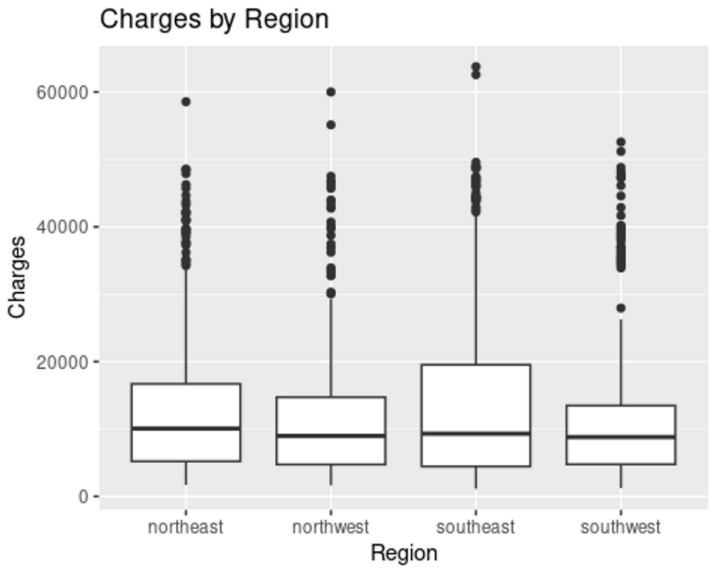

Figure 12: Hypothesis testing of Region

The provided R script output pertains to an ANOVA analysis that assesses whether there are statistically significant differences in medical charges across different regions, assuming 'region' is a categorical variable in the dataset. The ANOVA results show an F value of 2.97 with a p-value of 0.0309, which indicates there are some differences in charges between regions, albeit this difference is at a marginal significance level ($p < 0.05$). A boxplot is generated to visualize the distribution of charges across the regions, which would help in understanding the variation in charges geographically.

Analysis

As the health insurance charges are a continuous variable linear regression model was used to produce regression lines.

Regression line of Age vs Charges

```
# Assuming your data frame is named 'data'

# Creating a linear regression model with 'charges' as the dependent variable

# Load necessary libraries
library(ggplot2)

# Assuming your data frame 'data' is already loaded with appropriate columns
# data <- read.csv('your_data_file.csv')

# List of independent variables
independent_vars <- c('age', 'sex', 'bmi', 'children', 'smoker',
                      'regionnortheast', 'regionnorthwest',
                      'regionsoutheast', 'regionsouthwest')

# Perform pairwise linear regressions, print results, and show plot
for (var in independent_vars) {
  formula <- as.formula(paste('charges ~', var))
  model <- lm(formula, data = data)

  # Printing the summary
  cat("\nRegression Summary for", var, ":\n")
  print(summary(model))

  # Plotting the regression line
  plot_title <- paste("Regression Line for Charges vs", var)
  p <- ggplot(data, aes_string(x = var, y = 'charges')) +
       geom_point() +
       geom_smooth(method = "lm", col = "blue") +
       labs(title = plot_title) +
       theme_minimal()

  # Explicitly print the plot
  print(p)

  # Pause to view each plot
  if (var != tail(independent_vars, n=1)) {
    cat("\nPress [enter] to view the next plot...")
    readline()
  }
}
##
## Regression Summary for age :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8059  -6671  -5939   5440  47829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3165.9      937.1   3.378 0.000751 ***
## age            257.7       22.5  11.453  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 1336 degrees of freedom
## Multiple R-squared:  0.08941,    Adjusted R-squared:  0.08872
## F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## ℹ Please use tidy evaluation idioms with `aes()`.
## ℹ See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
## `geom_smooth()` using formula = 'y ~ x'
```

