21BDS0340

Abhinav Dinesh Srivatsa

Exploratory Data Analysis Lab

Assignment – II

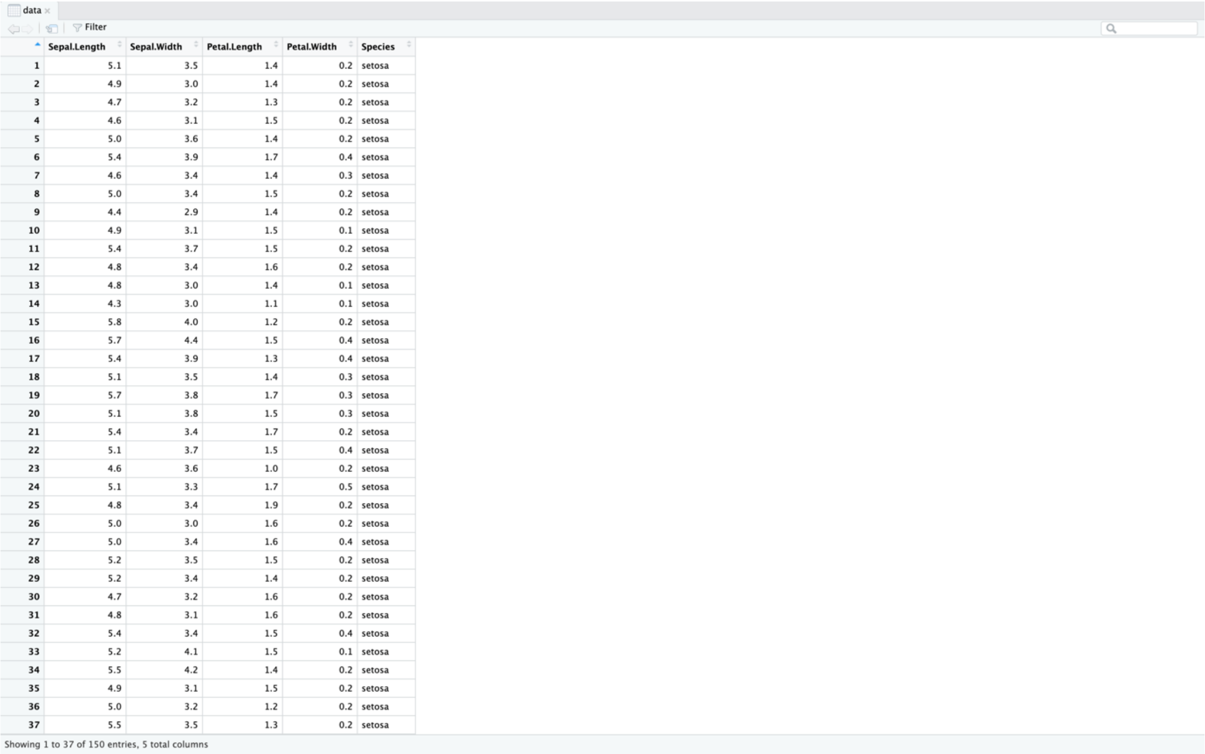
Experiment 5

Code:

```
library(dplyr)
library(missForest)
library(mice)
library(VIM)
library(ggplot2)
library(cowplot)

data = iris
View(data)
```

Output:



