TEST

Adib Habbou - Alae Khidour

2022-11-15

Setup

Data Import

```
load(file="project_park.RData")
```

Data Size

```
dim(geno.df)
```

```
[1] 36901   416
```

```
dim(pheno.df)
```

```
[1] 413  38
```

```
dim(Xmat)
```

```
[1]   413 36901
```

Y names

```
names(pheno.df)
```

```
 [1] "HybID"                          "NSFTVID"
 [3] "Flowering.time.at.Arkansas"     "Flowering.time.at.Faridpur"
 [5] "Flowering.time.at.Aberdeen"     "FT.ratio.of.Arkansas.Aberdeen"
 [7] "FT.ratio.of.Faridpur.Aberdeen"  "Culm.habit"
 [9] "Leaf.pubescence"                "Flag.leaf.length"
[11] "Flag.leaf.width"                "Awn.presence"
[13] "Panicle.number.per.plant"       "Plant.height"
[15] "Panicle.length"                 "Primary.panicle.branch.number"
[17] "Seed.number.per.panicle"        "Florets.per.panicle"
[19] "Panicle.fertility"              "Seed.length"
```

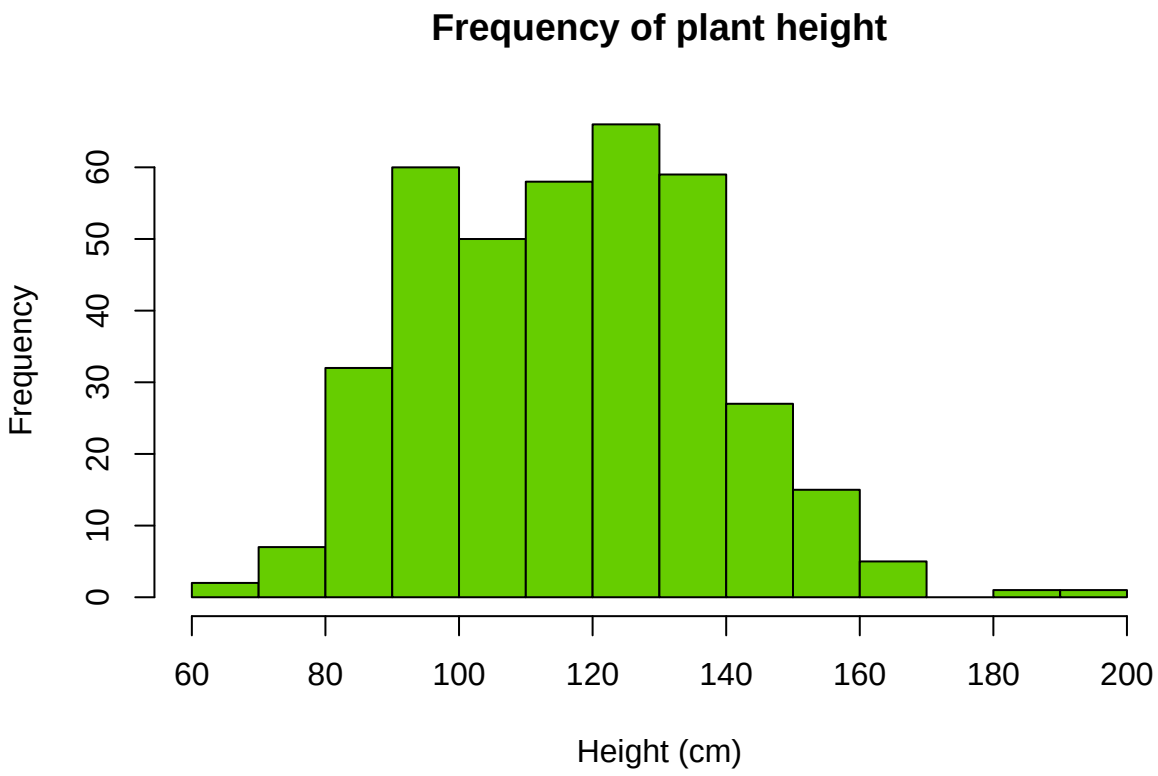
```
[21] "Seed.width"                          "Seed.volume"
[23] "Seed.surface.area"                   "Brown.rice.seed.length"
[25] "Brown.rice.seed.width"               "Brown.rice.surface.area"
[27] "Brown.rice.volume"                   "Seed.length.width.ratio"
[29] "Brown.rice.length.width.ratio"       "Seed.color"
[31] "Pericarp.color"                      "Straighthead.suseptability"
[33] "Blast.resistance"                    "Amylose.content"
[35] "Alkali.spreading.value"              "Protein.content"
[37] "Year07Flowering.time.at.Arkansas" "Year06Flowering.time.at.Arkansas"
```

