Project: Analysis on US Health Insurance Cost

Jenish, Suvid, Sakina

Introduction

This report provides an exploratory data analysis (EDA) of the US health insurance dataset to understand its structure, identify trends, and prepare it for further analysis.

Loading Required Libraries

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(summarytools)
```

```
## Warning in fun(libname, pkgname): couldn't connect to display ":0"

## system might not have X11 capabilities; in case of errors when using dfSummary(), set st_options(use
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(mplot)
```

Loading the Dataset

```r
# Load dataset (replace 'health_insurance.csv' with your actual file path)
data <- read.csv("insurance.csv")

# Display the first few rows
head(data)
```

```
##   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
```

Dataset Overview

Summary Statistics

```r
summary(data)
```

```
##       age            sex                 bmi           children    
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   Class  :character   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Mode   :character   Median :30.40   Median :1.000  
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095  
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000  
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000  
##     smoker             region             charges     
##  Length:1338        Length:1338        Min.   : 1122  
##  Class :character   Class :character   1st Qu.: 4740  
##  Mode  :character   Mode  :character   Median : 9382  
##                                        Mean   :13270  
##                                        3rd Qu.:16640  
##                                        Max.   :63770  
```

Checking Data Structure

```r
str(data)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Checking for Missing Values

```r
colSums(is.na(data))
```

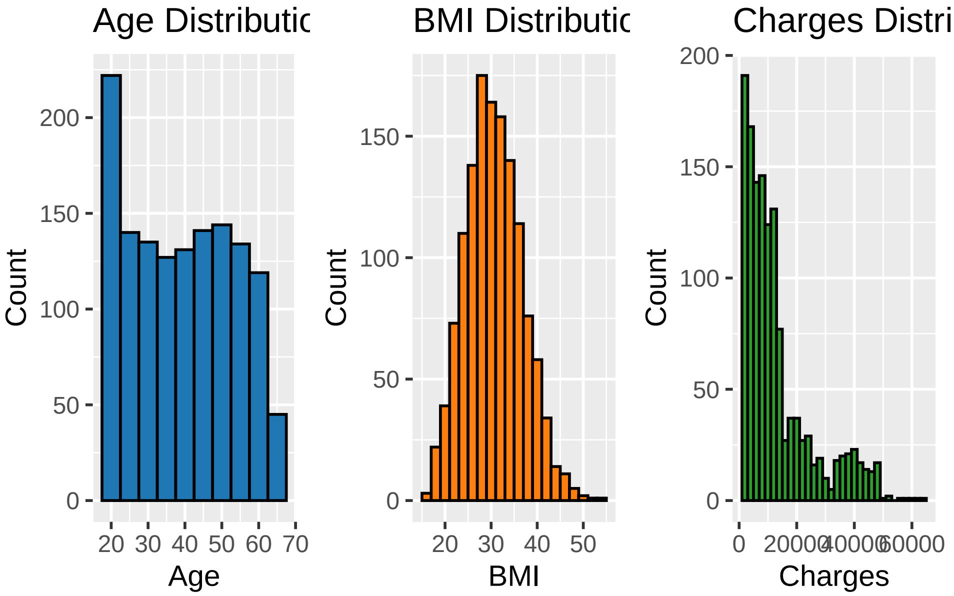
```
##      age      sex      bmi children   smoker   region  charges 
##        0        0        0        0        0        0        0 
```

Univariate Analysis

Combined Plots

```r
plot_age <- ggplot(data, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "#1f77b4", color = "black") +
  labs(title = "Age Distribution", x = "Age", y = "Count")
```