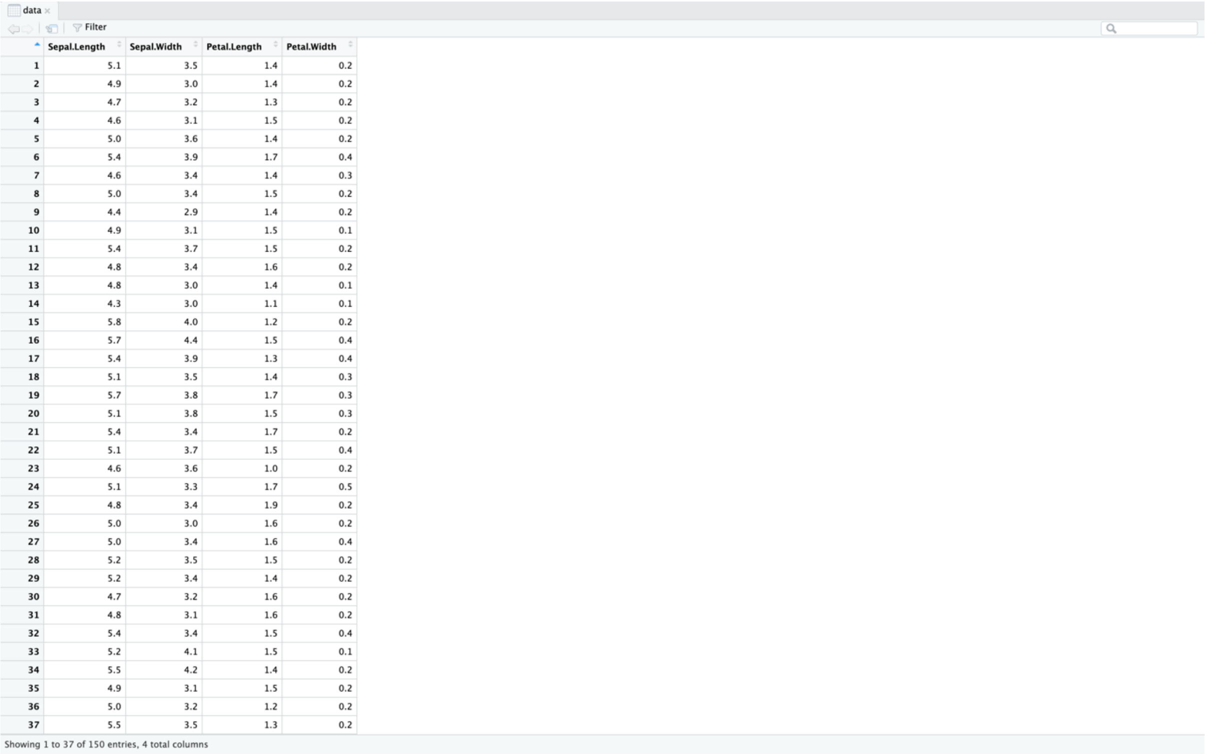
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa
26	5.0	3.0	1.6	0.2	setosa
27	5.0	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa
31	4.8	3.1	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
35	4.9	3.1	1.5	0.2	setosa
36	5.0	3.2	1.2	0.2	setosa
37	5.5	3.5	1.3	0.2	setosa

Showing 1 to 37 of 150 entries, 5 total columns

Code:

```
# dropping labels
data = data %>% select(-c("Species"))
View(data)
```

Output:



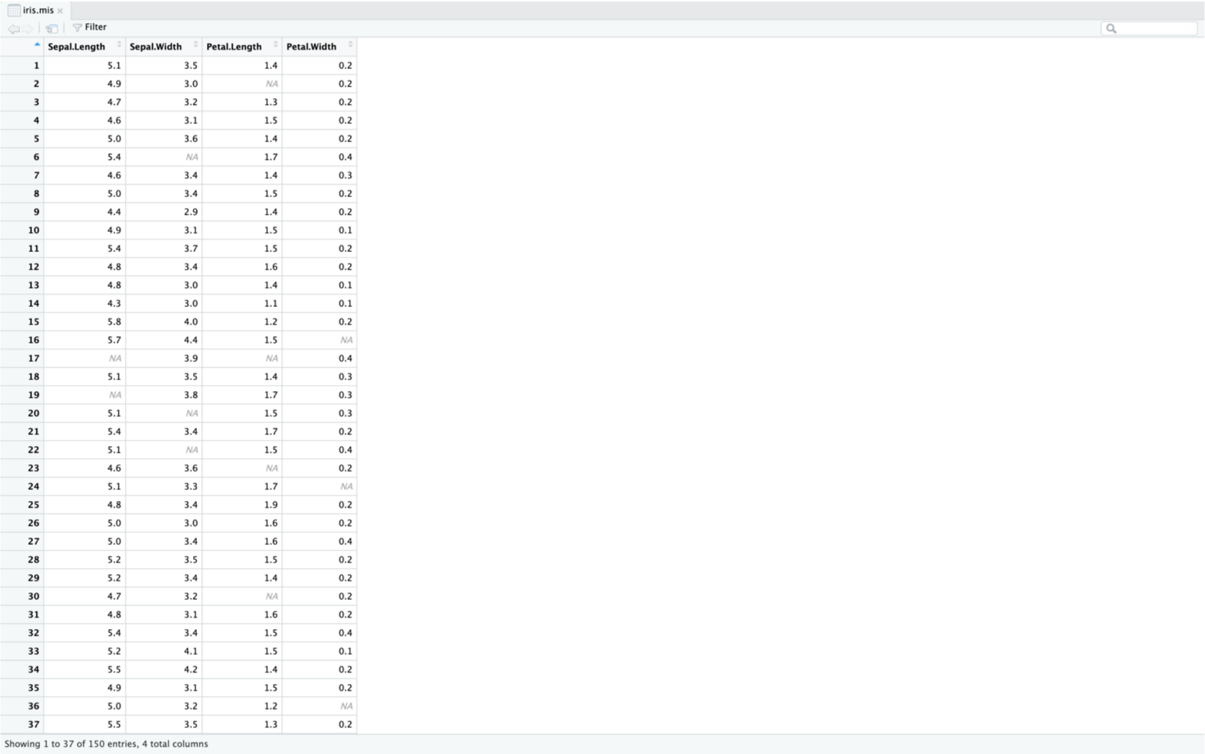
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3.0	1.4	0.1
14	4.3	3.0	1.1	0.1
15	5.8	4.0	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	5.1	3.8	1.5	0.3
21	5.4	3.4	1.7	0.2
22	5.1	3.7	1.5	0.4
23	4.6	3.6	1.0	0.2
24	5.1	3.3	1.7	0.5
25	4.8	3.4	1.9	0.2
26	5.0	3.0	1.6	0.2
27	5.0	3.4	1.6	0.4
28	5.2	3.5	1.5	0.2
29	5.2	3.4	1.4	0.2
30	4.7	3.2	1.6	0.2
31	4.8	3.1	1.6	0.2
32	5.4	3.4	1.5	0.4
33	5.2	4.1	1.5	0.1
34	5.5	4.2	1.4	0.2
35	4.9	3.1	1.5	0.2
36	5.0	3.2	1.2	0.2
37	5.5	3.5	1.3	0.2

Showing 1 to 37 of 150 entries, 4 total columns

Code:

```
# adding 10% random values
iris.mis <- prodNA(data, noNA = 0.1)
View(iris.mis)
```

Output:



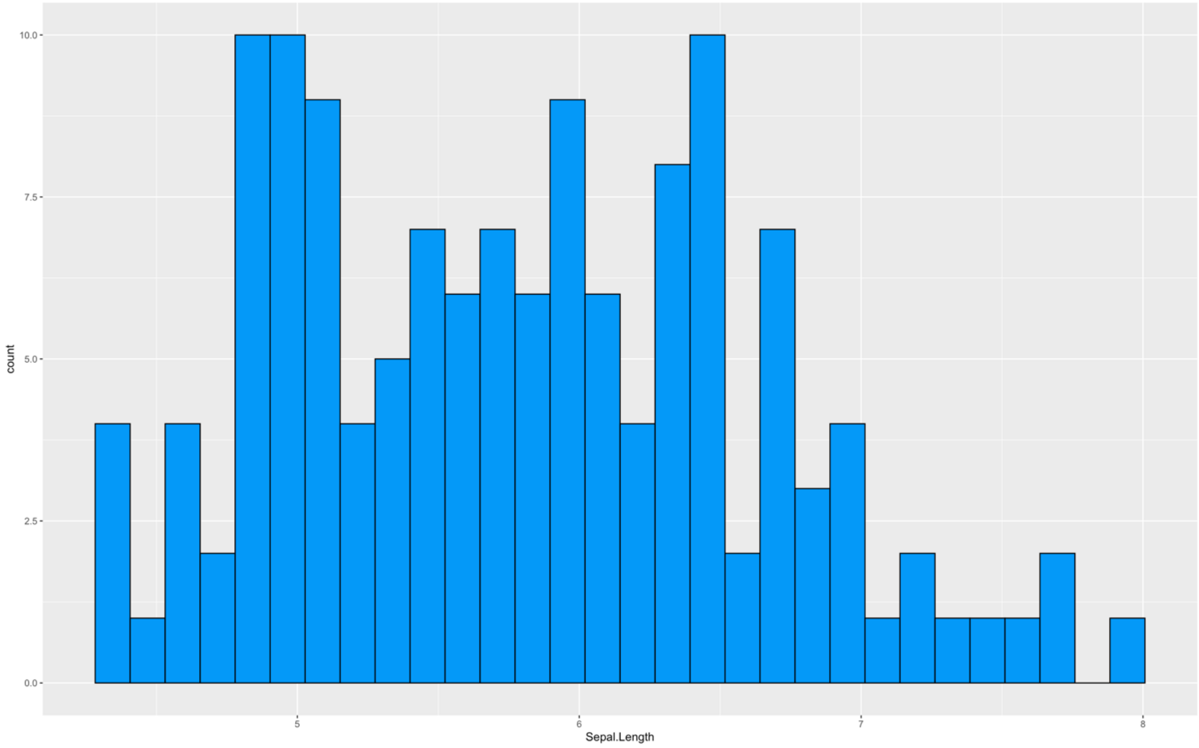
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	NA	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	NA	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3.0	1.4	0.1
14	4.3	3.0	1.1	0.1
15	5.8	4.0	1.2	0.2
16	5.7	4.4	1.5	NA
17	NA	3.9	NA	0.4
18	5.1	3.5	1.4	0.3
19	NA	3.8	1.7	0.3
20	5.1	NA	1.5	0.3
21	5.4	3.4	1.7	0.2
22	5.1	NA	1.5	0.4
23	4.6	3.6	NA	0.2
24	5.1	3.3	1.7	NA
25	4.8	3.4	1.9	0.2
26	5.0	3.0	1.6	0.2
27	5.0	3.4	1.6	0.4
28	5.2	3.5	1.5	0.2
29	5.2	3.4	1.4	0.2
30	4.7	3.2	NA	0.2
31	4.8	3.1	1.6	0.2
32	5.4	3.4	1.5	0.4
33	5.2	4.1	1.5	0.1
34	5.5	4.2	1.4	0.2
35	4.9	3.1	1.5	0.2
36	5.0	3.2	1.2	NA
37	5.5	3.5	1.3	0.2

Showing 1 to 37 of 150 entries, 4 total columns

Code:

```
ggplot(iris.mis, aes(x = Sepal.Length)) +
  geom_histogram(color="black", fill="#0099F8")
```

Output:



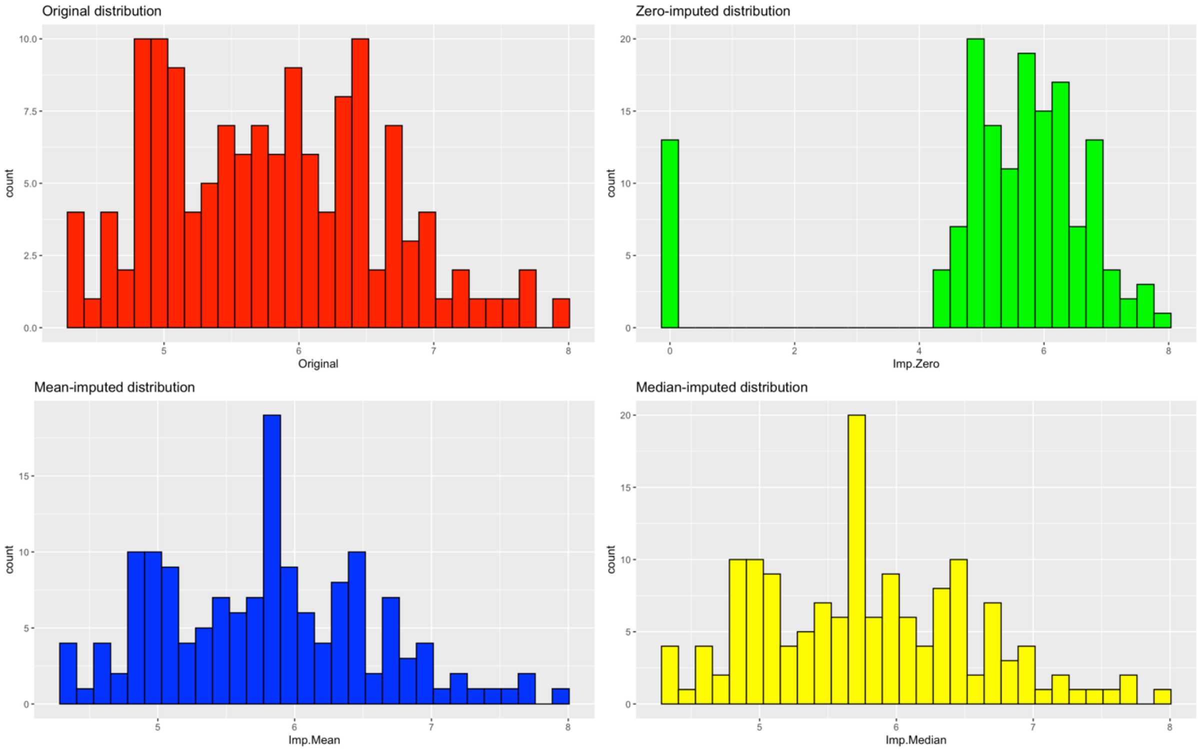
Code:

```
# simple imputations for Sepal.Length
imputed = data.frame(
  Original = iris.mis$Sepal.Length,
  Imp.Zero = replace(iris.mis$Sepal.Length, is.na(iris.mis$Sepal.Length), 0),
  Imp.Mean = replace(iris.mis$Sepal.Length, is.na(iris.mis$Sepal.Length),
mean(iris.mis$Sepal.Length, na.rm = TRUE)),
  Imp.Median = replace(iris.mis$Sepal.Length, is.na(iris.mis$Sepal.Length),
median(iris.mis$Sepal.Length, na.rm = TRUE))
)

# plotting the simple imputations
h1 = ggplot(imputed, aes(x=Original)) +
  geom_histogram(fill="red", color="black", position="identity") +
  ggtitle("Original distribution")
h2 = ggplot(imputed, aes(x=Imp.Zero)) +
  geom_histogram(fill="green", color="black", position="identity") +
  ggtitle("Zero-imputed distribution")
h3 = ggplot(imputed, aes(x=Imp.Mean)) +
  geom_histogram(fill="blue", color="black", position="identity") +
  ggtitle("Mean-imputed distribution")
h4 = ggplot(imputed, aes(x=Imp.Median)) +
  geom_histogram(fill="yellow", color="black", position="identity") +
  ggtitle("Median-imputed distribution")
```

```
plot_grid(h1, h2, h3, h4, nrow=2, ncol=2)
```

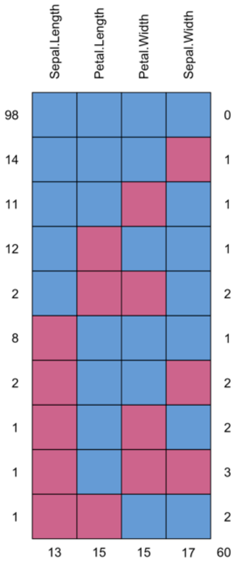
Output:



Code:

```
# viewing missing values
md.pattern(iris.mis, rotate.names=TRUE)
```

Output:



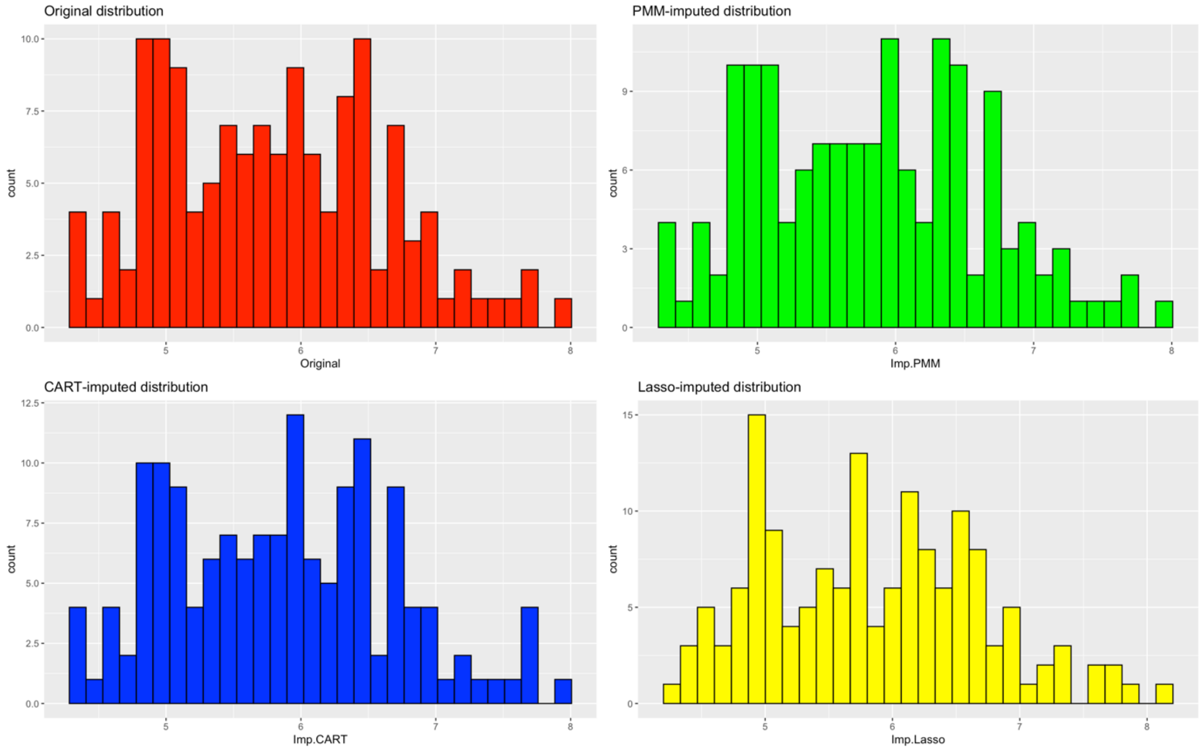
Code:

```
# performing imputations with mice algorithms
mice_imputed = data.frame(
  Original = iris.mis$Sepal.Length,
  Imp.PMM = complete(mice(iris.mis, method="pmm"))$Sepal.Length,
  Imp.CART = complete(mice(iris.mis, method="cart"))$Sepal.Length,
  Imp.Lasso = complete(mice(iris.mis, method="lasso.norm"))$Sepal.Length
)

# plotting the mice imputations
h1 = ggplot(mice_imputed, aes(x=Original)) +
  geom_histogram(fill="red", color="black", position="identity") +
  ggtitle("Original distribution")
h2 = ggplot(mice_imputed, aes(x=Imp.PMM)) +
  geom_histogram(fill="green", color="black", position="identity") +
  ggtitle("PMM-imputed distribution")
h3 = ggplot(mice_imputed, aes(x=Imp.CART)) +
  geom_histogram(fill="blue", color="black", position="identity") +
  ggtitle("CART-imputed distribution")
h4 = ggplot(mice_imputed, aes(x=Imp.Lasso)) +
  geom_histogram(fill="yellow", color="black", position="identity") +
  ggtitle("Lasso-imputed distribution")

plot_grid(h1, h2, h3, h4, nrow=2, ncol=2)
```

Output:



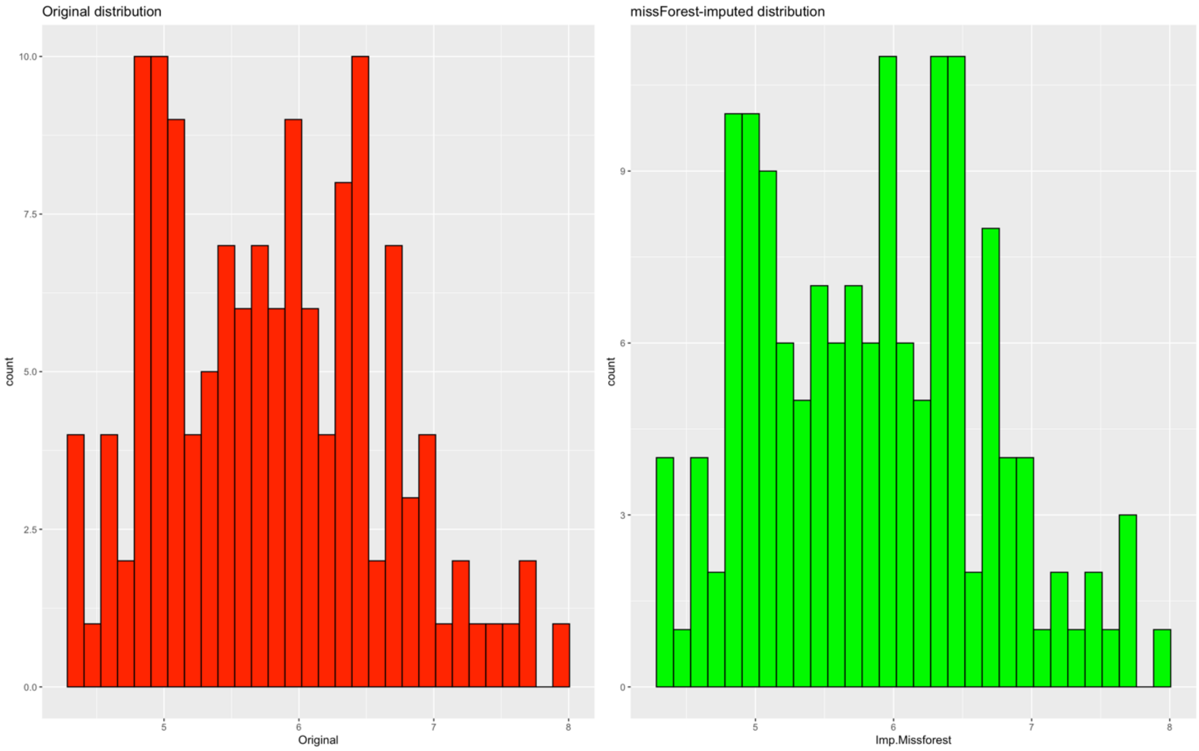
Code:

```
# imputations with missForest
missforest_imputed = data.frame(
  Original = iris.mis$Sepal.Length,
  Imp.Missforest = missForest(iris.mis)$ximp$Sepal.Length
)

# plotting the missForest imputations
h1 = ggplot(missforest_imputed, aes(x=Original)) +
  geom_histogram(fill="red", color="black", position="identity") +
  ggtitle("Original distribution")
h2 = ggplot(missforest_imputed, aes(x=Imp.Missforest)) +
  geom_histogram(fill="green", color="black", position="identity") +
  ggtitle("missForest-imputed distribution")

plot_grid(h1, h2, nrow=1, ncol=2)
```

Output:



Experiment 6

Code:

```
cov(data$Sepal.Length, data$Sepal.Width)
cor(data$Sepal.Length, data$Sepal.Width)
```

Output:

```
> cov(data$Sepal.Length, data$Sepal.Width)
[1] -0.042434
> cor(data$Sepal.Length, data$Sepal.Width)
[1] -0.1175698
```

Experiment 7

Code:

```
# z score method
data = iris$Sepal.Length

mean.data = mean(data)
std.data = sd(data)

z.scores = (data - mean.data) / std.data

# outliers have -3 < z.score < 3
outliers = data[abs(z.scores) > 3]
outliers
```

