Insurance Analysis

Introduction

Insurance is a critical industry that plays a pivotal role in managing financial risks and providing protection to individuals and businesses against unforeseen events. The ability to analyze insurance data effectively can offer valuable insights into customer behavior, risk assessment, pricing strategies, and overall industry trends. In this statistical analysis project, we will delve into a comprehensive insurance dataset to uncover patterns, relationships, and actionable insights that can inform decision-making within the insurance domain.

Dataset

The dataset at the heart of this project comprises various attributes related to insurance beneficiaries and their medical charges. Each entry includes demographic information such as age, sex, body mass index (BMI), number of children, smoking status, residential region, and individual medical charges billed by health insurance. The richness of this dataset provides ample opportunities for exploring correlations, making predictions, and drawing meaningful conclusions.

Project Goals

The primary goals of this statistical analysis project are as follows:

Exploratory Data Analysis (EDA): We will begin by conducting exploratory data analysis to gain a comprehensive understanding of the dataset. Visualization techniques such as histograms, scatter plots, and box plots will be employed to visualize the distribution of variables, identify outliers, and explore relationships among different attributes.

Descriptive Statistics: Calculating summary statistics like mean, median, standard deviation, and quartiles will provide us with a quantitative overview of the dataset. These statistics will aid in forming initial insights about the central tendencies and variability of variables.

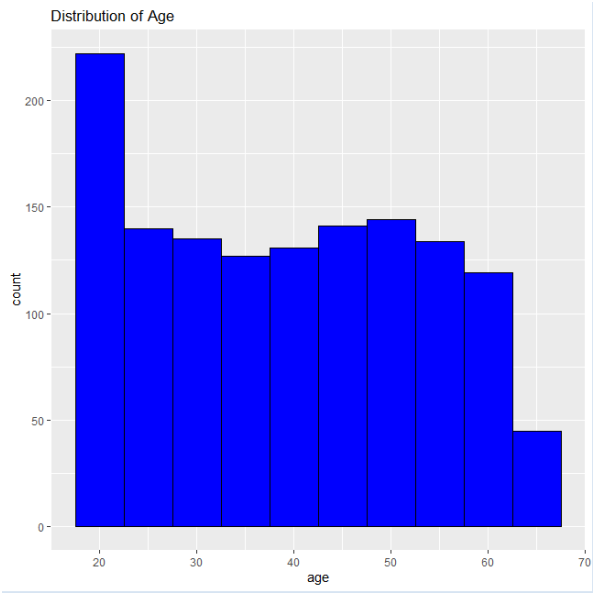
Hypothesis Testing: We will formulate and test hypotheses to answer specific questions about the data. For instance, we might investigate whether charges significantly differ between smokers and non-smokers, or if charges vary based on gender or region. Hypothesis testing will provide statistical evidence to support our conclusions.

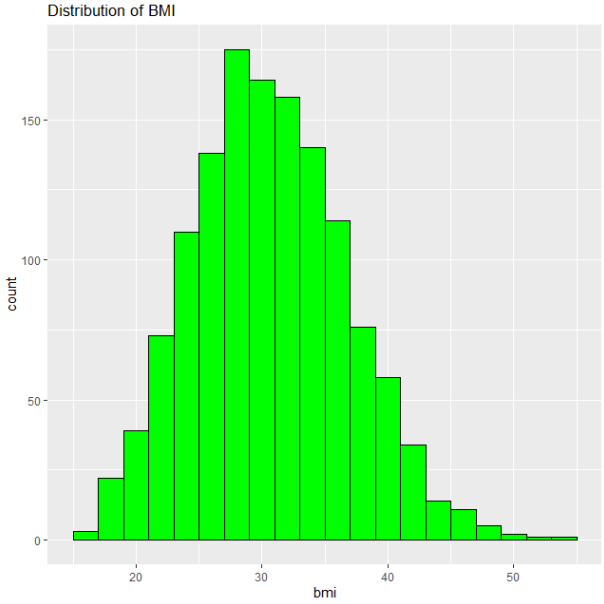
Correlation Analysis: By calculating correlation coefficients and creating scatter plots, we will explore the relationships between numerical variables such as age, BMI, and charges. Correlation analysis will help us identify potential associations and dependencies within the data.

