Data Description

The Iris dataset is a classic dataset from the field of machine learning and statistics. It contains data about 150 iris flowers from three different species: Setosa, Versicolor, and Virginica. There are 50 samples from each of three species. The dataset includes four features measured from each sample: the lengths and the widths of the sepals and petals in centimeters.

Variables

Sepal.Length:length of the sepals in centimeters.

Sepal.Width:width of the sepals in centimeters.

Petal.Length:length of the petals in centimeters.

Petal.Width:width of the petals in centimeters.

Species:species of the iris flower(Setosa, Versicolor, Virginica).

In [2]:
```
# Loading the dataset
data(iris)

head(iris)

#Getting summary of the dataset
summary(iris)
```

A data.frame: 6 × 5

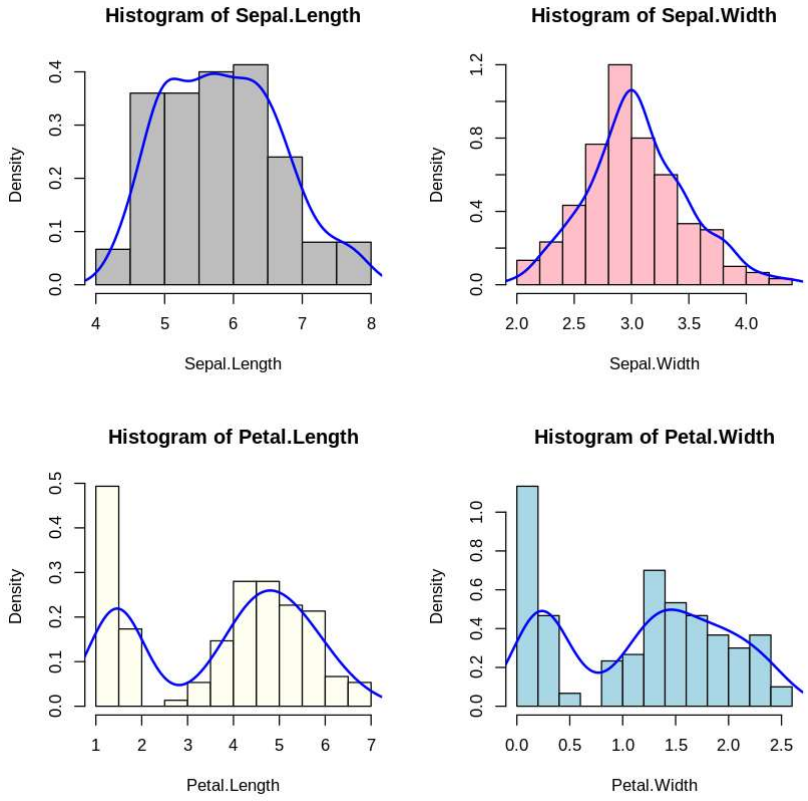
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

```
  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
```

Exploratory Data Analysis

Generating histograms for each of the Iris dataset's continuous variables- sepal length, sepal width, petal length, and petal width—in a 2x2 grid layout, with density lines to depict the distribution shape. This visual EDA approach shows insights into the data's distribution.

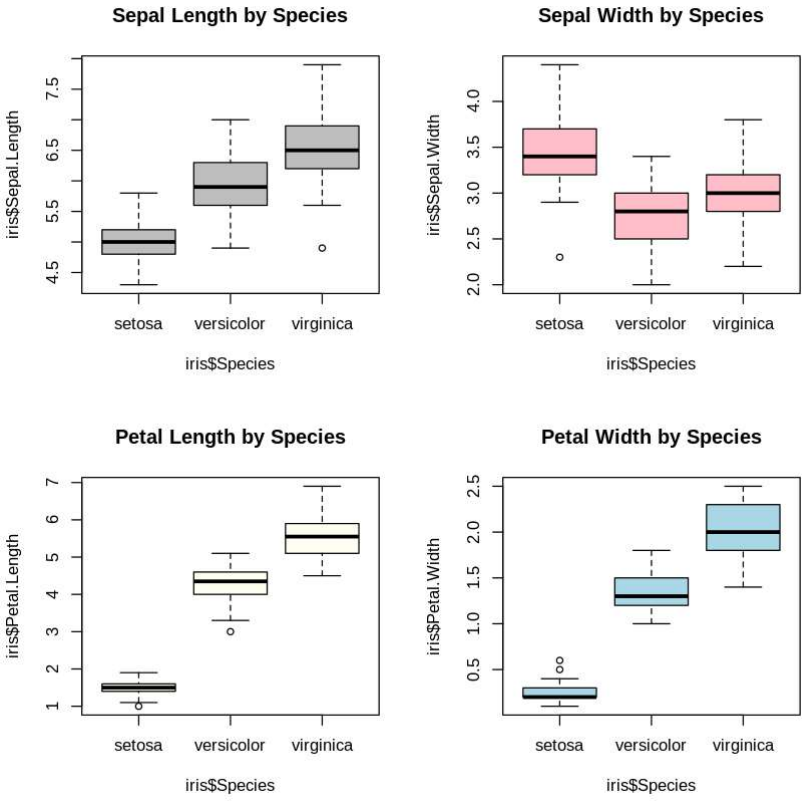
In [3]:
```
#Seting the layout for a 2x2 arrangement
par(mfrow=c(2,2))

colors <- c("grey", "pink", "ivory", "lightblue")
for(i in 1:4) {
  #xreating histogram
  hist(iris[,i], main=paste("Histogram of", names(iris)[i]),
       xlab=paste(names(iris)[i]), col=colors[i], freq=FALSE)
  #Also adding density line
  density_data <- density(iris[,i])
  lines(density_data, col="blue", lwd=2)
}
par(mfrow=c(1, 1))
```



Sepal length and sepal width shows approximate normal distributions indicated by the bell-shaped curves of the density plots. The sepal length shows a unimodal distribution with a slight skew towards larger lengths, while sepal width is also unimodal with a peak around 3 cm, suggesting most iris flowers have a sepal width within a moderate range.

In contrast, petal length and width exhibit bimodal distributions. Petal length's distribution suggests two distinct groups, potentially correlating with species differentiation—one group with shorter petals and another with significantly longer petals. Overall, the histograms and density plots provide a visual summary of the data's distribution offering insights into variations among the Iris species.

In [4]:
```
par(mfrow=c(2,2))
boxplot(iris$Sepal.Length ~ iris$Species, main="Sepal Length by Species", col="grey")
boxplot(iris$Sepal.Width ~ iris$Species, main="Sepal Width by Species", col="pink")
boxplot(iris$Petal.Length ~ iris$Species, main="Petal Length by Species", col="ivory")
boxplot(iris$Petal.Width ~ iris$Species, main="Petal Width by Species", col="lightblue
```



The box plots for the Iris dataset shows its different species variations and outliers in sepal and petal dimensions. Iris setosa is distinct with smaller petals and wider sepals, whereas Iris virginica and versicolor have longer petals, with virginica being the largest. Outliers are present particularly in sepal width for setosa and petal width for virginica, suggesting individual variation or measurement errors.

Research Question

How do Structural features correlate with response variable, and can we statistically validate the significance of these relationships? Specifically, which features strongly predict petal width?

Performing Linear Regression

Splitting some part of the dataset for training and testing process without including the categorical varaible(Species) for performing the linear regression model.

