Assignment3

Lorenzo Ricciardulli

2024-01-24

#Loading data

```
dati_senza_outlier=read.csv("dati_senza_outlier.csv", header = TRUE)
```

#Converting "Seed_Variety" cathegorical variable

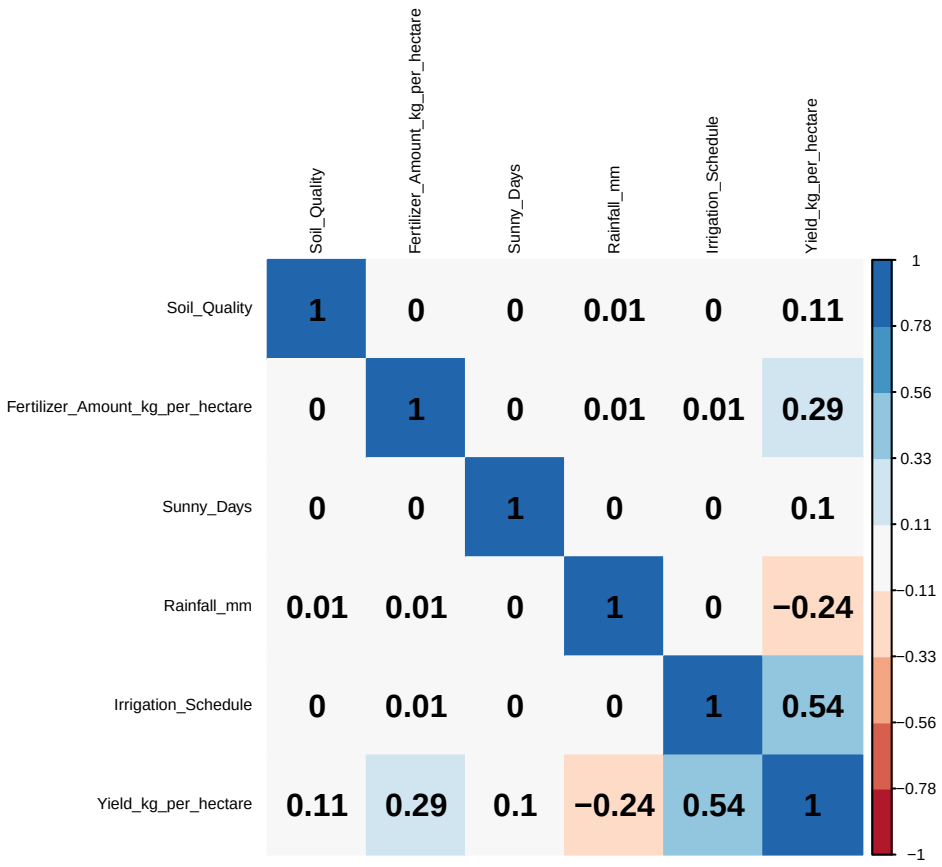
```
dati_senza_outlier$Seed_Variety=as.factor(dati_senza_outlier$Seed_Variety)
str(dati_senza_outlier)
```

```
## 'data.frame':    15509 obs. of  7 variables:
##  $ Soil_Quality                   : num  96.4 92.4 63.7 90.1 81.6 ...
##  $ Seed_Variety                   : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 1 ...
##  $ Fertilizer_Amount_kg_per_hectare: num  148 282 138 101 223 ...
##  $ Sunny_Days                     : num  94.6 90.5 97.3 113.4 83 ...
##  $ Rainfall_mm                    : num  444 518 420 548 435 ...
##  $ Irrigation_Schedule            : int  3 7 8 7 6 4 3 2 5 8 ...
##  $ Yield_kg_per_hectare           : num  684 679 935 906 898 ...
```

##(A) Consider all the variables collected in the previous assignment,with at least one categorical variable (hint: if you have only quantitative variables in your dataset you can consider to split one of them into categories). Provide a suitable graphical representation of the relationships among such variables.

A heatmap of all (continous) variable is shown below. It displays the values of correlation between the variables in the dataset. The variable "Seed_Variety" will not be shown in the graph below because it is a cathegorical variable.

