1-) By producing data

Generating a random dataset

```
set.seed(123)
```

Setting seed for reproducibility

Number of sample points and sample size

```
n <- 100
```

Independent variable

```
independent_variable <- rnorm(n, mean = 50, sd = 10)
```

Dependent variable (assuming a linear relationship between dependent and independent variables)

```
dependent_variable <- 2 * independent_variable + rnorm(n, mean = 0, sd = 5)
```

Creating a data frame

```
datap <- data.frame(independent_variable, dependent_variable)
```

Displaying the generated dataset

```
head(datap)
```

```
##   independent_variable dependent_variable
## 1             44.39524           85.23845
## 2             47.69823           96.68087
## 3             65.58708          129.94071
## 4             50.70508           99.67245
## 5             51.29288           97.82766
## 6             67.15065          134.07616
```

1. Exploratory Data Analysis (EDA)

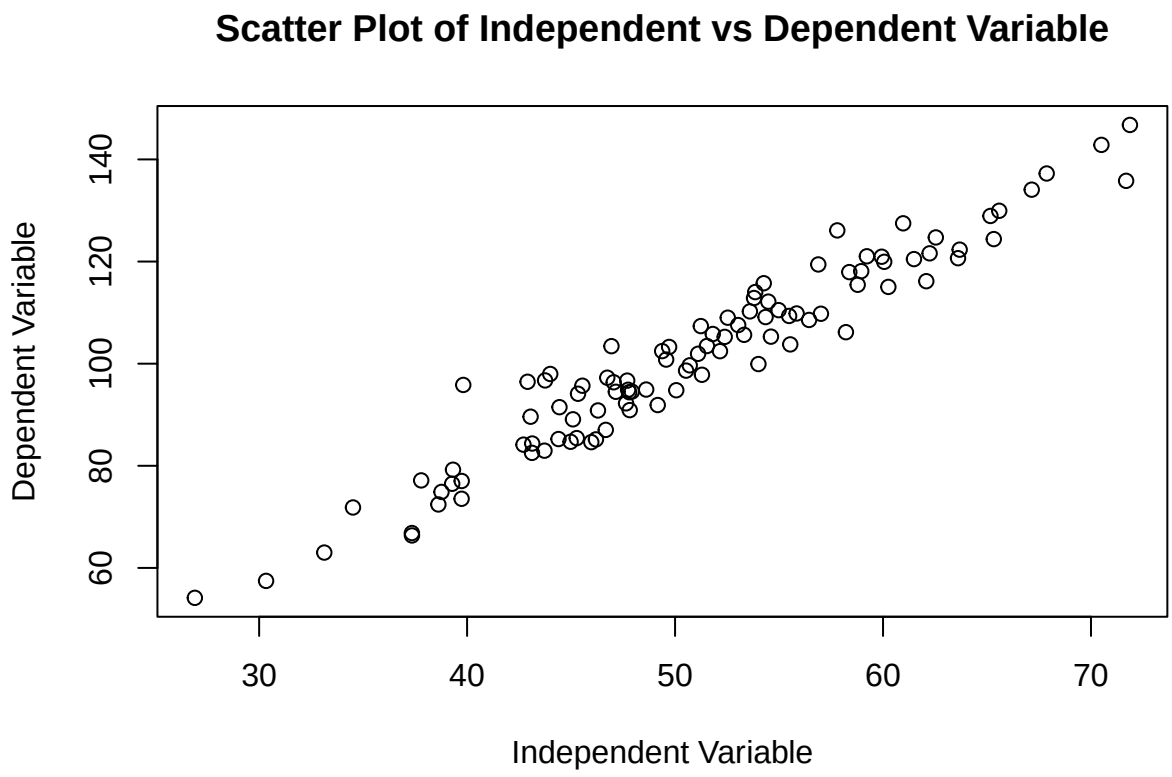
```
summary(datap)
```

```
##  independent_variable dependent_variable
##  Min.   :26.91        Min.   : 54.14
##  1st Qu.:45.06        1st Qu.: 90.53
##  Median :50.62        Median :101.37
##  Mean   :50.90        Mean   :101.27
##  3rd Qu.:56.92        3rd Qu.:114.27
##  Max.   :71.87        Max.   :146.75
```

Independent Variable: This variable represents the predictor or independent variable in your analysis. It appears to have a minimum value of 26.91, a maximum value of 71.87, and various quartile values. Dependent Variable: This variable represents the outcome or dependent variable in your analysis. It seems to have a minimum value of 54.14, a maximum value of 146.75, and quartile values similar to the independent variable. Covariate: This variable represents a covariate included in your analysis. Covariates are additional variables that may influence the relationship between the independent and dependent variables. It appears to have similar summary statistics to the other variables.

