# AMERICAN INTERNATIONAL UNIVERSITY - BANGLADESH

## Faculty of Science and Information Technology

# Assignment Cover Sheet

| | | | |
|---|---|---|---|
| Assign./Case Title: | MID TERM ASSIGNMENT | | |
| Assign./Case No: | 01 | Date of Submission: | 18 March 2024 |
| Course Title: | INTRODUCTION TO DATA SCIENCE | | |
| Course Code: | CSC 4180 | Section: | C |
| Semester: | Spring 2023-24 | Degree Program: | BSc [CSE] |
| Course Teacher: | TOHEDUL ISLAM | | |

**Declaration and Statement of Authorship:**

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaboration has been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand that Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to expulsion from the University. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.: 15

| No | Name | ID | Signature |
|---|---|---|---|
| 1 | IRTIZA AHSAN ABIR | 21-45009-2 | |
| 2 | AHNAF ABDULLAH ZAYAD | 21-45019-2 | |
| 3 | | | |
| 4 | | | |
| 5 | | | |
| 6 | | | |

| *Faculty use only* | | |
|---|---|---|
| FACULTY COMMENTS | **Marks Obtained** | |
| | | |
| | **Total Marks** | |
| | | |

**Short Note on the Maternal Health Risk Dataset :** The Maternal Health Risk dataset compiles vital health indicators from rural Bangladesh, focusing on maternal well-being. It encompasses 1013 instances, each comprising 6 key features: Age, Systolic and Diastolic Blood Pressure, Blood Sugar, Body Temperature, Heart Rate, and Risk Level assessment. These metrics play a pivotal role in identifying and predicting maternal health risks, directly contributing to the United Nations' Sustainable Development Goals, particularly in reducing maternal mortality rates. With no missing values, the dataset provides a clean and structured foundation for data analysis and machine learning tasks. Its primary objective is to accurately predict Risk Level, aiding in early intervention and effective healthcare management. Released under the Creative Commons Attribution 4.0 International license, it facilitates widespread use for academic research and development purposes. This dataset exemplifies the transformative potential of data-driven insights in enhancing maternal health outcomes, especially in resource-constrained settings.

1. The read.csv function reads a CSV file into R as a dataframe, specifying that the first row contains column headers (header=TRUE) and columns are separated by commas (sep=",").

```
dataset <- read.csv("D:/9th Semester/Mid/Data/project.csv", header=TRUE, sep=",")
```

