Required Library

1. library(ggplot2)
2. library(dplyr)

Data Set (sample data.csv)

There are three columns (Area, Length, Wheat_Varieties) in the dataset, which label Wheat varieties to represent Area and Length. Area refers to the extent to which the wheat variety occupies. Length refers to the length of the wheat variety. Wheat_Varieties represent varieties of wheat; there are 3 varieties (Kama, Rosa, Canandian) in the dataset

Project Overview

I will will use “sample data.csv” dataset. In the first, second, and third trials, four cases will be tested: the full dataset (df), the Kama variety (df_kama), the Rosa variety (df_rosa), and the Canadian variety. (df_canadian).

1. Anova test

I used the Anova test to determine whether there is a relationship between Area and Length in four distinct cases.

2. Independent t-test

I conducted an Independent t-test, a statistical test used to determine if there is a significant difference between the means of two independent groups. . In this case, I was interested in Length and Area.

3. Linear Regression

Next, I performed a Linear Regression analysis to investigate the linear association between Area and Length in those same four cases

4. Multi-class logistic regression.

I used multi-class logistic regression to predict the probabilities of multiple categories of a categorical dependent variable based on one or more independent variables. I have implemented multiclass logistic regression using gradient descent algorithm by myself. The model can help to classify when new data is added.

```
df <- read.csv("/Users/sanghyunkim/Desktop/ML (Github)/Data Set/sample data.csv")
head(df)
```

```
##    Area Length Wheat_Varieties
## 1 15.26  5.220           Kama
## 2 14.88  4.956           Kama
## 3 14.29  4.825           Kama
## 4 13.84  4.805           Kama
## 5 16.14  5.175           Kama
## 6 14.38  4.956           Kama
```

Main Research question: “Is there a correlation or speical relationship between length and Area?”

This question is of interest because it could help identify which variety of wheat has the most optimal length for certain agricultural purposes. To answer this question, we will use ANOVA test. Anova test is appropriate for this case because it tests for overall significance: ANOVA tests whether there is a significant difference among the means of the groups, rather than just comparing two groups at a time.

First of all, I divided the whole dataset by the Wheat_variables. And store them in a different dataframe.

```
df_kama <- subset(df, Wheat_Varieties == " Kama ") # Select data whose "Wheat_Varieties" is " Kama "
df_rosa <- subset(df, Wheat_Varieties == " Rosa ") # Select data whose "Wheat_Varieties" is " Rosa "
df_canadian <- subset(df, Wheat_Varieties == " Canadian ") # Select data whose "Wheat_Varieties" is " Canadian "
```

Data Exploration

```
head(df_kama)
```

```
##    Area Length Wheat_Varieties
## 1 15.26  5.220           Kama
## 2 14.88  4.956           Kama
## 3 14.29  4.825           Kama
## 4 13.84  4.805           Kama
## 5 16.14  5.175           Kama
## 6 14.38  4.956           Kama
```

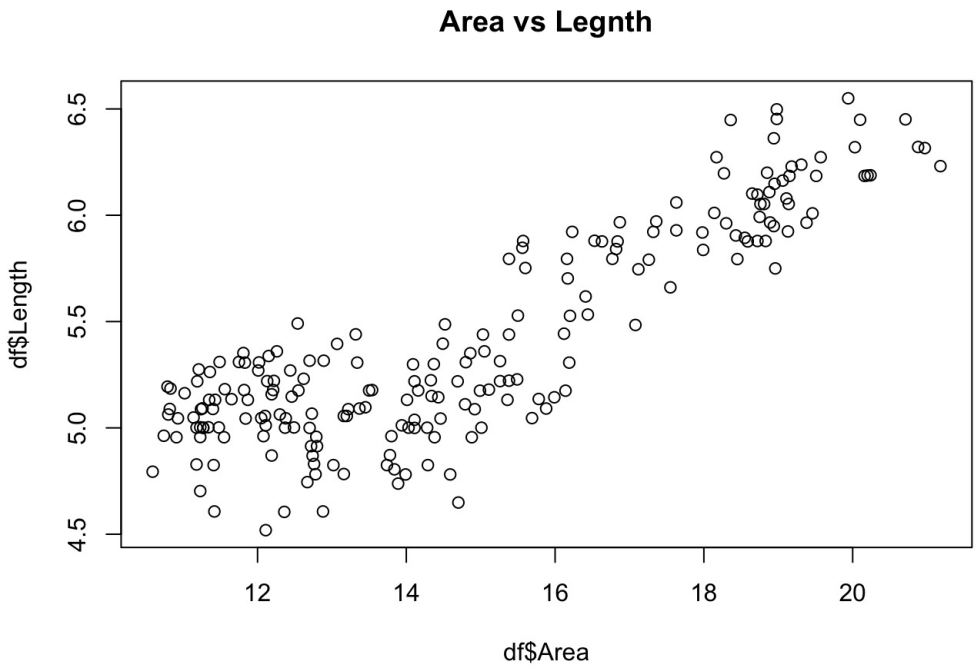
```
head(df_rosa)
```

```
##     Area Length Wheat_Varieties
## 71 17.63  6.060           Rosa
## 72 16.84  5.877           Rosa
## 73 17.26  5.791           Rosa
## 74 19.11  6.079           Rosa
## 75 16.82  5.841           Rosa
## 76 16.77  5.795           Rosa
```

