Data Mining and Discovery

Assignment 2: Data Imputation

Student ID: 20070328

Introduction.

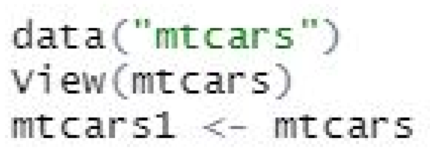
This report aims at understanding various imputation techniques on a built-in dataset of R. The dataset chosen is mtcars which has 32 obs. Of 11 Variables. The continuous columns chosen for imputation methods is ‘mpg’.

TASKS.

The Assignment is divided into five tasks.

TASK 1

Choose a dataset containing two or more continuous columns. Choose a column where we will be implementing the imputation techniques.

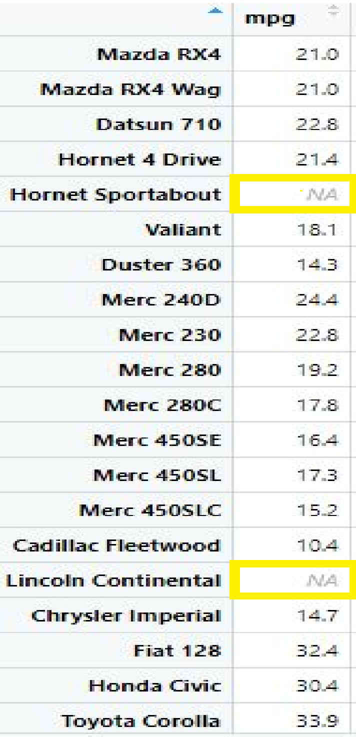


```
data("mtcars")
view(mtcars)
mtcars1 <- mtcars
```

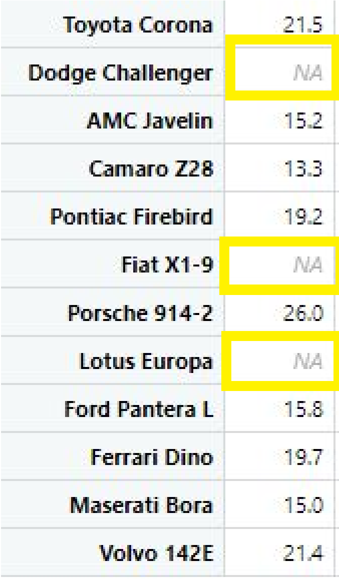
TASK 2

‘MCAR’ Mechanism implementations.

For implementing MCAR mechanism, 15% of the selected column ‘mpg’ of mtcars dataset was set to NA at random. Below screenshot shows the values converted to NA.

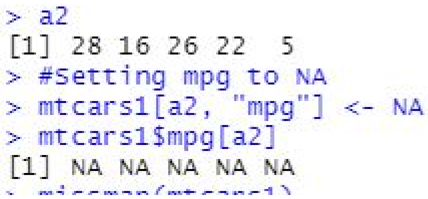


	mpg
Mazda RX4	21.0
Mazda RX4 Wag	21.0
Datsun 710	22.8
Hornet 4 Drive	21.4
Hornet Sportabout	NA
Valiant	18.1
Duster 360	14.3
Merc 240D	24.4
Merc 230	22.8
Merc 280	19.2
Merc 280C	17.8
Merc 450SE	16.4
Merc 450SL	17.3
Merc 450SLC	15.2
Cadillac Fleetwood	10.4
Lincoln Continental	NA
Chrysler Imperial	14.7
Fiat 128	32.4
Honda Civic	30.4
Toyota Corolla	33.9



Toyota Corona	21.5
Dodge Challenger	NA
AMC Javelin	15.2
Camaro Z28	13.3
Pontiac Firebird	19.2
Fiat X1-9	NA
Porsche 914-2	26.0
Lotus Europa	NA
Ford Pantera L	15.8
Ferrari Dino	19.7
Maserati Bora	15.0
Volvo 142E	21.4

