EXPERIMENT 5

Taniya Ahmed

21BDS0059

Aim: Estimation of missing data, global methods, class-based methods, multiple imputation methods.

Code:

1. Installing and loading packages.

```
data("airquality")
install.packages("ggplot2")
install.packages("missForest")
install.packages("VIM")
install.packages("Amelia")
install.packages("Hmisc")
install.packages("mi")
library(ggplot2)
library(cowplot)
library(mice)
library(missForest)
library(VIM)
library(Amelia)
library(Hmisc)
library(mi)
```

2. Number of missing data

```
df = airquality

missing_value = colSums(is.na(airquality))
missing_value
```

Output:

```
> missing_value
  Ozone Solar.R    Wind    Temp   Month     Day
     37       7       0       0       0       0
> print("Taniya Ahmed")
[1] "Taniya Ahmed"
>
```

3. Imputation of data

```
imputed_ozone = data.frame(
 original = df$Ozone,
 ozone_imputed_zero = replace(df$Ozone, is.na(df$Ozone), 0),
 ozone_imputed_mean = replace(df$Ozone, is.na(df$Ozone), mean(df$Ozone, na.rm = TRUE)),
 ozone_imputed_median = replace(df$Ozone, is.na(df$Ozone), median(df$Ozone, na.rm = TRUE))
)

imputed_ozone
print("Taniya Ahmed")
```

Output:

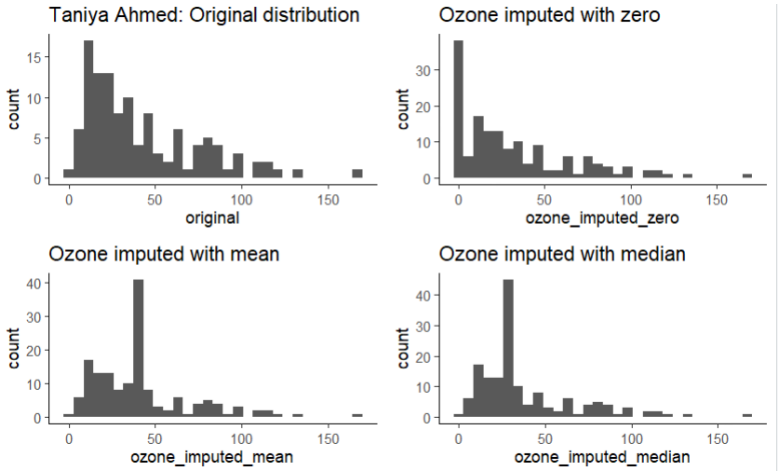
```
> head(imputed_ozone)
  original ozone_imputed_zero ozone_imputed_mean
1       41                 41            41.00000
2       36                 36            36.00000
3       12                 12            12.00000
4       18                 18            18.00000
5       NA                  0            42.12931
6       28                 28            28.00000
  ozone_imputed_median
1                 41.0
2                 36.0
3                 12.0
4                 18.0
5                 31.5
6                 28.0
> print("Taniya Ahmed")
[1] "Taniya Ahmed"
>
```

4. Checking the distribution of the imputed values

```
h1 = ggplot(imputed_ozone, aes(x = original)) +
 geom_histogram() +
 ggtitle("Taniya Ahmed: Original distribution") +
 theme_classic()

h2 = ggplot(imputed_ozone, aes(x = ozone_imputed_zero)) +
 geom_histogram() +
 ggtitle("Ozone imputed with zero") +
 theme_classic()

h3 = ggplot(imputed_ozone, aes(x = ozone_imputed_mean)) +
 geom_histogram() +
 ggtitle("Ozone imputed with mean") +
 theme_classic()

h4 = ggplot(imputed_ozone, aes(x = ozone_imputed_median)) +
```

```
 geom_histogram() +
 ggtitle("Ozone imputed with median") +
 theme_classic()

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

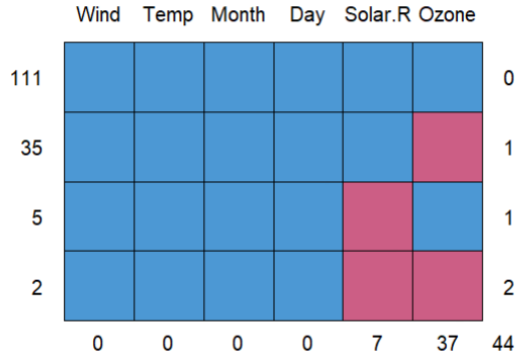
Output:



5. Visual representation of missing value