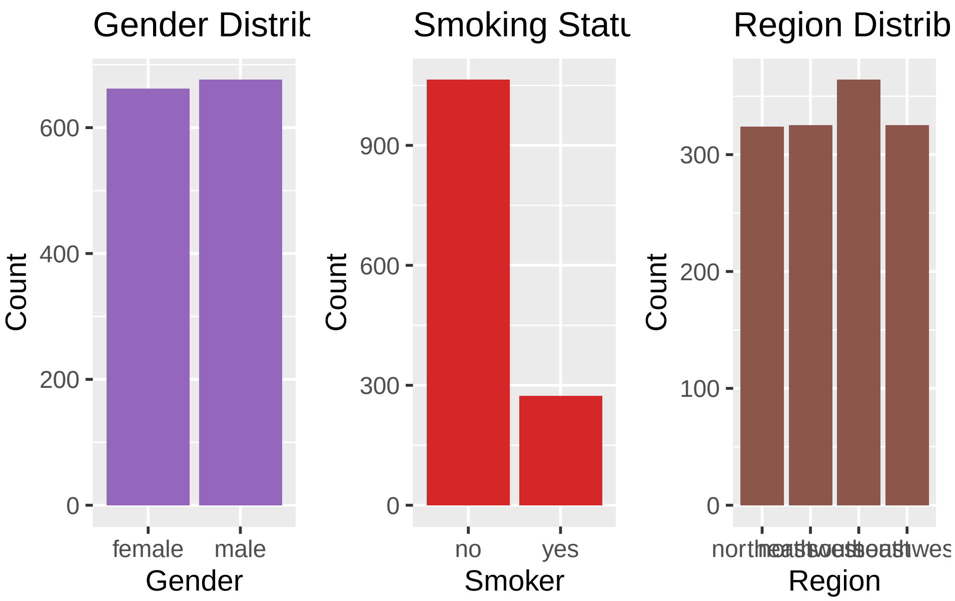
```r
plot_bmi <- ggplot(data, aes(x = bmi)) +
  geom_histogram(binwidth = 2, fill = "#ff7f0e", color = "black") +
  labs(title = "BMI Distribution", x = "BMI", y = "Count")

plot_charges <- ggplot(data, aes(x = charges)) +
  geom_histogram(binwidth = 2000, fill = "#2ca02c", color = "black") +
  labs(title = "Charges Distribution", x = "Charges", y = "Count")

grid.arrange(plot_age, plot_bmi, plot_charges, ncol = 3)
```



Categorical Variables

Combined Plots

```r
plot_gender <- ggplot(data, aes(x = sex)) +
  geom_bar(fill = "#9467bd") +
  labs(title = "Gender Distribution", x = "Gender", y = "Count")

plot_smoker <- ggplot(data, aes(x = smoker)) +
  geom_bar(fill = "#d62728") +
  labs(title = "Smoking Status", x = "Smoker", y = "Count")

plot_region <- ggplot(data, aes(x = region)) +
  geom_bar(fill = "#8c564b") +
  labs(title = "Region Distribution", x = "Region", y = "Count")

grid.arrange(plot_gender, plot_smoker, plot_region, ncol = 3)
```