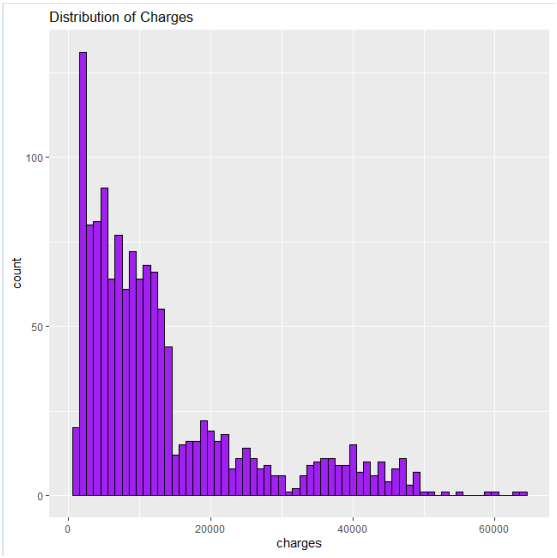
Categorical Analysis: Categorical variables like sex, smoking status, and region will be analyzed through frequency tables and bar plots. These analyses will shed light on the distribution of different categories and their impact on insurance charges.

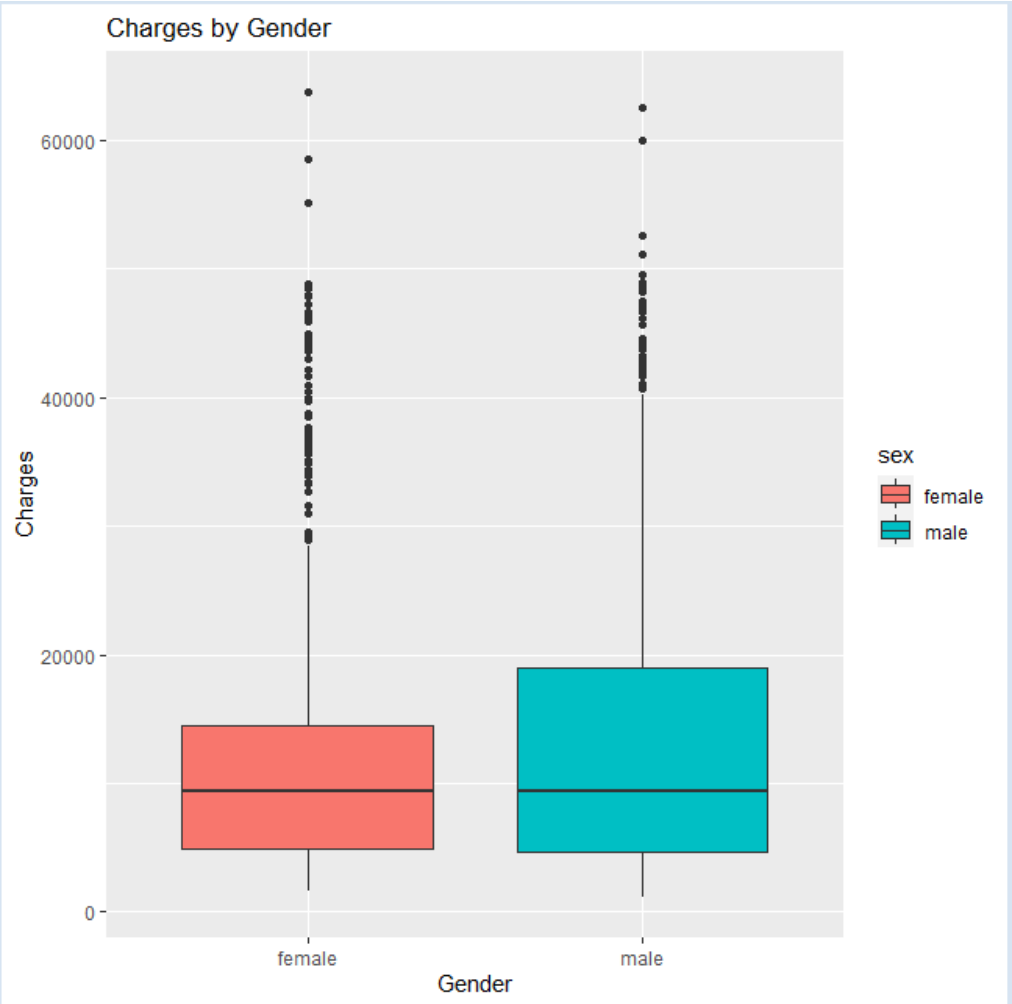
Regression Analysis: Regression techniques will be employed to predict medical charges based on various attributes. Simple linear regression and multiple regression models will be explored to understand how age, BMI, smoking status, and potentially other factors influence charges.