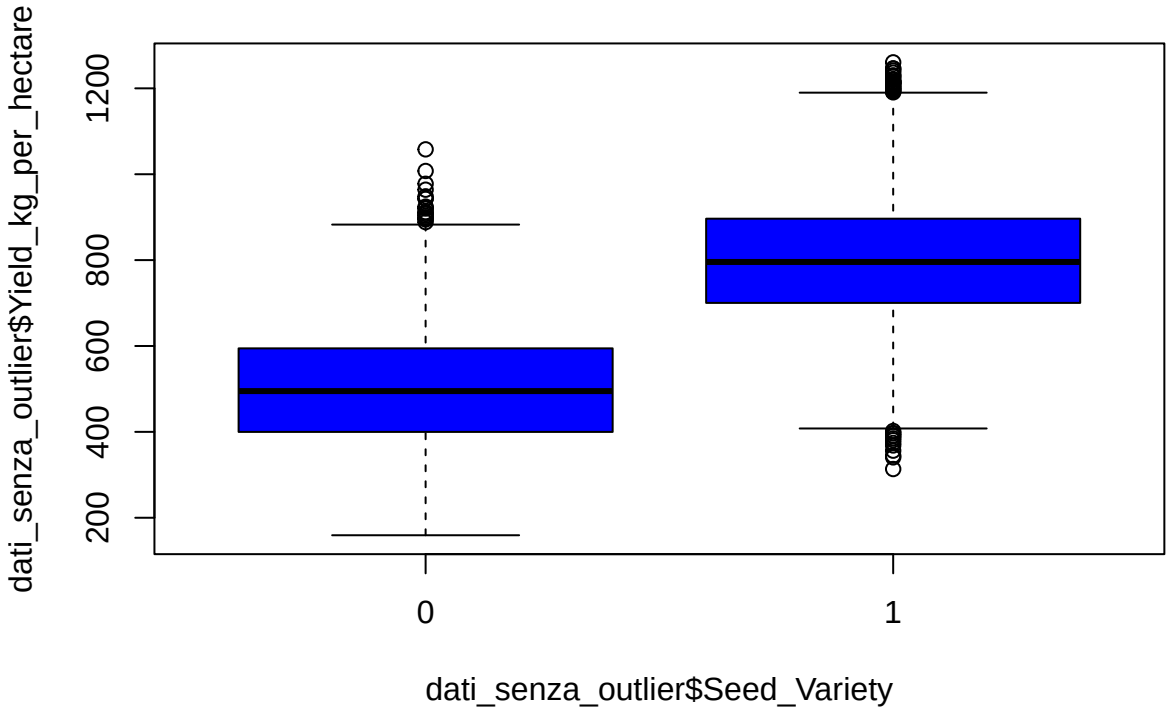
```
correlation_data =
cor(dati_senza_outlier[sapply(dati_senza_outlier, is.numeric)])
corrplot(correlation_data, method="color",
col=brewer.pal(n = 9, name = "RdBu"), addCoef.col = "black",
tl.cex = 0.5, cl.cex = 0.5, tl.col = "black")
```



Concerning the graph above, several considerations must be done: -the response variable "Yield_kg_per_hectare" is positively and moderately correlated with the variables "Irrigation_Schedule", "Fertilizer_Amount_kg_per_hectare". -the response variable Yield_kg_per_hectare" is positively and slightly correlated with "Soil_Quality" and "Sunny days". -the response variable Yield_kg_per_hectare" is negatively and slightly correlated with "Rainfall_mm".

Now, a further analysis must be done in order to investigate the relationship between the response variable Yield_kg_per_hectare" and the predictor "Seed_Variety". Considering that the predictor is a cathegorical variable, a comparison between the boxplot could be done to shed light on the possible correlation. The graph is shown below:

```
boxplot(dati_senza_outlier$Yield_kg_per_hectare ~
dati_senza_outlier$Seed_Variety, col = "blue")
```



Since the boxplot are misaligned, it could be said that there is a form of correlation between the two variables.

A quantitative analysis for the relationship between the cathegorical variable could be done using a regression model with Yield_kg_per_hectare as response variable and Seed_variety as the predictor, and analyzing the results.

```
ols = lm(dati_senza_outlier$Yield_kg_per_hectare ~ dati_senza_outlier$Seed_Variety,
data = dati_senza_outlier)
summary(ols)
```

```
##
## Call:
## lm(formula = dati_senza_outlier$Yield_kg_per_hectare ~ dati_senza_outlier$Seed_Variety,
##     data = dati_senza_outlier)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -485.63  -99.19   -3.97   96.69  557.95
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      499.962      2.090   239.2   <2e-16 ***
## dati_senza_outlier$Seed_Variety1  299.106      2.492   120.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 141.7 on 15507 degrees of freedom
## Multiple R-squared:  0.4817, Adjusted R-squared:  0.4817
## F-statistic: 1.441e+04 on 1 and 15507 DF,  p-value: < 2.2e-16
```