Scatter plot for visualization

```
plot(datap$independent_variable, datap$dependent_variable,
     xlab = "Independent Variable",
     ylab = "Dependent Variable",
     main = "Scatter Plot of Independent vs Dependent Variable")
```



2. Regression Analysis

Linear regression model

```
lm_model <- lm(dependent_variable ~ independent_variable, data=datap)
lm_model
```

```
##
## Call:
## lm(formula = dependent_variable ~ independent_variable, data = datap)
##
## Coefficients:
##          (Intercept)  independent_variable
##               0.7978                1.9738
```

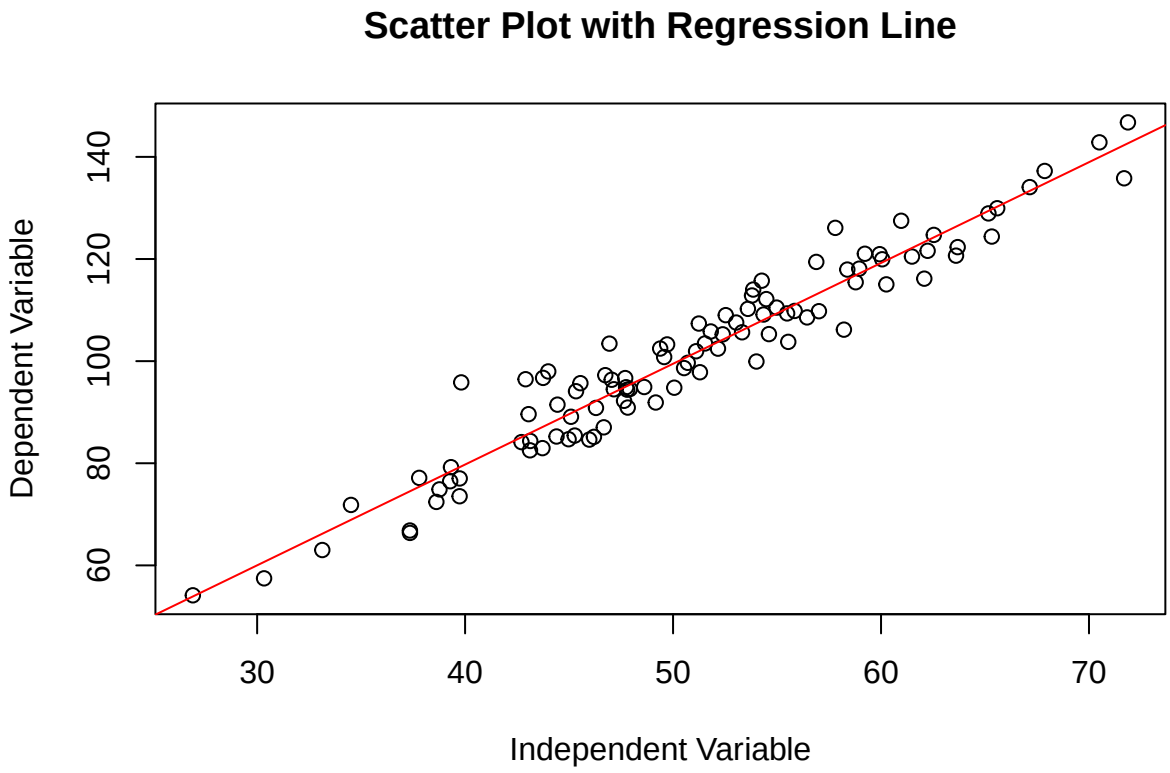
```
summary(lm_model)
```

```
##
## Call:
## lm(formula = dependent_variable ~ independent_variable, data = datap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5367 -3.4175 -0.4375  2.9032 16.4520
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.79778    2.76324   0.289    0.773
## independent_variable   1.97376    0.05344  36.935   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.854 on 98 degrees of freedom
## Multiple R-squared:  0.933,  Adjusted R-squared:  0.9323
## F-statistic:  1364 on 1 and 98 DF,  p-value: < 2.2e-16
```