In [5]:
```
#Splitting the data into train and test datasets
train <- iris[1:120,-5]
test <- iris[121:150,-5]

head(train)
```

A data.frame: 6 × 4

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

This code performs a series of linear regressions to predict Petal Width from other variables in the Iris dataset, excluding the Species. It calculates the Mean Squared Prediction Error (MSPE) for each model using the test set. The process iterates through each of the predictors combinations, each time removing one predictor, and identifies the best-performing model with the lowest MSPE, which is stored for future use.

In [6]:
```
# MSPE function to calculate MSPE values for each model
MSPE <- function(model, test) {
  actualPrices = test$Petal.Width
  predictedPrices = predict(model, test)
  squaredErrors = (actualPrices - predictedPrices)^2
  mspe = mean(squaredErrors)
  return(mspe)
}

model = lm(Petal.Width ~ ., data = train)
initialMSPE <- MSPE(model, test)
print(summary(model))
cat("\nInitial MSPE of full model:", initialMSPE, "\n")

#storing the best model
bestmodel <- model
lowestMSPE <- initialMSPE

#backward selection
variables <- names(coef(model))
while(length(variables) > 1) {
  # Excluding the intercept value from p-values
  p_values <- summary(model)$coefficients[-1, 4]
  max_p_value <- max(p_values)
  if(max_p_value > 0.3) {
    variable_to_remove <- names(p_values)[which.max(p_values)]
    model <- update(model, formula = paste(". ~ . - Species-", variable_to_remove))
    currentMSPE = MSPE(model, test)
    print(summary(model))
    cat("MSPE after removing", variable_to_remove, ":", currentMSPE, "\n")

    # Updating best model
    if (currentMSPE < lowestMSPE) {
      lowestMSPE <- currentMSPE
      bestmodel <- model
    }
```

```
    variables <- names(coef(model))

  } else {
    break

  }
}
cat("\nLowest MSPE observed:", lowestMSPE,"\n")
```

```
Call:
lm(formula = Petal.Width ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35041 -0.08847 -0.01330  0.08623  0.59398 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.17947    0.16050  -1.118    0.266    
Sepal.Length -0.18538    0.04396  -4.218 4.92e-05 ***
Sepal.Width   0.18243    0.04333   4.210 5.06e-05 ***
Petal.Length  0.49998    0.02267  22.057  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1528 on 116 degrees of freedom
Multiple R-squared:  0.9545,    Adjusted R-squared:  0.9533 
F-statistic: 810.8 on 3 and 116 DF,  p-value: < 2.2e-16


Initial MSPE of full model: 0.09499934

Lowest MSPE observed: 0.09499934
```

The optimal model using all continuous predictors which are `Sepal.Length`, `Sepal.Width` and `Petal.Length` as predictors achieved low MSPE of 0.09499934, which shows excellent prediction accuracy. This model got 95.45% of the variance (with an adjusted R-squared of 95.33%), highlighting its strong explanatory power. Also, with an F-statistic of 810.8 on 116 degrees of freedom and a highly significant p-value, this model proves to be exceptionally great for predicting Peta.Width values based on the specified predictors.

Now, We have achieved best model according to the MSPE, we will be perform some diagnostics on the data for which the best model has performed on.

Diagnostics

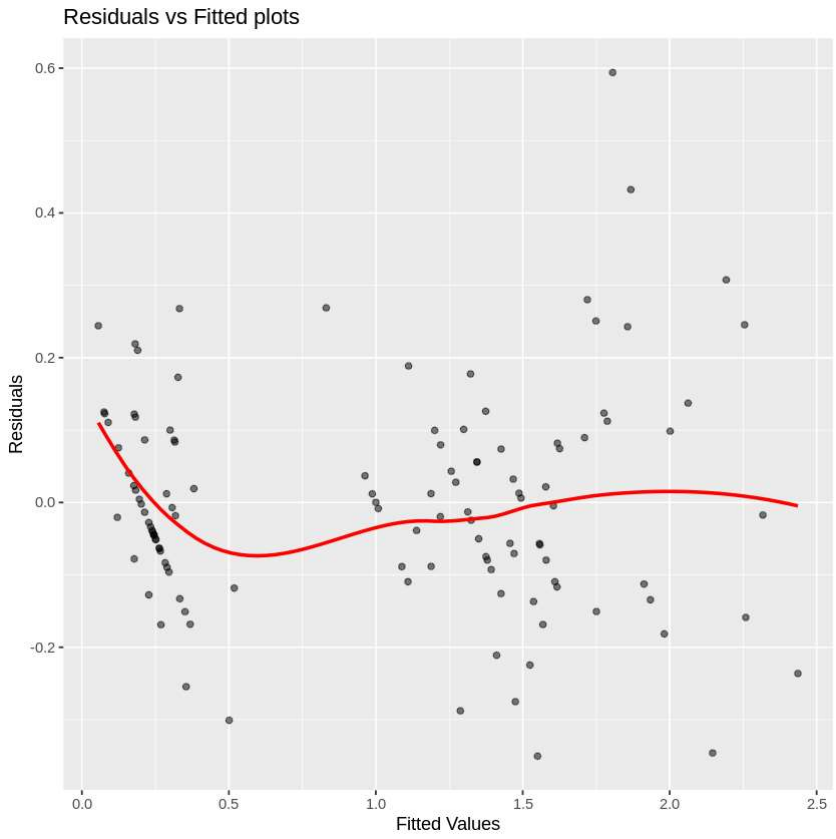
Diagnostics help verify if the model assumptions of linearity, independence, homoscedasticity (constant variance), and normality of residuals are met.

Creating a diagnostic dataframe to plot between residuals adn fitted values to check for violations

In [7]:
```
#creating a diagnostic dataframe
df_diagnostics = data.frame(yhat = fitted(bestmodel), r = resid(bestmodel), y = train$
```

In [8]:
```
#Residuals vs Fitted plot for the check for some violation on Non-Constant variance an
library(ggplot2)
ggplot(df_diagnostics, aes(x = yhat, y = r)) +
    geom_point(alpha = 0.5) +
    geom_smooth(se = F, col = "red") +
    xlab("Fitted Values") + ylab("Residuals")+
    ggtitle("Residuals vs Fitted plots")
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



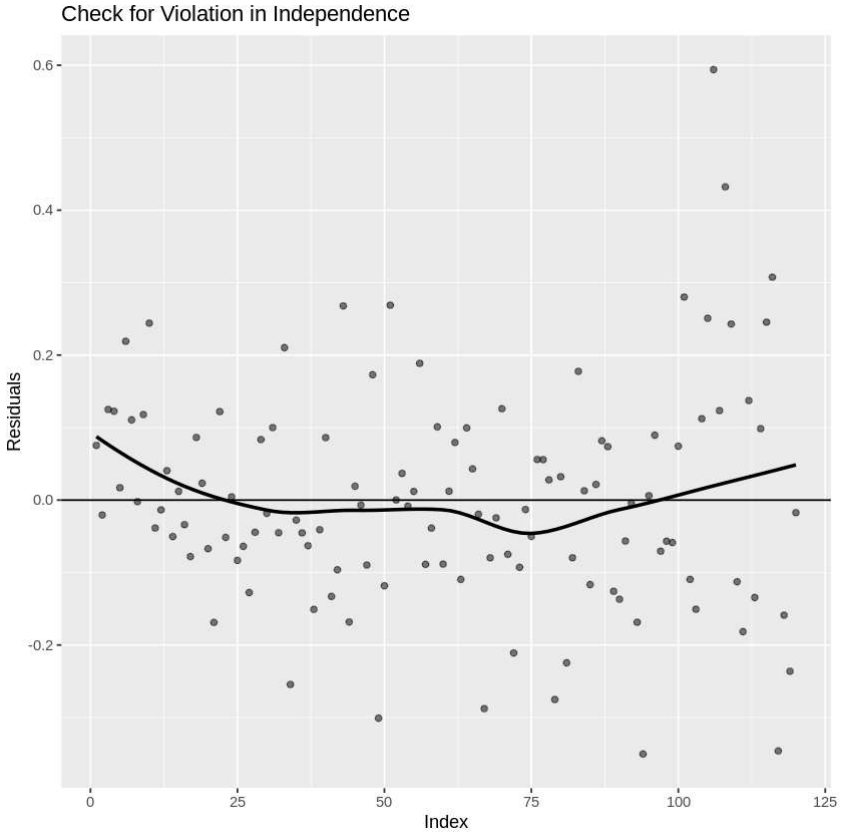
In [9]:
```
#test for violation of constant variation
install.packages("lmtest")
library(lmtest)