##(B) Fit a multiple linear regression with all the predictors. Your mean function should include at least one interaction term.

```
ols2 <- lm(dati_senza_outlier$Yield_kg_per_hectare ~
dati_senza_outlier$Rainfall_mm +
dati_senza_outlier$Sunny_Days +
dati_senza_outlier$Irrigation_Schedule +
dati_senza_outlier$Seed_Variety + dati_senza_outlier$Fertilizer_Amount_kg_per_hectare*
dati_senza_outlier$Soil_Quality, data = dati_senza_outlier)
summary(ols2)
```

```
##
## Call:
## lm(formula = dati_senza_outlier$Yield_kg_per_hectare ~ dati_senza_outlier$Rainfall_mm +
##     dati_senza_outlier$Sunny_Days + dati_senza_outlier$Irrigation_Schedule +
##     dati_senza_outlier$Seed_Variety + dati_senza_outlier$Fertilizer_Amount_kg_per_hectare *
##     dati_senza_outlier$Soil_Quality, data = dati_senza_outlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -200.357  -33.834    0.263   33.573  182.809
##
## Coefficients:
##                                                                                 Estimate
## (Intercept)                                                                    4.334e+01
## dati_senza_outlier$Rainfall_mm                                                -5.036e-01
## dati_senza_outlier$Sunny_Days                                                  1.992e+00
## dati_senza_outlier$Irrigation_Schedule                                         4.982e+01
## dati_senza_outlier$Seed_Variety1                                               2.999e+02
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                            8.403e-01
## dati_senza_outlier$Soil_Quality                                                1.616e+00
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare:dati_senza_outlier$Soil_Quality -4.164e-04
##                                                                               Std. Error
## (Intercept)                                                                    7.329e+00
## dati_senza_outlier$Rainfall_mm                                                 4.157e-03
## dati_senza_outlier$Sunny_Days                                                  4.136e-02
## dati_senza_outlier$Irrigation_Schedule                                         1.908e-01
## dati_senza_outlier$Seed_Variety1                                               8.774e-01
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                            2.918e-02
## dati_senza_outlier$Soil_Quality                                                7.282e-02
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare:dati_senza_outlier$Soil_Quality  3.838e-04
##                                                                                t value
## (Intercept)                                                                      5.913
## dati_senza_outlier$Rainfall_mm                                                -121.158
## dati_senza_outlier$Sunny_Days                                                   48.160
## dati_senza_outlier$Irrigation_Schedule                                         261.062
## dati_senza_outlier$Seed_Variety1                                               341.855
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                             28.791
## dati_senza_outlier$Soil_Quality                                                 22.186
```

```
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare:dati_senza_outlier$Soil_Quality   -1.085
##                                                                                      Pr(>|t|)
## (Intercept)                                                                          3.42e-09
## dati_senza_outlier$Rainfall_mm                                                        < 2e-16
## dati_senza_outlier$Sunny_Days                                                         < 2e-16
## dati_senza_outlier$Irrigation_Schedule                                                < 2e-16
## dati_senza_outlier$Seed_Variety1                                                      < 2e-16
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                                   < 2e-16
## dati_senza_outlier$Soil_Quality                                                       < 2e-16
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare:dati_senza_outlier$Soil_Quality    0.278
##
## (Intercept)                                                                          ***
## dati_senza_outlier$Rainfall_mm                                                       ***
## dati_senza_outlier$Sunny_Days                                                        ***
## dati_senza_outlier$Irrigation_Schedule                                               ***
## dati_senza_outlier$Seed_Variety1                                                     ***
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                                  ***
## dati_senza_outlier$Soil_Quality                                                      ***
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare:dati_senza_outlier$Soil_Quality
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.88 on 15501 degrees of freedom
## Multiple R-squared:  0.9358, Adjusted R-squared:  0.9357
## F-statistic: 3.226e+04 on 7 and 15501 DF,  p-value: < 2.2e-16
```

