# Project Pig Growth

Charles YANG

2024-11-03

```
rm(list=ls())
graphics.off()
```

## Exploratory data

```
train=read.csv("pig_data_proj/train1.csv",sep=",")
head(train)
```

```
##   Farm        Day ID Species Gender Age Weight Chest Length NumberID
## 1    1 2020-08-08  3       2      2   4    9.0    NA     NA        2
## 2    1 2020-08-15  3       2      2   5   11.5    NA     NA        2
## 3    1 2020-08-22  3       2      2   6   15.5    NA     NA        2
## 4    1 2020-08-29  3       2      2   7   20.0    NA     NA        2
## 5    1 2020-09-05  3       2      2   8   21.0    64     52        2
## 6    1 2020-09-12  3       2      2   9   24.0    67     50        2
```

```
summary(train)
```

```
##       Farm          Day                  ID           Species
##  Min.   :1.00   Length:2729        Min.   : 1.00   Min.   :1.000
##  1st Qu.:2.00   Class :character   1st Qu.: 7.00   1st Qu.:1.000
##  Median :4.00   Mode  :character   Median :13.00   Median :1.000
##  Mean   :3.93                      Mean   :12.78   Mean   :1.531
##  3rd Qu.:6.00                      3rd Qu.:19.00   3rd Qu.:2.000
##  Max.   :7.00                      Max.   :28.00   Max.   :3.000
##
##      Gender           Age             Weight           Chest
##  Min.   :1.000   Min.   : 3.00   Min.   :  5.30   Min.   : 57.00
##  1st Qu.:1.000   1st Qu.: 9.00   1st Qu.: 23.00   1st Qu.: 74.00
##  Median :1.000   Median :13.00   Median : 46.00   Median : 88.00
##  Mean   :1.387   Mean   :13.48   Mean   : 50.84   Mean   : 89.12
##  3rd Qu.:2.000   3rd Qu.:18.00   3rd Qu.: 77.50   3rd Qu.:101.00
##  Max.   :2.000   Max.   :25.00   Max.   :124.10   Max.   :138.00
##                                                   NA's   :1404
##      Length          NumberID
##  Min.   : 46.00   Min.   :  2.0
##  1st Qu.: 75.00   1st Qu.: 42.0
##  Median : 86.00   Median : 83.0
```

```
##  Mean   : 92.21   Mean   : 83.3
##  3rd Qu.:111.00   3rd Qu.:123.0
##  Max.   :148.00   Max.   :170.0
##  NA's   :1404
```