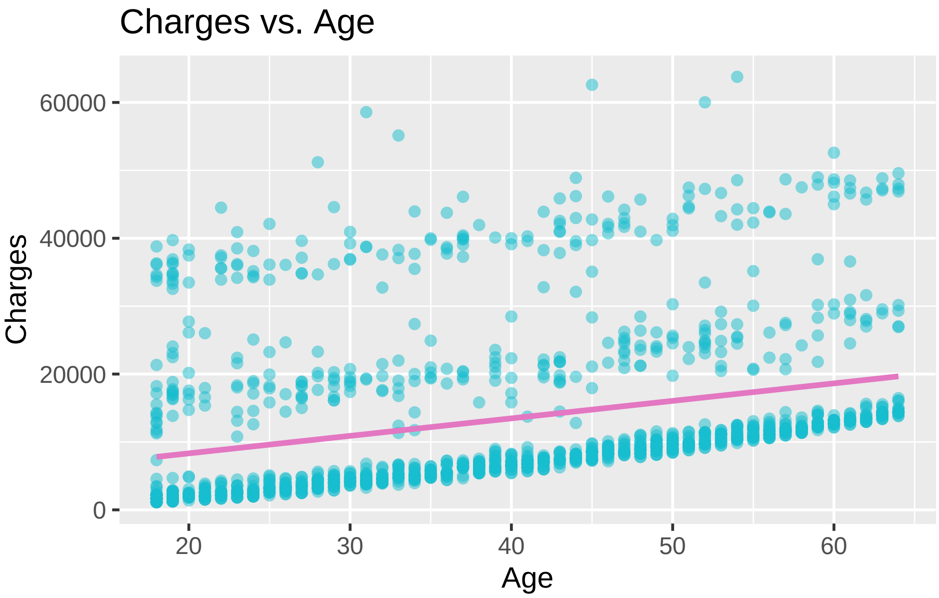


Bivariate Analysis

Charges vs. Age

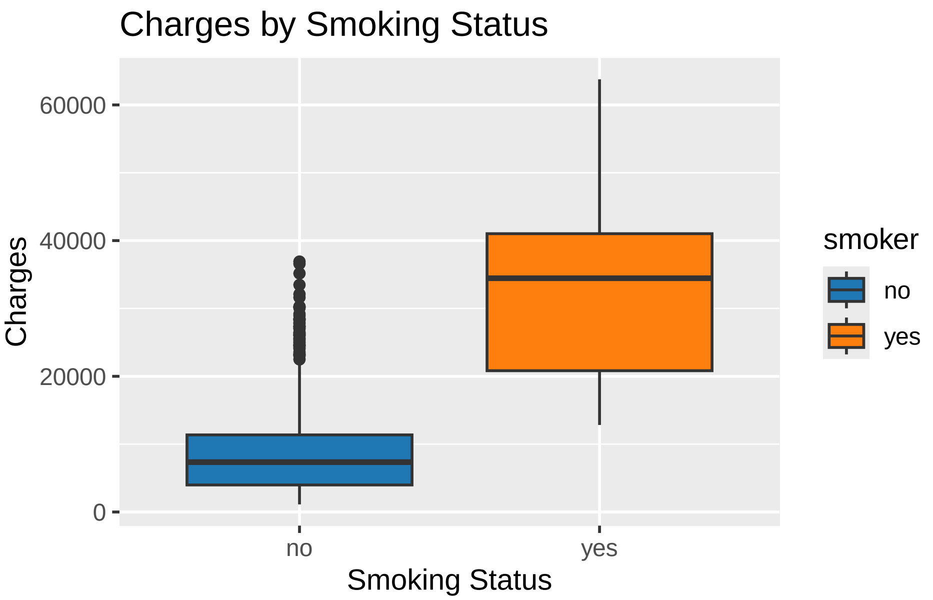
```
ggplot(data, aes(x = age, y = charges)) +
  geom_point(alpha = 0.5, color = "#17becf") +
  geom_smooth(method = "lm", color = "#e377c2", se = FALSE) +
  labs(title = "Charges vs. Age", x = "Age", y = "Charges")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



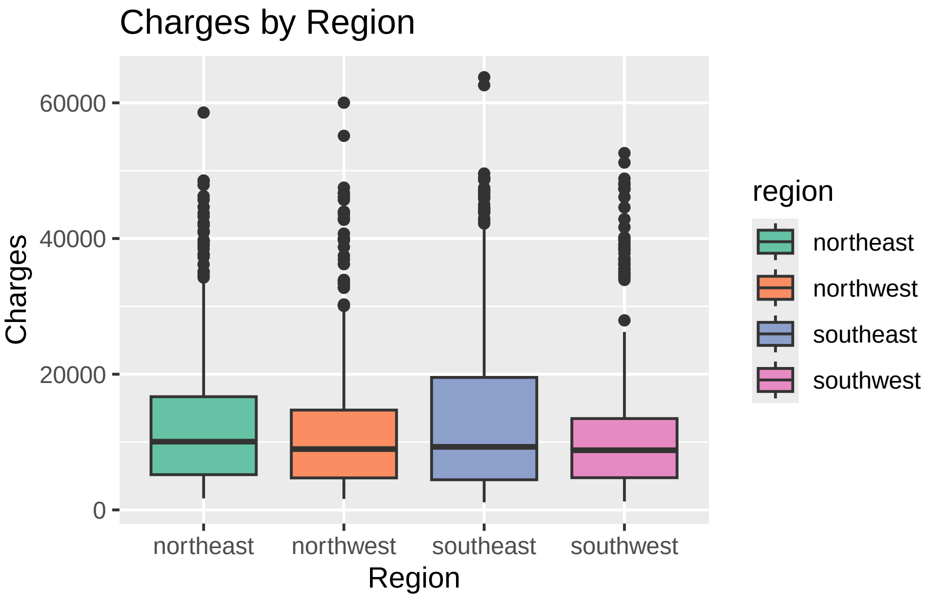
Charges by Smoking Status