Analysi


Distribution of Age
count
age
0
50
100
150
200
20
30
40
50
60
70

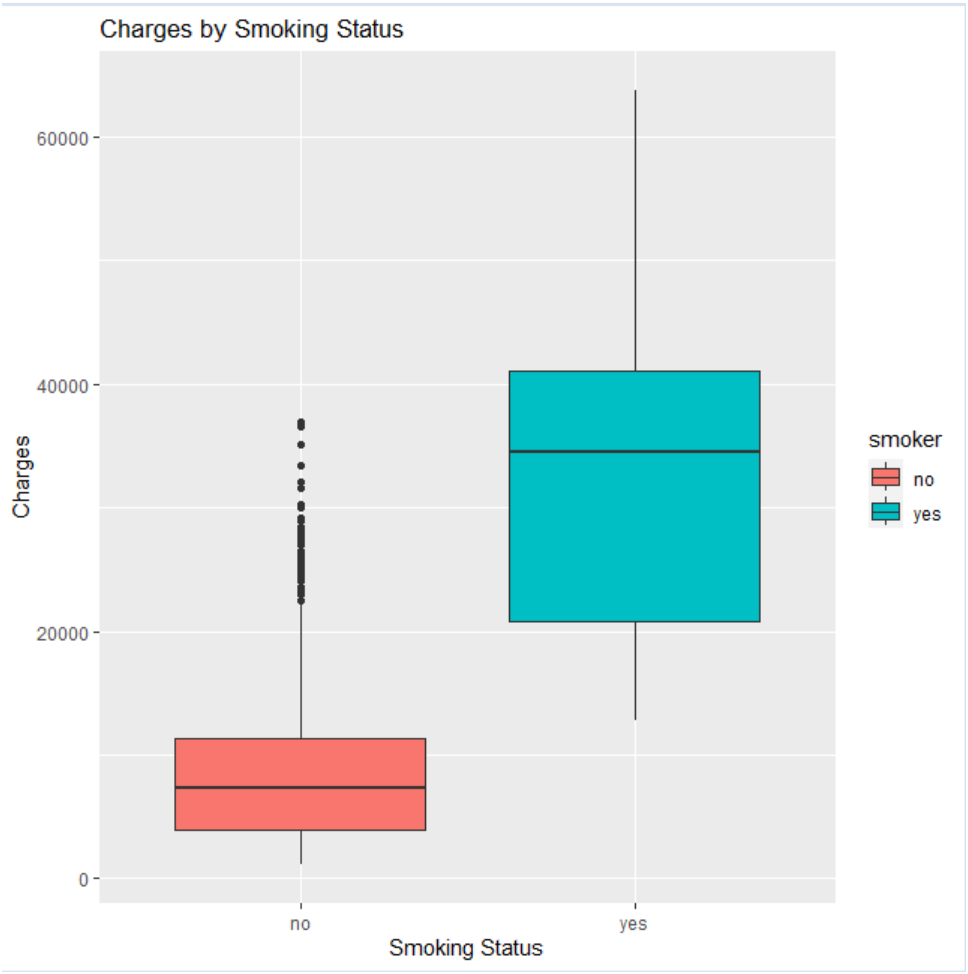

Distribution of BMI
count
bmi
0
50
100
150
20
30
40
50


Distribution of Charges
count
charges
0
50