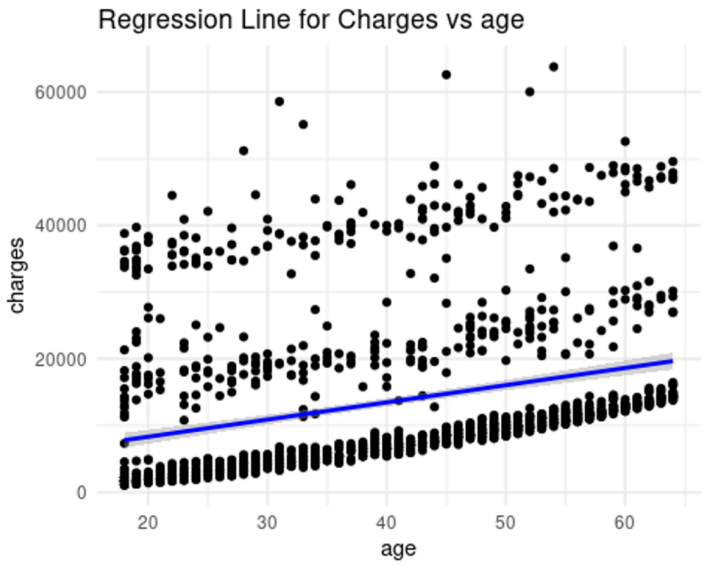

Figure 13: Regression line of Age vs Charges

The output for the regression with 'age' as the independent variable indicates a significant positive relationship. With each additional year of age, the charges increase by an estimated $257.7. The p-value for 'age' is less than 2e-16, which is highly significant, and the coefficient for 'age' is notably different from zero. This suggests a strong association between age and charges. The model has a relatively low R-squared value of approximately 0.08941, meaning that around 8.94% of the variability in charges is explained by age alone.

Regression line of SEX vs Charges

```
##
## Press [Enter] to view the next plot...
##
## Regression Summary for sex :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12835  -8435  -3980   3476  51201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12569.6      470.1  26.740   <2e-16
***
## sex           1387.2      661.3   2.098   0.0361
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
##
## Residual standard error: 12090 on 1336 degrees of
freedom
## Multiple R-squared:  0.003282,   Adjusted R-
squared:  0.002536
## F-statistic:   4.4 on 1 and 1336 DF,  p-value:
0.03613
## `geom_smooth()` using formula = 'y ~ x'
```

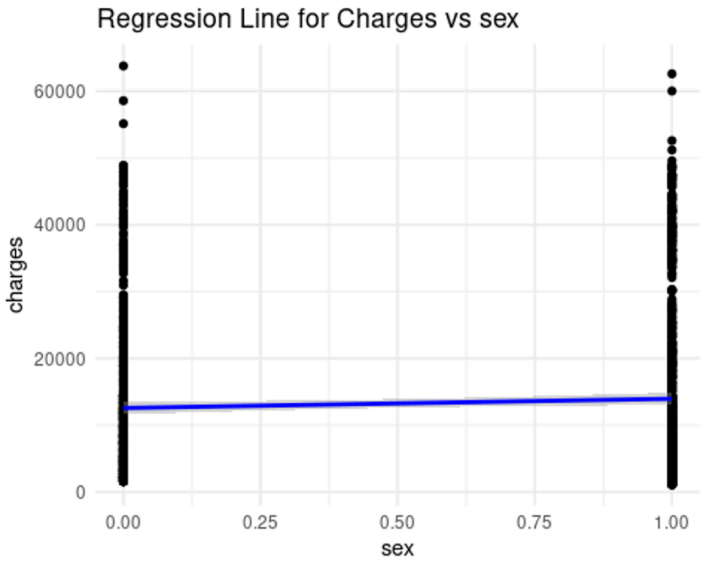

Figure 14: Regression line of SEX vs Charges

In examining the impact of gender on medical expenditure, a linear regression analysis revealed a significant association between the sex of the individual and the charges incurred. The regression coefficient for sex was found to be $1387.2 (p = 0.0361), indicating that males are likely to incur an average of $1387.2 more in medical charges than females, holding other factors constant. While statistically significant, the effect size was modest, with sex accounting for approximately 0.328% of the variance in charges, as denoted by an R-squared value of 0.003282. These findings suggest that while sex is a determinant of medical charges, it is a relatively minor one in the context of the model considered. The F-statistic of 4.4 further confirms the relevance of the model, despite the low explanatory power of this particular variable.

