Assignment1

Lorenzo Ricciardulli

17/1/2024

Dataset structure, data cleaning and applied goals

```r
dataset = read.csv(file = "yield.txt")
str(dataset)
```

```
## 'data.frame':    16000 obs. of  7 variables:
##  $ Soil_Quality                  : num  96.4 92.4 63.7 90.1 81.6 ...
##  $ Seed_Variety                  : int  1 0 1 1 1 1 1 1 1 0 ...
##  $ 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 ...
```

This synthetic dataset caters to those with an interest in agriculture, machine learning, and regression analysis. It emulates the dynamics of agricultural yield based on diverse environmental and management factors, offering a solid foundation for building predictive models. Agricultural yield prediction stands as a pivotal field benefiting from predictive modeling. Precise yield predictions aid in fine-tuning agricultural production, streamlining supply chains, and ensuring food security. This synthetic dataset is crafted to mirror real-world agricultural data.The dataset encompasses 16,000 entries featuring:

Soil_Quality: An index reflecting soil quality, ranging from 50 to 100. Seed_Variety: A binary indicator denoting seed variety, with 1 representing a high-yield variety. Fertilizer_Amount_kg_per_hectare: The quantity of fertilizer used per hectare, measured in kilograms. Sunny_Days: The count of sunny days during the growing season. Rainfall_mm: Total rainfall during the growing season, measured in millimeters. Irrigation_Schedule: The number of irrigation cycles during the growing season. Yield_kg_per_hectare: Agricultural yield per hectare, serving as the target variable for prediction. This dataset is synthetic and generated for machine learning practice. Since it does not derive from real-world data, no external acknowledgments are necessary.

Now it can be stressed the need to eliminate the outliers in order to obtain a better model:

```r
head(dataset)
```

```
##   Soil_Quality Seed_Variety Fertilizer_Amount_kg_per_hectare Sunny_Days
## 1     96.41566            1                         147.8530   94.59393
## 2     92.35263            0                         281.5654   90.50464
## 3     63.71479            1                         137.8649   97.32934
## 4     90.08426            1                         100.9467  113.40483
## 5     81.60034            1                         223.0889   83.04818
```

```
## 6     65.39434            1                         104.4849   95.92214
##   Rainfall_mm Irrigation_Schedule Yield_kg_per_hectare
## 1    444.2676                   3             683.7591
## 2    517.5855                   7             678.7149
## 3    420.3109                   8             934.6920
## 4    547.8176                   7             905.8425
## 5    434.7263                   6             897.5847
## 6    462.0362                   4             634.9782
```

```
rimuovi_outlier = function(dataset) {

  dati_senza_outlier = dataset


  for (colonna in names(dati_senza_outlier)) {

    box = boxplot(dati_senza_outlier[[colonna]], plot = FALSE)
    Q1 = box$stats[2]
    Q3 = box$stats[4]
    IQR = Q3 - Q1
    lower_limit = Q1 - 1.5 * IQR
    upper_limit = Q3 + 1.5 * IQR


    outliers <- dati_senza_outlier[[colonna]] < lower_limit | dati_senza_outlier[[colonna]] > upper_lim


    dati_senza_outlier = dati_senza_outlier[!outliers, ]
  }


  return(dati_senza_outlier)
}


dati_senza_outlier = rimuovi_outlier(dataset)

head(dati_senza_outlier)
```

```
##   Soil_Quality Seed_Variety Fertilizer_Amount_kg_per_hectare Sunny_Days
## 1     96.41566            1                         147.8530   94.59393
## 2     92.35263            0                         281.5654   90.50464
## 3     63.71479            1                         137.8649   97.32934
## 4     90.08426            1                         100.9467  113.40483
## 5     81.60034            1                         223.0889   83.04818
## 6     65.39434            1                         104.4849   95.92214
##   Rainfall_mm Irrigation_Schedule Yield_kg_per_hectare
## 1    444.2676                   3             683.7591
## 2    517.5855                   7             678.7149
## 3    420.3109                   8             934.6920
## 4    547.8176                   7             905.8425
## 5    434.7263                   6             897.5847
## 6    462.0362                   4             634.9782
```