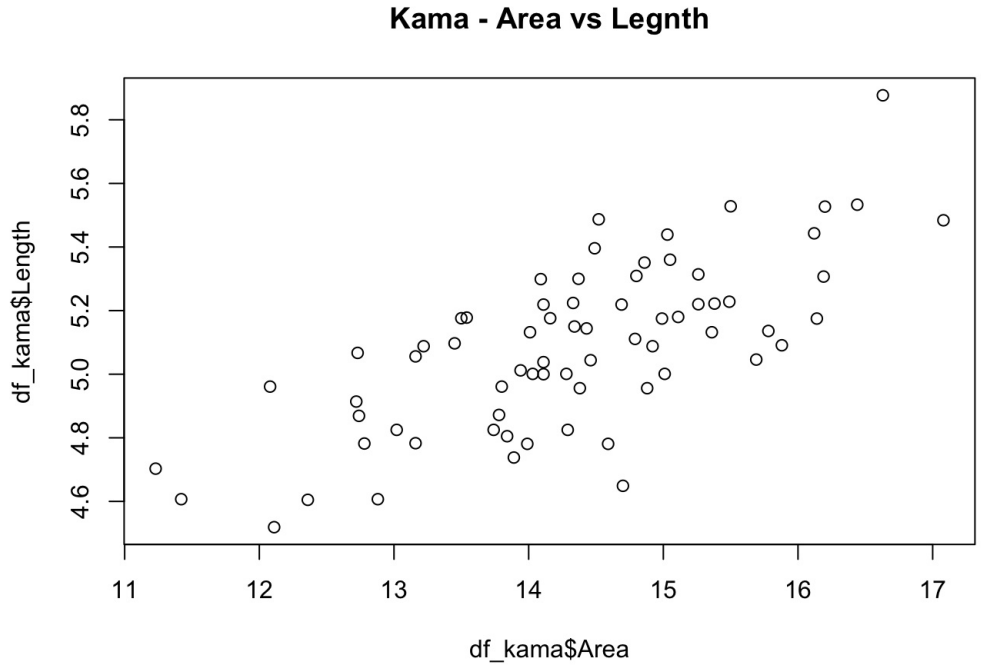
```
head(df_canadian)
```

```
##      Area Length Wheat_Varieties
## 141 13.07  5.395       Canadian
## 142 13.32  5.440       Canadian
## 143 13.34  5.307       Canadian
## 144 12.22  5.221       Canadian
## 145 11.82  5.178       Canadian
## 146 11.21  5.275       Canadian
```

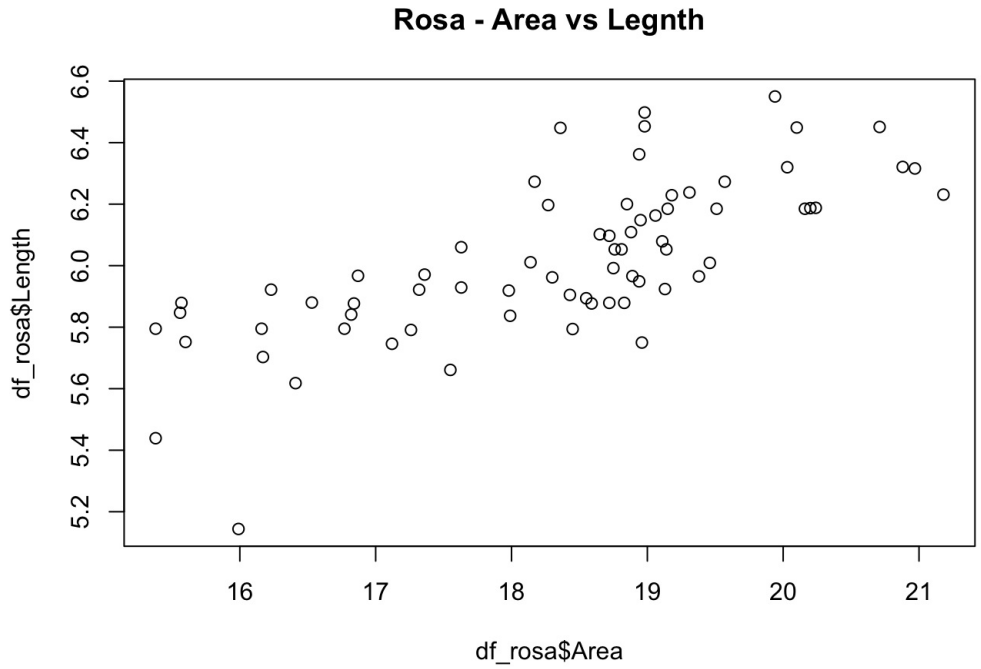
```
plot(df$Area, df$Length, main = "Area vs Legnth")
```



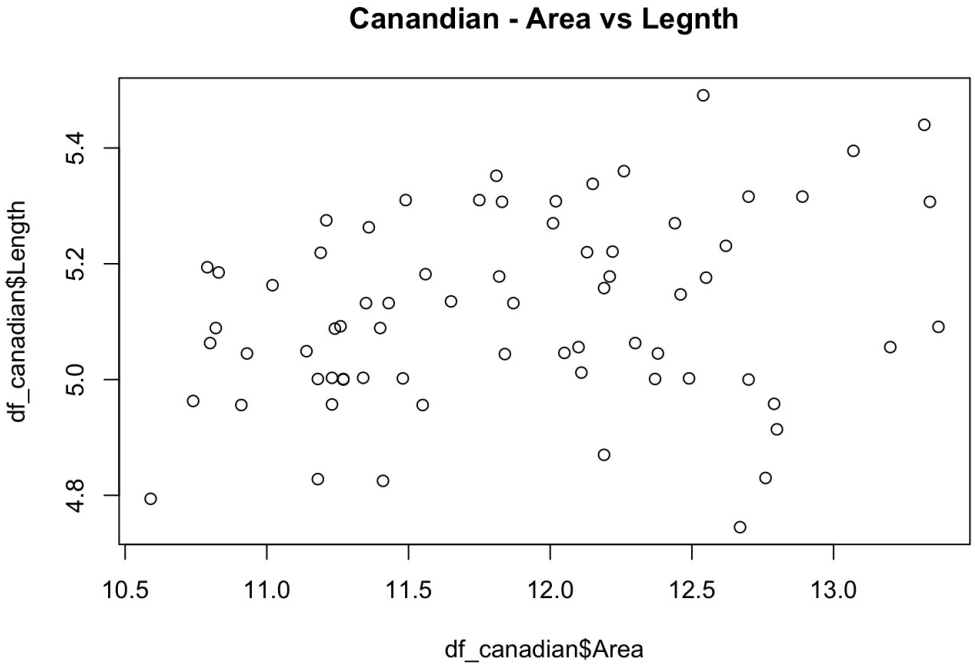

```
plot(df_kama$Area, df_kama$Length, main = "Kama - Area vs Legnth")
```




```
plot(df_rosa$Area, df_rosa$Length, main = "Rosa - Area vs Legnth")
```




```
plot(df_canadian$Area, df_canadian$Length, main = "Canandian - Area vs Legnth")
```