Regression line of BMI vs Charges

```
##
## Press [Enter] to view the next plot...
##
## Regression Summary for bmi :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20956  -8118  -3757   4722  49442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1192.94    1664.80   0.717    0.474
## bmi           393.87      53.25   7.397 2.46e-13
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of
freedom
## Multiple R-squared:  0.03934,    Adjusted R-
squared:  0.03862
## F-statistic: 54.71 on 1 and 1336 DF,  p-value:
2.459e-13
## `geom_smooth()` using formula = 'y ~ x'
```

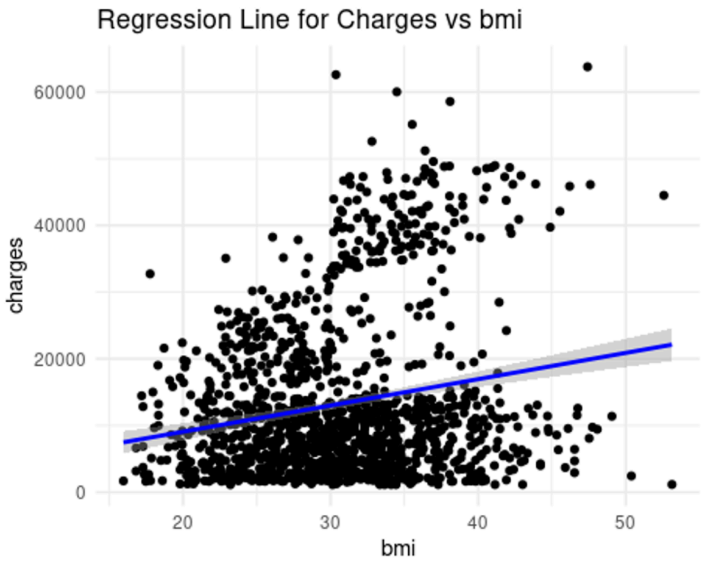

Figure 15: Regression line of BMI vs Charges

The regression analysis indicates a statistically significant positive relationship between BMI and insurance charges, with an increase in BMI associated with an increase of approximately $393.87 in charges. The significance of this relationship is supported by a p-value of 2.46e-13. However, BMI accounts for only about 3.934% of the variation in charges, as reflected by the R-squared value, suggesting that other factors also play a significant role in determining insurance charges.

Regression line of Number of Children vs Charges

```
##
## Press [Enter] to view the next plot...
##
## Regression Summary for children :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11585  -8759  -4071   3468  51248
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12522.5      446.5  28.049   <2e-16
***
## children       683.1      274.2   2.491   0.0129
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
' ' 0.1 ' ' 1
##
## Residual standard error: 12090 on 1336 degrees of
freedom
## Multiple R-squared:  0.004624,   Adjusted R-
squared:  0.003879
## F-statistic: 6.206 on 1 and 1336 DF,  p-value:
0.01285
## `geom_smooth()` using formula = 'y ~ x'
```

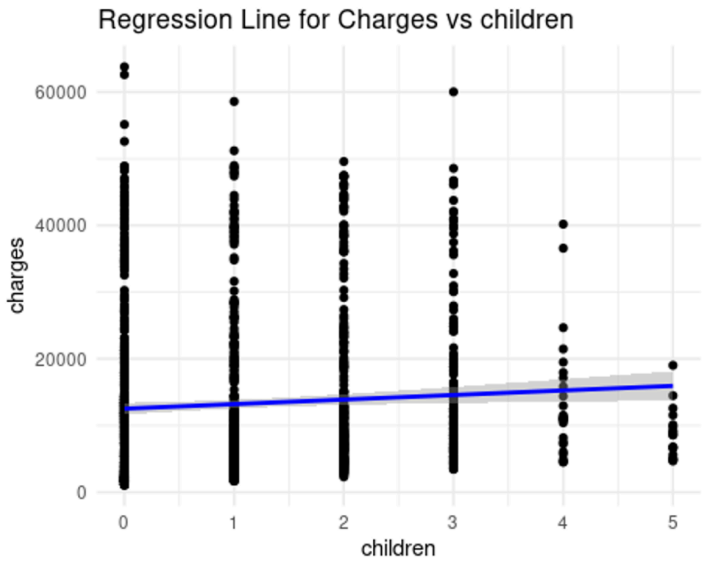

Figure 16: Regression line of Number of Children vs Charges

The regression summary for the variable 'children' indicates that the number of children is a statistically significant predictor of insurance charges. The model suggests that each additional child is associated with an increase of $683.1 in charges, holding other factors constant. This result is statistically significant at the 5% level with a p-value of 0.0129.