Several attempts have been done in order to find an interaction term among the variables that was significant, but none of them has been successful. Therefore, the interaction term should be removed from the model.

```
ols2 <- lm(dati_senza_outlier$Yield_kg_per_hectare ~
dati_senza_outlier$Rainfall_mm +
dati_senza_outlier$Sunny_Days +
dati_senza_outlier$Irrigation_Schedule +
dati_senza_outlier$Seed_Variety + dati_senza_outlier$Fertilizer_Amount_kg_per_hectare +
dati_senza_outlier$Soil_Quality, data = dati_senza_outlier)
summary(ols2)
```

```
##
## Call:
## lm(formula = dati_senza_outlier$Yield_kg_per_hectare ~ dati_senza_outlier$Rainfall_mm +
##     dati_senza_outlier$Sunny_Days + dati_senza_outlier$Irrigation_Schedule +
##     dati_senza_outlier$Seed_Variety + dati_senza_outlier$Fertilizer_Amount_kg_per_hectare +
##     dati_senza_outlier$Soil_Quality, data = dati_senza_outlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -200.340  -33.808    0.213   33.548  183.240
##
## Coefficients:
##                                                 Estimate Std. Error
## (Intercept)                                    48.836110   5.294748
## dati_senza_outlier$Rainfall_mm                 -0.503598   0.004157
## dati_senza_outlier$Sunny_Days                   1.991442   0.041358
```

```
## dati_senza_outlier$Irrigation_Schedule               49.823647   0.190850
## dati_senza_outlier$Seed_Variety1                    299.924982   0.877355
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare   0.809171   0.005574
## dati_senza_outlier$Soil_Quality                       1.542583   0.027564
##                                                      t value Pr(>|t|)
## (Intercept)                                            9.223   <2e-16 ***
## dati_senza_outlier$Rainfall_mm                      -121.153   <2e-16 ***
## dati_senza_outlier$Sunny_Days                         48.151   <2e-16 ***
## dati_senza_outlier$Irrigation_Schedule               261.062   <2e-16 ***
## dati_senza_outlier$Seed_Variety1                     341.852   <2e-16 ***
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare  145.169   <2e-16 ***
## dati_senza_outlier$Soil_Quality                       55.964   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.88 on 15502 degrees of freedom
## Multiple R-squared:  0.9358, Adjusted R-squared:  0.9357
## F-statistic: 3.764e+04 on 6 and 15502 DF,  p-value: < 2.2e-16
```

##(C) Interpret all the parameters and their uncertainties, also with the support of a graphical representation.

Coefficients (Estimates):

Intercept: The estimated intercept is 47.692579. This is the predicted value of the response variable when all predictor variables are zero.

Soil Quality: For a one-unit increase in Soil Quality, the predicted response variable (Yield_kg_per_hectare) increases by approximately 1.54 units.

Rainfall (mm): For a one-unit increase in Rainfall, the predicted response variable decreases by approximately 0.50 units.

Sunny Days: For a one-unit increase in Sunny Days, the predicted response variable increases by approximately 1.99 units.

Irrigation Schedule: For a one-unit increase in Irrigation Schedule, the predicted response variable increases by approximately 49.82 units.

Seed Variety 1: If Seed Variety is 1 (compared to other varieties), the predicted response variable increases by approximately 301.54 units.

Fertilizer Amount (kg per hectare): For a one-unit increase in Fertilizer Amount, the predicted response variable increases by approximately 0.82 units.

Standard Errors:

For instance, the standard error for Soil Quality is 0.027564, indicating relatively low uncertainty in the estimated effect of Soil Quality on the response variable.

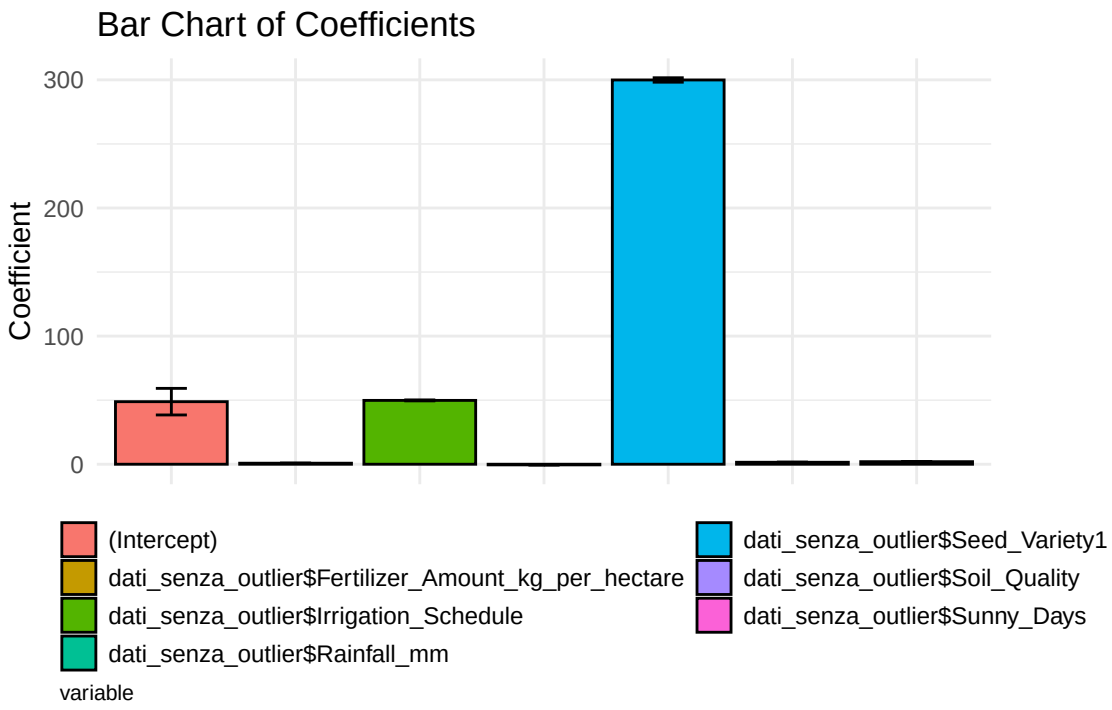
t-values and p-values:

All p-values are very close to zero, indicating that each predictor variable is statistically significant in predicting the response variable.

