

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

R

1 # Create the dataframe
2 print("K.Nichal haas,RegNo:22BCE9651")
3 df <- data.frame(
4 PatientID = c(101, 102, 103, 104, 105, 102, 106, 107, 108, 109),
5 Name = c("John Doe", "Jane Smith", NA, "Chris Evans", "Emily Davis", "Jane Smith",
6 "Mike Ross", "Sarah Lee", "David Kim", NA),
7 Age = c(25, 45, 60, 30, 50, 45, 70, 22, NA, 40),
8 Gender = c("M", "F", "M", "M", "F", "F", "M", "F", "M", "M"),
9 Disease = c("Flu", "Diabetes", "Flu", "Cancer", NA, "Diabetes", "Hypertension",
10 "Asthma", NA, "Covid-19"),
11 AdmissionDate = as.Date(c("2023-01-15", "2022-11-10", "2023-05-20", "2022-12-05",
12 "2023-07-30", "2022-11-10", "2021-06-18", "2023-02-14",
13 "2023-08-22", "2023-04-10")),
14 TreatmentCost = c(1000, 2500, NA, 5000, 3000, 2500, 7000, 4000, NA, 1500)
15)
16
17
18 print(df)
19

Program input

Output

```
[1] "K.Nichal haas,RegNo:22BCE9651"
```

	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
6	102	Jane Smith	45	F	Diabetes	2022-11-10	2500
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:

```
# 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).

```
# 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

	PatientID	Name	Age	Gender	Disease	AdmissionDate	TreatmentCost
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
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.

```
1 # Create the dataframe
2 print("K.Nichal haas,RegNo:22BCE9651")
3 df <- data.frame(
4   PatientID = c(101, 102, 103, 104, 105, 102, 106, 107, 108, 109),
5   Name = c("John Doe", "Jane Smith", NA, "Chris Evans", "Emily Davis", "Jane Smith",
6     "Mike Ross", "Sarah Lee", "David Kim", NA),
7   Age = c(25, 45, 60, 30, 50, 45, 70, 22, NA, 40),
8   Gender = c("M", "F", "M", "M", "F", "F", "M", "F", "M", "M"),
9   Disease = c("Flu", "Diabetes", "Flu", "Cancer", NA, "Diabetes", "Hypertension",
10     "Asthma", NA, "Covid-19"),
11   AdmissionDate = as.Date(c("2023-01-15", "2022-11-10", "2023-05-20", "2022-12-05",
12     "2023-07-30", "2022-11-10", "2021-06-18", "2023-02-14",
13     "2023-08-22", "2023-04-10")),
14   TreatmentCost = c(1000, 2500, NA, 5000, 3000, 2500, 7000, 4000, NA, 1500)
15 )
16
17
18 print(df)
19
20 df_sorted<-df[order(df$AdmissionDate),]
21 df_sorted
22
23 df_unique<-df[!duplicated(df$PatientID),]
24 df_unique
25
```

Program input

Output


	PatientID	Name	Age	Gender	Disease	AdmissionDate	TreatmentCost
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
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.

 R



```
2 print( k.nicnai naas,kegno:220LE901 )
3 df <- data.frame(
4   PatientID = c(101, 102, 103, 104, 105, 102, 106, 107, 108, 109),
5   Name = c("John Doe", "Jane Smith", NA, "Chris Evans", "Emily Davis", "Jane Smith",
6   "Mike Ross", "Sarah Lee", "David Kim", NA),
7   Age = c(25, 45, 60, 30, 50, 45, 70, 22, NA, 40),
8   Gender = c("M", "F", "M", "M", "F", "F", "M", "F", "M", "M"),
9   Disease = c("Flu", "Diabetes", "Flu", "Cancer", NA, "Diabetes", "Hypertension",
10  "Asthma", NA, "Covid-19"),
11  AdmissionDate = as.Date(c("2023-01-15", "2022-11-10", "2023-05-20", "2022-12-05",
12  "2023-07-30", "2022-11-10", "2021-06-18", "2023-02-14",
13  "2023-08-22", "2023-04-10")),
14  TreatmentCost = c(1000, 2500, NA, 5000, 3000, 2500, 7000, 4000, NA, 1500)
15 )
16
17
18 print(df)
19
20 df_sorted<-df[order(df$AdmissionDate),]
21 df_sorted
22
23 df_unique<-df[!duplicated(df$PatientID),]
24 df_unique
25
```

Program input


Output


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
	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

Code:

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

E) Convert AdmissionDate to Date format and TreatmentCost to numeric

 R



```
8 Gender = c("M", "F", "M", "M", "F", "F", "M", "F", "M", "M"),
9 Disease = c("Flu", "Diabetes", "Flu", "Cancer", NA, "Diabetes", "Hypertension",
10 "Asthma", NA, "Covid-19"),
11 AdmissionDate = as.Date(c("2023-01-15", "2022-11-10", "2023-05-20", "2022-12-05",
12 "2023-07-30", "2022-11-10", "2021-06-18", "2023-02-14",
13 "2023-08-22", "2023-04-10")),
14 TreatmentCost = c(1000, 2500, NA, 5000, 3000, 2500, 7000, 4000, NA, 1500)
15 )
16
17
18 print(df)
19
20 df_sorted<-df[order(df$AdmissionDate),]
21 df_sorted
22
23 df_unique<-df[!duplicated(df$PatientID),]
24 df_unique
25
26 df[df == "" | df == "NA"] <- NA
27 df
28 #E) Convert AdmissionDate to Date format and TreatmentCost to numeric
29 # Convert AdmissionDate to Date format
30 df$AdmissionDate <- as.Date(df$AdmissionDate, format = "%Y-%m-%d")
31 # Convert TreatmentCost to numeric
32 df$TreatmentCost <- as.numeric(df$TreatmentCost)
33 df
34
```

Program input

Output

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
	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

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+).

```
24 df_unique
25
26
27 #E) Convert AdmissionDate to Date format and TreatmentCost to numeric
28
29
30
31
32
33 # Convert TreatmentCost to numeric
34 df$TreatmentCost <- as.numeric(df$TreatmentCost)
35
36 df
37
38
39
40 df$AgeGroup <- cut(
41   df$Age,
42   breaks = c(-Inf, 30, 60, Inf),
43   labels = c("Young", "Middle-aged", "Senior")
44 )
45
46
47 print(df)
48
49
50
51
```

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")
)
```