# Assuming `model` is your linear model object
bptest(model)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependency ‘zoo’


Loading required package: zoo


Attaching package: ‘zoo’


The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric
```

```
        studentized Breusch-Pagan test

data:  model
BP = 14.395, df = 3, p-value = 0.002414
```

In [10]:
```
#Residuals vs Index plot to check for the violation of Independence
library(tidyverse)
df_diagnostics_order = arrange(df_diagnostics, petal_len)
ggplot(df_diagnostics_order, aes(x = 1:length(train$Petal.Width), y = r)) +
    geom_point(alpha = 0.5) +
    geom_abline(slope = 0, intercept = 0) +
    xlab("Index") +
    geom_smooth(se = F, col = "black") +
    ylab("Residuals") +
    ggtitle("Check for Violation in Independence")
```

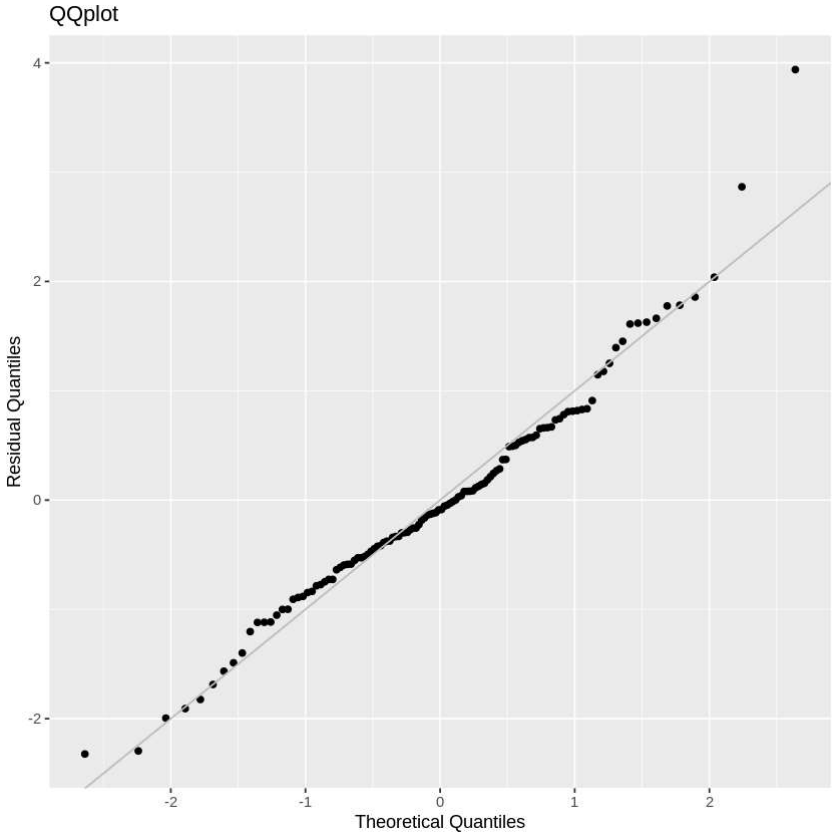
```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ lubridate 1.9.3     ✓ tibble    3.2.1
✓ purrr     1.0.2     ✓ tidyr     1.3.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ purrr::%||%()   masks base::%||%()
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



In [11]:
```
#test for violation of independence
dwtest(model)
```

```
        Durbin-Watson test

data:  model
DW = 1.6613, p-value = 0.02386
alternative hypothesis: true autocorrelation is greater than 0
```

In [12]:
```
#Q-Q plot to check for the violation of Normality
ggplot(df_diagnostics, aes(sample = (r - mean(r))/sd(r))) +
    stat_qq() + geom_abline(slope = 1, intercept = 0, col = "grey") +
    xlab("Theoretical Quantiles") +
    ylab("Residual Quantiles") +
    ggtitle("QQplot")
```



Constant Variation: The test results and residual plots together provide a mixed view on homoscedasticity. While the residual vs fitted plot is not strongly indicating heteroscedasticity, the Breusch-Pagan test with a p-value of 0.002414 suggests a significant deviation from constant variance. This discrepancy implies that there is enough evidence to confirm the presence of heteroscedasticity.

Independence: The Durbin-Watson test's p-value of 0.02386, along with the pattern in the residuals which says the spread of residuals are sparsed not only around zero, points to a violation of the independence assumption.

Linearity: From the Residuals vs Fitted graph, we can say that residuals vs fitted plot suggests that the relationship between the predictors and the response may not be perfectly linear. So we have violation of linearity in the model.

Normality: - From the QQ plot in the diagnostics it can be seen that the points did not follow a straight line y=x line at the extreme levels. So, we can say that it violates the Normality.

MODEL SELECTION (AIC, BIC, Adjusted R^2)

Even if we have a best model based on one criterion(MSPE), checking against others(AIC, BIC, Adjusted R^2) is valuable as they can offer different perspectives on model quality.

Performing Model Selction to get other best models on our IRIS dataset based on Alkaline In formation Criterion, Bayesian Information Criterion, Adjusted R^2 by Performing Best subset

regression which is a method of examining all the possible combinations of predictors to determine which best predicts the Outcome Variable (Petal.Width).

In [13]:
```
#Installing leaps to perform subset regression
install.packages('leaps')
library(leaps)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

In [14]:
```
#Taking subset of max 4 as we only have 4 total variables after excluding the 'Species
reg_subsets <- regsubsets(Petal.Width ~ ., data = train, nvmax = 4)
summary_subset <- summary(reg_subsets)
print(summary_subset$which)
#Identifying the best models according to BIC and adjusted R^2
bic <- which.min(summary_subset$bic)
Adjr2a <- which.max(summary_subset$adjr2)
cat("\nThe model which is best based on BIC has", bic, "predictors.\n")
cat("The model which is best based on Adjusted R^2 has", Adjr2a, "predictors.\n")
```

```
  (Intercept) Sepal.Length Sepal.Width Petal.Length
1        TRUE        FALSE       FALSE         TRUE
2        TRUE         TRUE       FALSE         TRUE
3        TRUE         TRUE        TRUE         TRUE

The model which is best based on BIC has 3 predictors.
The model which is best based on Adjusted R^2 has 3 predictors.
```

In the Above output, it indicates that the best model according to both BIC and adjusted R^2 includes 3 predictors. The $which table shows TRUE for Sepal.Length, Sepal.Width, and Petal.Length which says that these are the predictors included in the best model according to both criterions.

In [15]:
```
n <- nrow(train)
#initialising a vector to store AIC values
aic_models <- numeric(length = 3)
#calculating AIC for each model
for (i in 1:3) {
    k <- sum(summary_subset$which[i,])
    rss <- summary_subset$rss[i]
    aic_models[i] <- n * log(rss/n) + 2 * k
}
# Determining the best model based on minimum AIC
best_aic <- which.min(aic_models)
cat("The best model based on AIC has", best_aic, "predictors.\n")
```