The random values in mpg were changed to NA as shown below in the image.



```
> a2
[1] 28 16 26 22  5
> #Setting mpg to NA
> mtcars1[a2, "mpg"] <- NA
> mtcars1$mpg[a2]
[1] NA NA NA NA NA
> missmap(mtcars1)
```

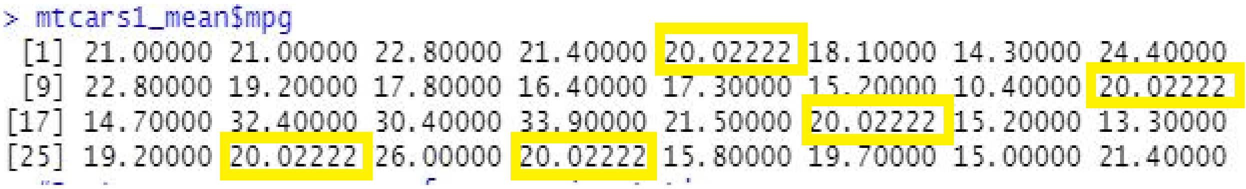
Different imputations method like (Mean imputation, kNN-Imputation, Amelia imputation) were applied to check how the values were replaced by another value which were changed to NA.

TASK 3

#APPLYING IMPUTATION METHODS ON MCAR MECHANISM.

Mean-Imputation.

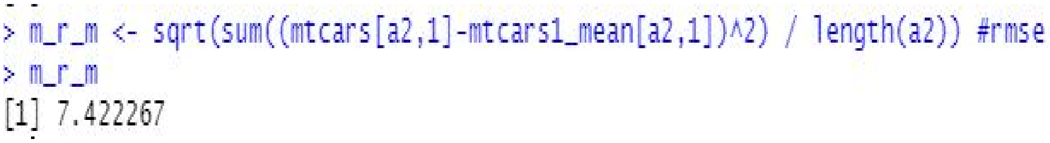
Mean imputation also known as Mean substitution, replaces the missing values of imputed values by calculating the mean of non-missing cases of that variable. For the column’s mpg, the values changed to NA were replaced with the mean of the values in this column. The NA values that were replaced by mean is 20.02222 as shown below.



```
> mtcars1_mean$mpg
 [1] 21.00000 21.00000 22.80000 21.40000 20.02222 18.10000 14.30000 24.40000
 [9] 22.80000 19.20000 17.80000 16.40000 17.30000 15.20000 10.40000 20.02222
[17] 14.70000 32.40000 30.40000 33.90000 21.50000 20.02222 15.20000 13.30000
[25] 19.20000 20.02222 26.00000 20.02222 15.80000 19.70000 15.00000 21.40000
```

To check the performance of the imputed values we use the ‘Root mean squared error’. which is calculated by the formula - root of sum of the squared differences between the predicted and observed values divided by the number of observations. The lower the RMSE value, the better is the model.

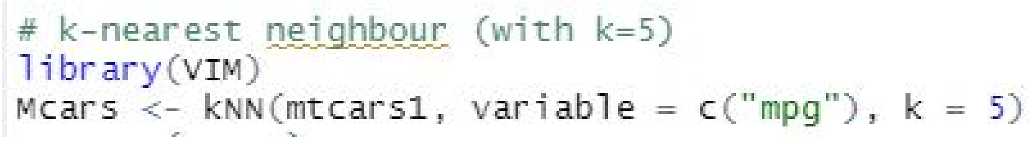
Root mean squared error (RMSE) was performed on the ‘mpg’ column after replacement of NA values and it was found that the RMSE value was higher.



```
> m_r_m <- sqrt(sum((mtcars[a2,1]-mtcars1_mean[a2,1])^2) / length(a2)) #rmse
> m_r_m
[1] 7.422267
```

Knn-imputation.