Exploration data analysis

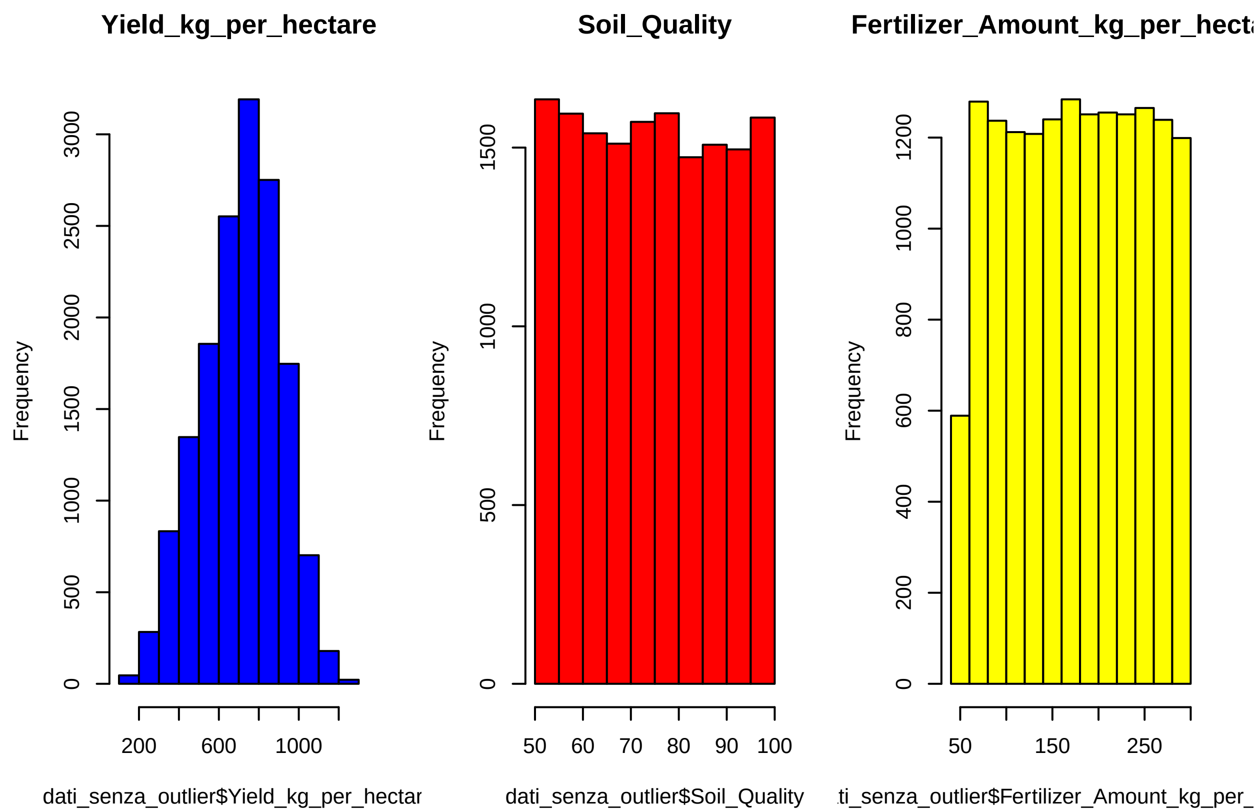
The summary statistics of the variables are shown below:

```
summary(dati_senza_outlier)
```