Choice of Y

```
# Plant.height
i = 14
y = matrix(pheno.df[,i])
yname = names(pheno.df)[i]
summary(y)
```

```
       V1
 Min.   : 67.75
 1st Qu.: 99.75
 Median :117.50
 Mean   :116.58
 3rd Qu.:131.39
 Max.   :194.33
 NA's   :30
```

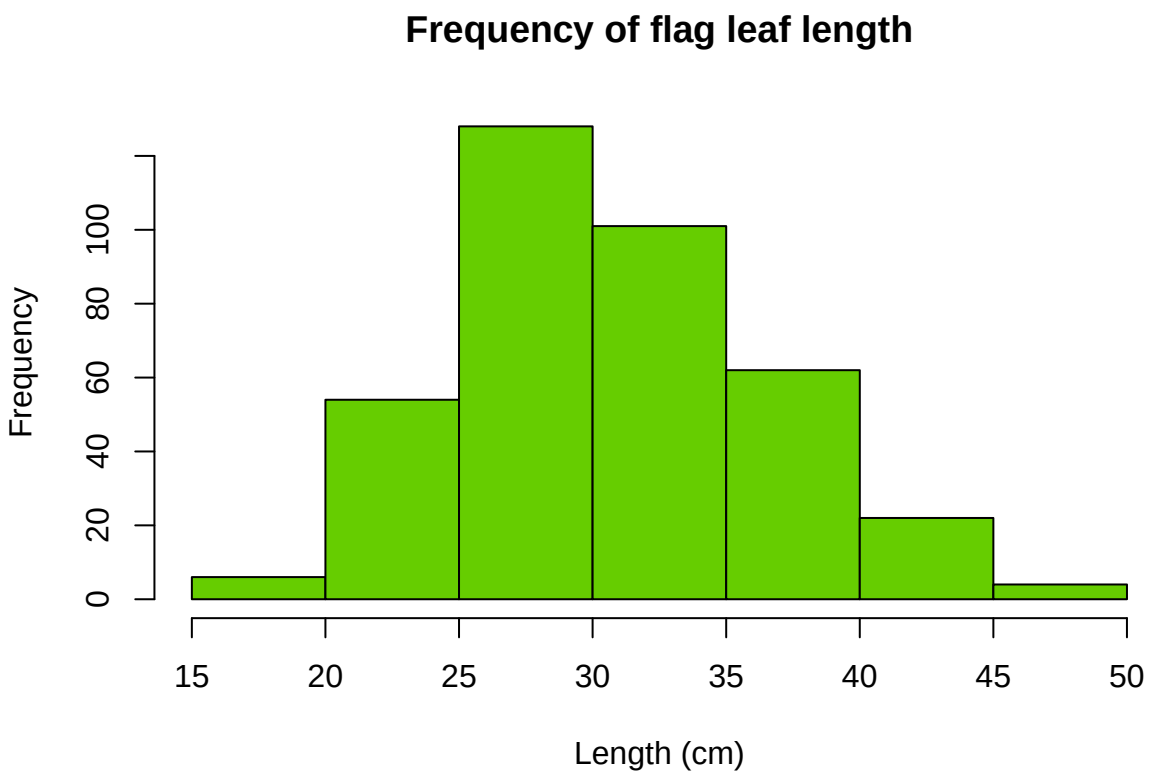
```
hist(y, main = "Frequency of plant height", xlab = "Height (cm)", col = "chartreuse3")
```




```
# Flag.leaf.length
i = 10
y = matrix(pheno.df[,i])
yname = names(pheno.df)[i]
summary(y)
```

```
       V1
 Min.   :15.42
 1st Qu.:26.62
 Median :30.05
 Mean   :30.63
 3rd Qu.:34.55
 Max.   :49.44
 NA's   :36
```