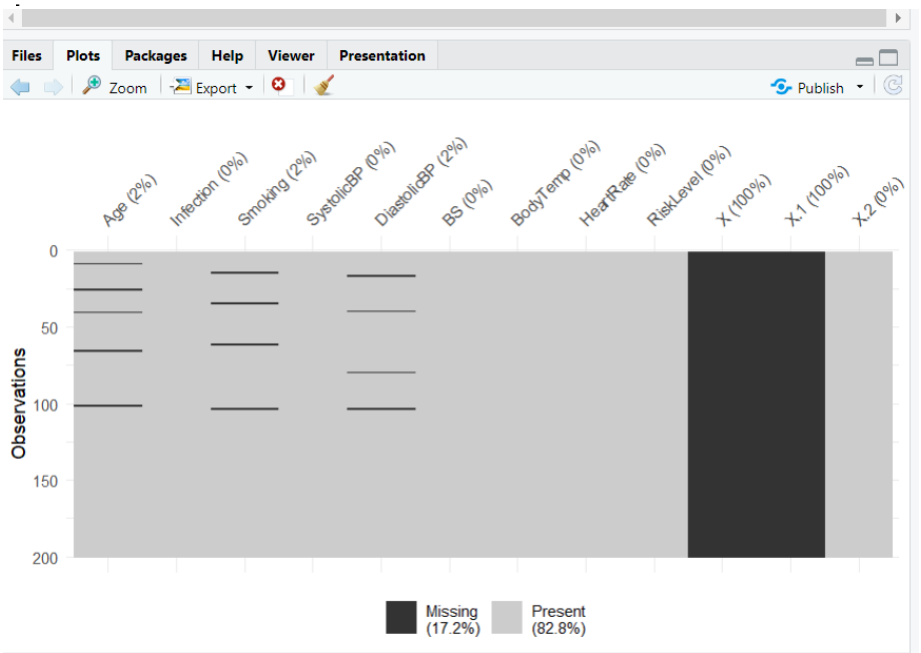
Filter

| | Age | Infection | Smoking | SystolicBP | DiastolicBP | BS | BodyTemp | HeartRate | RiskLevel | X | X.1 | X.2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 25 | yes | 1 | 130 | 80 | 15.00 | 98 | 86 | high risk | NA | NA | |
| 2 | 35 | yes | 1 | 140 | 90 | 13.00 | 98 | 70 | high risk | NA | NA | Sn |
| 3 | 29 | yes | 1 | 90 | 70 | 8.00 | 100 | 80 | high risk | NA | NA | 1= |
| 4 | 30 | yes | 1 | 140 | 85 | 7.00 | 98 | 70 | high risk | NA | NA | 2= |
| 5 | 35 | no | 3 | 120 | 60 | 6.10 | 98 | 76 | low risk | NA | NA | 3= |
| 6 | 23 | yes | 1 | 140 | 80 | 7.01 | 98 | 70 | high risk | NA | NA | |
| 7 | 23 | | 2 | 130 | 70 | 7.01 | 98 | 78 | mid risk | NA | NA | |
| 8 | NA | yes | 1 | 85 | 60 | 11.00 | 102 | 86 | high risk | NA | NA | |
| 9 | 32 | marginal | 2 | 120 | 90 | 6.90 | 98 | 70 | mid risk | NA | NA | |
| 10 | 42 | yes | 1 | 130 | 80 | 18.00 | 98 | 70 | high risk | NA | NA | |
| 11 | 23 | no | 3 | 90 | 60 | 7.01 | 98 | 76 | low risk | NA | NA | |
| 12 | 19 | marginal | 2 | 120 | 80 | 7.00 | 98 | 70 | mid risk | NA | NA | |
| 13 | 25 | no | 3 | 110 | 89 | 7.01 | 98 | 77 | low risk | NA | NA | |

2. The vis_miss function from the naniar library visualizes missing values in the dataset, helping to identify patterns or clusters of missing data.

```
if (!require("naniar")) install.packages("naniar")
library(naniar)

# Visualizing missing values
vis_miss(dataset)
```



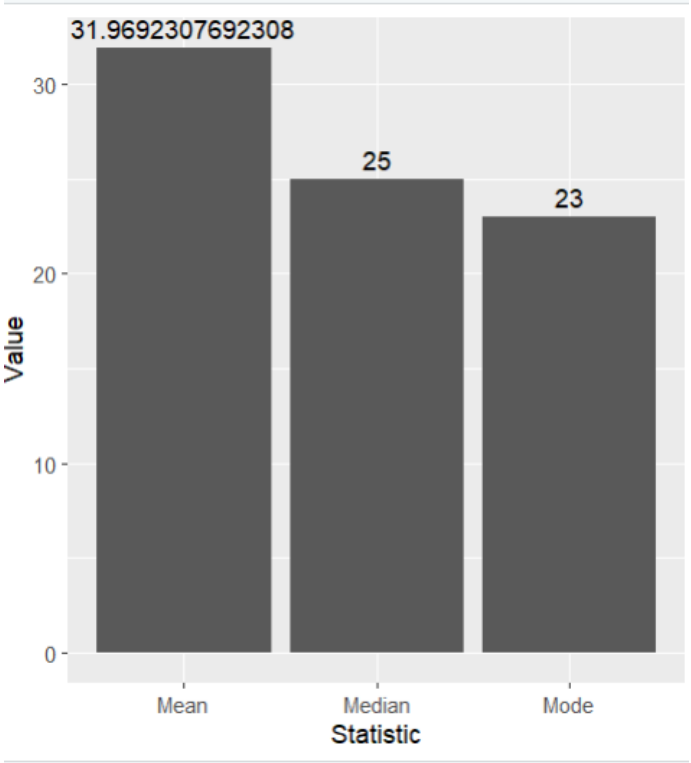
3. This code snippet creates a summary_stats data frame summarizing the Age column from a dataset. It calculates the mean (average age), median (middle age value), and mode (most frequently occurring age) while ignoring missing values. The results are stored in a data frame with two columns: one for the type of statistic and one for its value, providing a quick overview of the age distribution's central tendencies.

```
# Assuming 'Age' is the column
summary_stats <- data.frame(
  Statistic = c("Mean", "Median", "Mode"),
  Value = c(mean(dataset$Age, na.rm = TRUE), median(dataset$Age, na.rm = TRUE), as.numeric(names(sort(-table(dataset$Age)))[1]))
)
```

| | Statistic | Value |
|---|---|---|
| 1 | Mean | 31.96923 |
| 2 | Median | 25.00000 |
| 3 | Mode | 23.00000 |

4. This code checks if the ggplot2 package is installed and installs it if not. Then, it uses ggplot2 to create a bar chart from the summary_stats data frame, displaying Statistic on the x-axis and Value on the y-axis. Additionally, it adds text labels above each bar to show the numerical value, providing a visual comparison of mean, median, and mode for the Age column.

```
if (!require("ggplot2")) install.packages("ggplot2")
library(ggplot2)
ggplot(summary_stats, aes(x = Statistic, y = Value)) + geom_col() + geom_text(aes(label = Value), vjust = -0.5)
```



5. This line of code adds a new column named "AgeCategory" to the dataset, based on the values in the existing "Age" column. It categorizes individuals into three age groups:

   "Young" for ages below 18,

   "Middle-aged" for ages between 18 and 65 (inclusive),

   "Senior" for ages above 65.