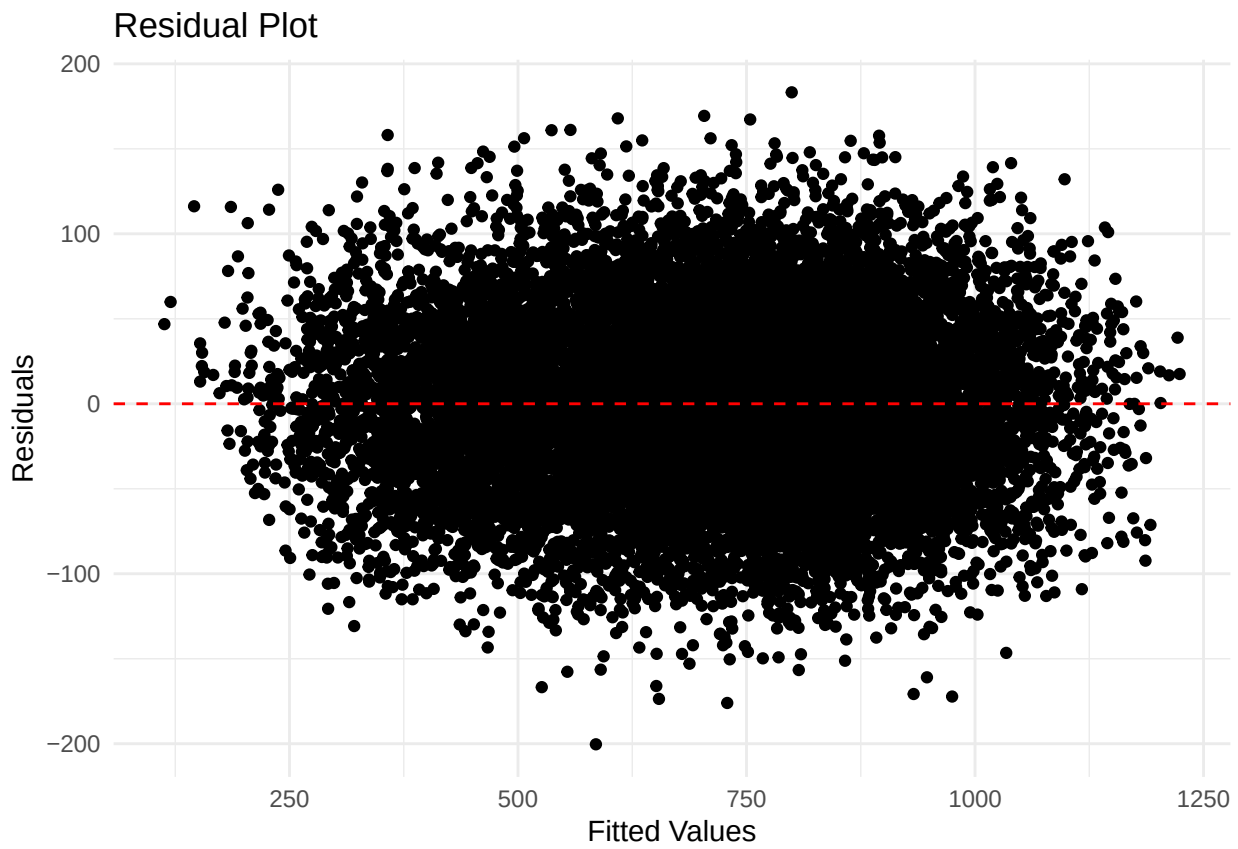
Residual Standard Error: 49.88

This is an estimate of the standard deviation of the errors (residuals) in the model. It provides a measure of how well the model fits the data.In this case, a value of 49.88 is relatively low, indicating that the model is able to explain a substantial portion of the variability in the response variable.

```r
coefficienti = coef(ols2)
errori_standard = summary(ols2)$coefficients[, "Std. Error"]
variabili = names(coefficienti)
intervalli_di_confidenza = 1.96 * errori_standard
plot_dei_dati = data.frame(
variable = variabili,
coefficient = coefficienti,
lower_bound = coefficienti - intervalli_di_confidenza,
upper_bound = coefficienti + intervalli_di_confidenza)
library(ggplot2)
bar_chart <- ggplot(plot_dei_dati, aes(x = variable, y = coefficient, fill = variable)) +
geom_col(position = "dodge", color = "black") +
geom_errorbar(
aes(ymin = lower_bound, ymax = upper_bound),
position = position_dodge(0.7),
width = 0.25) +
labs(title = "Bar Chart of Coefficients",
x = NULL,
y = "Coefficient") +
theme_minimal() +
theme(axis.text.x = element_blank(),
axis.title.x = element_blank(),
legend.position = "bottom",
legend.title = element_blank(),
legend.key.size = unit(0.5, "cm"),
plot.margin = margin(5, 50, 60, 50, "pt"),
legend.margin = margin(0, 0, 0, 30)) +
guides(fill = guide_legend(title.position = "bottom", title.theme = element_text(size = 8), ncol = 2))
print(bar_chart)
```




```
residual_plot = ggplot(data = ols2, aes(x = .fitted, y = .resid)) +
geom_point() + geom_hline(yintercept = 0, linetype = "dashed", color = "red") + labs(title = "Residual
y = "Residuals") +theme_minimal()
print(residual_plot)
```



##(D) Find and comment sigma and R2

The residual standard error is provided in the output as 49.88. This value represents the standard deviation of the residuals, which are the differences between the observed and predicted values of the dependent variable. It measures the average amount by which the observed values deviate from the predicted values.

The Adjusted R-squared is given as 0.9357. This means that approximately 93.57% of the variability in the dependent variable (Yield_kg_per_hectare) is explained by the independent variables (Rainfall_mm, Sunny_Days, Irrigation_Schedule, Seed_Variety, Fertilizer_Amount_kg_per_hectare, Soil_Quality) included in the model.

The low residual standard error suggests that the model's predictions are, on average, close to the actual observed values. The high adjusted R-squared (0.9357) indicates that the independent variables in the model collectively explain a substantial portion of the variability in the dependent variable.

##(E) Identify potential collinearity issues and propose possible solutions

```
cor(model.matrix(ols2)[,-1])
```

```
##                                                 dati_senza_outlier$Rainfall_mm
## dati_senza_outlier$Rainfall_mm                                      1.0000000000
## dati_senza_outlier$Sunny_Days                                      -0.0011775389
## dati_senza_outlier$Irrigation_Schedule                             -0.0010658907
## dati_senza_outlier$Seed_Variety1                                    0.0002848234
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                 0.0058748552