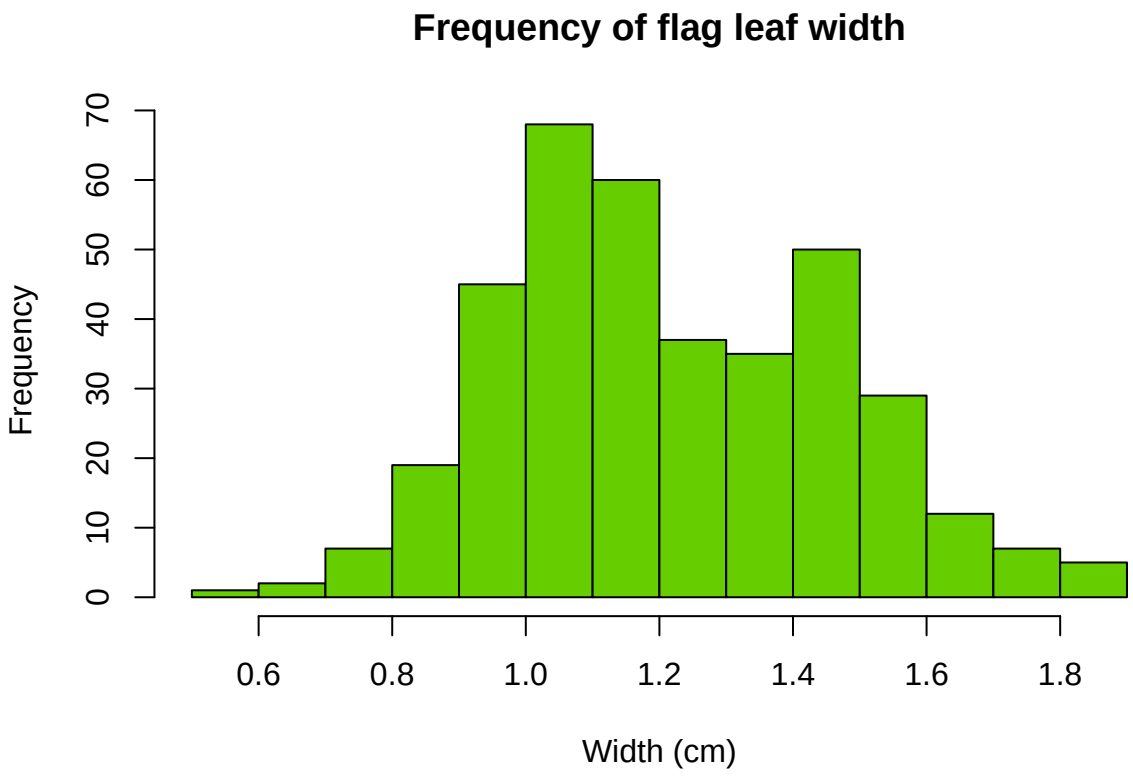
```
hist(y, main = "Frequency of flag leaf length", xlab = "Length (cm)", col = "chartreuse3")
```




```
# Flag.leaf.width
i = 11
y = matrix(pheno.df[,i])
yname = names(pheno.df)[i]
summary(y)
```

```
       V1
 Min.   :0.5917
 1st Qu.:1.0400
 Median :1.1833
 Mean   :1.2217
 3rd Qu.:1.4111
 Max.   :1.8917
 NA's   :36
```

```
hist(y, main = "Frequency of flag leaf width", xlab = "Width (cm)", col = "chartreuse3")
```



