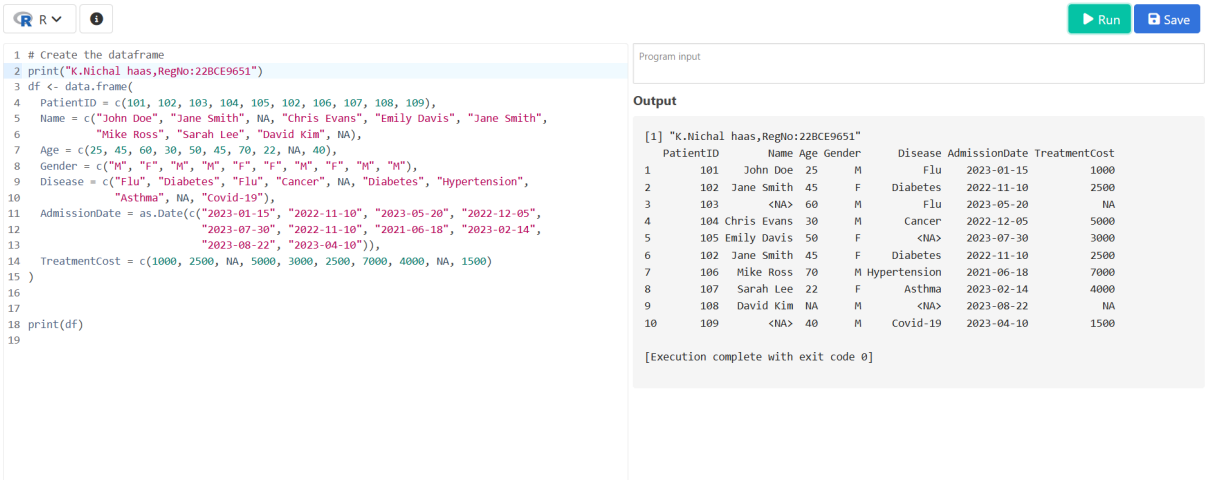
Name : K.Nichal haas
RegNo: 22BCE9651
Slot : **L22+L23**
Prof :Posham Uppamma

Vellore Institute of Technology
SCOPE
FDA (CSE1006) - Slot - L22+L23

DA -II

A)Create the DataFrame df



Code:

```
# Create the dataframe
print("K.Nichal haas,RegNo:22BCE9651")
df <- data.frame(
  PatientID = c(101, 102, 103, 104, 105, 102, 106, 107, 108, 109),
  Name = c("John Doe", "Jane Smith", NA, "Chris Evans", "Emily Davis", "Jane Smith",
           "Mike Ross", "Sarah Lee", "David Kim", NA),
  Age = c(25, 45, 60, 30, 50, 45, 70, 22, NA, 40),
  Gender = c("M", "F", "M", "M", "F", "F", "M", "F", "M", "M"),
  Disease = c("Flu", "Diabetes", "Flu", "Cancer", NA, "Diabetes", "Hypertension",
              "Asthma", NA, "Covid-19"),
  AdmissionDate = as.Date(c("2023-01-15", "2022-11-10", "2023-05-20", "2022-12-05",
                            "2023-07-30", "2022-11-10", "2021-06-18", "2023-02-14",
                            "2023-08-22", "2023-04-10")),
  TreatmentCost = c(1000, 2500, NA, 5000, 3000, 2500, 7000, 4000, NA, 1500)
)


print(df)
```

B)Sort the dataset by AdmissionDate (oldest to newest).

```r
# Create the dataframe
print("K.Nichal haas,RegNo:22BCE9651")
df <- data.frame(
  PatientID = c(101, 102, 103, 104, 105, 102, 106, 107, 108, 109),
  Name = c("John Doe", "Jane Smith", NA, "Chris Evans", "Emily Davis", "Jane Smith",
           "Mike Ross", "Sarah Lee", "David Kim", NA),
  Age = c(25, 45, 60, 30, 50, 45, 70, 22, NA, 40),
  Gender = c("M", "F", "M", "M", "F", "F", "M", "F", "M", "M"),
  Disease = c("Flu", "Diabetes", "Flu", "Cancer", NA, "Diabetes", "Hypertension",
              "Asthma", NA, "Covid-19"),
  AdmissionDate = as.Date(c("2023-01-15", "2022-11-10", "2023-05-20", "2022-12-05",
                            "2023-07-30", "2022-11-10", "2021-06-18", "2023-02-14",
                            "2023-08-22", "2023-04-10")),
  TreatmentCost = c(1000, 2500, NA, 5000, 3000, 2500, 7000, 4000, NA, 1500)
)


print(df)

df_sorted<-df[order(df$AdmissionDate),]
df_sorted
```