1.a) Anova test for whole dataset df

Null hypothesis: There’s no significant difference between the “Length” and “Area” variables.

Alternative hypothesis: There’s a significant difference between the “Length” and “Area” variables.

Interpretation

The ANOVA test conducted on the entire dataset indicates that there is a statistically significant relationship between the “Length” and “Area” variables. The very low p-value of 2e-16 suggests strong evidence to reject the null hypothesis that there is no relationship between these variables. Therefore, it can be concluded that there is a significant relationship between the “Length” and “Area” variables.

```
fit <- aov(Area ~ Length, data = df)
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Length        1 1320.0  1320.0   610.8 <2e-16 ***
## Residuals   208  449.5     2.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.b) Anova test for df_kama

Null hypothesis: There’s no significant difference between the “Length” and “Area” variables of wheat varieties Kama.

Alternative hypothesis: There’s a significant difference between the “Length” and “Area” variables of wheat varieties Kama.

Interpretation

The p-value (Pr(>F)) is 1.95e-12, which is much smaller than the significance level of 0.05, indicating that we can reject the null hypothesis that there is no relationship between the “Length” and “Area” variables.

Therefore, we can conclude that there is a significant relationship between the “Length” and “Area” variables in the “df_kama” dataset.

```
fit <- aov(Area ~ Length, data = df_kama)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Length       1  53.01   53.01    73.6 1.95e-12 ***
## Residuals   68  48.97    0.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.c) Anova test for df_rosa

Null hypothesis: There's no significant difference between the "Length" and "Area" variables of wheat varieties Rosa

Alternative hypothesis: There's a significant difference between the "Length" and "Area" variables of wheat varieties Rosa

Interpretation

Based on the ANOVA test conducted on the "df_rosa" dataset, the p-value (Pr(>F)) is found to be 9.1e-13, which is significantly smaller than the commonly used significance level of 0.05. This suggests that the null hypothesis, which assumes no relationship between the "Length" and "Area" variables, should be rejected. Therefore, it can be concluded that there is a significant relationship between the "Length" and "Area" variables in the "df_rosa" dataset.

```
fit <- aov(Area ~ Length, data = df_rosa)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Length       1  75.82   75.82   76.77 9.1e-13 ***
## Residuals   68  67.16    0.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.d) Anova test for df_canadian

Null hypothesis: There's no significant difference between the "Length" and "Area" variables of wheat varieties Canadian

Alternative hypothesis: There's a significant difference between the "Length" and "Area" variables of wheat varieties Candandian

Interpretation

The p-value (Pr(>F)) is 0.0308 , which is smaller than the significance level of 0.05, indicating that we can reject the null hypothesis that there is no relationship between the "Length" and "Area" variables.

Therefore, we can conclude that there is a significant relationship between the "Length" and "Area" variables in the "df_canadian" dataset. However, compared to the previous two datasets, it does not have a strong correlation as strong as the other two datasets, because the p-value of the other two datasets is overwhelmingly lower than 0,05, while the current dataset is slightly smaller than 0,05. In other words, we can support the alternative hypothesis that there's a significant difference between the "Length" and "Area" variables.

```
fit <- aov(Area ~ Length, data = df_canadian)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Length       1   2.41   2.408   4.865 0.0308 *
## Residuals   68  33.66   0.495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Independent t-test

Research Question: Is there a significant difference between the mean values of the "Area" and "Length" columns? And Is there a variation if I specify the wheat variety.

The independent t-test is appropriate when we have two independent groups of continuous data that follow a normal distribution. In this case, we are comparing the means of two different variables, "Area" and "Length", so an independent t-test is appropriate.

Assumptions

Assume that the two groups being compared are independent and normally distributed. Also assume that the variances of the two groups are equal.

Hypothesis

Null hypothesis: There is no significant difference between the mean values of the two groups. Alternative Hypothesis: There is significant difference between the mean values of the two groups.

Interpretation for dataset “df”

The independent t-test results indicate that there is a significant difference between the mean values of the “Area” and “Length” columns in the “df” dataframe (t = 46.355, df = 220.92, p < 2.2e-16).

The confidence interval (9.038141 to 9.840763) does not include zero, which indicates that the true difference in means is likely not equal to zero.

The sample estimates of the mean values for “Area” and “Length” are 14.847524 and 5.408071, respectively.