```
ggplot(data, aes(x = smoker, y = charges, fill = smoker)) +
  geom_boxplot() +
  scale_fill_manual(values = c("#1f77b4", "#ff7f0e")) +
  labs(title = "Charges by Smoking Status", x = "Smoking Status", y = "Charges")
```



Charges by Region

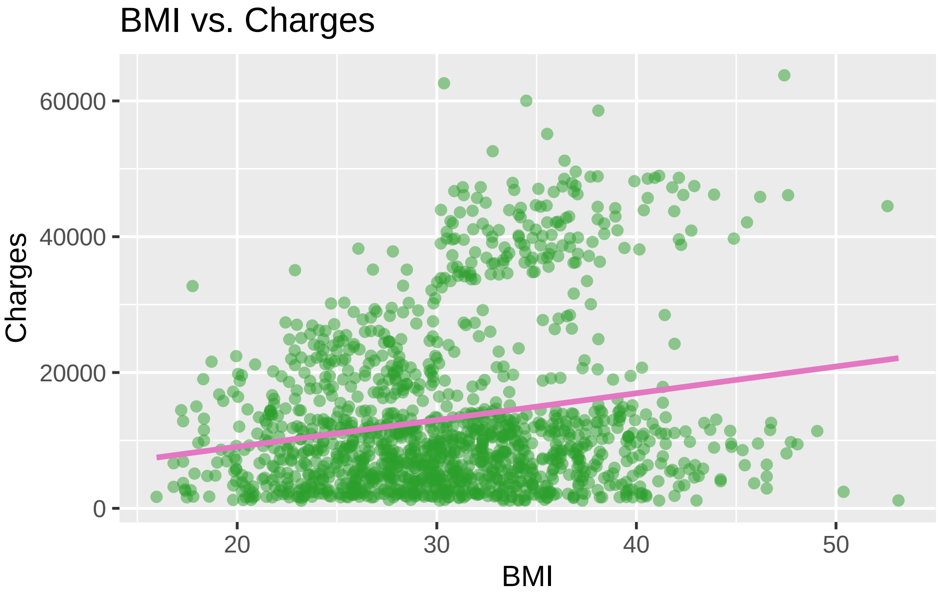
```
ggplot(data, aes(x = region, y = charges, fill = region)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Set2") +
  labs(title = "Charges by Region", x = "Region", y = "Charges")
```



BMI vs. Charges

```
ggplot(data, aes(x = bmi, y = charges)) +
  geom_point(alpha = 0.5, color = "#2ca02c") +
  geom_smooth(method = "lm", color = "#e377c2", se = FALSE) +
  labs(title = "BMI vs. Charges", x = "BMI", y = "Charges")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Correlation Analysis