100
0
20000
40000
60000



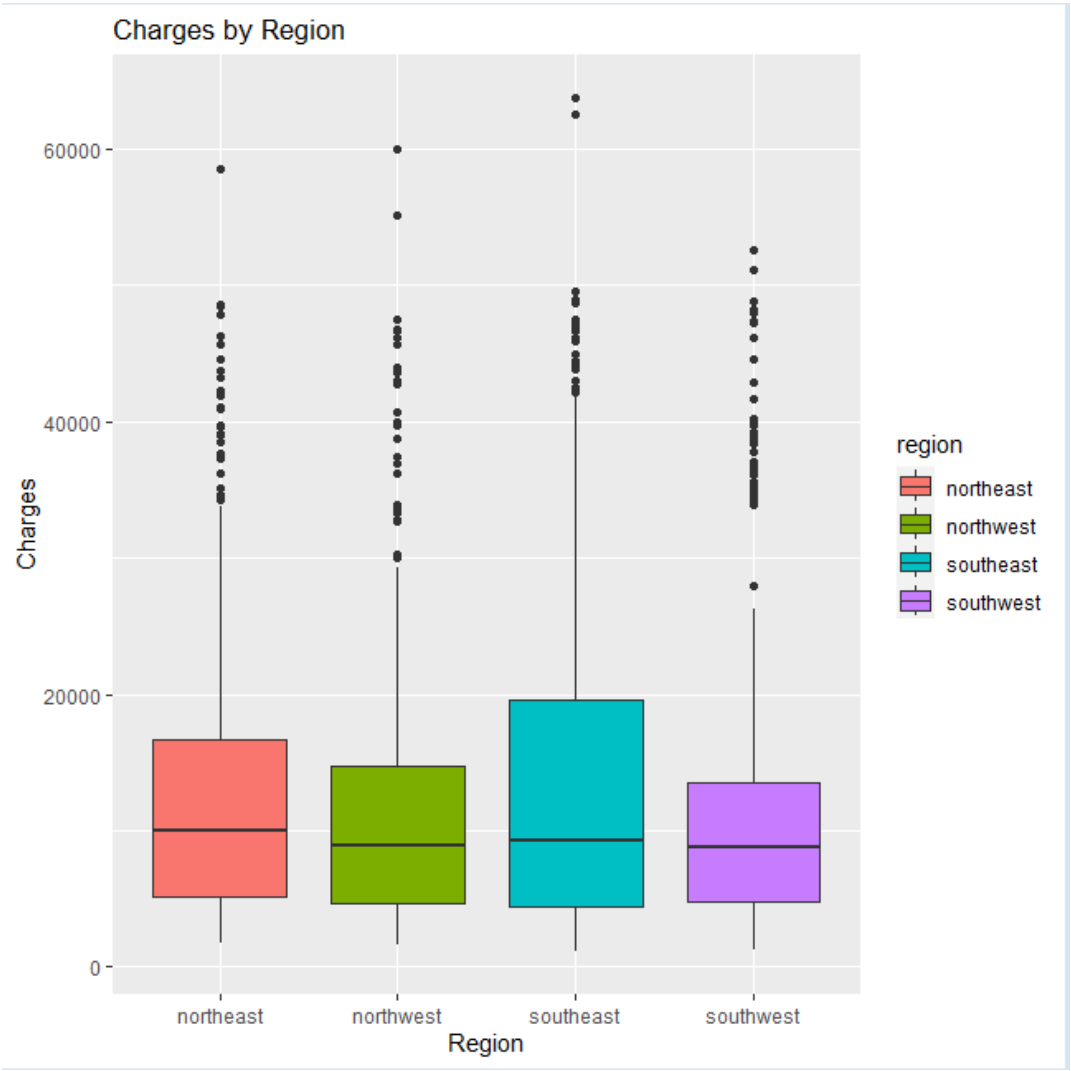
The box plot shows the distribution of medical charges for males and females. The median charge for females is higher than the median charge for males.