```
md.pattern(df)
```

Output:

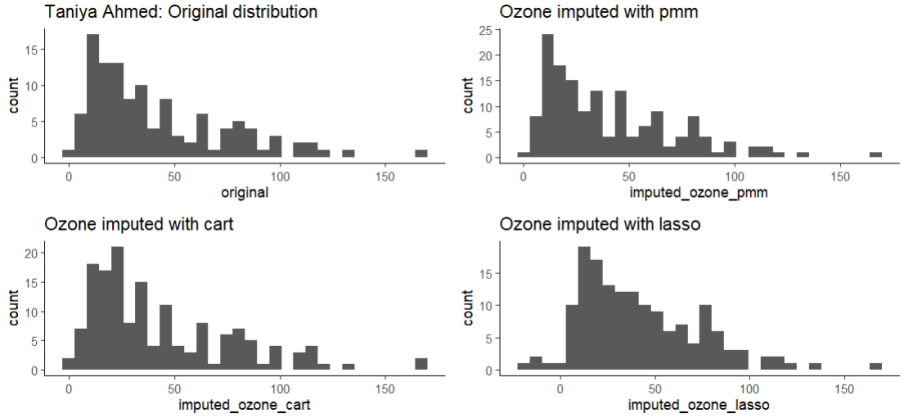


6. Mice imputation methods

```
imputed_ozone_mice = data.frame(
 original = df$Ozone,
 imputed_ozone_pmm = complete(mice(df, method = "pmm"))$Ozone,
 imputed_ozone_cart = complete(mice(df, method = "cart"))$Ozone,
 imputed_ozone_lasso = complete(mice(df, method = "lasso.norm"))$Ozone
)

imputed_ozone_mice
print("Taniya Ahmed")
```

```
h1 = ggplot(imputed_ozone_mice, aes(x = original)) +
 geom_histogram() +
 ggtitle("Original distribution") +
 theme_classic()

h2 = ggplot(imputed_ozone_mice, aes(x = imputed_ozone_pmm)) +
 geom_histogram() +
 ggtitle("Ozone imputed with pmm") +
 theme_classic()

h3 = ggplot(imputed_ozone_mice, aes(x = imputed_ozone_cart)) +
 geom_histogram() +
 ggtitle("Ozone imputed with cart") +
 theme_classic()

h4 = ggplot(imputed_ozone_mice, aes(x = imputed_ozone_lasso)) +
 geom_histogram() +
 ggtitle("Ozone imputed with lasso") +
 theme_classic()

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

Output:



7. MissForest imputation methods

```
imputed_ozone_missforest = data.frame(
 original = df$Ozone,
 imputed_ozone_missforest = missForest(df)$ximp$Ozone
)

head(imputed_ozone_missforest)
print("Taniya Ahmed")
```