Correlation Matrix

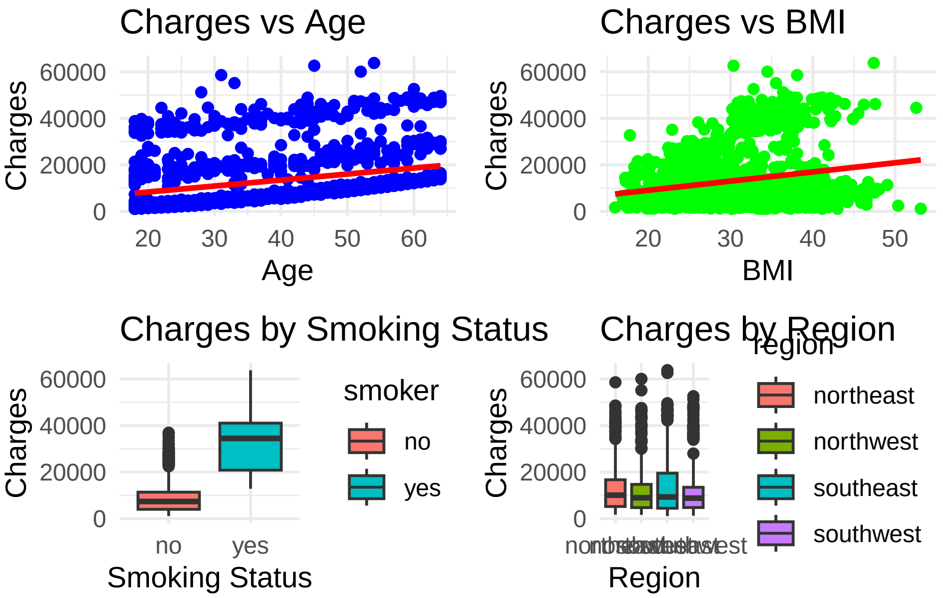
```
# Correlation matrix for numerical variables
cor_matrix <- cor(data %>% select_if(is.numeric))
cor_matrix
```

```
##                age       bmi   children    charges
## age      1.0000000 0.1092719 0.04246900 0.29900819
## bmi      0.1092719 1.0000000 0.01275890 0.19834097
## children 0.0424690 0.0127589 1.00000000 0.06799823
## charges  0.2990082 0.1983410 0.06799823 1.00000000
```

RQ1:

```
library(ggplot2)
library(gridExtra)

# Plot 1: Charges vs Age with a fitted line
p1 <- ggplot(data, aes(x = age, y = charges)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) + # Add linear regression line
  labs(title = "Charges vs Age", x = "Age", y = "Charges") +
  theme_minimal()

# Plot 2: Charges vs BMI with a fitted line
p2 <- ggplot(data, aes(x = bmi, y = charges)) +
  geom_point(color = "green") +
  geom_smooth(method = "lm", color = "red", se = FALSE) + # Add linear regression line
  labs(title = "Charges vs BMI", x = "BMI", y = "Charges") +
  theme_minimal()

# Plot 3: Charges by Smoking Status (boxplot)
p3 <- ggplot(data, aes(x = smoker, y = charges, fill = smoker)) +
  geom_boxplot() +
  labs(title = "Charges by Smoking Status", x = "Smoking Status", y = "Charges") +
  theme_minimal()

# Plot 4: Charges by Region (boxplot)
p4 <- ggplot(data, aes(x = region, y = charges, fill = region)) +
  geom_boxplot() +
  labs(title = "Charges by Region", x = "Region", y = "Charges") +
  theme_minimal()

# Adjust plotting device size
options(repr.plot.width = 20, repr.plot.height = 15) # Adjust dimensions for better layout

# Arrange all plots in a grid
grid.arrange(p1, p2, p3, p4, ncol = 2, heights = c(1, 1))
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



```
model <- lm(charges ~ age + smoker + bmi + bmi*smoker, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = charges ~ age + smoker + bmi + bmi * smoker, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14595.4  -2015.2  -1319.2   -290.5  29313.7
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -2290.008    831.999  -2.752    0.006 ** 
## age               266.758      9.617  27.739   <2e-16 ***
## smokeryes      -20093.508   1666.827 -12.055   <2e-16 ***
## bmi                 7.109     25.058   0.284    0.777    
## smokeryes:bmi    1430.920     53.217  26.888   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4907 on 1333 degrees of freedom
## Multiple R-squared:  0.8363, Adjusted R-squared:  0.8358 
## F-statistic:  1702 on 4 and 1333 DF,  p-value: < 2.2e-16
```

RQ2:

```
# Filter data for smokers and non-smokers
smokers <- data %>% filter(smoker == "yes")
non_smokers <- data %>% filter(smoker == "no")