   This categorization is achieved using the cut() function, which segments the numeric values in the "Age" column into discrete intervals defined by the breaks spec

```
dataset$AgeCategory <- cut(dataset$Age, breaks = c(-Inf, 18, 65, Inf), labels = c("Young", "Middle-aged", "Senior"))
```

| AgeCategory |
|---|
| Middle-aged |
| Young |
| Young |
| Middle-aged |
| Young |
| Young |
| Young |
| Young |
| Young |
| Middle-aged |
| Young |
| Senior |
| Young |
| Middle-aged |

6. This line normalizes the "SystolicBP" column values within the range [0, 1] and stores them in a new column called "SystolicBP_norm".

```
dataset$SystolicBP_norm <- (dataset$SystolicBP - min(dataset$SystolicBP, na.rm = TRUE)) / (max(dataset$SystolicBP, na.rm = TRUE) - min(dataset$SystolicBP, na.rm = TRUE))
```

| SystolicBP_norm |
|---|
| 0.66666667 |
| 0.77777778 |
| 0.22222222 |
| 0.77777778 |
| 0.55555556 |
| 0.77777778 |
| 0.66666667 |
| 0.16666667 |
| 0.55555556 |
| 0.66666667 |
| 0.22222222 |
| 0.55555556 |
| 0.44444444 |
| 0.55555556 |

7.This line shows the missing values in the row

```
missing_age_rows <- which(is.na(dataset$Age))
print(missing_age_rows)
```

```
> print(missing_age_rows)
[1]   8  25  40  65 101
```

8.This line converts categorical value into numerical value

```
dataset$Infection<-factor(dataset$Infection,levels=c("yes","no","marginal"),labels=c(1,2,3))
```

| Infection |
|---|
| 1 |
| 1 |
| 1 |
| 1 |
| 2 |
| 1 |
| NA |
| 1 |
| 3 |
| 1 |
| 2 |
| 3 |
| 2 |
| 3 |
| 3 |

9. This line handles missing value with replacing NA values with median for age

```
dataset$Age<-as.inteager(as.character(dataset$Age))
Age_median<-round(median(dataset$Age,na.rm=TRUE))
dataset$Age[is.na(dataset$Age)]<-Age_median
```

| Age |
|---|
| 25 |
| 35 |
| 29 |
| 30 |
| 35 |
| 23 |
| 23 |
| 25 |
| 32 |
| 42 |
| 23 |

10. These line of codes handle invalid data from infection. It replaces invalid value with NA

```
invalid_values <- !(dataset$Infection %in% c("yes", "no", "marginal"))
dataset$Infection[invalid_values] <- NA
```

| Infection |
|---|
| yes |
| yes |
| yes |
| yes |
| no |
| yes |
| *NA* |
| yes |
| marginal |

11. These line of codes detects outliers

```
Q1 <- quantile(dataset$Age, 0.25, na.rm = TRUE)
Q3 <- quantile(dataset$Age, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR
outliers <- dataset$Age[which(dataset$Age < lower_bound | dataset$Age > upper_bound)]
print(outliers)
```

```
> print(outliers)
[1] 148 161 170
```

12.These line of codes handles outliers

```
Q1 <- quantile(dataset$Age, 0.25, na.rm = TRUE)
Q3 <- quantile(dataset$Age, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
dataset <- dataset[dataset$Age >= (Q1 - 1.5 * IQR) & dataset$Age <= (Q3 + 1.5 * IQR), ]
```

```
> print(outliers)
integer(0)
```

13. This line of codes omits the dataset with values NA

```
cleaned_dataset <- na.omit(dataset)
```

| Age | Infection | Smoking | SystolicBP | DiastolicBP | BS | BodyTemp | HeartRate | RiskLevel |
|---|---|---|---|---|---|---|---|---|
| 25 | yes | 1 | 130 | 80 | 15.00 | 98 | 86 | high risk |
| 35 | yes | 1 | 140 | 90 | 13.00 | 98 | 70 | high risk |
| 29 | yes | 1 | 90 | 70 | 8.00 | 100 | 80 | high risk |
| 30 | yes | 1 | 140 | 85 | 7.00 | 98 | 70 | high risk |
| 35 | no | 3 | 120 | 60 | 6.10 | 98 | 76 | low risk |
| 23 | yes | 1 | 140 | 80 | 7.01 | 98 | 70 | high risk |
| 32 | marginal | 2 | 120 | 90 | 6.90 | 98 | 70 | mid risk |
| 42 | yes | 1 | 130 | 80 | 18.00 | 98 | 70 | high risk |
| 23 | no | 3 | 90 | 60 | 7.01 | 98 | 76 | low risk |