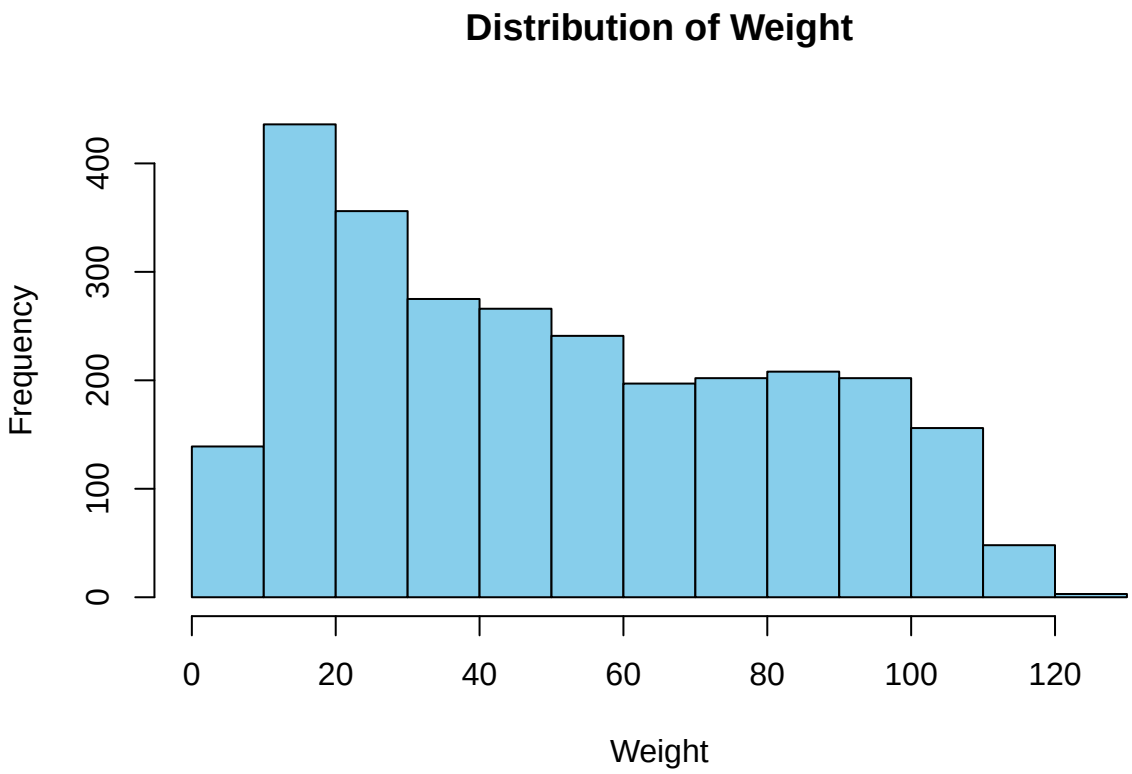
```
str(train)
```

```
## 'data.frame':    2729 obs. of  10 variables:
##  $ Farm    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Day     : chr  "2020-08-08" "2020-08-15" "2020-08-22" "2020-08-29" ...
##  $ ID      : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Species : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ Gender  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ Age     : int  4 5 6 7 8 9 10 11 12 13 ...
##  $ Weight  : num  9 11.5 15.5 20 21 24 27 31.5 36 41 ...
##  $ Chest   : num  NA NA NA NA 64 67 67 69 70 72 ...
##  $ Length  : num  NA NA NA NA 52 50 55 59 59 63 ...
##  $ NumberID: int  2 2 2 2 2 2 2 2 2 2 ...
```

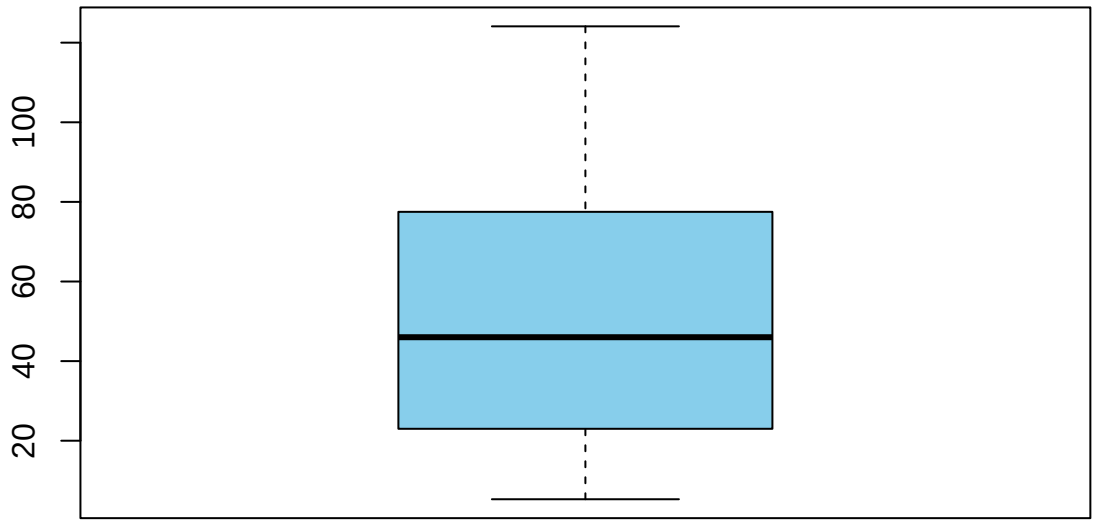
```
colSums(is.na(train))
```

```
##     Farm      Day       ID  Species   Gender      Age   Weight    Chest
##        0        0        0        0        0        0        0     1404
##   Length NumberID
##     1404        0
```

```
hist(train$Weight, main="Distribution of Weight", xlab="Weight", col="skyblue")
```




```
boxplot(train$Weight, main="Boxplot of Weight", col="skyblue")
```




```r
mean(train$Weight, na.rm = TRUE)
```

```
## [1] 50.83611
```