# Create summary statistics for smokers and non-smokers
summary_stats <- data %>%
  group_by(smoker) %>%
  summarise(
    mean_charges = mean(charges),
    median_charges = median(charges),
    sd_charges = sd(charges),
    count = n()
  )

# Display the summary statistics
summary_stats
```

```
## # A tibble: 2 x 5
##   smoker mean_charges median_charges sd_charges count
##   <chr>         <dbl>          <dbl>      <dbl> <int>
## 1 no            8434.          7345.      5994.  1064
## 2 yes          32050.         34456.     11542.   274
```

```
# Perform a t-test to compare charges between smokers and non-smokers
t_test_result <- t.test(charges ~ smoker, data = data)

# View the results
t_test_result
```

```
## 
##  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
```

RQ3: Is there a significant difference in the insurance costs between male and female smokers?

Hypotheses

- **Null Hypothesis** (H_0): The mean insurance costs for male and female smokers are equal.
- **Alternative Hypothesis** ($H_1$):The mean insurance costs for male and female smokers are not equal.

To answer this question, we will use a two-sample t-test.

```
# Filter data for smokers only
smokers <- data %>%
  filter(smoker == "yes")

# Create separate groups for male and female smokers
male_smokers <- smokers %>% filter(sex == "male") %>% pull(charges)
female_smokers <- smokers %>% filter(sex == "female") %>% pull(charges)

# Display summary statistics
summary_stats <- smokers %>%
  group_by(sex) %>%
  summarise(
    mean_charges = mean(charges),
    median_charges = median(charges),
    sd_charges = sd(charges),
    count = n()
  )
```

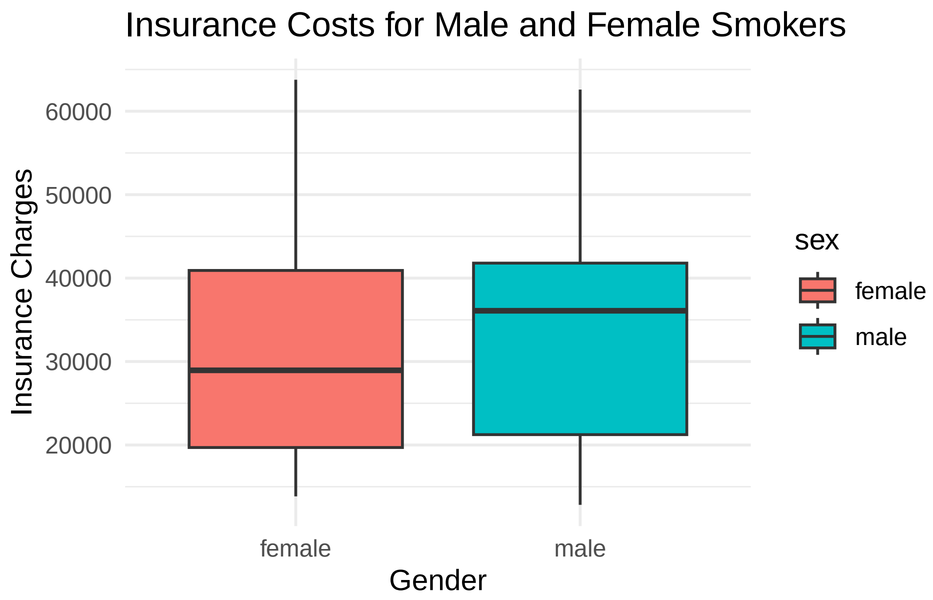
```
# Display the summary table
knitr::kable(summary_stats, caption = "Summary Statistics for Male and Female Smokers")
```

Table 1: Summary Statistics for Male and Female Smokers

sex	mean_charges	median_charges	sd_charges	count
female	30679.00	28950.47	11907.54	115
male	33042.01	36085.22	11202.67	159

Boxplot to visualize the distribution of insurance charges

```
ggplot(smokers, aes(x = sex, y = charges, fill = sex)) +
  geom_boxplot() +
  labs(
    title = "Insurance Costs for Male and Female Smokers",
    x = "Gender",
    y = "Insurance Charges"
  ) +
  theme_minimal()
```



Perform a two-sample t-test