Since the p-value (2.2e-16) is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is a significant difference between the mean values of the “Area” and “Length” columns.

```
t_test <- t.test(df$Area, df$Length, paired = FALSE)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Area and df$Length
## t = 46.355, df = 220.92, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  9.038141 9.840763
## sample estimates:
## mean of x mean of y
## 14.847524  5.408071
```

Interpretation for dataset “df_kama”

This dataset is df_kama, a dataframe whose wheat variety is kama. The independent t-test results indicate that there is a significant difference between the mean values of the “Area” and “Length” columns in the “df” dataframe (t = 62.194, df = 75.479, p-value < 2.2e-16). The confidence interval (8.951053 to 9.543376) does not include zero, which indicates that the true difference in means is likely not equal to zero. The sample estimates of the mean values for “Area” and “Length” are 8.951053 and 9.543376, respectively. Since the p-value (2.2e-16) is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is a significant difference between the mean values of the “Area” and “Length” columns.

```
t_test <- t.test(df_kama$Area, df_kama$Length, paired = FALSE)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  df_kama$Area and df_kama$Length
## t = 62.194, df = 75.479, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  8.951053 9.543376
## sample estimates:
## mean of x mean of y
## 14.334429  5.087214
```

Interpretation for dataset “df_rosa”

This dataset is df_rosa, a dataframe whose wheat variety is rosa The independent t-test results indicate that there is a significant difference between the mean values of the “Area” and “Length” columns in the “df” dataframe (t = 70.481, df = 73.29, p-value < 2.2e-16). The confidence interval (11.96551 to 12.66186) does not include zero, which indicates that the true difference in means is likely not equal to zero. The sample estimates of the mean values for “Area” and “Length” are 18.33429 and 6.02060, respectively. Since the p-value (2.2e-16) is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is a significant difference between the mean values of the “Area” and “Length” columns.

```
t_test <- t.test(df_rosa$Area, df_rosa$Length, paired = FALSE)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  df_rosa$Area and df_rosa$Length
## t = 70.481, df = 73.29, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.96551 12.66186
## sample estimates:
## mean of x mean of y
##  18.33429   6.02060
```

Interpretation for dataset “df_canadian”

This dataset is df_canadian, a dataframe whose wheat variety is canadian The independent t-test results indicate that there is a significant difference between the mean values of the “Area” and “Length” columns in the “df” dataframe (t = 76.304, df = 75.917, p-value < 2.2e-16). The confidence interval (6.581072 to 6.933843) does not include zero, which indicates that the true difference in means is likely not equal to zero. The sample estimates of the mean values for “Area” and “Length” are 11.87386 and 5.11640 , respectively. Since the p-value (2.2e-16) is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is a significant difference between the mean values of the “Area” and “Length” columns.

```
t_test <- t.test(df_canadian$Area, df_canadian$Length, paired = FALSE)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  df_canadian$Area and df_canadian$Length
## t = 76.304, df = 75.917, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.581072 6.933843
## sample estimates:
## mean of x mean of y
##  11.87386   5.11640
```

3 Linear Regression Analysis

Research question: Is there a significant linear relationship between the Area and Length of the wheat fields?

The research question is of interest because it can provide insights into the growth patterns of wheat fields. To answer this question,I used paired linear regression analysis to examine the relationship between Area and Length.

The paired linear regression analysis is appropriate because it allows us to model the relationship between two continuous variables.

3.a) Linear Regression Analysis for whole dataset df

Interpretation and Conclusion

The intercept term is -12.806, which represents the estimated value of the response variable (Area) when the predictor variable (Length) is equal to zero.

The coefficient is 5.113, which indicates that for every one unit increase in “Area”, the predicted value of “Length” will increase by 5.113 units, all other things being equal.

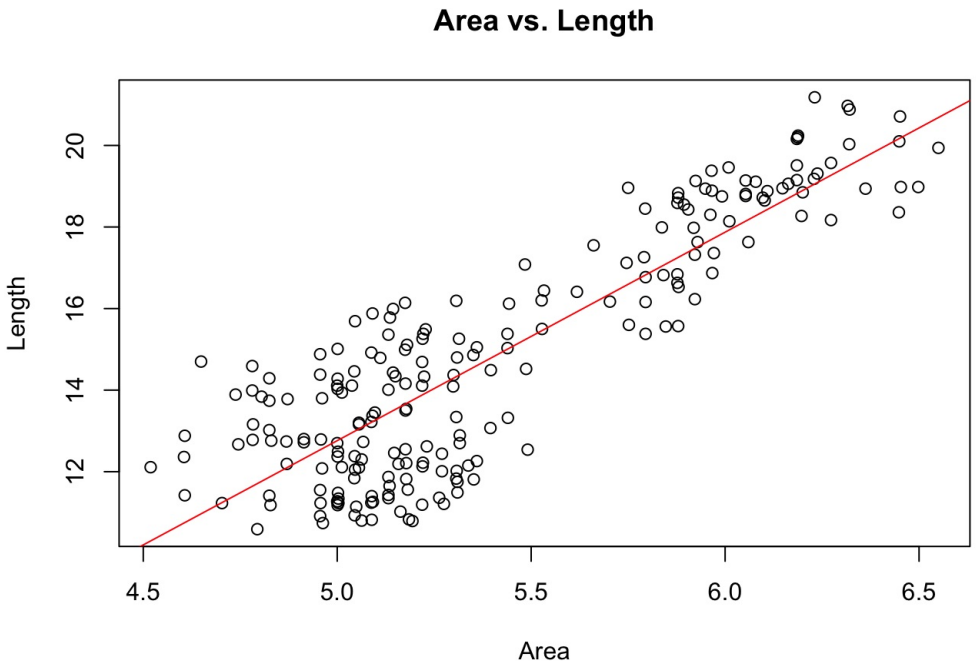
In summary, the equation for this linear regression model is: Length = -12.806 + 5.113 * Area. This equation can be used to make predictions of “Length” based on values of “Area” in the dataset or in new data.