The interquartile range (IQR) for females is also higher than the IQR for males. This suggests that there is more variation in the medical charges for females than for males. There are also outliers in the female data, but not in the male data.



The median charge for smokers is higher than the median charge for non-smokers. The interquartile range (IQR) for smokers is also higher than the IQR for non-smokers. This suggests that there is more variation in the medical charges for smokers than for non-smokers.

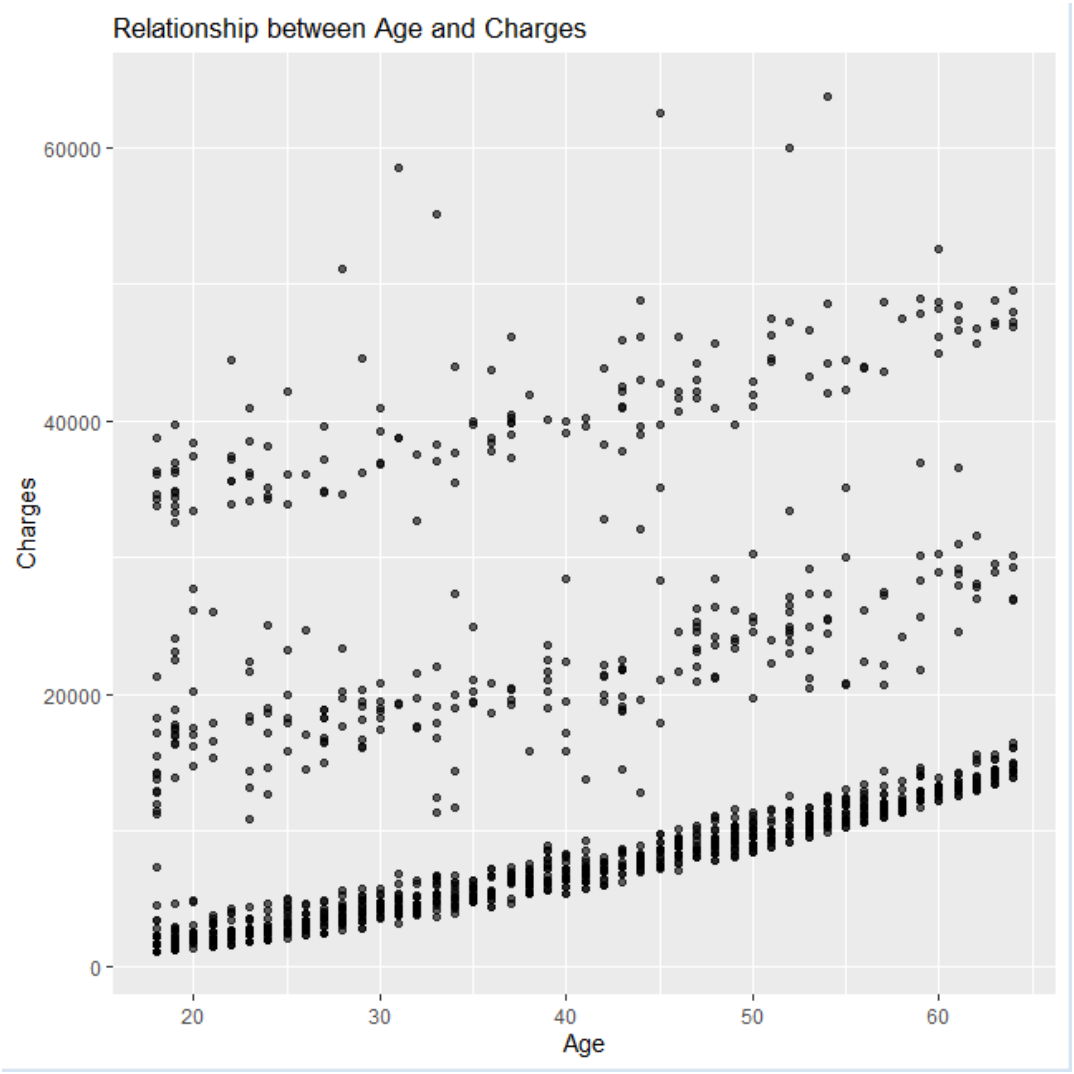
There are also outliers in the smoker data, but not in the non-smoker data.