## dati_senza_outlier$Soil_Quality                                     0.0116089692
##                                                 dati_senza_outlier$Sunny_Days
## dati_senza_outlier$Rainfall_mm                                  -1.177539e-03
```

```
## dati_senza_outlier$Sunny_Days                                    1.000000e+00
## dati_senza_outlier$Irrigation_Schedule                           9.081766e-05
## dati_senza_outlier$Seed_Variety1                                -4.239522e-03
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare              2.279502e-03
## dati_senza_outlier$Soil_Quality                                 -4.349405e-03
##                                                  dati_senza_outlier$Irrigation_Schedule
## dati_senza_outlier$Rainfall_mm                                          -1.065891e-03
## dati_senza_outlier$Sunny_Days                                            9.081766e-05
## dati_senza_outlier$Irrigation_Schedule                                   1.000000e+00
## dati_senza_outlier$Seed_Variety1                                         4.662744e-03
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                      6.026376e-03
## dati_senza_outlier$Soil_Quality                                          3.362139e-03
##                                                  dati_senza_outlier$Seed_Variety1
## dati_senza_outlier$Rainfall_mm                                       0.0002848234
## dati_senza_outlier$Sunny_Days                                       -0.0042395222
## dati_senza_outlier$Irrigation_Schedule                               0.0046627441
## dati_senza_outlier$Seed_Variety1                                     1.0000000000
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                 -0.0119890193
## dati_senza_outlier$Soil_Quality                                     -0.0030556809
##                                                  dati_senza_outlier$Fertilizer_Amount_kg_per_hecta
## dati_senza_outlier$Rainfall_mm                                                                0.0058748
## dati_senza_outlier$Sunny_Days                                                                 0.0022795
## dati_senza_outlier$Irrigation_Schedule                                                        0.0060263
## dati_senza_outlier$Seed_Variety1                                                             -0.0119890
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                                           1.0000000
## dati_senza_outlier$Soil_Quality                                                              -0.0045999
##                                                  dati_senza_outlier$Soil_Quality
## dati_senza_outlier$Rainfall_mm                                       0.011608969
## dati_senza_outlier$Sunny_Days                                       -0.004349405
## dati_senza_outlier$Irrigation_Schedule                               0.003362139
## dati_senza_outlier$Seed_Variety1                                    -0.003055681
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare                 -0.004599919
## dati_senza_outlier$Soil_Quality                                      1.000000000
```