```
The best model based on AIC has 3 predictors.
```

Akaike Information Criterion (AIC) concluded that the best model for predicting Petal.Width from the training dataset includes three predictors. This model selection was determined by evaluating various combinations where the model with three predictors gave the lowest AIC score.

Now, I will be Individually performing linear models by taking some combination of predictors to predict Petal.Width inorder to perform and validate some tests and Confidence Intervals

In [16]:
```
#Randomly performing linear regression using different set of predictors
#for further analysis
model1<- lm(Petal.Width~., data= train)
model2<- lm(Petal.Width~ Petal.Length+Sepal.Length, data= train)
model3<- lm(Petal.Width~ Sepal.Width+Sepal.Length, data= train)
```

Confidence Intervals

In [17]:
```
coef1 <- coef(summary(model1))
coef2 <- coef(summary(model2))

#difference in coefficients and their standard errors
diff_coef <- coef1[, "Estimate"] - coef2[, "Estimate"]
se_diff <- sqrt(coef1[, "Std. Error"]^2 + coef2[, "Std. Error"]^2)

#Calcultating 95% confidence intervals for the difference
lower_bound <- diff_coef - 1.96 * se_diff
upper_bound <- diff_coef + 1.96 * se_diff

data.frame(
  Lower_Bound = lower_bound,
  Upper_Bound = upper_bound
)
```

```
Warning message in coef1[, "Estimate"] - coef2[, "Estimate"]:
“longer object length is not a multiple of shorter object length”
Warning message in coef1[, "Std. Error"]^2 + coef2[, "Std. Error"]^2:
“longer object length is not a multiple of shorter object length”
```

A data.frame: 4 × 2

	Lower_Bound	Upper_Bound
	<dbl>	<dbl>
(Intercept)	-0.6100982	0.2977274
Sepal.Length	-0.7094957	-0.5248224
Sepal.Width	0.1439385	0.3685558
Petal.Length	0.1930509	0.8534879

The confidence intervals for the coefficients from the above processsuggest the following:

- The Intercept interval shifts from negative to positive, which means it's not statistically significant at this confidence level.
- Sepal.Length has a confidence interval entirely below zero, indicating that it has a statistically significant negative effect on Petal.Width.
- Sepal.Width has a significant level which is totally above zero, a positive interval suggesting a positive relationship with Petal.Width.
- Petal.Length has a confidence interval fully above zero, indicating a strong and statistically significant positive impact on Petal.Width.

These results helps us to identify which features of the flower are significant predictors of Petal.Width in the Iris dataset.

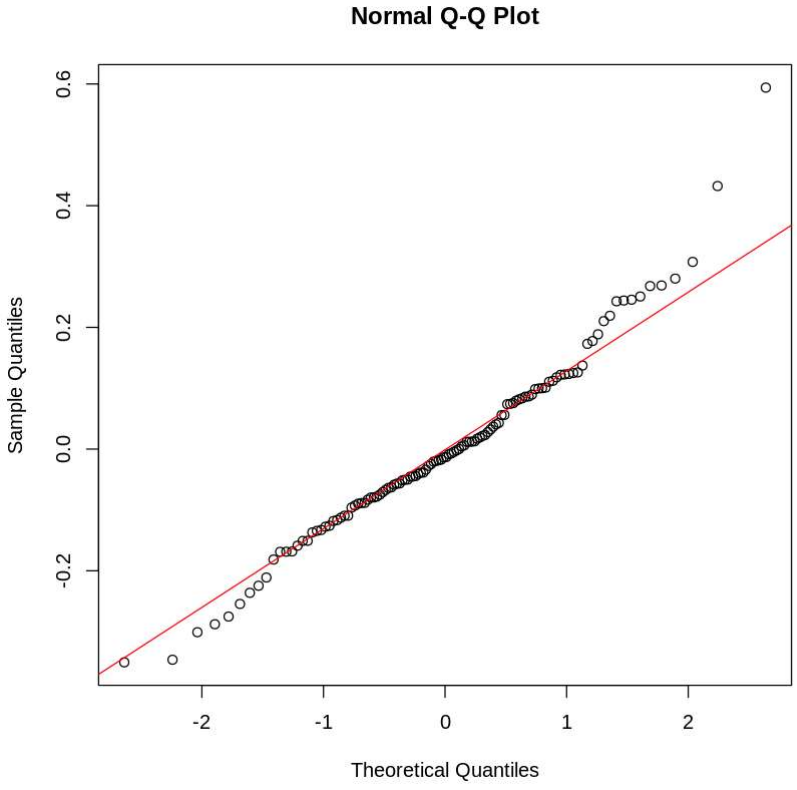
- Performing **Shapiro-Wilk** test which is used to test the normality of residuals in regression analysis for using anova tests. So, the assumed hypothesis statements would be:

  Null Hypothesis (H0): The model's data is normally distributed.

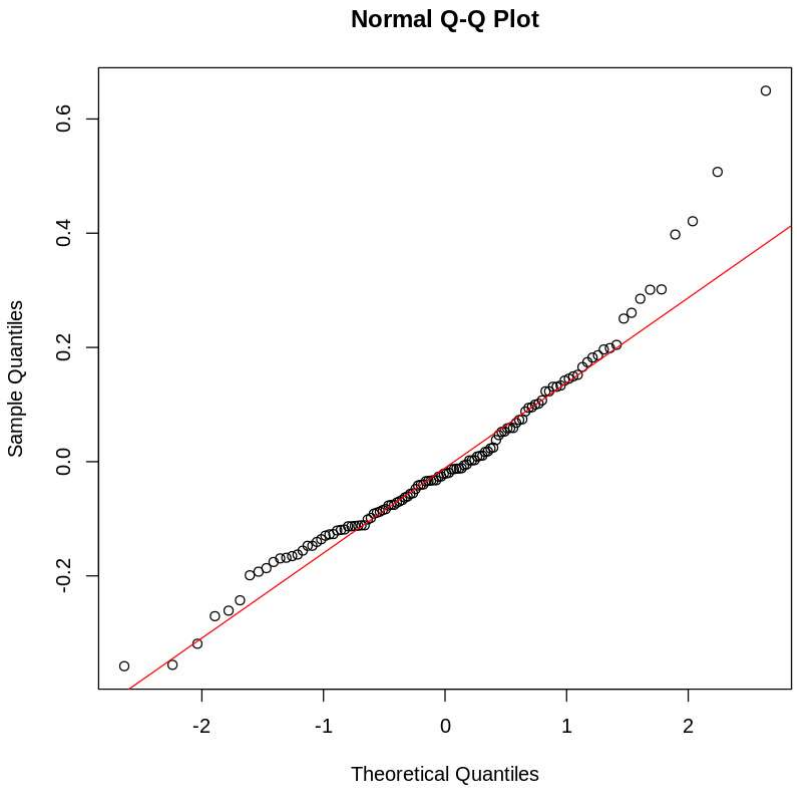
  Alternative Hypothesis (HA): The model's data is not normally distributed.

In [23]:
```
# Checking Normality for Model 1 using Shapiro-Wilk test
qqnorm(residuals(model1))
qqline(residuals(model1), col = "red")
shapiro.test(residuals(model1))