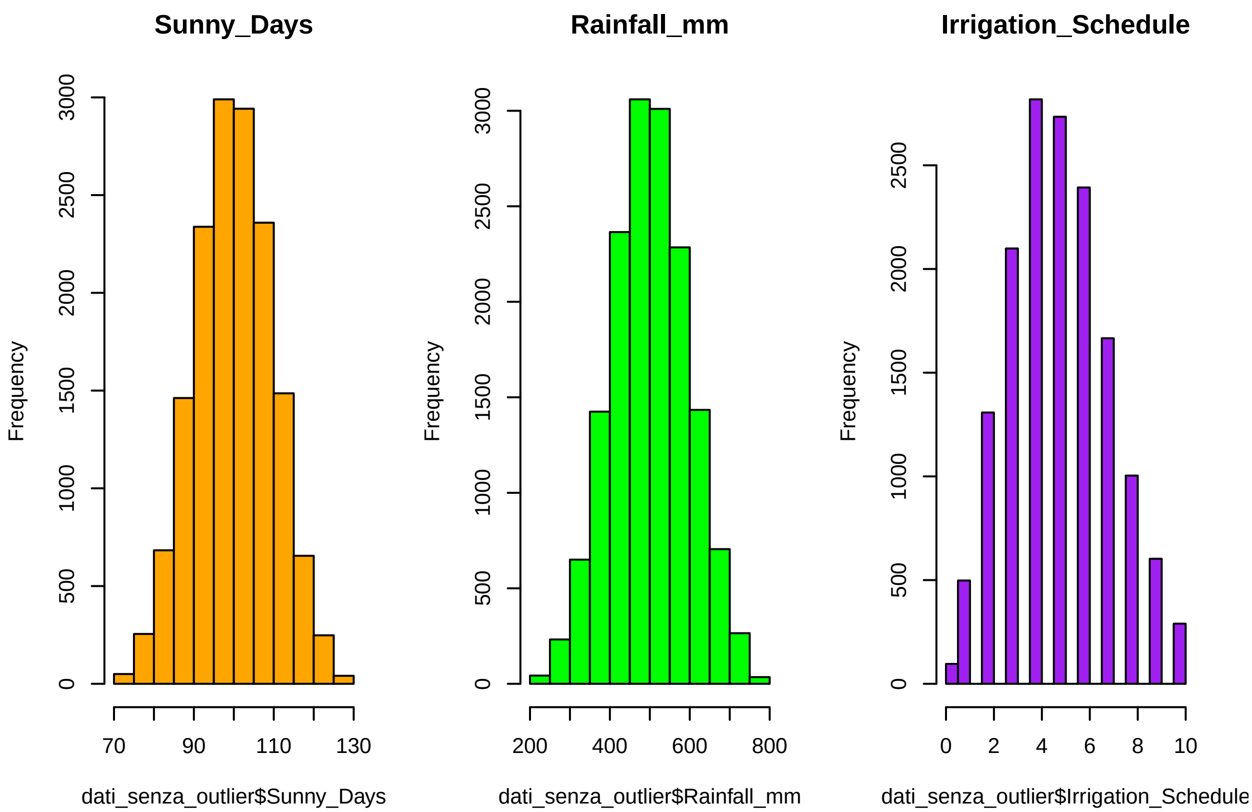
```
##   Soil_Quality     Seed_Variety    Fertilizer_Amount_kg_per_hectare
##  Min.   : 50.01   Min.   :0.0000   Min.   : 50.05
##  1st Qu.: 62.17   1st Qu.:0.0000   1st Qu.:112.61
##  Median : 74.67   Median :1.0000   Median :175.73
##  Mean   : 74.77   Mean   :0.7039   Mean   :175.13
##  3rd Qu.: 87.41   3rd Qu.:1.0000   3rd Qu.:237.43
##  Max.   :100.00   Max.   :1.0000   Max.   :299.99
##    Sunny_Days      Rainfall_mm    Irrigation_Schedule Yield_kg_per_hectare
##  Min.   : 72.91   Min.   :232.8   Min.   : 0.000      Min.   : 159.3
##  1st Qu.: 93.26   1st Qu.:433.9   1st Qu.: 3.000      1st Qu.: 576.1
##  Median : 99.97   Median :499.7   Median : 5.000      Median : 726.9
##  Mean   : 99.94   Mean   :500.5   Mean   : 4.948      Mean   : 710.5
##  3rd Qu.:106.64   3rd Qu.:566.5   3rd Qu.: 6.000      3rd Qu.: 853.1
##  Max.   :126.83   Max.   :767.5   Max.   :10.000      Max.   :1260.5
```