The correlation matrix shows that apparentely there are no collineartity issues, but further analysis are required, as shown below.

```
library(car)
```

```
## Caricamento del pacchetto richiesto: carData
```

```
vif(ols2)
```

```
##                       dati_senza_outlier$Rainfall_mm
##                                             1.000173
##                        dati_senza_outlier$Sunny_Days
##                                             1.000043
##                dati_senza_outlier$Irrigation_Schedule
##                                             1.000072
##                      dati_senza_outlier$Seed_Variety
##                                             1.000194
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare
##                                             1.000243
```

```
##                       dati_senza_outlier$Soil_Quality
##                                              1.000197
```

The majority of VIF values is close to 1. It can be stated that there are not collinearity issues among the variables in the model.

##(F)Test each of the beta j to be 0 against one-sided or two-sided alternatives (motivate your choices). Explicitly state and discuss the hypotheses tested and the conclusions

Looking at the graph in point (C), it is reasonable that the coefficients of the intercept, Seed_Variety are Irrigation_Schedule are way larger than 0; therefore an one sided test could be implemented, where the null hypothesis states that the coefficient is 0, while the alternative hypothesis states that the coefficient is greater than 0.

```
summary(ols2)
```

```
##
## Call:
## lm(formula = dati_senza_outlier$Yield_kg_per_hectare ~ dati_senza_outlier$Rainfall_mm +
##     dati_senza_outlier$Sunny_Days + dati_senza_outlier$Irrigation_Schedule +
##     dati_senza_outlier$Seed_Variety + dati_senza_outlier$Fertilizer_Amount_kg_per_hectare +
##     dati_senza_outlier$Soil_Quality, data = dati_senza_outlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -200.340  -33.808    0.213   33.548  183.240
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                     48.836110   5.294748
## dati_senza_outlier$Rainfall_mm                  -0.503598   0.004157
## dati_senza_outlier$Sunny_Days                    1.991442   0.041358
## dati_senza_outlier$Irrigation_Schedule          49.823647   0.190850
## dati_senza_outlier$Seed_Variety1               299.924982   0.877355
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare   0.809171   0.005574
## dati_senza_outlier$Soil_Quality                  1.542583   0.027564
##                                                 t value Pr(>|t|)
## (Intercept)                                       9.223   <2e-16 ***
## dati_senza_outlier$Rainfall_mm                 -121.153   <2e-16 ***
## dati_senza_outlier$Sunny_Days                    48.151   <2e-16 ***
## dati_senza_outlier$Irrigation_Schedule          261.062   <2e-16 ***
## dati_senza_outlier$Seed_Variety1                341.852   <2e-16 ***
## dati_senza_outlier$Fertilizer_Amount_kg_per_hectare  145.169   <2e-16 ***
## dati_senza_outlier$Soil_Quality                  55.964   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.88 on 15502 degrees of freedom
## Multiple R-squared:  0.9358, Adjusted R-squared:  0.9357
## F-statistic: 3.764e+04 on 6 and 15502 DF,  p-value: < 2.2e-16
```

```
p_values_intercept = summary(ols2)$coefficients[1, 4]
alpha <- 0.05
p_value_right <- p_values_intercept / 2
```

```r
significant_right <- p_value_right < alpha
cat("Unilateral right test (positive):\n")
```

```
## Unilateral right test (positive):
```

```r
cat("Significant coefficient:", significant_right, "\n")
```

```
## Significant coefficient: TRUE
```

The result means that, regarding the test for the intercept with the hypothesis displayed above, the null hypothesis (that states that the coefficient is equal to zero) could be rejected.

```r
p_values_Irrigation_Schedule = summary(ols2)$coefficients[4, 4]
alpha <- 0.05
p_value_right <- p_values_intercept / 2
significant_right <- p_value_right < alpha
cat("Unilateral right test (positive):\n")
```

```
## Unilateral right test (positive):
```

```r
cat("Significant coefficient:", significant_right, "\n")
```

```
## Significant coefficient: TRUE
```

The result means that, regarding the test for the variable "Irrigation_Schedule" with the hypothesis displayed above, the null hypothesis (that states that the coefficient is equal to zero) could be rejected.

```r
p_values_intercept = summary(ols2)$coefficients[5, 4]
alpha <- 0.05
p_value_right <- p_values_intercept / 2
significant_right <- p_value_right < alpha
cat("Unilateral right test (positive):\n")
```

```
## Unilateral right test (positive):
```

```r
cat("Significant coefficient:", significant_right, "\n")
```

```
## Significant coefficient: TRUE
```

The result means that, regarding the test for the variable "Seed_Variety" with the hypothesis displayed above, the null hypothesis (that states that the coefficient is equal to zero) could be rejected.