# Checking Normality for Model 1 using Shapiro-Wilk test
qqnorm(residuals(model2))
qqline(residuals(model2), col = "red")
shapiro.test(residuals(model2))

qqnorm(residuals(model3))
qqline(residuals(model3), col = "red")
shapiro.test(residuals(model3))

# Repeat for other models
```

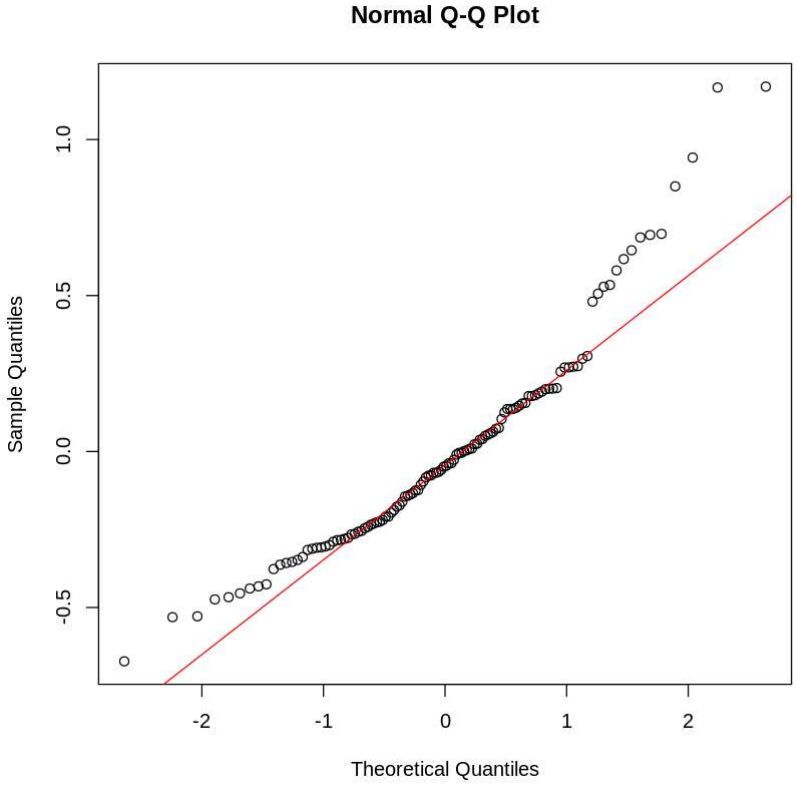
```
        Shapiro-Wilk normality test

data:  residuals(model1)
W = 0.97173, p-value = 0.01248
```



```
        Shapiro-Wilk normality test

data:  residuals(model2)
W = 0.95054, p-value = 0.0002369
```




```
        Shapiro-Wilk normality test

data:  residuals(model3)
W = 0.93106, p-value = 1.121e-05
```



The above tests proves us the all the p values from the shakiro-wil tests are less than the siginificant leve(0.05). It suggests thatt each adn every model are violating the assumptions of norrmality by rejecting the null hypothesis

- Succesively, performing the constant variance(homogeneity) test by using the levene test. This is important to validate before performing ANOVA to ensure that the variances in petal width predictions across the different models are not significantly different

In [29]:
```
install.packages('car')
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘MatrixModels’, ‘quantreg’


Warning message in install.packages("car"):
“installation of package ‘quantreg’ had non-zero exit status”
Warning message in install.packages("car"):
“installation of package ‘car’ had non-zero exit status”
```

In [37]:
```
library(car)
#Levenes test for assumption of constant variances
leveneTest(residuals(model1), group = model1$fitted.values)
leveneTest(residuals(model2), group = model2$fitted.values)
leveneTest(residuals(model3), group = model3$fitted.values)
```

```
Warning message in leveneTest.default(residuals(model1), group = model1$fitted.value
s):
“model1$fitted.values coerced to factor.”
Warning message in anova.lm(lm(resp ~ group)):
“ANOVA F-tests on an essentially perfect fit are unreliable”
```

A anova: 2 × 3

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	117	7.818693e+28	1.278986e-29
	2	NA	NA

```
Warning message in leveneTest.default(residuals(model2), group = model2$fitted.value
s):
“model2$fitted.values coerced to factor.”
```

A anova: 2 × 3

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	99	1.888384	0.05257758
	20	NA	NA

```
Warning message in leveneTest.default(residuals(model3), group = model3$fitted.value
s):
“model3$fitted.values coerced to factor.”
```

A anova: 2 × 3

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	98	0.9104464	0.6371091
	21	NA	NA

From the above Levene's test results, we can definitely say that 2 out of 3 the p-values for 3 models are greater than 0.05 suggests that there is no violation of constant variance different models and the other model has a p-value of 1.278986e-29 which is way less than 0.05.

So, from both of the test results for the assumptions of normality and constant variance, it is evident that performing ANOVA on these models is not suggested.

Despite the violation of key assumptions such as normality and homogeneity of variances, we performed with the ANOVA analysis for one reason- the nature of ANOVA is valuable in our context as it allows us to evaluate the relative performance of different models in explaining the variation in petal width. Also this analysis is important for identifying which model irrespective of the assumption violations provides the most statistically significant explanation of the data.

ANOVA Testing between 3 Models

Performing ANOVA with the above three models will help us understand whether the predictors included in each model have a statistically significant effect on predicting petal width which helps us identify- which structural features correlate strongly with petal width in Iris flowers.

In [19]:
```
#Comparing model1, model2, and model3 using ANOVA
anova_results <- anova(model1, model2, model3)
print(anova_results)
```

```
Analysis of Variance Table

Model 1: Petal.Width ~ Sepal.Length + Sepal.Width + Petal.Length
Model 2: Petal.Width ~ Petal.Length + Sepal.Length
Model 3: Petal.Width ~ Sepal.Width + Sepal.Length
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    116  2.7070
2    117  3.1206 -1   -0.4136 17.724 5.064e-05 ***
3    117 14.0597  0  -10.9391
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table above suggests that there is a statistically significant difference in the ability of these models to predict `Petal.Width`. Model 1, which includes all predictors `Sepal.Length`, `Sepal.Width`, and `Petal.Length` are compared against Model 2 and Model 3, each with a subset of these predictors. The F-statistic and its associated p-value (very less than 0.05) indicate that Model 1 fits the data significantly better than Model 2. There iss no comparison presented directly between Model 1 and Model 3 or Model 2 and Model 3 in the

results, but the p-value associated with Model 2 comparison suggests that the predictors in Model 1 provide additional explanation.