```
t_test_result <- t.test(male_smokers, female_smokers, var.equal = TRUE)

# Display the test results
t_test_result
```

```
##
##  Two Sample t-test
##
## data:  male_smokers and female_smokers
## t = 1.6781, df = 272, p-value = 0.09448
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -409.2696 5135.2890
## sample estimates:
## mean of x mean of y
##  33042.01  30679.00
```

Summary Since the p-value (0.09448) is greater than 0.05 and the confidence interval includes zero, we fail to reject the null hypothesis. This suggests that there is no statistically significant difference in the mean insurance charges between male smokers and female smokers at the 5% significance level. However, there might still be a practical difference in the mean charges, as the confidence interval is quite wide, indicating variability in the data.

RQ4: Is there a significant difference in charges for individuals in different regions?

Is there a significant difference in insurance charges among individuals living in different regions (northeast, northwest, southeast, southwest)?

Hypotheses

- **Null Hypothesis** (H_0): The mean insurance charges are equal across all regions.
- **Alternative Hypothesis** (H_1): At least one region has a significantly different mean insurance charge.

We will use a one-way ANOVA test to answer this research question.

```r
# Summarize charges by region
summary_by_region <- data %>%
  group_by(region) %>%
  summarise(
    mean_charges = mean(charges),
    median_charges = median(charges),
    sd_charges = sd(charges),
    count = n()
  )
```

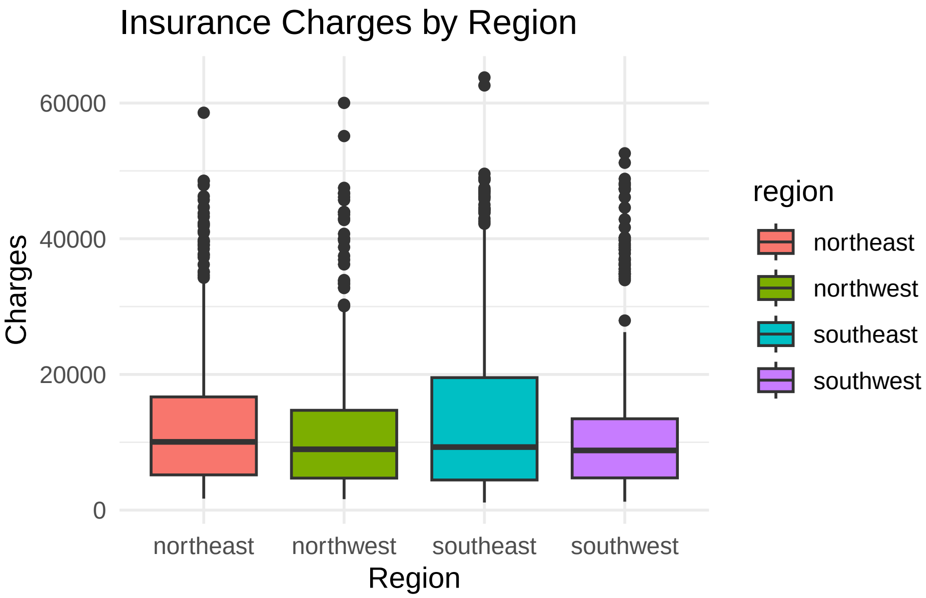
```r
# Display the summary table
knitr::kable(summary_stats, caption = "Summary Statistics for Male and Female Smokers")
```

Table 2: Summary Statistics for Male and Female Smokers

sex	mean_charges	median_charges	sd_charges	count
female	30679.00	28950.47	11907.54	115
male	33042.01	36085.22	11202.67	159

Boxplot to visualize the distribution of charges by region

```r
ggplot(data, aes(x = region, y = charges, fill = region)) +
  geom_boxplot() +
  labs(
    title = "Insurance Charges by Region",
    x = "Region",
    y = "Charges"
  ) +
  theme_minimal()
```



Perform one-way ANOVA

```r
anova_result <- aov(charges ~ region, data = data)