However, the variable 'children' has a relatively small effect on the variability of insurance charges, with an R-squared value of 0.004624. This implies that the number of children a policyholder has explains only about 0.4624% of the variance in insurance charges according to this model. Despite the small effect size, the F-statistic of 6.206 indicates that the number of children is a significant factor in the model overall.

Regression line of Smoker Status vs Charges

```
##
## Press [Enter] to view the next plot...
##
## Regression Summary for smoker :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19221  -5042   -919   3705  31720
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8434.3      229.0   36.83   <2e-16
***
## smoker       23616.0      506.1   46.66   <2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
##
## Residual standard error: 7470 on 1336 degrees of
freedom
## Multiple R-squared:  0.6198, Adjusted R-squared:
0.6195
## F-statistic:  2178 on 1 and 1336 DF,  p-value: <
2.2e-16
## `geom_smooth()` using formula = 'y ~ x'
```

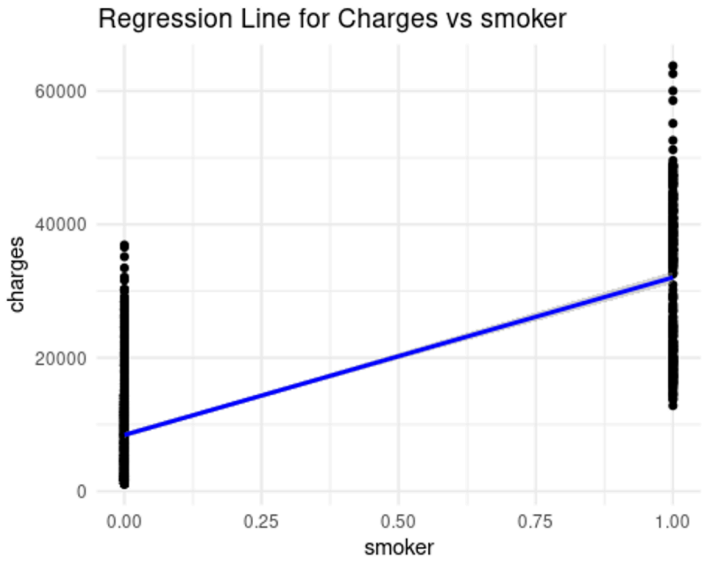

Figure 17: Regression line of Smoker Status vs Charges

The regression analysis results for the 'smoker' variable show a highly significant impact on insurance charges. The model estimates that smokers are charged an additional $23,616 on average compared to non-smokers. This is supported by a very high t-value of 46.66 and an extremely significant p-value of less than 2e-16.

The model's Multiple R-squared value is 0.6198, indicating that approximately 61.98% of the variation in insurance charges can be explained by whether the individual is a smoker. This high R-squared value, along with the significant F-statistic of 2178, demonstrates that smoking status is a powerful predictor of insurance charges in this data set.

Regression line of North East Region vs Charges

```
##
## Press [Enter] to view the next plot...
##
## Regression Summary for regionnortheast :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12105  -8531  -3951   3357  50543
##
## Coefficients:
##                  Estimate Std. Error t value
Pr(>|t|)
## (Intercept)       13227.0      380.4  34.768
<2e-16 ***
## regionnortheast     179.4      773.1   0.232
0.817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
##
## Residual standard error: 12110 on 1336 degrees of
freedom
## Multiple R-squared:  4.031e-05,  Adjusted R-
squared:  -0.0007082
## F-statistic: 0.05385 on 1 and 1336 DF,  p-value:
0.8165
## `geom_smooth()` using formula = 'y ~ x'
```