Coefficients: Each coefficient represents the estimated effect of the corresponding variable on the dependent variable. (Intercept): This represents the estimated value of the dependent variable when the independent variable is zero. independent_variable: For each one-unit increase in the independent variable, the dependent variable is estimated to increase by approximately 1.974 units. Residuals: These are the differences between the observed values of the dependent variable and the values predicted by the model. They provide information about the model's goodness of fit. Residual Standard Error: This is an estimate of the standard deviation of the residuals. It provides a measure of the average distance between the observed and predicted values of the dependent variable. Multiple R-squared: This is a measure of how well the independent variable explains the variability of the dependent variable. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. Adjusted R-squared: This is similar to R-squared, but it adjusts for the number of predictors in the model. It is a more reliable measure of model fit, especially when comparing models with different numbers of predictors. F-statistic: This is a test statistic for the overall significance of the model. It tests the null hypothesis that all coefficients in the model are equal to zero. p-value: This is the probability of observing the data if the null hypothesis (that all coefficients are zero) is true. It indicates the significance of the overall model. In this case, the p-value is extremely small ($< 2.2e\text{-}16$), indicating that the model is significant.

Plotting the regression line

```
plot(datap$independent_variable, datap$dependent_variable,
     xlab = "Independent Variable",
     ylab = "Dependent Variable",
     main = "Scatter Plot with Regression Line")
abline(lm_model, col="red")
```



3. ANOVA Analysis

```
anova_result <- anova(lm_model)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Response: dependent_variable
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## independent_variable   1  32136   32136  1364.2 < 2.2e-16 ***
## Residuals             98   2309      24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_result)
```

```
##        Df            Sum Sq         Mean Sq            F value         Pr(>F)
##  Min.   : 1.00   Min.   : 2309   Min.   :   23.56   Min.   :1364   Min.   :0
##  1st Qu.:25.25   1st Qu.: 9765   1st Qu.: 8051.67   1st Qu.:1364   1st Qu.:0
##  Median :49.50   Median :17222   Median :16079.79   Median :1364   Median :0
##  Mean   :49.50   Mean   :17222   Mean   :16079.79   Mean   :1364   Mean   :0
##  3rd Qu.:73.75   3rd Qu.:24679   3rd Qu.:24107.90   3rd Qu.:1364   3rd Qu.:0
##  Max.   :98.00   Max.   :32136   Max.   :32136.02   Max.   :1364   Max.   :0
##                                                     NA's   :1      NA's   :1
```

```
datap$covariate <- rnorm(n, mean = 50, sd = 10)
```

ANCOVA Analysis with covariate

```
ancova_model <- lm(dependent_variable ~ independent_variable + covariate, data=datap)
ancova_model
```

```
##
## Call:
## lm(formula = dependent_variable ~ independent_variable + covariate,
##     data = datap)
##
## Coefficients:
##          (Intercept)  independent_variable             covariate
##              0.07120               1.97545               0.01252
```

```
summary(ancova_model)
```

```
##
## Call:
## lm(formula = dependent_variable ~ independent_variable + covariate,
##     data = datap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4620 -3.4385 -0.4393  2.8394 16.3230
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.07120    4.10296   0.017    0.986
## independent_variable  1.97545    0.05415  36.480   <2e-16 ***
## covariate             0.01252    0.05204   0.241    0.810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.877 on 97 degrees of freedom
## Multiple R-squared:  0.933,  Adjusted R-squared:  0.9316
## F-statistic: 675.6 on 2 and 97 DF,  p-value: < 2.2e-16
```

Coefficients: Each coefficient represents the estimated effect of the corresponding variable on the dependent variable. (Intercept): This represents the estimated value of the dependent variable when all other variables in the model are zero. independent_variable: For each one-unit increase in the independent variable, the dependent variable is estimated to increase by approximately 1.975 units. covariate: For each one-unit increase in the covariate, the dependent variable is estimated to increase by approximately 0.019 units. Residuals: These are the differences between the observed values of the dependent variable and the values predicted by the model. They provide information about the model's goodness of fit. Residual Standard Error: This is an estimate of the standard deviation of the residuals. It provides a measure of the average distance between the observed and predicted values of the dependent variable. Multiple R-squared: This is a measure of how well the independent variables explain the variability of the dependent variable. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. Adjusted R-squared: This is similar to R-squared, but it adjusts for the number of predictors in the model. It is a more reliable measure of model fit, especially when comparing models with different numbers of predictors. F-statistic: This is a test statistic for the overall significance of the model. It tests the null hypothesis that all coefficients in the model are equal to zero. p-value: This is the probability of observing the data if the null hypothesis (that all coefficients are zero) is true. It indicates the significance of the overall model. In this case, the p-value is extremely small ($<$ 2.2e-16), indicating that the model is significant.