The median charge for the Northeast region is the highest, followed by the Northwest region, then the Southeast region, and finally the Southwest region.

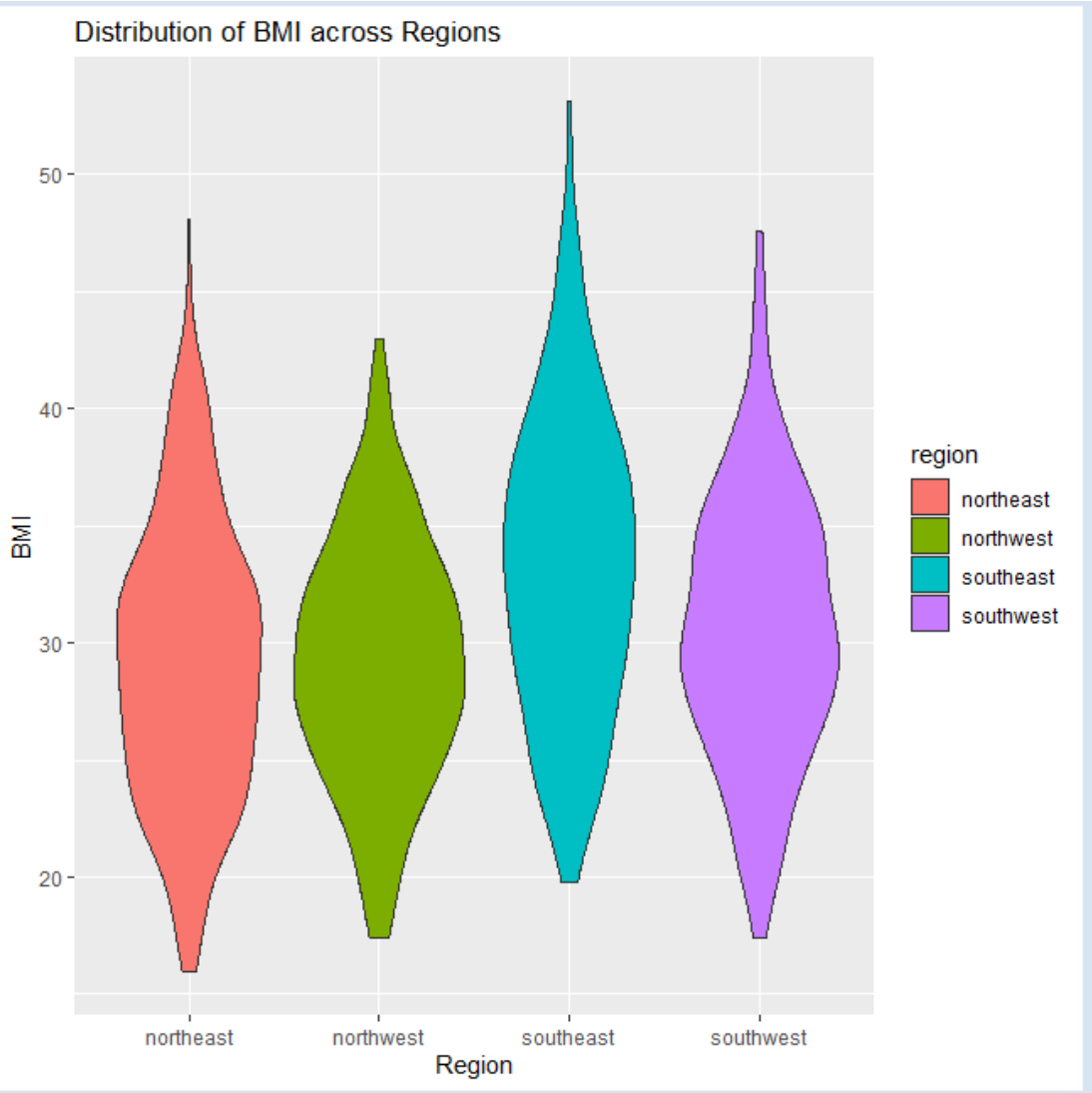
The interquartile range (IQR) for the Northeast region is also the highest, followed by the Northwest region, then the Southeast region, and finally the Southwest region.