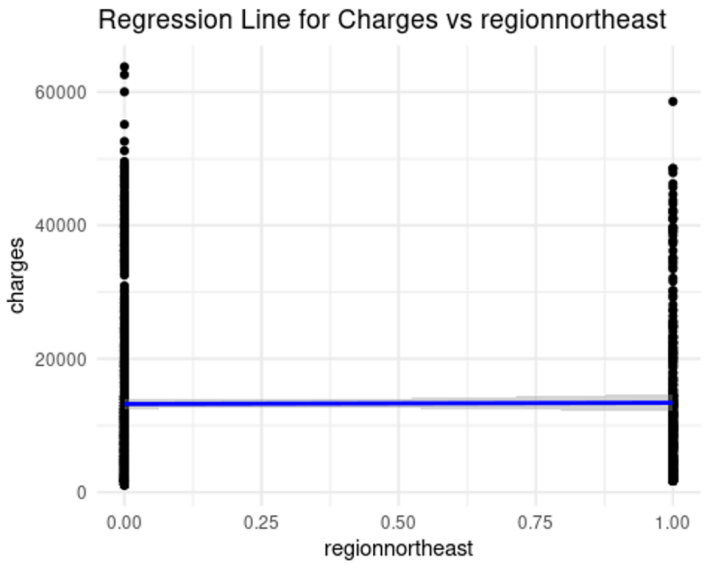

Figure 18: Regression line of North East Region vs Charges

The regression analysis for the variable 'regionnortheast' shows that being in the Northeast region of the U.S. does not significantly affect insurance charges. The coefficient for this region is not statistically significant with a p-value of 0.817, indicating that geographic location in this context is not a predictive factor for insurance charges. The model's R-squared values are near zero, confirming the variable's negligible explanatory power for the variation in charges.

Regression line of North West Region vs Charges

```
##
## Press [Enter] to view the next plot...
##
## Regression Summary for regionnorthwest :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12422  -8540  -3985   3386  50226
##
## Coefficients:
##                  Estimate Std. Error t value
Pr(>|t|)
## (Intercept)       13544.0      380.3   35.61
<2e-16 ***
## regionnorthwest  -1126.5      771.7   -1.46
0.145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
##
## Residual standard error: 12100 on 1336 degrees of
freedom
## Multiple R-squared:  0.001592,   Adjusted R-
squared:  0.0008451
## F-statistic: 2.131 on 1 and 1336 DF,  p-value:
0.1446
## `geom_smooth()` using formula = 'y ~ x'
```

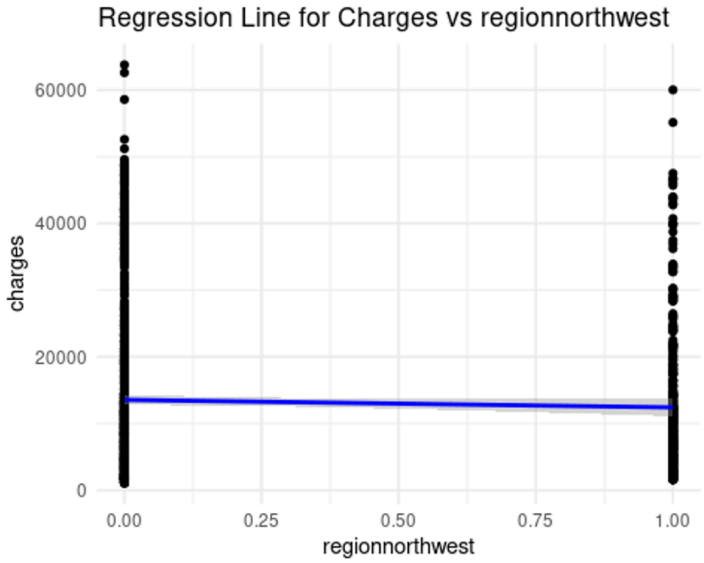

Figure 19: Regression line of North West Region vs Charges

The regression analysis for the 'regionnorthwest' variable indicates a non-significant negative effect on insurance charges, with a p-value of 0.145. This suggests that being in the Northwest region does not significantly differ in charges compared to the reference region, and the variable explains a very small portion of the variance in insurance charges, as evidenced by a low R-squared value.