We can colude that the results of the analysis indicate that there is a significant positive linear relationship between Area and Length, meaning that as the Area of the wheat field increases, so does the Length of the wheat plants.

```
model <- lm(Area ~ Length, data = df)
model
```

```
##
## Call:
## lm(formula = Area ~ Length, data = df)
##
## Coefficients:
## (Intercept)       Length
##     -12.806        5.113
```

```
plot(df$Length,df$Area,  main = "Area vs. Length", xlab = "Area", ylab = "Length")
abline(model, col = "red")
```



3.b) Linear Regression Analysis for dataset df_kama

Interpretation and Conclusion

The intercept term is -2.574, which represents the estimated value of the response variable (Area) when the predictor variable (Length) is equal to zero.

The coefficient is 3.324 , which indicates that for every one unit increase in “Area”, the predicted value of “Length” will increase by 3.324 units, all other things being equal.

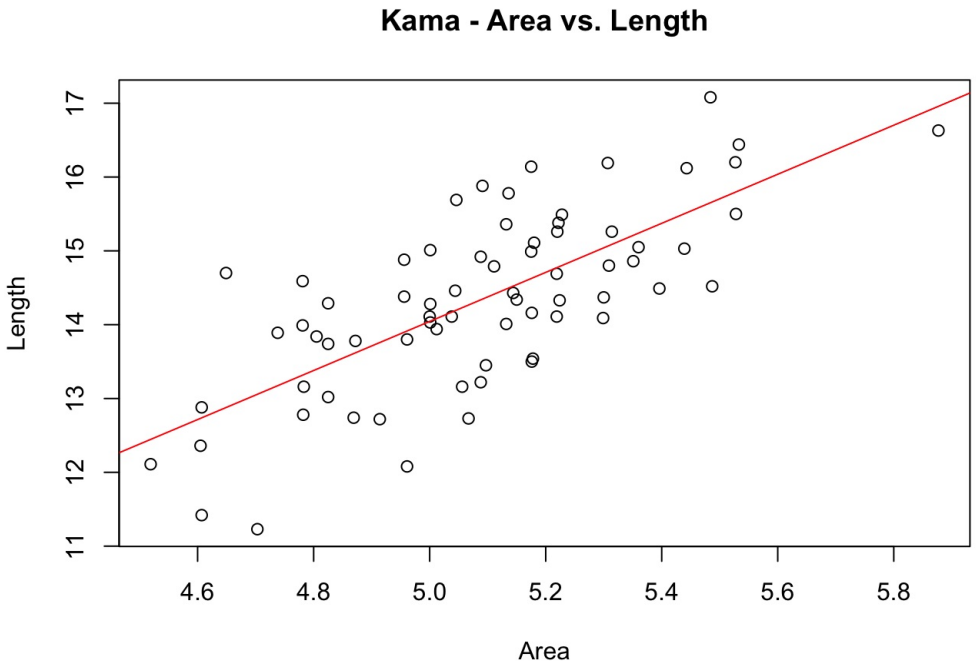
In summary, the equation for this linear regression model is: Length = -2.574 + 3.324 * Area. This equation can be used to make predictions of “Length” based on values of “Area” in the dataset or in new data.

Based on the analysis results of the dataset, it can be concluded that there exists a meaningful positive correlation between the Area and Length variables. This implies that when the wheat field’s Area increases, the Length of the wheat plants also tends to increase.

```
model <- lm(Area ~ Length, data = df_kama)
model
```

```
##
## Call:
## lm(formula = Area ~ Length, data = df_kama)
##
## Coefficients:
## (Intercept)       Length
##      -2.574        3.324
```

```
plot(df_kama$Length,df_kama$Area,  main = "Kama - Area vs. Length", xlab = "Area", ylab = "Length")
abline(model, col = "red")
```



3.c) Linear Regression Analysis for dataset df_rosa

Interpretation and Conclusion

The intercept term is -6.519, which represents the estimated value of the response variable (Area) when the predictor variable (Length) is equal to zero.

The coefficient is 4.128, which indicates that for every one unit increase in “Area”, the predicted value of “Length” will increase by 4.128 units, all other things being equal.

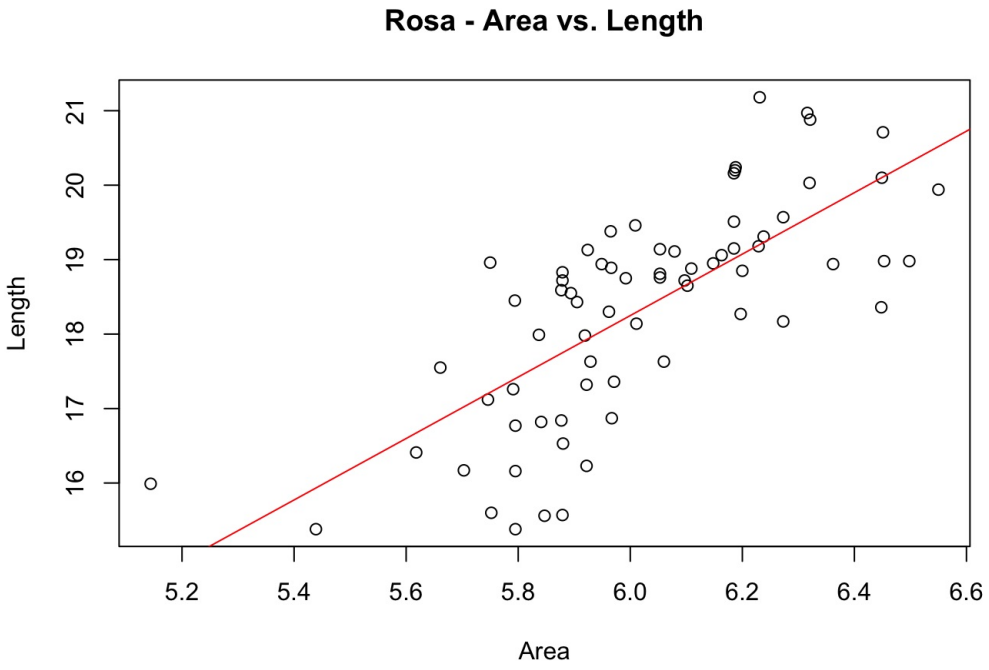
In summary, the equation for this linear regression model is: Length = -6.519 + 4.128 * Area. This equation can be used to make predictions of “Length” based on values of “Area” in the dataset or in new data.