THE kNN imputation also known as k-nearest neighbors algorithm can be used for imputing missing data by finding k-closest neighbors of the missing data and then imputing them based on the non-missing values around neighbor values of the missing data. In this data kNN imputation was performed on the ‘mpg’ which had missing data. The K value was set to 5. Below is the attached code which was used.



```
# k-nearest neighbour (with k=5)
library(VIM)
Mcars <- kNN(mtcars1, variable = c("mpg"), k = 5)
```

Root mean squared error was performed on this column after replacement of NA values and it was found that the RMSE value is **2.615339**.

Amelia Imputation.

Amelia assumes that your data is distributed as multivariate normal. Amelia uses the following algorithm:

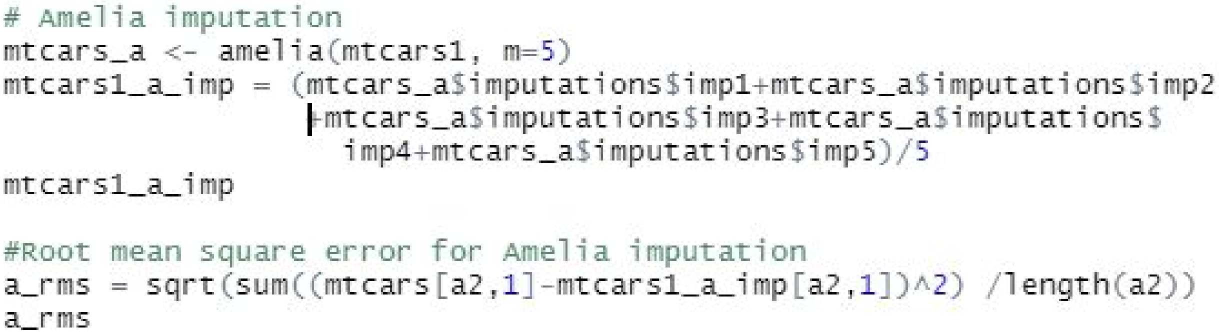
EM – expectation-maximization (Amelia I)

EM (expectation-maximization) algorithm alternates between an expectation (E-step) and maximization (M-step) steps until convergence. Convergence is reached when the current and previous values are close enough to each other (usually set by the analyst).

EMB – expectation-maximization with bootstrapping (Amelia II)

Multiple imputation involves imputing m values for each missing cell in your data matrix and creating m “completed” data sets. Across these completed data sets, the observed values are the same, but the missing values are filled in with a distribution of imputations that reflect the uncertainty about the missing data.

The code used to implement Amelia:



```
# Amelia imputation
mtcars_a <- amelia(mtcars1, m=5)
mtcars1_a_imp = (mtcars_a$imputations$imp1+mtcars_a$imputations$imp2
                 +mtcars_a$imputations$imp3+mtcars_a$imputations$
                   imp4+mtcars_a$imputations$imp5)/5
mtcars1_a_imp

#Root mean square error for Amelia imputation
a_rms = sqrt(sum((mtcars[a2,1]-mtcars1_a_imp[a2,1])^2) /length(a2))
a_rms
```

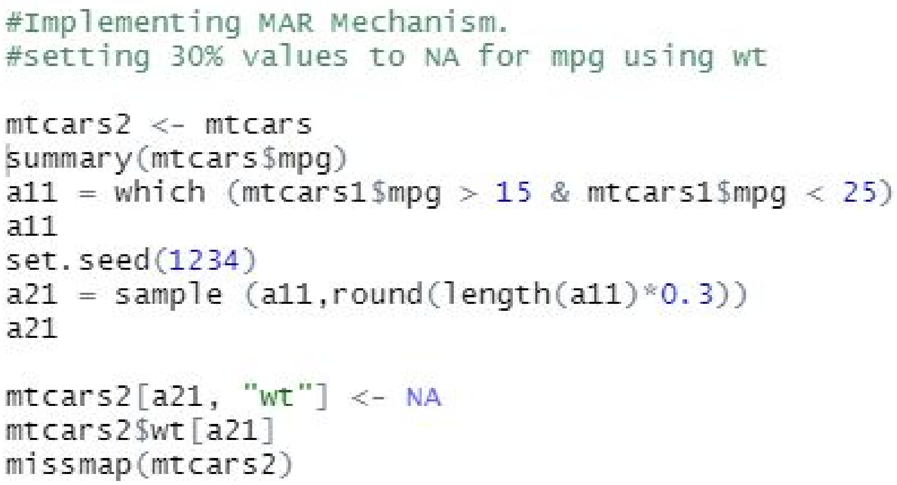
Overall, the RMSE value was the least for Amelia imputation stating that its performance of calculating/imputing the missing data is more accurate. The mean method is least reliable, having the highest RMSE value.