Regarding the other variables in the model, since their coefficients (as shown in the summary of the model) are close to zero, it may be reasonable to implement a two sided test. Therefore, the null hypotesis states that the coefficient is equal to 0 and the alternative hypothesis states that the coefficient is not equal to 0. The two sided test is done automatically by the function "lm" used in the precious points and the results can be seen with the function summary. The p_values (with level of significance of 0.05) of the remaining variables are shown below:

```r
p_values_Rainfall_mm = summary(ols2)$coefficients[2, 4]
p_values_Sunny_days = summary(ols2)$coefficients[3, 4]
p_values_Fertilizer = summary(ols2)$coefficients[6, 4]
p_values_Soil_Quality = summary(ols2)$coefficients[7, 4]
print(p_values_Rainfall_mm)
```

```
## [1] 0
```

```r
print(p_values_Sunny_days)
```

```
## [1] 0
```

```r
print(p_values_Fertilizer)
```

```
## [1] 0
```

```r
print(p_values_Soil_Quality)
```

```
## [1] 0
```

The print function returns values as 0 due to the very low values of the p-values. Finally, it can be stated that the p-values associated with the bilateral tests above are very close to 0 and therefore the null hypothesis could be rejected.

##(G) Test a group of regressors (motivate your choices) and all the regressors. Explicitly state and discuss the hypotheses tested and the conclusions.

As seen in the point C, it may be clear that the coefficients of the variables Soil_Quality, Sunny_days, Fertilizer_Amount_kg_per_hectare and Rainfall_mm are low compared to the other variables'ones. This raises a doubt, namely, whether the variables contribute to explaining the variance of the model or not. An F-test could be done in order to answer this question, with the following hypotheses. Null Hypothesis: The coefficients associated with the group of regressors are jointly equal to zero. Alternative Hypothesis: At least one coefficient in the group is non-zero, indicating that the group of regressors is collectively significant. If the p-value associated with the F-statistic is below a chosen significance level (e.g., 0.05),the null hypothesis could be rejected, indicating that at least one variable in the model is significantly different from zero. This implies that the model as a whole is significant in predicting the dependent variable.

```r
reduced_model = lm(dati_senza_outlier$Yield_kg_per_hectare ~
dati_senza_outlier$Irrigation_Schedule +
dati_senza_outlier$Seed_Variety,
data = dati_senza_outlier)
f_test = anova(reduced_model, ols2)
print(f_test)
```

```
## Analysis of Variance Table
##
## Model 1: dati_senza_outlier$Yield_kg_per_hectare ~ dati_senza_outlier$Irrigation_Schedule +
##     dati_senza_outlier$Seed_Variety
## Model 2: dati_senza_outlier$Yield_kg_per_hectare ~ dati_senza_outlier$Rainfall_mm +
##     dati_senza_outlier$Sunny_Days + dati_senza_outlier$Irrigation_Schedule +
##     dati_senza_outlier$Seed_Variety + dati_senza_outlier$Fertilizer_Amount_kg_per_hectare +
```

```
##     dati_senza_outlier$Soil_Quality
##   Res.Df       RSS Df Sum of Sq     F    Pr(>F)
## 1  15506 140030248
## 2  15502  38563383  4 101466866 10197 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As seen above, the p-value associeted with the F-statistic is very low, indicating that the null-hypothesis (that states that the coefficients associated with the group of regressors are jointly equal to zero) should be rejected.

##(H)Suppose you have information about a new observation of your regressors. Provide the prediction of your response variable with the associated uncertainty.

```
new_observation = data.frame(
Rainfall_mm = 500,
Sunny_Days = 100,
Irrigation_Schedule = 7,
Seed_Variety1 = 1,
Fertilizer_Amount_kg_per_hectare = 150,
Soil_Quality = 95)
prediction = predict(ols2, newdata = new_observation, interval = "prediction")
```

```
## Warning: 'newdata' ha 1 riga ma la variabile trovata ha 15509 righe
```

```
cat("Predicted Yield:", prediction[1], "kg per hectare\n")
```

```
## Predicted Yield: 731.2453 kg per hectare
```

```
uncertainty = prediction[3] - prediction[2]
cat("Uncertainty:", uncertainty, "kg per hectare\n")
```

```
## Uncertainty: 251.8725 kg per hectare
```