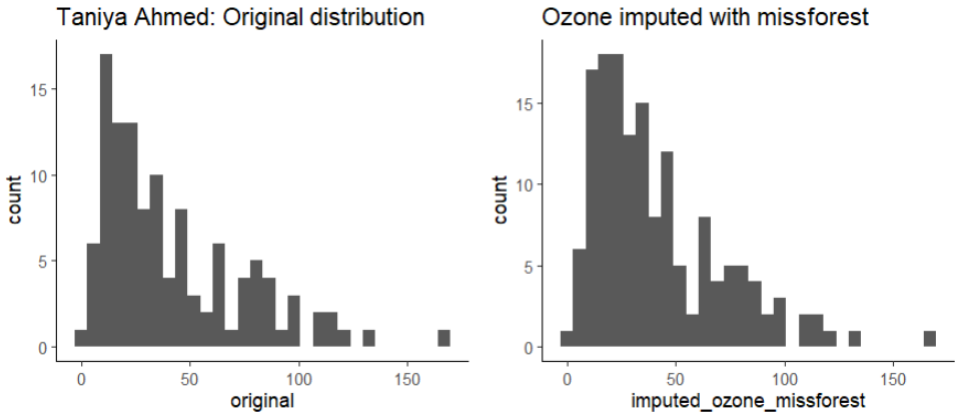
Output:

```
> head(imputed_ozone_missforest)
  original imputed_ozone_missforest
1       41                     41.00
2       36                     36.00
3       12                     12.00
4       18                     18.00
5       NA                     17.42
6       28                     28.00
> print("Taniya Ahmed")
[1] "Taniya Ahmed"
>
```

8. Visual representation of missing value

```
h1 = ggplot(imputed_ozone_missforest, aes(x = original)) +
 geom_histogram() +
 ggtitle("Taniya Ahmed: Original distribution") +
 theme_classic()

h2 = ggplot(imputed_ozone_missforest, aes(x = imputed_ozone_missforest)) +
 geom_histogram() +
 ggtitle("Ozone imputed with missforest") +
 theme_classic()

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

Output:



9. Creating missing values for practice

```
df.miss = prodNA(df, noNA = 0.1)
summary(df.miss)
print("Taniya Ahmed")
```

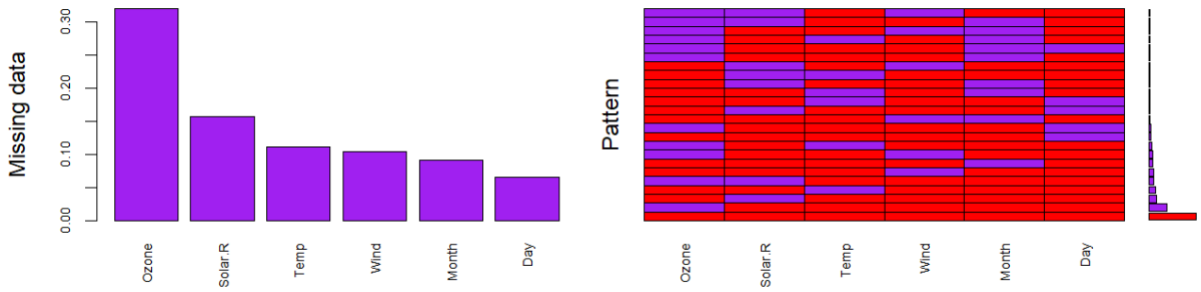
Output:

```
> df.miss = prodNA(df, noNA = 0.1)
> summary(df.miss)
     Ozone           Solar.R          Wind             Temp           Month      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000  
 1st Qu.: 18.00   1st Qu.:116.5   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000  
 Median : 33.00   Median :201.0   Median : 9.700   Median :79.00   Median :7.000  
 Mean   : 43.36   Mean   :184.2   Mean   : 9.882   Mean   :77.89   Mean   :7.036  
 3rd Qu.: 65.25   3rd Qu.:258.5   3rd Qu.:11.500   3rd Qu.:84.50   3rd Qu.:8.000  
 Max.   :168.00   Max.   :332.0   Max.   :20.700   Max.   :97.00   Max.   :9.000  
 NA's   :53       NA's   :18      NA's   :19       NA's   :18      NA's   :13     
      Day       
 Min.   : 1.00  
 1st Qu.: 8.00  
 Median :16.00  
 Mean   :15.77  
 3rd Qu.:24.00  
 Max.   :31.00  
 NA's   :12     
> print("Taniya Ahmed")
[1] "Taniya Ahmed"
> 
```

10. Mice imputation methods