Plots of the variables, in order to observe the distributions, are displayed as it follows:

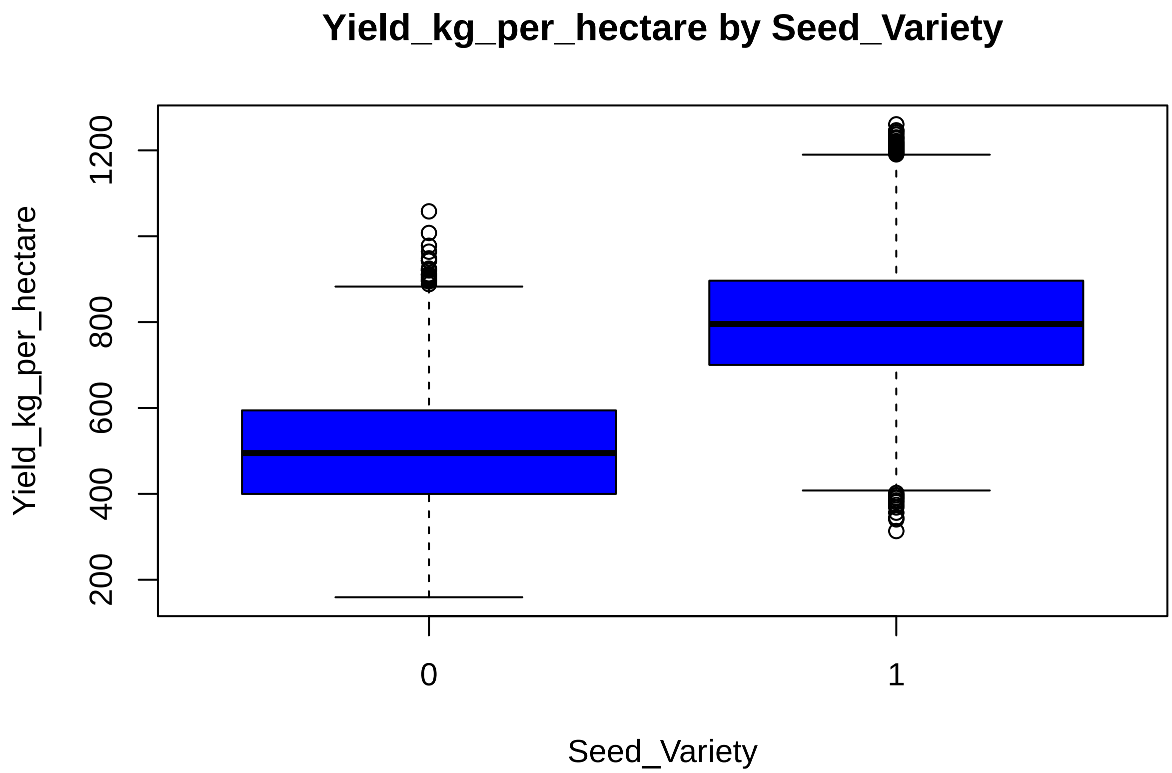
```
par(mfrow=c(1,3))
hist(dati_senza_outlier$Yield_kg_per_hectare, main="Yield_kg_per_hectare", col = "blue")
hist(dati_senza_outlier$Soil_Quality, main="Soil_Quality", col = "red")
hist(dati_senza_outlier$Fertilizer_Amount_kg_per_hectare, main="Fertilizer_Amount_kg_per_hectare", col =
```




```r
par(mfrow=c(1,3))
hist(dati_senza_outlier$Sunny_Days, main="Sunny_Days", col = "orange")
hist(dati_senza_outlier$Rainfall_mm, main="Rainfall_mm", col = "green")
hist(dati_senza_outlier$Irrigation_Schedule, main="Irrigation_Schedule", col = "purple")
```