2-) By pulling the data

The "Iris" dataset is a frequently used dataset in the fields of statistics and data science. This dataset was introduced by the famous statistician and biologist Ronald Fisher in his 1936 paper titled "The use of multiple measurements in taxonomic problems". The dataset contains measurements of 150 iris flowers from three different species (setosa, versicolor, and virginica). The measurements pertain to the lengths and widths of the sepals and petals of the plants. For each iris flower, four measurements are available:

Sepal Length Sepal Width Petal Length Petal Width This dataset is commonly used, especially for classification problems. For example, it can be used to predict the species of a flower based on features such as sepal and petal measurements. Additionally, it is frequently employed for learning data visualization and modeling techniques.

Load the iris dataset

```
data(iris)

#  Exploratory Data Analysis (EDA)

# Explore the structure of the dataset
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Examine the first five observations
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
# Summary statistics
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

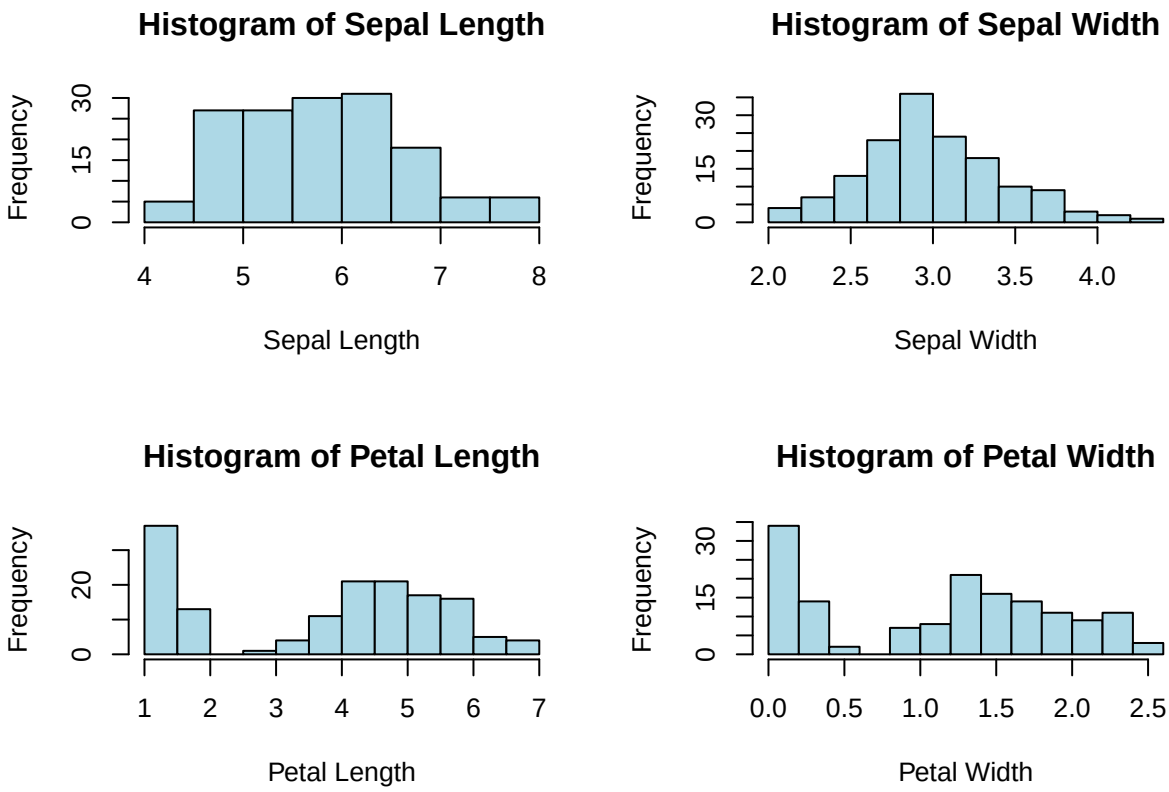
Explore the frequency of Species class