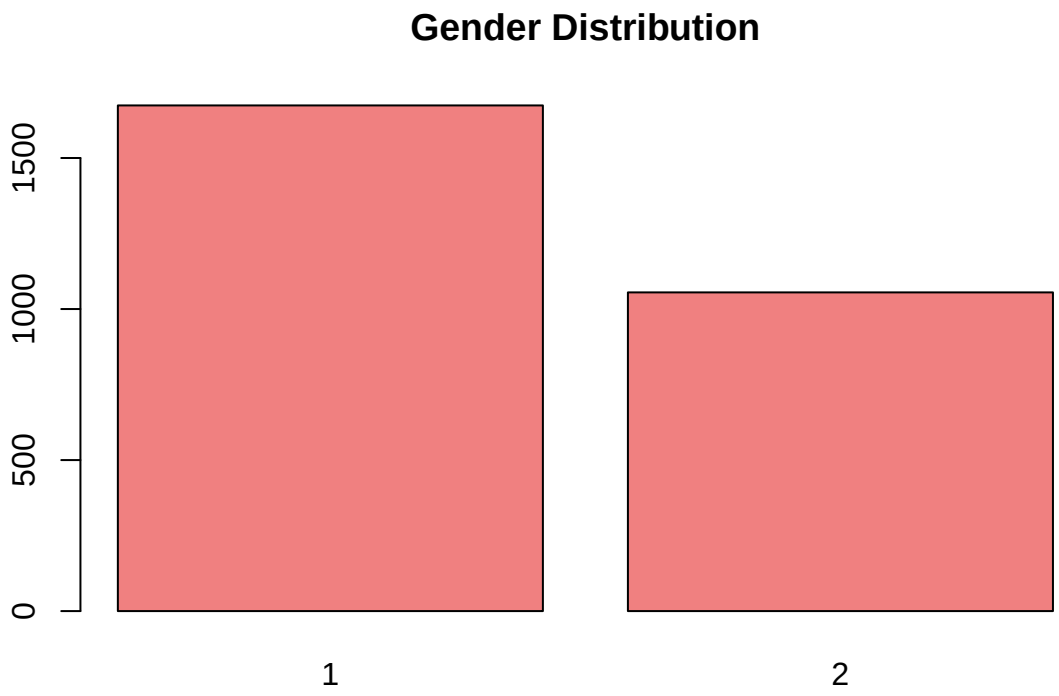
```r
sd(train$Weight, na.rm = TRUE)
```

```
## [1] 31.07581
```