Generalised Linear Modelling

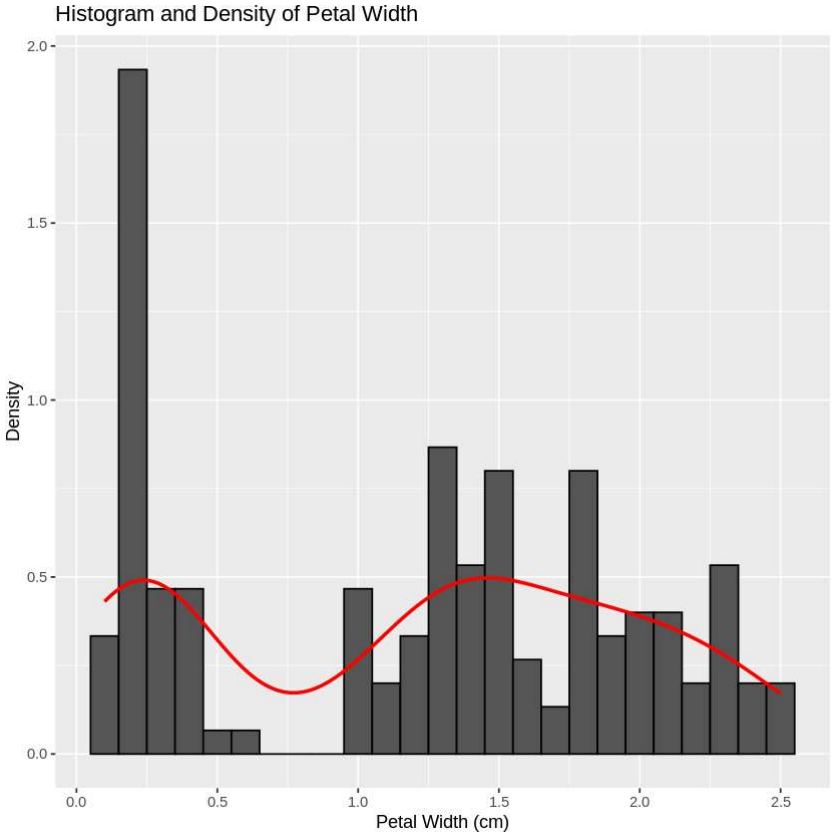
As our linear model violates the normality and linearity, it does mean that it attain non-normal and non-linear in nature. Thus, performing GLM on the same set of predictors to predict the response variable(Petal.Width) in order to check the performance which directly supports our investigation into the structural features their relationships to petal width.

GLM allow the dependent variable, Y, to be generated by any distribution function belonging to the exponential family. So, First I have to check the repsonse variable's distribution. Based on that, we have to chose the link functions and families which has to be performd in GLM.

In [20]:
```
library(ggplot2)
```

In [21]:
```
#Checking for distribution of the response variable
ggplot(iris, aes(x = Petal.Width)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.1, color="black") +
  geom_density(color="red", size=1) +
  ggtitle("Histogram and Density of Petal Width") +
  xlab("Petal Width (cm)") +
  ylab("Density")
```

```
Warning message:
“Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.”
Warning message:
“The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
ℹ Please use `after_stat(density)` instead.”
```



It appears petal width is follwoing a non-normal distribution, this suggests that the relationship between the predictors and the petal width may not be approximately linear, and the residuals of the model might also be non-normally distributed. So, Fitting a GLM with Gamma family along with the 'log' link function

In [22]:
```
#Fitting a GLM with a log link function
glm_model <- glm(Petal.Width ~ Sepal.Length + Sepal.Width + Petal.Length, family = Gam
summary(glm_model)
#Predicting the response with GLM model
predicted_values <- predict(glm_model, newdata = test, type = "response")

#Calculating MSPE
actual_values <- test$Petal.Width
mspe <- mean((actual_values - predicted_values)^2)
cat("\nMSPE for the GLM Model is:", mspe)
```

```
Call:
glm(formula = Petal.Width ~ Sepal.Length + Sepal.Width + Petal.Length,
    family = Gamma(link = "log"), data = train)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.37100    0.32125  -4.268 4.05e-05 ***
Sepal.Length -0.12751    0.08798  -1.449    0.150
Sepal.Width  -0.05085    0.08673  -0.586    0.559
Petal.Length  0.57336    0.04537  12.637  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.09348123)

    Null deviance: 92.102  on 119  degrees of freedom
Residual deviance: 11.401  on 116  degrees of freedom
AIC: -23.112

Number of Fisher Scoring iterations: 6
MSPE for the GLM Model is: 0.3253174
```

The initial linear regression model yielded an MSPE of 0.09499934, indicating a relatively good predictive accuracy for petal width. However recognizing the non-normal distribution with bimodel nature of the response variable, a GLM with a Gamma family and log link was applied on to the data resulting in a higher MSPE of 0.3253174. This increase in MSPE could tell us that that despite the use of GLM, it may not capture the relationships between response and the predictors as effectively as the linear model.

REPORT

INTRODUCTION

Our project explores the intriguing relationships between various structural features of flowers, like petal width, and how these connects with one another. We're particularly focused on discovering the key factors that determine the width of a petal. By using statistical methods, we're seeking to validate which of these features truly matter and how they're interlinked. Our **Research question** states that- How do Structural features correlate with response variable, and can we statistically validate the significance of these relationships? Specifically, which features strongly predict petal width?

My interest in this problem arises from a desire to understand the patterns of plant growth. By studying the relationship between various structural patterns and their effect on the Petal.Width feature, we can uncover the language of flowers and how differently they come together to shape their identity. And, this can be done by applying various statistical methods on to the IRIS data which i took it directly from the R directory and it can also be available from different sources in the web. If we were to explain our project to a layman- In our project, we are studying the Iris flower, seeing how its features like petal and sepal sizes influence the petal width. We used statistics to test our theories to ensure the assumptions are actually true.

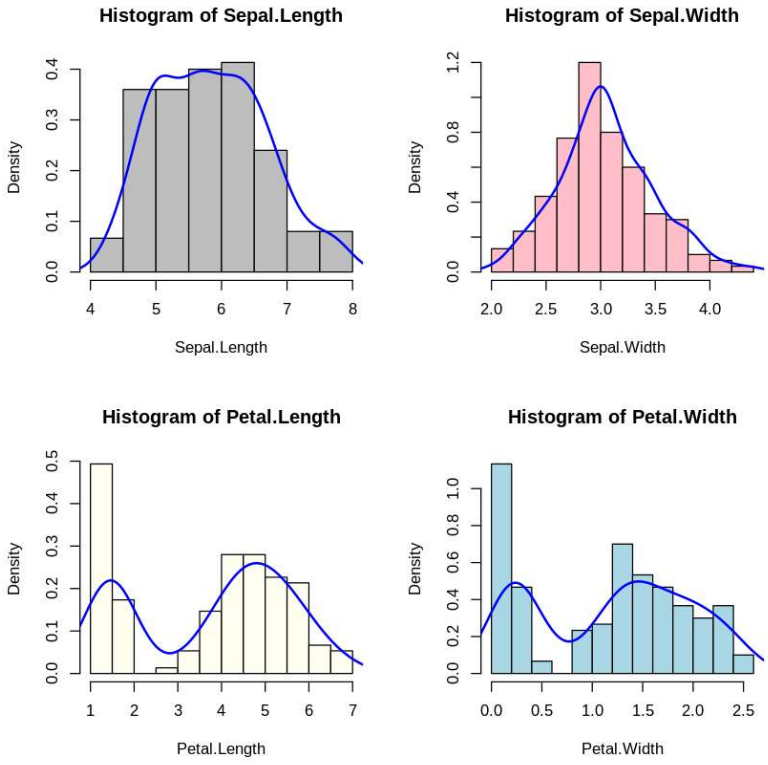
Predicting the size of Petal Width using othr structural featuers and how they are correlated with the output. It's all about finding the hidden connections in flower designs.