Program input

Output

```
 8        107   Sarah Lee  22      F       Asthma    2023-02-14          4000
 9        108   David Kim  NA      M         <NA>    2023-08-22            NA
 10       109        <NA>  40      M     Covid-19    2023-04-10          1500
    PatientID        Name Age Gender      Disease AdmissionDate TreatmentCost
 7        106   Mike Ross  70      M Hypertension    2021-06-18          7000
 2        102  Jane Smith  45      F     Diabetes    2022-11-10          2500
 6        102  Jane Smith  45      F     Diabetes    2022-11-10          2500
 4        104 Chris Evans  30      M       Cancer    2022-12-05          5000
 1        101    John Doe  25      M          Flu    2023-01-15          1000
 8        107   Sarah Lee  22      F       Asthma    2023-02-14          4000
 10       109        <NA>  40      M     Covid-19    2023-04-10          1500
 3        103        <NA>  60      M          Flu    2023-05-20            NA
 5        105 Emily Davis  50      F         <NA>    2023-07-30          3000
 9        108   David Kim  NA      M         <NA>    2023-08-22            NA

[Execution complete with exit code 0]
```

Code:

df_sorted<-df[order(df$AdmissionDate),]
df_sorted

C)Find and remove duplicate patient records based on PatientID.

```r
# Create the dataframe
print("K.Nichal haas,RegNo:22BCE9651")
df <- data.frame(
  PatientID = c(101, 102, 103, 104, 105, 102, 106, 107, 108, 109),
  Name = c("John Doe", "Jane Smith", NA, "Chris Evans", "Emily Davis", "Jane Smith",
           "Mike Ross", "Sarah Lee", "David Kim", NA),
  Age = c(25, 45, 60, 30, 50, 45, 70, 22, NA, 40),
  Gender = c("M", "F", "M", "M", "F", "F", "M", "F", "M", "M"),
  Disease = c("Flu", "Diabetes", "Flu", "Cancer", NA, "Diabetes", "Hypertension",
              "Asthma", NA, "Covid-19"),
  AdmissionDate = as.Date(c("2023-01-15", "2022-11-10", "2023-05-20", "2022-12-05",
                            "2023-07-30", "2022-11-10", "2021-06-18", "2023-02-14",
                            "2023-08-22", "2023-04-10")),
  TreatmentCost = c(1000, 2500, NA, 5000, 3000, 2500, 7000, 4000, NA, 1500)
)


print(df)

df_sorted<-df[order(df$AdmissionDate),]
df_sorted

df_unique<-df[!duplicated(df$PatientID),]
df_unique

```

Program input

Output

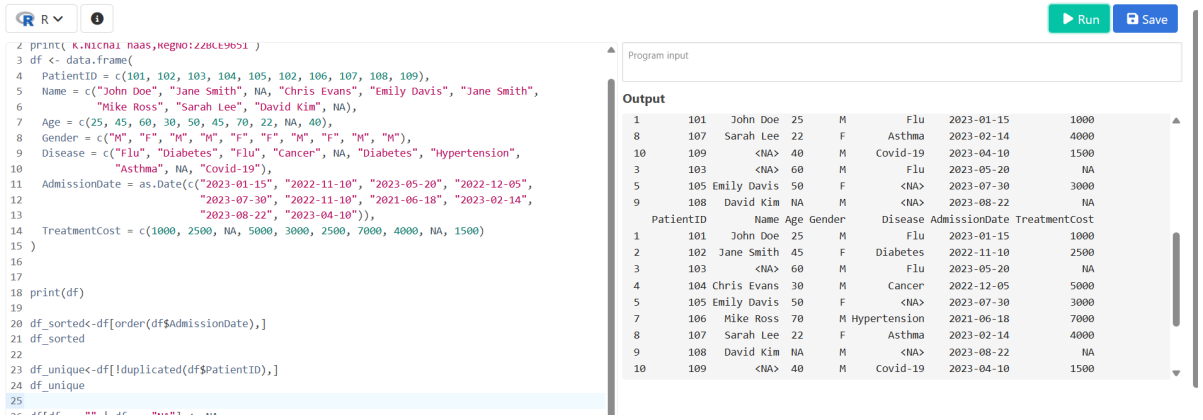
```
 3        103        <NA>  60      M          Flu    2023-05-20            NA
 5        105 Emily Davis  50      F         <NA>    2023-07-30          3000
 9        108   David Kim  NA      M         <NA>    2023-08-22            NA
    PatientID        Name Age Gender      Disease AdmissionDate TreatmentCost
 1        101    John Doe  25      M          Flu    2023-01-15          1000
 2        102  Jane Smith  45      F     Diabetes    2022-11-10          2500
 3        103        <NA>  60      M          Flu    2023-05-20            NA
 4        104 Chris Evans  30      M       Cancer    2022-12-05          5000
 5        105 Emily Davis  50      F         <NA>    2023-07-30          3000
 7        106   Mike Ross  70      M Hypertension    2021-06-18          7000
 8        107   Sarah Lee  22      F       Asthma    2023-02-14          4000
 9        108   David Kim  NA      M         <NA>    2023-08-22            NA
 10       109        <NA>  40      M     Covid-19    2023-04-10          1500

[Execution complete with exit code 0]
```