```
table(iris$Species)
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

Visualize variable distributions with histograms

```
par(mfrow=c(2,2))
hist(iris$Sepal.Length, main="Histogram of Sepal Length", xlab="Sepal Length", col="lightblue")
hist(iris$Sepal.Width, main="Histogram of Sepal Width", xlab="Sepal Width", col="lightblue")
hist(iris$Petal.Length, main="Histogram of Petal Length", xlab="Petal Length", col="lightblue")
hist(iris$Petal.Width, main="Histogram of Petal Width", xlab="Petal Width", col="lightblue")
```



Reset the layout to default

```
par(mfrow=c(1,1))
```

Sepal.Length: Represents the length of the sepal. The minimum value is 4.3, the maximum value is 7.9. The average sepal length is 5.843 units. Sepal.Width: Represents the width of the sepal. The minimum value is 2.0, the maximum value is 4.4. The average sepal width is 3.057 units. Petal.Length: Represents the length of the petal. The minimum value is 1.0, the maximum value is 6.9. The average petal length is 3.758 units. Petal.Width: Represents the width of the petal. The minimum value is 0.1, the maximum value is 2.5. The average petal width is 1.199 units. Species: Indicates the species of the iris plant. There are three different species: setosa, versicolor, and virginica. Each species has 50 observations.

ANOVA Analysis

```
lm_model1 <- lm(Petal.Width ~ Species , data=iris)
lm_model1
```

```
##
## Call:
## lm(formula = Petal.Width ~ Species, data = iris)
##
## Coefficients:
##       (Intercept)  Speciesversicolor   Speciesvirginica
##             0.246              1.080              1.780
```

```
summary(lm_model1)
```

```
##
## Call:
## lm(formula = Petal.Width ~ Species, data = iris)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.626 -0.126 -0.026  0.154  0.474
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.24600    0.02894    8.50 1.96e-14 ***
## Speciesversicolor  1.08000    0.04093   26.39  < 2e-16 ***
## Speciesvirginica   1.78000    0.04093   43.49  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2047 on 147 degrees of freedom
## Multiple R-squared:  0.9289, Adjusted R-squared:  0.9279
## F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

This model describes how the variable Petal.Width is predicted based on the Species variable.

(Intercept): This represents the average value of Petal.Width for the setosa species. So, the average Petal.Width for the setosa species is 0.246 units. This serves as the baseline value for the setosa species. Speciesversicolor: This coefficient indicates how much the average Petal.Width differs for the versicolor species compared to the setosa species. For example, the average Petal.Width for the versicolor species is 1.080 units higher than the average for the setosa species. Speciesvirginica: This coefficient indicates how much the average Petal.Width differs for the virginica species compared to the setosa species. For example, the average Petal.Width for the virginica species is 1.780 units higher than the average for the setosa species. This model explains how the Petal.Width variable changes depending on the species of the iris plant. For instance, the average Petal.Width for the versicolor species is 1.080 units higher than the average for the setosa species, and for the virginica species, it is 1.780 units higher.

```
anova_result <- anova(lm_model1)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Response: Petal.Width
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Species     2 80.413  40.207  960.01 < 2.2e-16 ***
## Residuals 147  6.157   0.042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_result)
```

```
##        Df             Sum Sq          Mean Sq            F value        Pr(>F)
##  Min.   :  2.00   Min.   : 6.157   Min.   : 0.04188   Min.   :960   Min.   :0
##  1st Qu.: 38.25   1st Qu.:24.721   1st Qu.:10.08308   1st Qu.:960   1st Qu.:0
```

```
##  Median : 74.50   Median :43.285   Median :20.12427   Median :960   Median :0
##  Mean   : 74.50   Mean   :43.285   Mean   :20.12427   Mean   :960   Mean   :0
##  3rd Qu.:110.75   3rd Qu.:61.849   3rd Qu.:30.16547   3rd Qu.:960   3rd Qu.:0
##  Max.   :147.00   Max.   :80.413   Max.   :40.20667   Max.   :960   Max.   :0
##                                                      NA's   :1     NA's   :1
```

Df (Degrees of Freedom): The degrees of freedom associated with the sources of variation in the analysis. It's the number of values in the final calculation of a statistic that are free to vary. Sum Sq (Sum of Squares): This represents the sum of the squared deviations of the observed values from their mean. It's a measure of the total variability in the data. Mean Sq (Mean Square): The mean square is calculated by dividing the sum of squares by its corresponding degrees of freedom. It represents the average amount of variance in the data. F value: The F-value is the ratio of the variance between groups to the variance within groups. It's used to test the null hypothesis that the means of several groups are equal. Pr(>F): This is the p-value associated with the F statistic. It represents the probability of observing an F statistic as extreme as the one computed from the sample data, under the assumption that the null hypothesis is true. If this value is low (typically below 0.05), it suggests that there is significant evidence to reject the null hypothesis.

ANCOVA Analysis