Based on the analysis of the dataset, it can be concluded that there is a strong correlation between the Area and Length variables. Specifically, as the Area of the wheat field increases, there is a corresponding increase in the Length of the wheat plants. The Correlation of this dataset is stronger than the that of dataset (df_kama).

```
model <- lm(Area ~ Length, data = df_rosa)
model
```

```
##
## Call:
## lm(formula = Area ~ Length, data = df_rosa)
##
## Coefficients:
## (Intercept)       Length
##      -6.519        4.128
```

```
plot(df_rosa$Length,df_rosa$Area,  main = "Rosa - Area vs. Length", xlab = "Area", ylab = "Length")
abline(model, col = "red")
```



3.d) Linear Regression Analysis for dataset df_canadian

Interpretation and Conclusion

The intercept term is 5.976, which represents the estimated value of the response variable (Area) when the predictor variable (Length) is equal to zero.

The coefficient is 1.153, which indicates that for every one unit increase in “Area”, the predicted value of “Length” will increase by 1.153 units, all other things being equal.

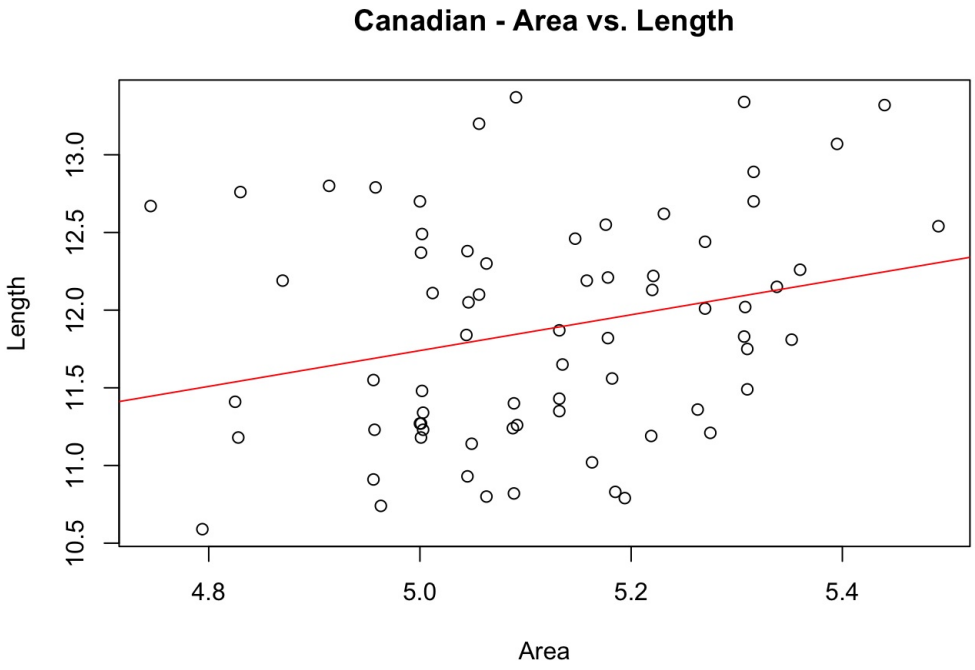
In summary, the equation for this linear regression model is: Length = 5.976 + 1.153 * Area. This equation can be used to make predictions of “Length” based on values of “Area” in the dataset or in new data.

Through the analysis, I could reach the conclusion: there is a perfect positive linear relationship between Area and Length. there is a corresponding unit increase in the dependent variable (usually plotted on the y-axis).”

```
model <- lm(Area ~ Length, data = df_canadian)
model
```

```
##
## Call:
## lm(formula = Area ~ Length, data = df_canadian)
##
## Coefficients:
## (Intercept)       Length
##       5.976        1.153
```

```
plot(df_canadian$Length,df_canadian$Area,  main = "Canadian - Area vs. Length", xlab = "Area", ylab = "Length")
abline(model, col = "red")
```



4. Multi Class Logistic Regression

Big Idea

Extend the binary logistic regression model to handle classification problems with more than two classes. Using gradient descent algorithm, find the parameters which minize the cost function.

Method

Sigmoid Function: An activation function to produce the predicted probabilities for each class.

g(z) = 1 / (1 + exp(-z))

Hypothesis Function: A set of K separate binary logistic regression models, where K is the number of classes.

h(x) = g($w^T x$)

Cost Function

cost(h(x), y) = -log(h(x)) if y = 1 (y is the result of classification) cost(h(x), y) = -log(1 - h(x)) if y = 0

Comnbined Cost Function

cost(h(x), y) = -y * log(h(x)) - (1 - y) * log(1 - h(x))

Gradient Descent Algorithm

w_j = w_j - (Learning Rate) * $\sum_{i=1}^{m}$(h($x^{(i)}$) - y^i) * x^i

Optimization

Gradient Descent is an optimization algorithm for finding first-order approximations. The basic idea is to obtain the slope of the function and keep moving the slope to the lower side until it reaches the extremum.

Since the value of the differential coefficient (= gradient) decreases as it approaches the minimum value, a value proportional to the differential coefficient is used for the travel distance.

Then, when it is far from the minimum value, it moves a lot, and when it is close to the minimum value, it can move little by little. x_{i+1} = x_i - (Learning Rate) * (df/dx) * x_i (example)

Learning rate

The learning rate is used to control the step size at each iteration towards the optimal solution.

If the learning rate is too high, the algorithm may overshoot the minimum of the loss function and fail to converge.