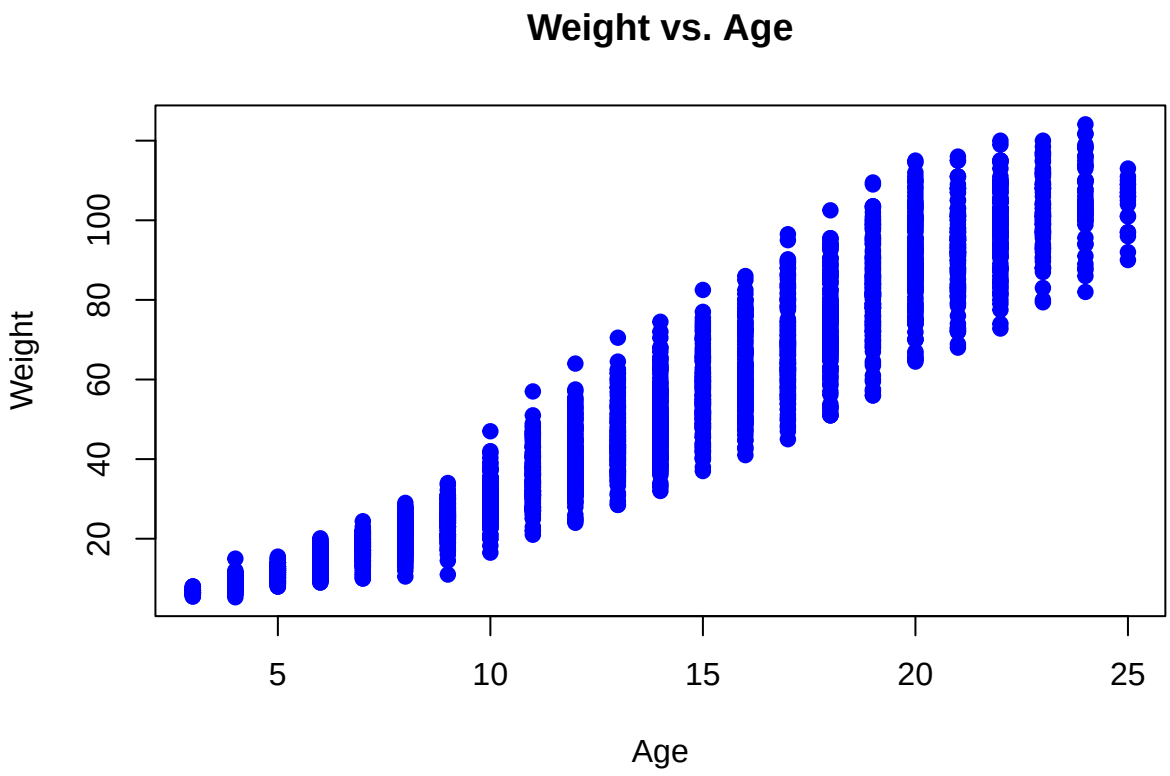
```r
table(train$Species)
```

```
##
##    1    2    3
## 1525  960  244
```

```r
barplot(table(train$Gender), main="Gender Distribution", col="lightcoral")
```