Related Work

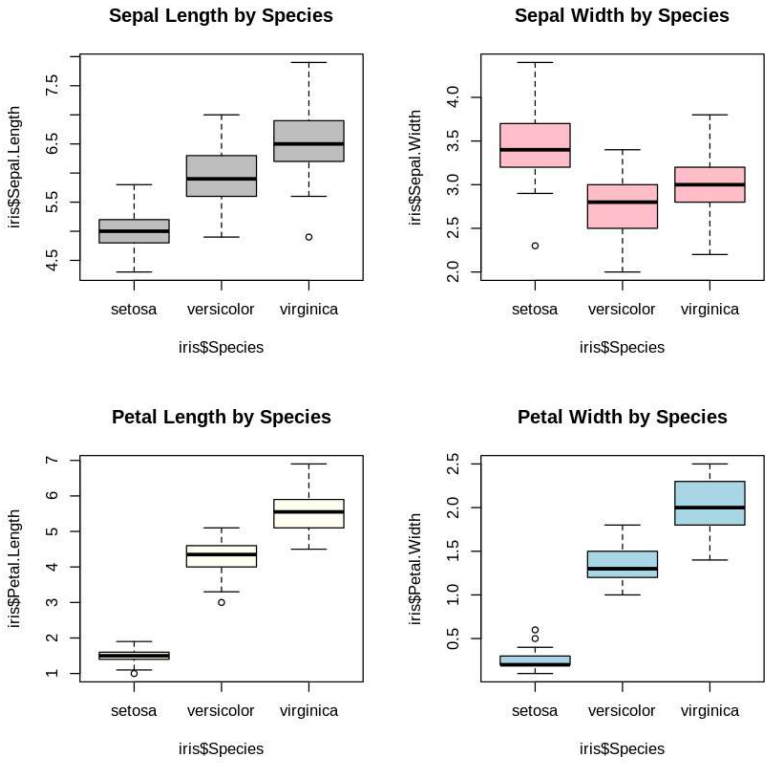
Various researchers have utilized the dataset to demonstrate the use of linear regression and other statistical methods. One of which is the analysis by Cecilia Lee, titled "Iris — Linear Regression," available on RPubs. Lee's work utilizes the Iris dataset to demonstrate the application of linear regression techniques. Her analysis serves as an informative example of how predictive modeling can be applied to IRIS data, providing insights into the statistical relationships between the attributes of Iris species.

Exploratory Data Analyses

Through histograms, we have ve examined Iris flower measurements and their distributions, revealing sepal length and width to typically follow a bell-shaped pattern, suggesting at a standard range across the flowers. In contrast, petal lengths and widths distinctly has bimodal distributions, suggesting potential differences in Iris species.



The boxplots distinguish between Iris species based on sepal and petal sizes. Setosa is markedly smaller in petal size yet broader in sepal width. Virginica and Versicolor exhibit larger petal sizes, with Virginica generally outsizeing Versicolor, highlighting the diversity in Iris species' physical characteristics.



Statistical Analysis (In Predicting Response)

Linear Regression

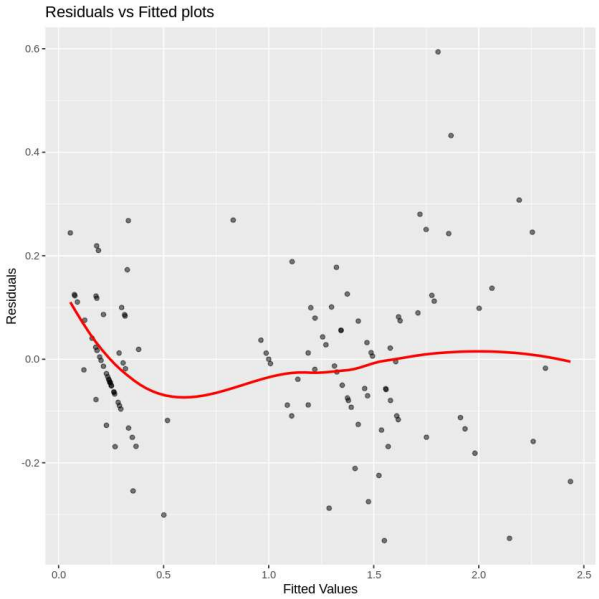
To address a part of research question on structural feature with petal width in Iris flowers, linear regression seems to be most appropriate analyses. The model's performance was evaluated using the Mean Squared Prediction Error (MSPE) to determine accuracy.

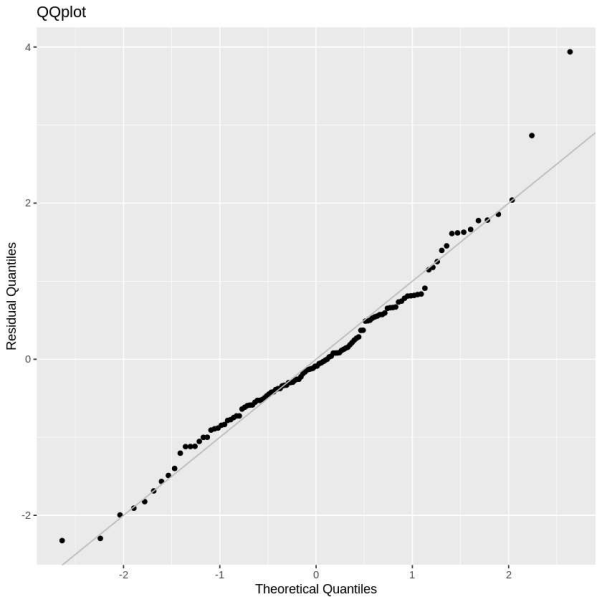
From the regression analysis it became evident that petal length, sepal length and sepal width, had a postive effect in predicting the petal width, with petal length emerging as the strongest predictor based on it coefficient value. The model demonstrated high predictive accuracy with an MSPE of 0.09499934, capturing 95.45% of the variance with an adjusted R-squared of 95.33%. These results confirms the significant influence of these structural features on petal width and showcases the models ability in making reliable predictions. Results are shown in the below image.

```
Call:
lm(formula = Petal.Width ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35041 -0.08847 -0.01330  0.08623  0.59398

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.17947    0.16050  -1.118    0.266
Sepal.Length -0.18538    0.04396  -4.218 4.92e-05 ***
Sepal.Width   0.18243    0.04333   4.210 5.06e-05 ***
Petal.Length  0.49998    0.02267  22.057  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1528 on 116 degrees of freedom
Multiple R-squared:  0.9545,    Adjusted R-squared:  0.9533
F-statistic: 810.8 on 3 and 116 DF,  p-value: < 2.2e-16


Initial MSPE of full model: 0.09499934

Lowest MSPE observed: 0.09499934
```

Diagnostics:

The diagnostic checks on the best-performing linear regression model present some concerns. Despite the residual plot not clearly showing heteroscedasticity, a statistical test indicates we might have non-constant variance. The test for independence, supported by the spread of residuals, suggests the data points may not be entirely independent. Moreover, the linearity assumption is questioned by the shape of the residuals versus fitted values plot. Lastly, the normality of residuals, as judged by a QQ plot, appears to be violated at the tails, indicating that the data may not be normally distributed. These findings suggest that while the linear model is a good fit, it's not perfect, and adjustments or alternative models might be needed to better meet the assumptions of linear regression.

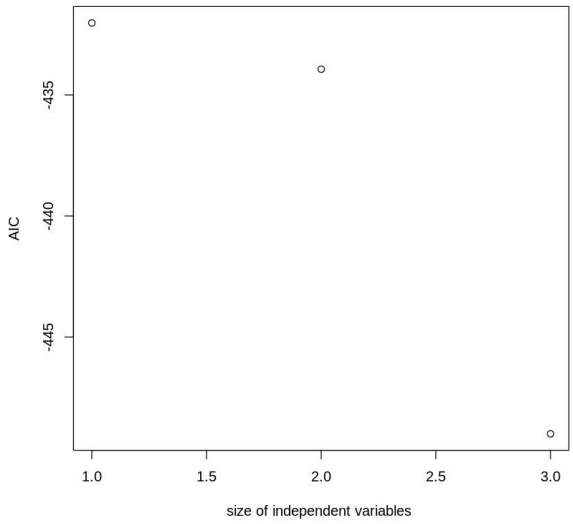




Model Selection based on AIC, BIC, Adjusted R^2

In our search to pinpoint the perfect model for predicting petal width, we went beyond our initial MSPE criterion. We explored the Best Subset Regression, a method that evaluates all possible predictor combinations. This revealed that the models with three predictors—Sepal Length, Sepal Width, and Petal Length— yields as the best models according to BIC and Adjusted R^2, metrics that balance model fit and complexity. These models stand out not only for their prediction accuracy but also for their statistical efficiency, underlining the robustness of our statistical approach in uncovering the features that influences petal width in Iris flowers.