TASK 4

Implementing MAR mechanism.

For implementing MAR mechanism, a range of values (15 to 25) for mpg column from the dataset mtcars were picked, then randomly 30% of values from this range were chosen randomly from mpg and were set to NA.

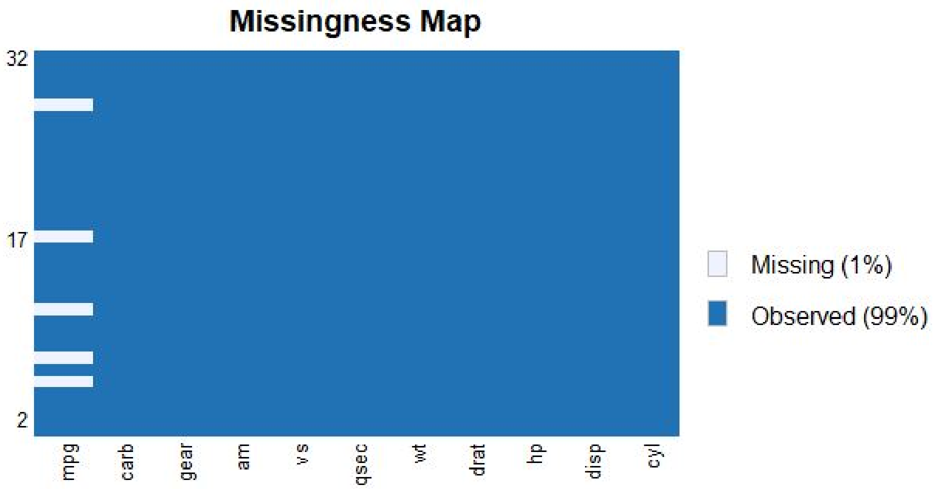
CODE FOR IMPLEMENTING MAR IS SHOWN BELOW.



```
#Implementing MAR Mechanism.
#setting 30% values to NA for mpg using wt

mtcars2 <- mtcars
summary(mtcars$mpg)
a11 = which (mtcars1$mpg > 15 & mtcars1$mpg < 25)
a11
set.seed(1234)
a21 = sample (a11,round(length(a11)*0.3))
a21

mtcars2[a21, "wt"] <- NA
mtcars2$wt[a21]
missmap(mtcars2)
```



TASK 5

#APPLYING IMPUTATION METHODS ON MAR DATASET

Mean Imputation.

Mean imputation on the ‘mpg’ column having missing data. The NA values were replaced by the mean of the non-misSING values of this column.

Root mean squared error (RMSE) was performed on this column after replacement of NA values and it was found that the RMSE value was **0.4540915**.

Knn Imputation.

Knn imputation was performed on MAR mechanism with RMSE after replacement of NA values and it was found that RMSE value was **0.2813006**.

Amelia imputation.

Amelia imputation performed on MAR mechanism with RMSE after replacement of NA values and it was found that RMSE value was **0.4011719**.

Conclusion:

The missing data imputation using various imputation methods were evaluated on their performance was evaluated using the RMSE method.

References:

impute.knn function - RDocumentation.

Amelia https://gking.harvard.edu/amelia