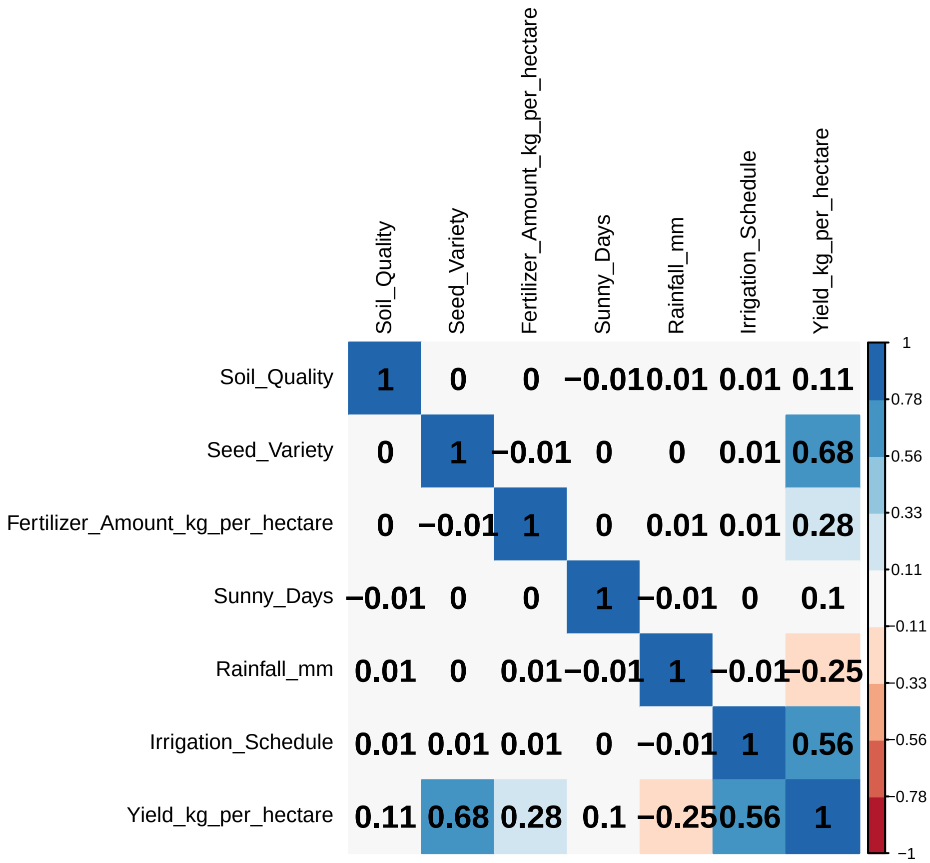

```
boxplot(dati_senza_outlier$Yield_kg_per_hectare ~ dati_senza_outlier$Seed_Variety, col ="blue", main =
        ylab = "Yield_kg_per_hectare")
```



From the boxplot comparison above, at first glance a correlation could be observed between the target (response) variable Yield_kg_per_hectare and Seed_Variety.

Further correlations could be stressed looking at the heatmap below:

```
correlation_data <- cor(dataset[sapply(dataset, is.numeric)])
corrplot(correlation_data, method="color", col=brewer.pal(n = 9, name = "RdBu"), addCoef.col = "black",
```



Looking at the Yield_kg_per_hectare, the target variable, it can be stated that: -It has a strong linear correlation with the seed variety -It has an average linear correlation with the irrigation schedule -It has a weak correlation (in absolute terms) with rainfall and fertilizer amount used on the lands -It has not relevant correlation with the other variables