AIC model also identified the same model with Sepal Length, Sepal Width, and Petal Length as best, aligning with BIC and Adjusted R^2 outcomes, highlighting its predictive strength. And here is a graph depicting the AIC for all the combination of predictors.



Statistical Significance of Predictors

To clearup the connections between each features/predictors and their influence on petal width, we delved into confidence interval analysis and ANOVA. These statistical methods enabled us to validate which structural features significantly contributes to petal width.

Confidence Interval Analysis:

The confidence interval analysis was important in determining the predictors that have a meaningful impact on Petal.Width. Notably, the negative confidence interval for Sepal.Length and the positive ones for Sepal.Width and Petal.Length provided clear evidence of their significant roles. The non-significant intercept suggests that the Petal.Width when all predictors are zero, is not informative or not significant at all. With these confidence intervals, we have discovered which features matter when it comes to the width of a petal.

The below image is the Confidence Interval table

	Lower_Bound	Upper_Bound
	<dbl>	<dbl>
(Intercept)	-0.6100982	0.2977274
Sepal.Length	-0.7094957	-0.5248224
Sepal.Width	0.1439385	0.3685558
Petal.Length	0.1930509	0.8534879

ANOVA Results Examination:

Performing anova on 3 different models allows researchers to statistically test the incremental value of each predictor added to the models, guiding decisions on model selection based on

statistical evidence of improved performance.

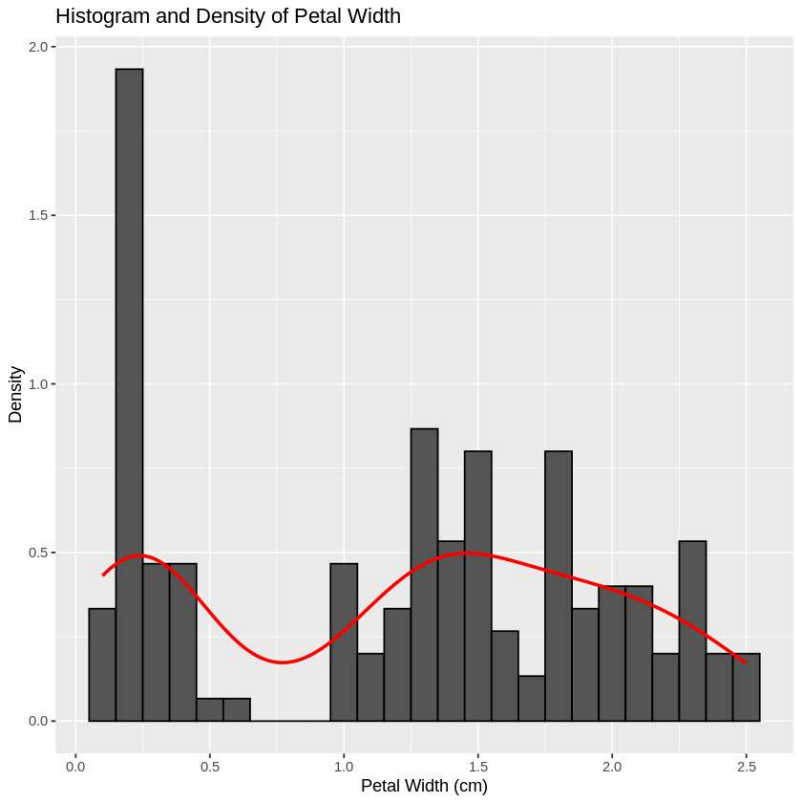
The ANOVA test gave further clarity to our findings. By comparing three distinct models, each with different combinations of predictors, ANOVA showed us that the model with all the predictors Sepal.Length, Sepal.Width, and Petal.Length provided the best fit. The significant F-statistic (17.724) and the associated p-value(5.064e-05) confirmed that this full model, including all chosen predictors, had higher predictive ability compared to the reduced models. This result from ANOVA highlights the combined strength of all the predictors and confirms the robustness of our chosen model in explaining petal width variations.

Later,The Generalized Linear model(GLM) was performed on the data even after linear regression due to the its inability to satisfy the assumption of normality and linearity in the data. This statistical change allows for a flexible approach, adjusting for the distributional characteristics of the response variable(Petal.Width), potentially leading to a even better model.

Generalized Linear Modeling

The exploration of petal width distribution through a density plot and histogram was a crucial step for performing the GLM.

The below image demonstrate about the distribution of our response variable(Petal.Width).



The distribution of Petal Width showcases a bimodal trend, diverging from a normal distribution.So, for our GLM approach we selected the Gamma distribution and the logarithmic link function. This combination is particularly effective for modeling response variables that showcase skewness and non-normality just as we observed in the petal width of the Iris. Such a

choice ensures that our statistical model is well-aligned with the actual data structure, enhancing the reliability of our findings.

Initially, linear regression showcased a promising model for petal width prediction, with an MSPE of 0.09499934. Yet, the bimodal distribution of our responde variable Petal Width leads us to apply a GLM using the Gamma distribution with a log link. But surprisingly, the GLMs MSPE rose to 0.3253174, suggesting that this more complex model didn't quite grasped the predictor-response relationship as well as the simpler linear regression did.

Conclusions

Throughout the project, we precisely applied statistical models and analyses to the Iris dataset to understand how structural features correlate with petal width. We started with exploratory data analysis to visualize the distributions and relationships. Linear regression provided an initial predictive model, identifying key features influencing petal width. Diagnostic checks revealed some assumptions were not fully met, hence leading us to explore Generalized Linear Models, which adjusted for the non-normal distribution of petal width. Model selection criteria such as AIC, BIC, and Adjusted R-squared were applied to refine our model choice. Confidence intervals and ANOVA further validated the significance of the relationships between features and petal width. This comprehensive approach brought clarity to which structural features most strongly predict petal width.

From this analytical processes, I learned that even when data suggests complex relationships, simpler models can sometimes offer the best insights. The Iris dataset revealed that its variations in petal width could be effectively captured by linear relationships, reassuting the importance of matching the model to the data various patterns for insightful conclusions.

Future work

To further explore the Iris dataset using some other statistical methods in R, we could enhance our approach by integrating interaction terms in our linear regression models to assess the combined effects of predictors on petal width. Additionally, expanding the ANOVA analyses more deeply to examine variations within species could provide much clearer insights into species differences. Simple polynomial regression could be tested to evaluate improvements in model fit through non-linear relationships.

References

1. https://rpubs.com/cecilialee/iris
2. https://medium.com/@elsasaji02/an-exploration-of-the-iris-dataset-through-fundamental-statistical-analysis-with-r-programming-9e0ed52f2acd
3. https://rpubs.com/Tanzir/Statistical-Analysis_IRIS-Data
4. https://xiaorui.site/Data-Mining-R/lecture/2.A_ExploratoryAnalyses.html

5. Diagnostic_plots.ipynb file from professor

6. Homework-5(which was submitted by me)