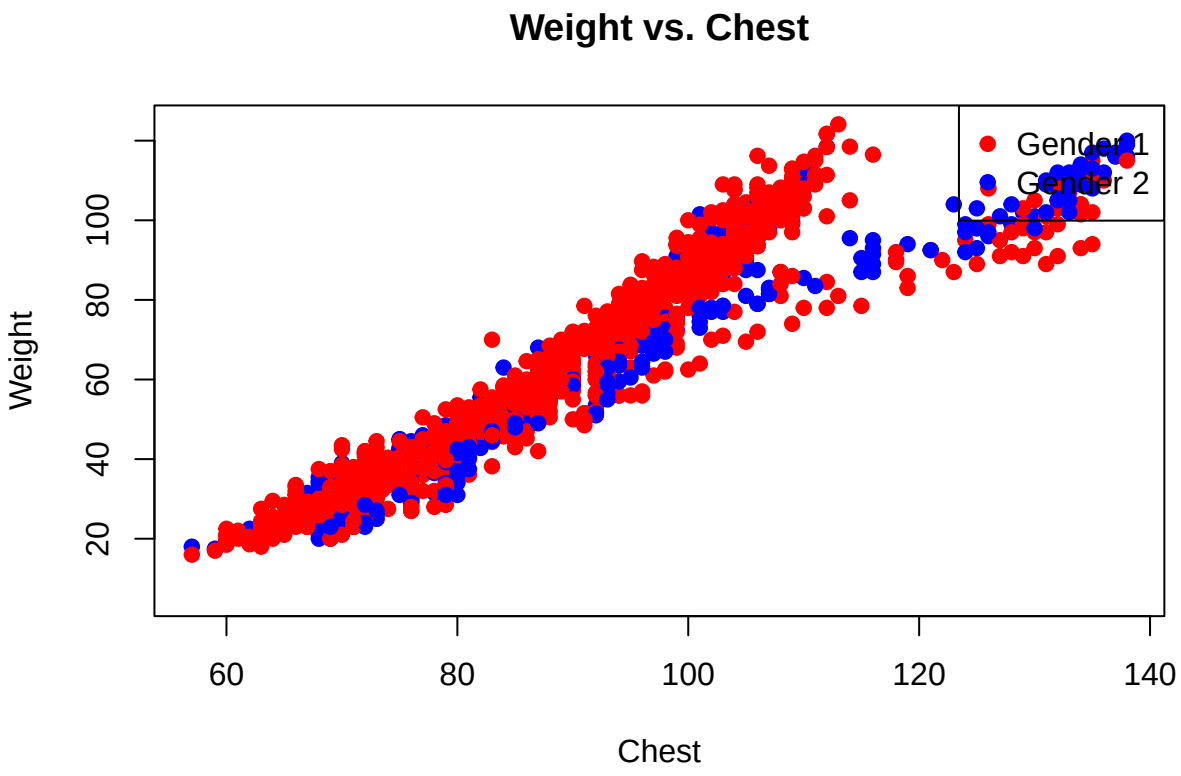



```
plot(train$Age, train$Weight, main="Weight vs. Age", xlab="Age", ylab="Weight", col="blue", pch=19)
```




```r
# Scatter plot with conditional colors based on Gender
plot(train$Chest, train$Weight,
     main = "Weight vs. Chest",
     xlab = "Chest",
     ylab = "Weight",
     col = ifelse(train$Gender == 1, "red", "blue"),
     pch = 19)

# Add legend
legend("topright", legend = c("Gender 1", "Gender 2"),
       col = c("red", "blue"), pch = 19)
```

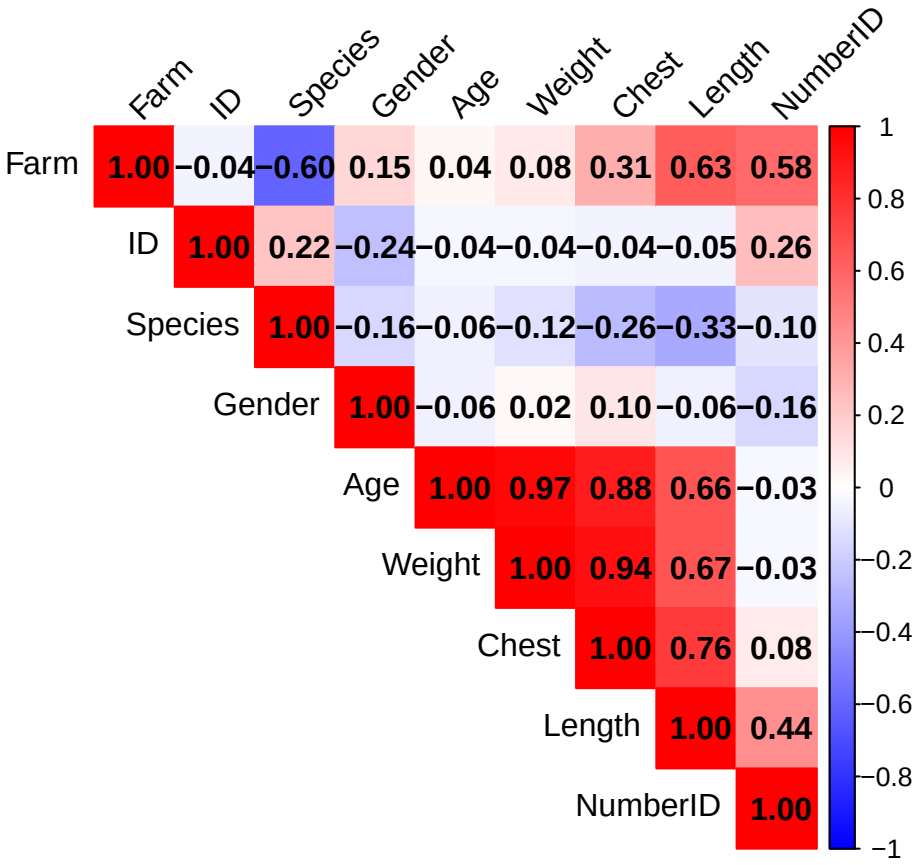



```
# Calculate correlation matrix for numeric columns only
numeric_data <- train[, sapply(train, is.numeric)]
cor_matrix <- cor(numeric_data, use = "complete.obs")
```