Regression line of South East Region vs Charges

```
##
## Press [Enter] to view the next plot...
##
## Regression Summary for regionsoutheast :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13614  -8417  -3775   3410  49035
##
## Coefficients:
##                 Estimate Std. Error t value
Pr(>|t|)
## (Intercept)      12722.9      387.1  32.866  <
2e-16 ***
## regionsoutheast   2012.5      742.2   2.712
0.00678 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
##
## Residual standard error: 12080 on 1336 degrees of
freedom
## Multiple R-squared:  0.005473,   Adjusted R-
squared:  0.004729
## F-statistic: 7.353 on 1 and 1336 DF,  p-value:
0.006783
## `geom_smooth()` using formula = 'y ~ x'
```

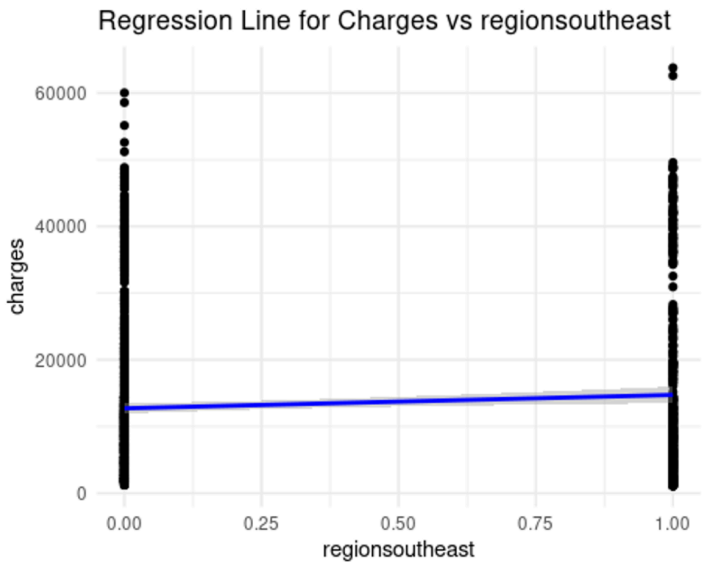

Figure 20: Regression line of South East Region vs Charges

The regression analysis for 'regionsoutheast' shows a statistically significant positive effect on insurance charges, with charges being on average $2012.5 higher in the Southeast compared to the baseline region. This effect is significant (p = 0.00678), but the 'regionsoutheast' variable accounts for only about 0.547% of the variance in insurance charges, as indicated by a low R-squared value.

Regression line of South West Region vs Charges

```
##
## Press [Enter] to view the next plot...
##
## Regression Summary for regionsouthwest :
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12445  -8583  -3847   3476  50204
##
## Coefficients:
##                  Estimate Std. Error t value
Pr(>|t|)
## (Intercept)       13566.7      380.3  35.676
<2e-16 ***
## regionsouthwest  -1219.8      771.6  -1.581
0.114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
'.' 0.1 ' ' 1
##
## Residual standard error: 12100 on 1336 degrees of
freedom
## Multiple R-squared:  0.001867,   Adjusted R-
squared:  0.00112
## F-statistic: 2.499 on 1 and 1336 DF,  p-value:
0.1141
## `geom_smooth()` using formula = 'y ~ x'
```