This suggests that there is more variation in the medical charges for the Northeast region than for the other regions. There are also outliers in the Northeast region data, but not in the other regions data.



First, the medical charges increase with age. This is to be expected, as people tend to need more medical care as they get older.

Third, there is a group of points that are outliers. These points are located above the main trend of the graph. These points may represent people who have experienced unusually high medical costs, such as those who have had serious illnesses or injuries.



First, the BMIs are generally higher in the Southeast and Southwest regions than in the Northeast and Northwest regions. This suggests that people in these regions are more likely to be overweight or obese.

Second, there is a clear trend of increasing BMI with increasing region. This suggests that the risk of being overweight or obese increases as you move from the Northeast to the Southwest region.

Descriptive StatAge:

Age:

The average age of the primary beneficiaries in the dataset is approximately 39.21 years.

The median age is 39 years, indicating that the distribution is relatively balanced around the median.

The standard deviation of age is about 14.05, which suggests moderate variability in ages.

The first quartile (Q1) is 27 years, and the third quartile (Q3) is 51 years, indicating that 25% of the ages are below 27 and 75% are below 51.

BMI:

The average BMI (Body Mass Index) is approximately 30.66, which falls into the overweight range.

The median BMI is 30.4, which further confirms the prevalence of higher BMIs.

The standard deviation of BMI is about 6.10, indicating some variability in body weights.

The first quartile (Q1) of BMI is 26.30, and the third quartile (Q3) is 34.69, showing the spread of BMI values.

Charges:

The average individual medical cost charged by health insurance is around $13,270.42.

The median charges are $9,382.03, which is considerably lower than the mean, indicating a possible right-skewed distribution.

The standard deviation of charges is about $12,110.01, indicating a wide range of charges.

The first quartile (Q1) of charges is $4,740.29, and the third quartile (Q3) is $16,639.91, illustrating the spread of charges.

Conclusions:

Age is relatively normally distributed with moderate variability.

BMI has a distribution skewed towards higher values, indicating that many individuals have higher BMIs.

Charges vary significantly, with a wide range of values and a potential skew towards higher charges.

These descriptive statistics provide initial insights into the characteristics of the insurance dataset, which can guide further analyses and decision-making.

Hypothesis Testing:

T-Test for Charges (Smokers vs. Non-Smokers):

- The p-value obtained from the t-test is extremely small (5.889464e-103), much smaller than 0.05.
- The conclusion is: The mean charges for smokers and non-smokers are significantly different.

ANOVA for Charges based on Gender:

- The ANOVA results show that the p-value is 0.0361, which is less than 0.05.
- The conclusion is: There is a significant difference in charges based on gender.

Kruskal-Wallis Test for Charges based on Region:

- The Kruskal-Wallis test results in a p-value of 0.1923291, which is greater than 0.05.
- The conclusion is: There is no significant difference in charges among different regions.

Keep in mind that a p-value less than 0.05 is often considered statistically significant, while a p-value greater than 0.05 suggests that you do not have enough evidence to reject the null hypothesis. These conclusions provide insights into how charges vary based on smoking status, gender, and region in your dataset.

Correlation Analysis:

```
> # Print the correlation matrix
> print(correlation_matrix)
               age       bmi   charges
age      1.0000000 0.1092719 0.2990082
bmi      0.1092719 1.0000000 0.1983410
charges  0.2990082 0.1983410 1.0000000
```

The correlation matrix you've provided shows the correlation coefficients between the numerical variables "age," "bmi," and "charges" in your dataset. Here's how to interpret the matrix:

The correlation between "age" and "charges" is 0.299. This positive correlation indicates that, generally, as age increases, medical charges tend to increase as well. However, the correlation is not extremely strong, suggesting that while there is a trend, other factors also influence charges.

The correlation between "bmi" and "charges" is 0.198. This positive correlation suggests that there is a tendency for higher BMI values to be associated with higher medical charges. Again, the correlation is not very strong, indicating that other factors may also play a role.