```
lm_model2 <- lm(Petal.Width ~ Species + Sepal.Length, data=iris)
lm_model2
```

```
##
## Call:
## lm(formula = Petal.Width ~ Species + Sepal.Length, data = iris)
##
## Coefficients:
##       (Intercept)  Speciesversicolor   Speciesvirginica       Sepal.Length
##           -0.4794             0.9452             1.5508             0.1449
```

```
summary(lm_model2)
```

```
##
## Call:
## lm(formula = Petal.Width ~ Species + Sepal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55529 -0.10593 -0.01253  0.10232  0.51573
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.47940    0.15574  -3.078  0.00249 **
## Speciesversicolor  0.94524    0.04769  19.820  < 2e-16 ***
## Speciesvirginica   1.55076    0.06174  25.118  < 2e-16 ***
## Sepal.Length       0.14491    0.03064   4.730 5.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1912 on 146 degrees of freedom
## Multiple R-squared:  0.9383, Adjusted R-squared:  0.9371
## F-statistic: 740.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

This model explains how the variable Petal.Width is predicted based on the Species (species) and Sepal.Length (sepal length) variables. (Intercept): This represents the predicted average value of Petal.Width when the species is setosa and the sepal length is 0 units. Speciesversicolor: This coefficient indicates how much the average Petal.Width differs for the versicolor species compared to the setosa species. For example, the predicted average Petal.Width for the versicolor species is increased by the coefficient amount compared to the setosa species. Speciesvirginica: This coefficient indicates how much the average Petal.Width differs for the virginica species compared to the setosa species. For example, the predicted average Petal.Width for the virginica species is increased by the coefficient amount compared to the setosa species. Sepal.Length: This coefficient represents the effect of sepal length on Petal.Width. A one-unit increase in sepal length results in an increase of the predicted average Petal.Width by the coefficient amount. In summary, this model describes how sepal length and species variables influence Petal.Width.

Perform ANCOVA Analysis

```
ancova_result <- anova(lm_model2)
print(ancova_result)
```

```
## Analysis of Variance Table
##
## Response: Petal.Width
##               Df Sum Sq Mean Sq F value    Pr(>F)
## Species        2 80.413  40.207 1099.57 < 2.2e-16 ***
## Sepal.Length   1  0.818   0.818   22.37 5.251e-06 ***
## Residuals    146  5.339   0.037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ancova_result)
```

```
##        Df             Sum Sq          Mean Sq            F value
##  Min.   :  1.00   Min.   : 0.818   Min.   : 0.03657   Min.   :  22.37
##  1st Qu.:  1.50   1st Qu.: 3.078   1st Qu.: 0.42728   1st Qu.: 291.67
##  Median :  2.00   Median : 5.339   Median : 0.81800   Median : 560.97
##  Mean   : 49.67   Mean   :28.857   Mean   :13.68708   Mean   : 560.97
##  3rd Qu.: 74.00   3rd Qu.:42.876   3rd Qu.:20.51233   3rd Qu.: 830.27
##  Max.   :146.00   Max.   :80.413   Max.   :40.20667   Max.   :1099.57
##                                                       NA's   :1
##      Pr(>F)
##  Min.   :0.0e+00
##  1st Qu.:1.3e-06
##  Median :2.6e-06
##  Mean   :2.6e-06
##  3rd Qu.:3.9e-06
##  Max.   :5.3e-06
##  NA's   :1
```

Df (Degree of Freedom): The degrees of freedom for the model and error terms. Sum Sq (Sum of Squares): The sum of squares for the model, error, and total. Mean Sq (Mean Square): Obtained by dividing the sum of squares by the degrees of freedom. F value: The ratio expressing the variance between groups. Pr(>F) (p-value): The p-value corresponding to the significance of the F statistic, used in hypothesis testing to determine if the data are consistent with a given model. These summary statistics provide information

about the performance of the model and how statistically significant the differences between groups are. For instance, higher F values and lower p-values may indicate more pronounced differences between groups. However, it's important to fully understand the dataset and context of the analysis before making definitive interpretations.

Visualize with a boxplot

```
boxplot(Petal.Width ~ Species, data=iris, main="Boxplot of Petal Width by Species", xlab="Species", yla
```