Code:

df_unique<-df[!duplicated(df$PatientID),]
df_unique

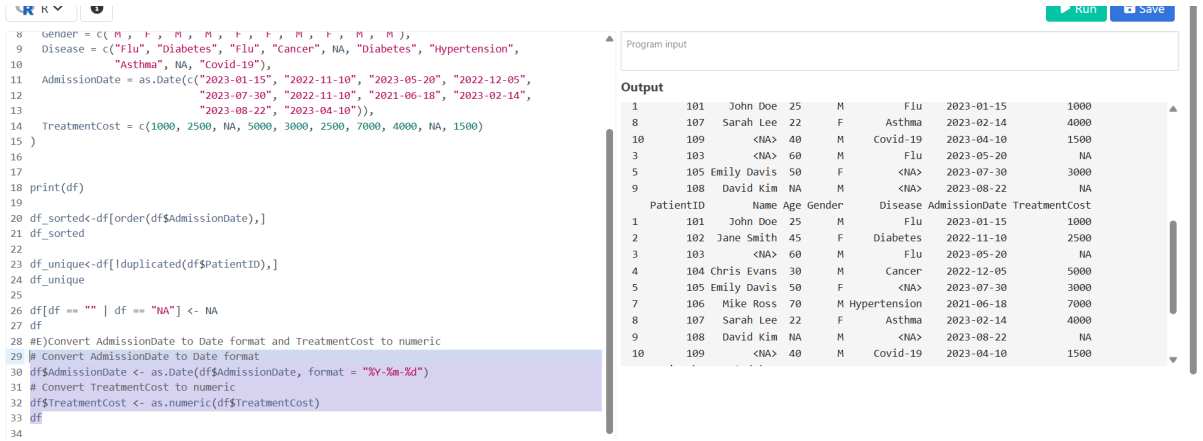
D)Replace empty strings and "NA" as text with actual NA values.



Code:

df[df==" "|df=="NA"]<-NA
df

E)Convert AdmissionDate to Date format and TreatmentCost to numeric



Code:

Convert AdmissionDate to Date format
df$AdmissionDate <- as.Date(df$AdmissionDate, format = "%Y-%m-%d")
Convert TreatmentCost to numeric
df$TreatmentCost <- as.numeric(df$TreatmentCost)
df

F)Convert Age column to categorize patients into Young (0-30), Middle- aged (31-60), and Senior (61+).

```
df_unique


#E)Convert AdmissionDate to Date format and TreatmentCost to numeric




# Convert TreatmentCost to numeric
df$TreatmentCost <- as.numeric(df$TreatmentCost)

df



df$AgeGroup <- cut(
  df$Age,
  breaks = c(-Inf, 30, 60, Inf),
  labels = c("Young", "Middle-aged", "Senior")
)


print(df)
```

Program input

Output

```
9        108   David Kim  NA      M          <NA>    2023-08-22           NA
10       109        <NA>  40      M     Covid-19     2023-04-10         1500
      AgeGroup
1        Young
2  Middle-aged
3  Middle-aged
4        Young
5  Middle-aged
6  Middle-aged
7       Senior
8        Young
9         <NA>
10 Middle-aged

[Execution complete with exit code 0]
```

df$AgeGroup<-cut(
df$Age,
break=c(-Inf,30,60,Inf),
Label=c(“Young”,”Middle-aged”,”Senior”)
)