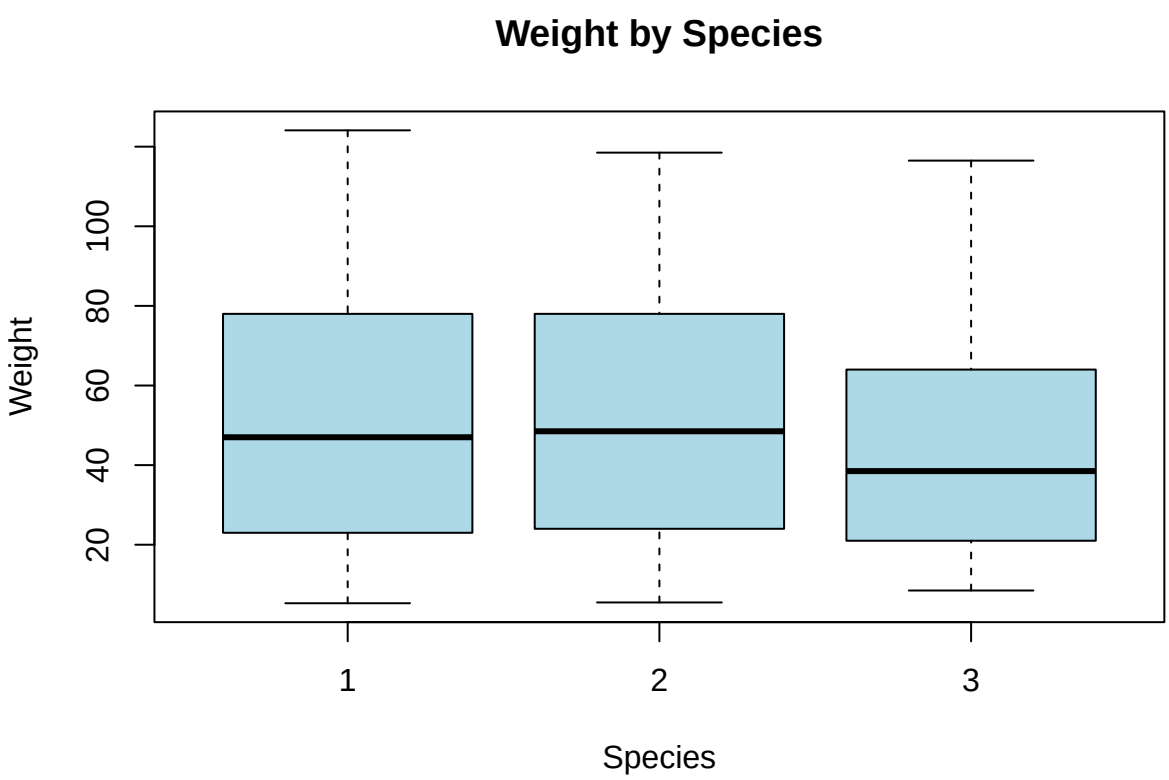
```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
# Plot the correlation matrix as a heatmap
corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", # Shows correlation coefficient values
         col = colorRampPalette(c("blue", "white", "red"))(200)) # Color scale
```




```
boxplot(train$Weight ~ train$Species, main = "Weight by Species", xlab = "Species", ylab = "Weight", co
```


Weight by Species
Weight
20
40
60
80
100
1
2
3
Species