On the other hand, if the learning rate is too small, the algorithm may take a long time to converge or get stuck in a suboptimal solution.

Result

(Result can be differ everytime we run the code. But, there will be no big difference)

Parameter vector for Kama: [44.848877, 1.745317, -13.127459]

Logistic regression line for Kama: y = $\frac{44.848877}{-13.127459} + \frac{1.745317}{-13.127459}x$

Parameter vector for Rosa: [-30.5199448, 0.9497909, 2.7286191]

Logistic regression line for Rosa: y = $\frac{-30.5199448}{2.7286191} + \frac{0.9497909}{2.7286191}x$

Parameter vector for Canadian: [2.022084, -3.498555, 8.5516453]

Logistic regression line for Canadian: $\frac{2.022084}{8.5516453} + \frac{-3.498555}{8.5516453}x$

Conclusion

Able predict the probability of a categorical dependent variable with more three categories (Kama, Rosa, and Canadian).

For example, to predict the probability of a sample belonging to the Kama category given an input value of x, we can plug in the x value into the Kama equation: y = $\frac{44.819922}{-13.181489} + \frac{1.576196}{-13.181489}x$.

The output probabilities from the logistic regression can be interpreted as the likelihood of a sample belonging to each category. The category with the highest probability is the predicted category for that sample.

The probabilities obtained from the logistic regression represent the chances or likelihood of a given sample belonging to each category. By comparing these probabilities, we can determine which category the sample is most likely to belong to. The category with the highest probability is considered the predicted category for that sample.

```
# Required libraries
library(ggplot2)
library(dplyr)
```

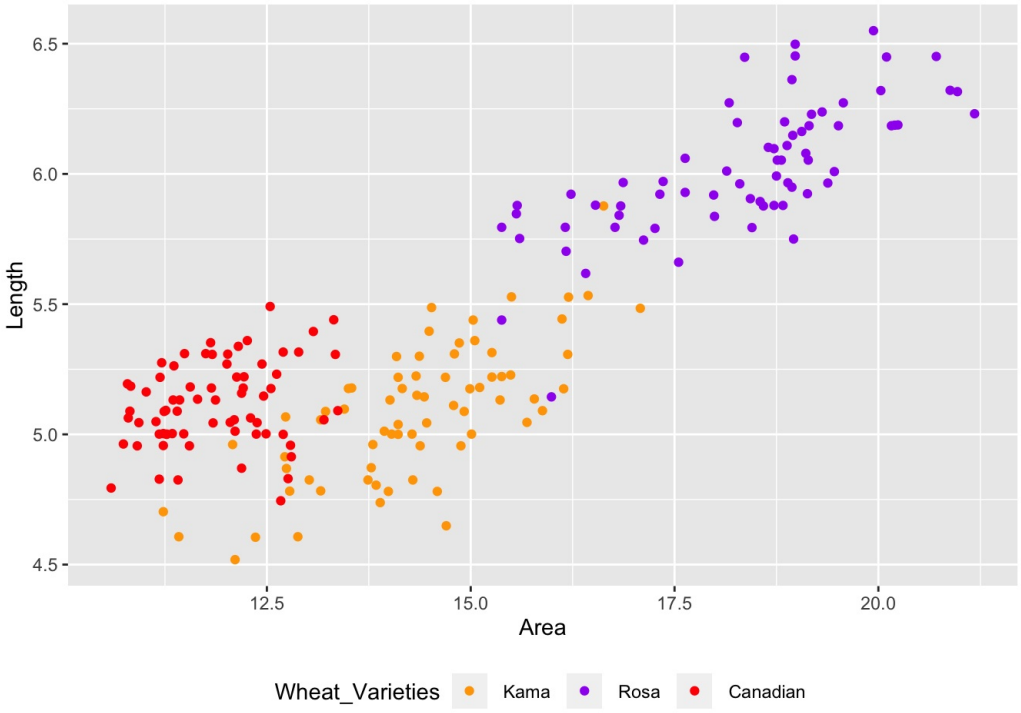
```
##
## Attaching package: 'dplyr'
```

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

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

```
# Drawing plot function
draw_plot <- function(){
  grouped <- group_by(df, Wheat_Varieties)

  wheat_color <- c(" Kama " = "orange", " Rosa " = "purple", " Canadian " = "red")

  ggplot(df, aes(x = Area, y = Length, color = Wheat_Varieties)) +
    geom_point() +
    scale_color_manual(values = wheat_color) +
    labs(x = "Area", y = "Length")
}

# Plot label function
plot_label <- function(){
  draw_plot() +
    labs(color = "Wheat_Varieties") +
    theme(legend.position = "bottom")
}

# Sigmoid function
sigmoid <- function(x){
  1 / (1 + exp(-x))
}
```

```
plot_label()
```