Sample of data

```
Xmat[1:5,1:5]
```

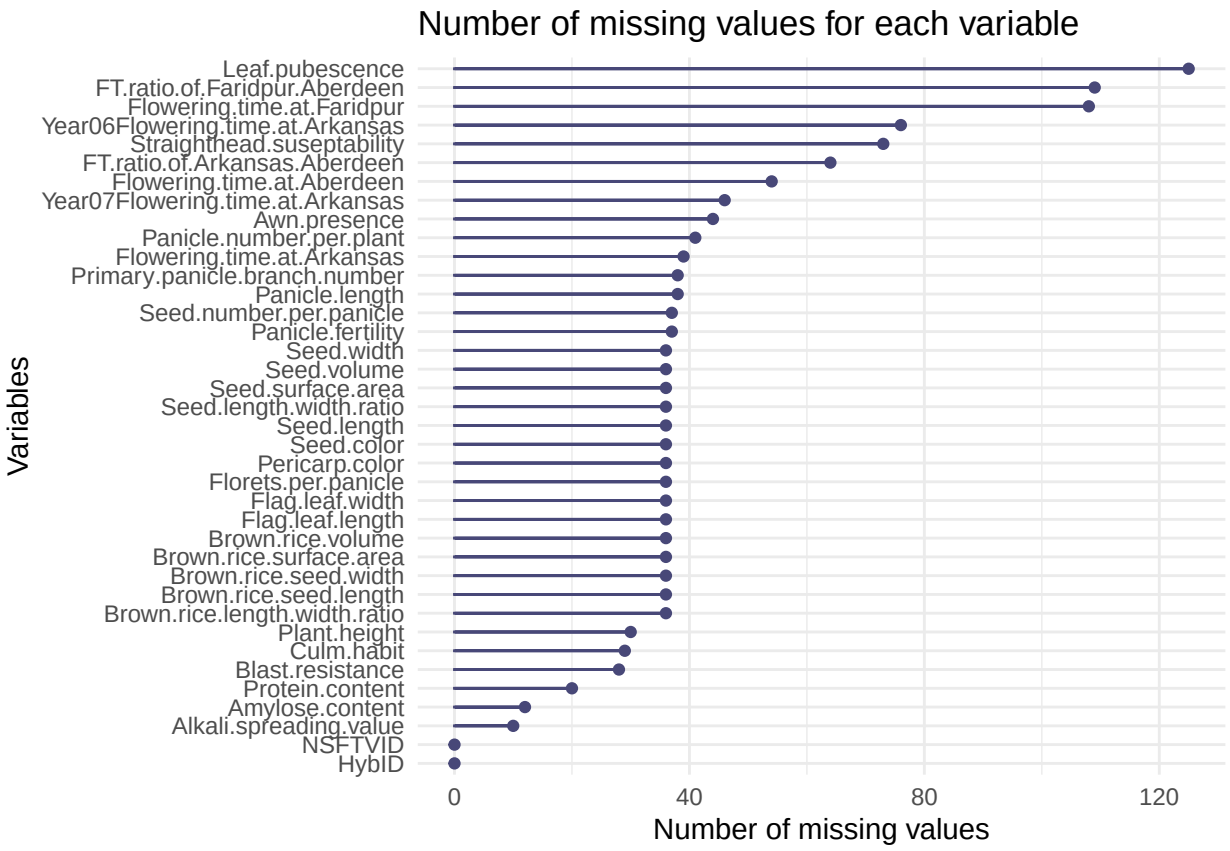
```
  id1000001 id1000003 id1000005 id1000007 id1000008
1         0         0         0         0         0
3         2         2         0         2         2
4         2         2         0         2         2
5         2         2         2         0         2
6         2         2         0         2         2
```

```
geno.df[1:5,1:10]
```

```
     marker chrom   pos 1 2 3 4 5 6 7
1 id1000001     1 13147 0 2 2 2 2 0 0
2 id1000003     1 73192 0 2 2 2 2 0 0
3 id1000005     1 74969 0 0 0 2 0 0 0
4 id1000007     1 75852 0 2 2 0 2 0 0
5 id1000008     1 75953 0 2 2 2 2 0 0
```

Plot of NA

```
library(naniar)
library(ggplot2)
gg_miss_var(pheno.df) + labs(y = "Number of missing values") + ggtitle("Number of missing values for ea
```



Imputation of NA in Y

```
Y <- pheno.df["Plant.height"]
mean <- mean(Y[!is.na(Y)])
Plant.height.is.missing <- vector(length = nrow(Y))
for (i in 1:nrow(Y))
{
  if (is.na(Y[i,1]))
  {
    Plant.height.is.missing[i] <- 1
    Y[i,1] <- mean
  }
}
Plant.height <- cbind(Y,Plant.height.is.missing)
knitr::kable(Plant.height[10:18,], , caption = "Sample of our final target variable")
```

Table 1: Sample of our final target variable

	Plant.height	Plant.height.is.missing
10	116.5826	1
11	135.1667	0
12	117.8889	0
13	116.5826	1
14	161.8333	0
15	88.0000	0
16	132.0000	0
17	130.5000	0
18	120.6667	0

```
new_X <- read.csv("bio_data.csv")
```

```
knitr::kable(new_X[1:10,1:5], caption = "Sample of our final dataset")
```

Table 2: Sample of our final dataset

X	id1000001	id1000003	id1000005	id1000007
3	2	2	0	2
4	2	2	0	2
5	2	2	2	0
6	2	2	0	2
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0
13	2	2	0	2