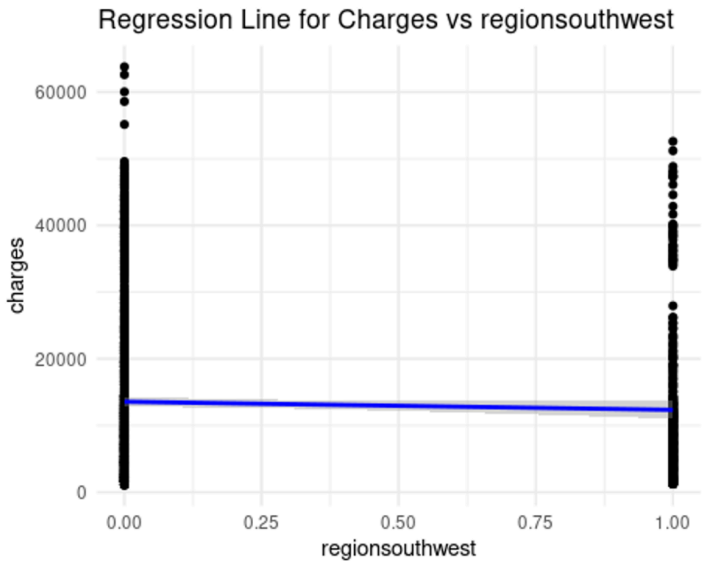

Figure 21: Regression line of South West Region vs Charges

The regression analysis for 'regionsouthwest' suggests a non-significant negative association with insurance charges, with a p-value of 0.114 indicating that living in the Southwest region does not significantly alter charges compared to the reference region. The variable accounts for only 0.1867% of the variance in charges, as shown by the R-squared value.

Correlation Matrix

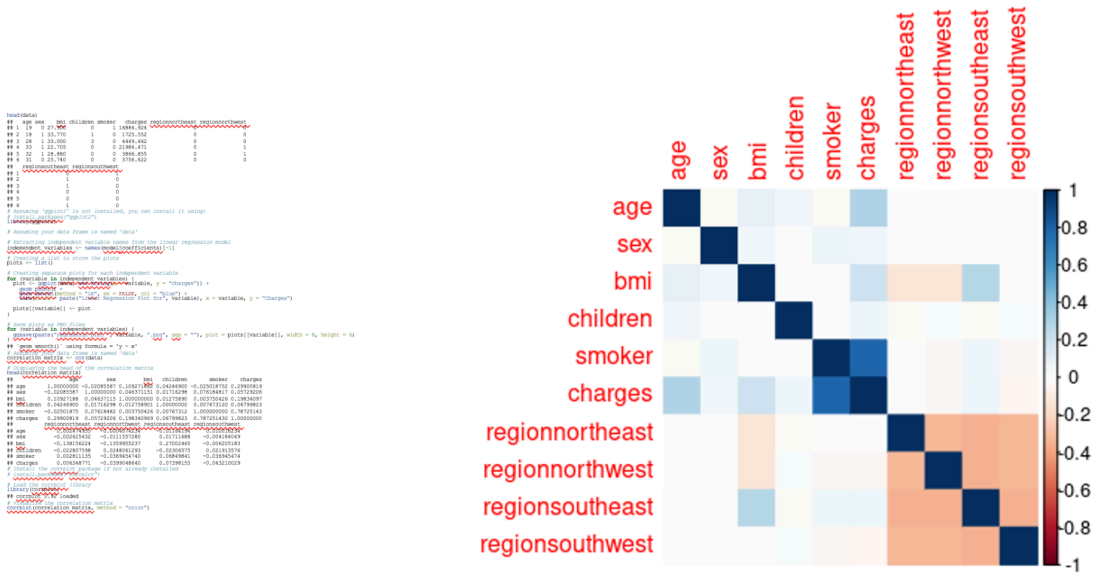

Figure 22: Correlation Matrix

The correlation matrix summary highlights the pairwise correlations among various health insurance variables. Notably, there is a strong positive correlation between 'smoker' status and insurance 'charges' (0.787), indicating smoking is a key predictor of higher charges. Age also shows a moderate positive correlation with charges (0.299), while 'bmi' has a weaker positive relationship (0.198). Correlations involving 'sex', 'children', and regional categories are relatively low, suggesting these factors have a less pronounced linear relationship with insurance charges.

Conclusion

The analysis of health insurance data indicates that smoking status and age are the most influential factors in determining insurance charges, with smokers and older individuals facing higher premiums. Body Mass Index (BMI) also contributes to higher charges, albeit to a lesser extent. Other factors such as sex, number of children, and region demonstrate minimal correlation with charges, suggesting their impact on insurance costs is comparatively limited. These insights could be pivotal for insurers in premium calculations, individuals when considering insurance options, and policymakers focusing on health-related fiscal measures.