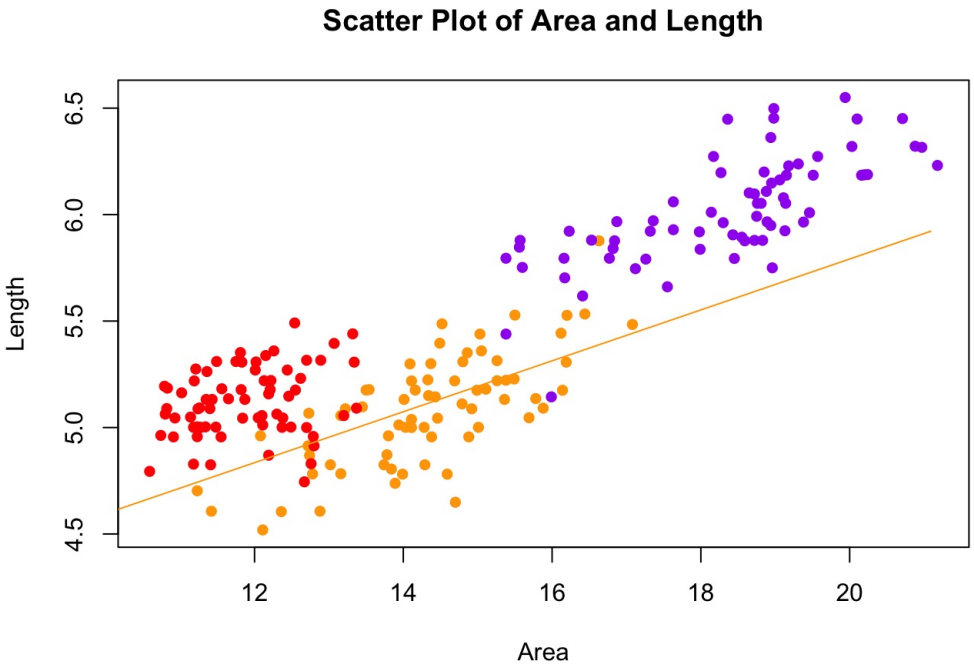

```
 # number of data
m <- nrow(df)
# Parameter Vector
w <- c(runif(1), runif(1), runif(1))
# Number of repitition for training
echo <- 5e4
# Learning rate
learning_rate <- 0.1
print(w)
```

```
## [1] 0.07950266 0.45506899 0.75537536
```

```
# initialize w as a vector of three 0's
w <- c(0, 0, 0)

# Traing process
for (iter in 1:echo) {
  temp <- matrix(c(rep(1, nrow(df)), df$Area, df$Length), nrow = 3, byrow = TRUE)

  # dot product of w and temp
  # sigma(w0*1 + w1*x1 + w2*x2)
  w_dot <- w %*% temp

  # cost function
  h <- sigmoid(w_dot)

  y <- ifelse(df$Wheat_Varieties == " Kama ", 1, 0) # Binary Classification

  # update the cost function
  h <- h - y

  # update the parameter vector > gradient descent algorithm
  delta_w0 <- learning_rate * sum(h)
  delta_w1 <- learning_rate * sum(h * df$Area)
  delta_w2 <- learning_rate * sum(h * df$Length)

  # update w0, w1, w2
  w[1] <- w[1] - delta_w0 / nrow(df)
  w[2] <- w[2] - delta_w1 / nrow(df)
  w[3] <- w[3] - delta_w2 / nrow(df)
}

print(w)
```

```
## [1]  44.816356   1.575875 -13.180305
```

```
# Binary classification
# group the data by wheat varieties
grouped <- split(df, df$Wheat_Varieties)

# create a color dictionary for wheat varieties
wheat_color <- c(' Kama ' = 'orange', ' Rosa ' = 'purple', ' Canadian ' = 'red')

# create a scatter plot with different colors for each wheat variety
plot(Length ~ Area, data = df, pch = 16, col = wheat_color[df$Wheat_Varieties],
     main = "Scatter Plot of Area and Length",
     xlab = "Area", ylab = "Length")

# add a line for the decision boundary
x_range <- seq(min(df$Area) - 2, max(df$Area), by = 0.1)
y_range <- -(w[1] + w[2] * x_range) / w[3]
lines(x_range, y_range, col = "orange")
```




```
m <- nrow(df) # number of data

w <- list() # Parameter Vector

for (i in 1:3) {
  w[[i]] <- c(runif(1), runif(1), runif(1))
}

print(w[[1]])
```

```
## [1] 0.3614363 0.8193925 0.2870458
```

```
print(w[[2]])
```

```
## [1] 0.8651871 0.4341020 0.2802583
```

```
print(w[[3]])
```

```
## [1] 0.7568752 0.8048972 0.6695680
```

```
# Multi class binary classification
variety_name <- c(" Kama ", " Rosa ", " Canadian ")

# loop through the iterations
for (y_idx in 1:3) {
  for (iter in 1:echo) {
    temp <- matrix(c(rep(1, nrow(df)), df$Area, df$Length), nrow = 3, byrow = TRUE)

    # dot product of w and temp
    # sigma(w0*1 + w1*x1 + w2*x2)
    w_dot <- w[[y_idx]] %*% temp

    # cost function
    h <- sigmoid(w_dot)

    y <- ifelse(df$Wheat_Varieties == variety_name[y_idx], 1, 0)

    h <- h - y

    delta_w0 <- learning_rate * sum(h)
    delta_w1 <- learning_rate * sum(h * df$Area)
    delta_w2 <- learning_rate * sum(h * df$Length)

    # update w0, w1, w2
    w[[y_idx]][1] <- w[[y_idx]][1] - delta_w0 / nrow(df)
    w[[y_idx]][2] <- w[[y_idx]][2] - delta_w1 / nrow(df)
    w[[y_idx]][3] <- w[[y_idx]][3]- delta_w2 / nrow(df)

  }
  print(w[[y_idx]])
}
```

```
## [1]  44.824826   1.576638 -13.183117
## [1] -30.3610196   0.9507747   2.6971106
## [1]  1.812837 -3.500269  8.597199
```

```
# Orange dot and line are for Kama
# Purple dot and line are for Rosa
# Red dot and line are for Canadian

wheat_color <- c(' Kama ' = 'orange', ' Rosa ' = 'purple', ' Canadian ' = 'red')

x_range <- seq(min(df$Area) - 2, max(df$Area), length.out = 50)
line_color <- c("orange", "purple", "red")

plot(df$Area, df$Length, xlab = "Area", ylab = "Price",  col = wheat_color[df$Wheat_Varieties])
for (i in 1:3) {
  y_range <- -(w[[i]][1] + w[[i]][2]*x_range) / w[[i]][3]
  lines(x_range, y_range, col = line_color[i])
}
points(df$Area, df$Price, pch = 20, col = "black")
```