```
df.miss = aggr(df.miss, col = c("red", "purple"),
               numbers = TRUE, sortVars = TRUE,
               labels = names(df.miss), cex.axis = .7,
               gap = 3, ylab = c("Missing data", "Pattern"))

imputed_df = mice(df.miss, m = 5, maxit = 50, method = "pmm", seed = 500)
summary(imputed_df)
print("Taniya Ahmed")
```

Output:



11. Hmisc imputation methods

```
df.miss$imputed_ozone_random = with(iris.miss, impute(Ozone, "random"))
df.miss$imputed_ozone_min = with(iris.miss, impute(Ozone, "min"))
df.miss$imputed_ozone_max = with(iris.miss, impute(Ozone, "max"))

impute_arg = aregImpute(~Ozone + Solar.R + Wind + Temp + Month + Day, data =
df.miss, n.impute = 5)
impute_arg
```

Output:

```
> impute_arg = aregImpute(~Ozone + Solar.R + Wind + Temp + Month + Day, data = df.miss, n.impute =
5)
Iteration 8
There were 50 or more warnings (use warnings() to see the first 50)
> head(impute_arg)
$call
aregImpute(formula = ~Ozone + Solar.R + Wind + Temp + Month +
    Day, data = df.miss, n.impute = 5)

$formula
~Ozone + Solar.R + Wind + Temp + Month + Day

$match
[1] "weighted"

$fweighted
[1] 0.2

$pmmtype
[1] 1

$constraint
NULL

> print("Taniya Ahmed")
[1] "Taniya Ahmed"
>
```

12. Mi imputation methods

df.miss = prodNA(df, noNA = 0.1)

mi_data = mi(df.miss)
mi_data
summary(mi_data)
print("Taniya Ahmed")

Output:

```
> mi_data = mi(df.miss)
> mi_data
Object of class mi with 4 chains, each with 30 iterations.
Each chain is the evolution of an object of missing_data.frame class with 153 observations on 6 variables.
> summary(mi_data)
$Ozone
$Ozone$is_missing
missing
FALSE  TRUE
  102    51

$Ozone$imputed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.84717 -0.42202 -0.09569 -0.09425  0.23080  1.26259

$Ozone$observed
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.6186 -0.3649 -0.1634  0.0000  0.3216  1.8736


$Solar.R
$Solar.R$is_missing
missing
FALSE  TRUE
  130    23
```

```
$Solar.R$imputed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.20314 -0.39134 -0.04855 -0.04371  0.33968  1.30836

$Solar.R$observed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.00909 -0.39195  0.06804  0.00000  0.40232  0.85946


$Wind
$Wind$is_missing
missing
FALSE  TRUE
  134    19

$Wind$imputed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.14328 -0.27925  0.05574  0.07461  0.39146  1.69132

$Wind$observed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.19522 -0.36515 -0.03021  0.00000  0.23192  1.57168


$Temp
$Temp$is_missing
missing
FALSE  TRUE

$Temp$imputed
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.3799 -0.5418 -0.1610 -0.1510  0.2463  0.9624

$Temp$observed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.18847 -0.32974  0.04595  0.00000  0.36797  1.01201


$Month
$Month$crosstab

    observed imputed
  5      112      10
  6      112       9
  7      108      11
  8      112       6
  9      120      12


$Day
$Day$is_missing
missing
FALSE  TRUE
  140    13

$Day$imputed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.08904 -0.36007 -0.04993 -0.02970  0.29870  1.01263

$Day$observed
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.82600 -0.42744  0.02806  0.00000  0.42663  0.88213


> print("Taniya Ahmed")
[1] "Taniya Ahmed"
>
```