The correlation between "age" and "bmi" is 0.109. This correlation is relatively weak, indicating a mild positive relationship between age and BMI

Categorical Analysis:

```
> sex_freq <- table(insurance$sex)
> print(sex_freq)

female   male
   662    676
>
> # Create frequency table for smoker
> smoker_freq <- table(insurance$smoker)
> print(smoker_freq)

  no  yes
1064  274
>
> # Create frequency table for region
> region_freq <- table(insurance$region)
> print(region_freq)

northeast northwest southeast southwest
      324       325       364       325
```

Regression Analysis:

```
> summary(lm_age_charges)

Call:
lm(formula = charges ~ age, data = insurance)

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
```

The output you've provided is from the simple linear regression analysis, specifically for predicting charges based on age. Here's how to interpret the key parts of the output:

Call: This indicates the formula used for the regression.

Residuals: These are the differences between the actual charges and the predicted charges. They show the errors of the model's predictions.

Coefficients: This table provides information about the estimated coefficients of the model. In this case, the model's equation is: Charges = 3165.9 + 257.7 * Age. The "Estimate" column provides the coefficient values, and the "Std. The " Error" column gives the standard errors of the estimates. The "t value" and "Pr(>|t|)" columns represent the t-statistic and the associated p-value for testing if each coefficient is significantly different from zero.

Residual standard error: This is an estimate of the standard deviation of the residuals. It represents the average distance between the observed charges and the predicted charges by the model.

Multiple R-squared: This is the coefficient of determination, indicating the proportion of the variance in charges that can be explained by the independent variable (age) in the model. In this case, about 8.94% of the variance in charges is explained by age.

Adjusted R-squared: This is the adjusted version of the R-squared that accounts for the number of predictors in the model. It's useful when comparing models with different numbers of predictors.

F-statistic: This is the test statistic for the overall significance of the model. It tests whether at least one coefficient in the model is different from zero. The associated p-value indicates whether the model as a whole is significant.

The p-value associated with the "age" coefficient (Pr(>|t|)) is very close to zero, indicating that age is a statistically significant predictor of charges.

```
> summary(lm_multiple)

Call:
lm(formula = charges ~ age + bmi + smoker, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-12415.4  -2970.9   -980.5   1480.0  28971.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11676.83     937.57  -12.45   <2e-16 ***
age            259.55      11.93   21.75   <2e-16 ***
bmi            322.62      27.49   11.74   <2e-16 ***
smokeryes    23823.68     412.87   57.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6092 on 1334 degrees of freedom
Multiple R-squared:  0.7475,    Adjusted R-squared:  0.7469
F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

The output you've provided is from the multiple regression analysis, specifically for predicting charges based on age, BMI, and smoking status. Here's how to interpret the key parts of the output:

Call: This indicates the formula used for the regression.

Residuals: Similar to before, these are the differences between the actual charges and the predicted charges. They represent the errors of the model's predictions.

Coefficients: This table provides information about the estimated coefficients of the model. The model's equation is: Charges = -11676.83 + 259.55 * Age + 322.62 * BMI + 23823.68 * Smoker_Yes. Here, "Smoker_Yes" is an indicator variable representing whether the individual is a smoker. The "Estimate" column provides the coefficient values, and the other columns provide relevant statistical information.

Residual standard error: This is an estimate of the standard deviation of the residuals, similar to the simple linear regression case.

Multiple R-squared: This coefficient of determination indicates that the combination of age, BMI, and smoking status explains about 74.75% of the variance in charges. This is a significant improvement compared to the simple linear regression with only age.

Adjusted R-squared: Similar to before, this is the adjusted version of the R-squared that accounts for the number of predictors in the model.

F-statistic: This is the test statistic for the overall significance of the model. It tests whether at least one coefficient in the model is different from zero. The low p-value (close to zero) indicates that the model as a whole is significant.

The p-values associated with each coefficient (Pr(>|t|)) are very close to zero, indicating that all three predictors (age, BMI, and smoking status) are statistically significant predictors of charges.