# Display the ANOVA table
summary(anova_result)
```

```
##               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
```

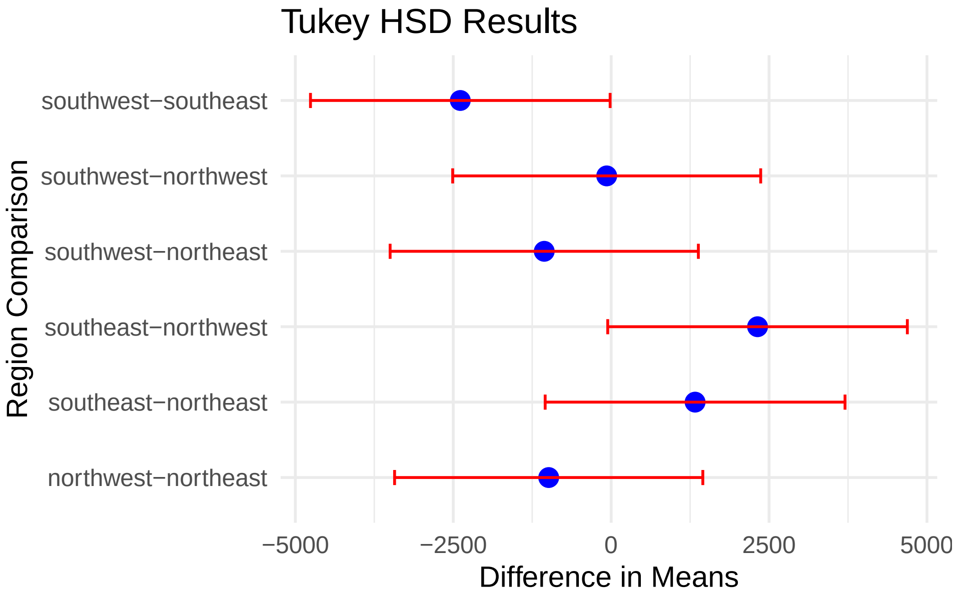
Summary The results of the one-way ANOVA test indicate a statistically significant difference in mean insurance charges across the four regions (northeast, northwest, southeast, southwest). The F-statistic is 2.97, with a p-value of 0.0309, which is below the significance level of 0.05. Thus, we reject the null hypothesis that the mean insurance charges are the same across all regions. This finding suggests that regional differences in insurance charges exist. However, the ANOVA test does not specify which regions differ from each other. To determine the specific pairwise differences, we perform a post hoc analysis such as Tukey's Honest Significant Difference (HSD) test.

Tukey's Honest Significant Difference (HSD) test

```r
post_hoc <- TukeyHSD(anova_result)

# Display the results
post_hoc
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = charges ~ region, data = data)
##
## $region
##                            diff         lwr        upr     p adj
## northwest-northeast  -988.8091 -3428.93434 1451.31605 0.7245243
## southeast-northeast  1329.0269 -1044.94167 3702.99551 0.4745046
## southwest-northeast -1059.4471 -3499.57234 1380.67806 0.6792086
## southeast-northwest  2317.8361   -54.19944 4689.87157 0.0582938
## southwest-northwest   -70.6380 -2508.88256 2367.60656 0.9998516
## southwest-southeast -2388.4741 -4760.50957  -16.43855 0.0476896
```

```r
# Convert Tukey HSD results to a data frame
tukey_df <- as.data.frame(post_hoc$region)
tukey_df$Comparison <- rownames(tukey_df)

# Plot the results
ggplot(tukey_df, aes(x = Comparison, y = diff)) +
  geom_point(size = 3, color = "blue") +
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.2, color = "red") +
  labs(
    title = "Tukey HSD Results",
    x = "Region Comparison",
    y = "Difference in Means"
  ) +
  theme_minimal() +
  coord_flip()
```



Summary From this plot, we can observe which region pairs show significant differences. For region pairs where the confidence interval does not include zero (southwest-southeast), we conclude that the insurance charges differ significantly. However, rest of the regions there is insufficient evidence to suggest a significant difference in insurance charges.