Output:

```
> # z score method
> data = iris$Sepal.Length
> mean.data = mean(data)
> std.data = sd(data)
> z.scores = (data - mean.data) / std.data
> # outliers have -3 < z.score < 3
> outliers = data[abs(z.scores) > 3]
> outliers
numeric(0)
```

Code:

```
# inter quartile range method
data = iris$Sepal.Length

q1 = quantile(data, 0.25)
q3 = quantile(data, 0.75)
iqr = q3 - q1

# outliers lie outside of the inter quartile range
outliers <- data[data < q1 | data > q3]
outliers
```

Output:

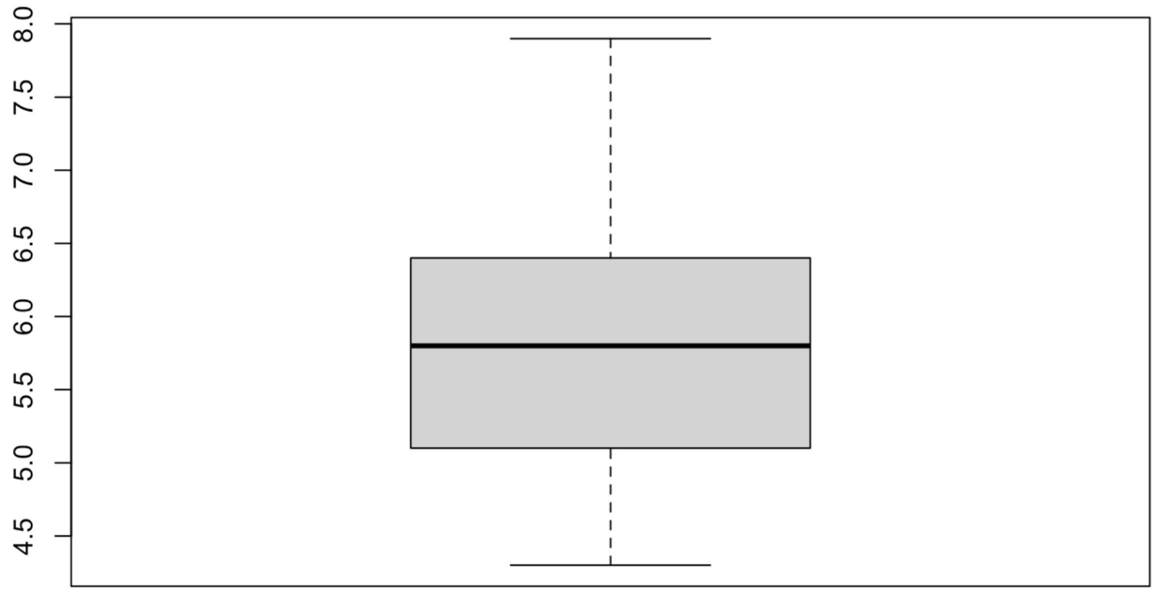
```
> # inter quartile range method
> # inter quartile range method
> data = iris$Sepal.Length
> q1 = quantile(data, 0.25)
> q3 = quantile(data, 0.75)
> iqr = q3 - q1
> # outliers lie outside of the inter quartile range
> outliers <- data[data < q1 | data > q3]
> outliers
 [1] 4.9 4.7 4.6 5.0 4.6 5.0 4.4 4.9 4.8 4.8 4.3 4.6 4.8 5.0 5.0 4.7 4.8 4.9 5.0
4.9 4.4 5.0
[23] 4.5 4.4 5.0 4.8 4.6 5.0 7.0 6.9 6.5 4.9 6.6 5.0 6.7 6.6 6.8 6.7 6.7 5.0 7.1
6.5 7.6 4.9
```

```
[45] 7.3 6.7 7.2 6.5 6.8 6.5 7.7 7.7 6.9 7.7 6.7 7.2 7.2 7.4 7.9 7.7 6.9 6.7 6.9
6.8 6.7 6.7
[67] 6.5
```

Code:

```
# boxplot method (purely visualisation)
data = iris$Sepal.Length
boxplot(data)
```

Output:


8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5