1. GETTING AND READING DATA SET

```
df <- read.csv("/content/Dataset_Exam_RSheet1.csv")
head(df)
```

	ResponseId	Gender	PRN.Number	Admitted.Year	Admission.Type	Caste	Are
	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	
1	1	Male	2020016402315553	AY 2020-21	Diploma	Open	
2	2	Male	2020016402315553	AY 2020-21	Diploma	OPEN	
3	3	Male	NA	AY 2021-22	Diploma	Brahman	
4	4	Male	2019016401511262	AY 2019-20	First Year	Obc	
5	5	Female	2020016402315777	AY 2021-22	Diploma	OBC	
6	6	Male	2019016401511784	AY 2019-20	First Year	OBC	

2. UNDERSTANDING DATA CONTENTS

```
dim(df) #dimensions
```

77 · 42

```
str(df) #structure
```

```
'data.frame':   77 obs. of  42 variables:
 $ ResponseId                           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                               : chr  "Male" "Male" "Male" "Male" ...
 $ PRN.Number                           : chr  "2020016402315553" "2020016402315553"
 $ Admitted.Year                        : chr  "AY 2020-21" "AY 2020-21" "AY 2021-22"
 $ Admission.Type                       : chr  "Diploma" "Diploma" "Diploma" "First Y
 $ Caste                                : chr  "Open" "OPEN" "Brahman " "Obc" ...
 $ Are.you.DSE..Direct.2nd.Year.Student.: chr  "yes" "yes" "yes" "No" ...
 $ Semester1                            : chr  "" "" "" "Semester 1" ...
 $ Exam.Type                            : chr  "" "" "" "C' Scheme" ...
 $ Passing.Month.and.Year               : chr  "" "" "" "05  Jan 2021" ...
 $ Exam.Seat.Number                     : int  NA NA NA 6036696 NA 6036697 4028747 NA
 $ Total.Marks.Obtained                 : num  NA NA NA 455 NA 441 445 NA NA NA ...
 $ Total.Out.of.Marks                   : num  NA NA NA 675 NA 675 675 NA NA NA ...
 $ SGPI                                 : num  NA NA NA 7.61 NA 7.58 7 NA NA NA ...
 $ Semester2                            : chr  "" "" "" "Semester 2" ...
 $ Exam.Type.1                          : chr  "" "" "" "C' Scheme" ...
 $ Passing.Month.and.Year.1             : chr  "" "" "" "29 Jan  2021" ...
 $ Exam.Seat.Number.1                   : int  NA NA NA 6021567 NA 6021568 6021569 NA
 $ Total.Marks.Obtained.1               : num  NA NA NA 483 NA 490 509 NA NA NA ...
 $ Total.Out.of.Marks.1                 : int  NA NA NA 725 NA 675 725 NA NA NA ...
 $ SGPI.1                               : num  NA NA NA 7.4 NA 7.63 8.03 NA NA NA ...
```

```
$ Semester.3                    : chr  "Semester 3" "Semester 3" "Semester 3"
$ Exam.Type.2                   : chr  "C' Scheme" "C' Scheme" "C' Scheme" "C
$ Passing.Month.and.Year.2      : chr  "2020-21" "2020" "April 2021" "24.03.2
$ Exam.Seat.Number.2            : chr  "20B3630" "20B3630" "20B 3631" "20B 36
$ Total.Marks.Obtained.2        : chr  "576" "578" "524" "574" ...
$ Total.Out.of.Marks.2          : chr  "775" "776" "775" "725" ...
$ SGPI.2                        : num  8.45 8.45 7.52 8.57 8.52 8.83 9.17 8.7
$ Semester.4                    : chr  "Semester 4" "Semester 4" "Semester 4"
$ Exam.Type.3                   : chr  "C' Scheme" "C' Scheme" "C' Scheme" "C
$ Passing.Month.and.Year.3      : chr  "2020-21" "June 2021" "June 2021" "30.
$ Exam.Seat.Number.3            : chr  "21A4601" "21A4601" "21A4602" "21A 460
$ Total.Marks.Obtained.3        : chr  "582" "582" "528" "587" ...
$ Total.Out.of.Marks.3          : chr  "775" "776" "775" "725" ...
$ SGPI.3                        : num  8.57 8.57 7.61 8.74 9.3 8.78 8.96 9.35
$ Semester.5                    : chr  "Semester 5" "Semester 5" "Semester 5"
$ Exam.Type.4                   : chr  "C' Scheme" "C' Scheme" "C' Scheme" "C
$ Passing.Month.and.Year.4      : chr  "2021-22" "NOV 2021" "November 2021" "
$ Exam.Seat.Number.4            : chr  "21B5601" "21B5601" "21B 5602" "21B 56
$ Total.Marks.Obtained.4        : chr  "618" "618" "527" "607" ...
$ Total.Out.of.Marks.4          : chr  "800" "800" NA "800" ...
$ SGPI.4                        : num  8.83 8.83 7.3 8.52 8.83 8.96 9.26 8.83
```

```
# getting the overall summary of data
summary(df)
```

```
  ResponseId    Gender           PRN.Number        Admitted.Year
Min.   : 1   Length:77          Length:77          Length:77
1st Qu.:20   Class :character   Class :character   Class :character
Median :39   Mode  :character   Mode  :character   Mode  :character
Mean   :39
3rd Qu.:58
Max.   :77

Admission.Type        Caste           Are.you.DSE..Direct.2nd.Year.Student.
Length:77          Length:77          Length:77
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character



 Semester1          Exam.Type         Passing.Month.and.Year
Length:77          Length:77          Length:77
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character



Exam.Seat.Number   Total.Marks.Obtained Total.Out.of.Marks      SGPI
Min.   : 4028747   Min.   :  6.7        Min.   :  6.7      Min.   :5.470
1st Qu.: 6036698   1st Qu.:371.2        1st Qu.:675.0      1st Qu.:6.610
Median : 6036705   Median :435.0        Median :675.0      Median :7.070
Mean   :10784685   Mean   :395.5        Mean   :644.0      Mean   :7.051
3rd Qu.: 6041420   3rd Qu.:452.0        3rd Qu.:675.0      3rd Qu.:7.555
Max.   :82100415   Max.   :518.0        Max.   :675.0      Max.   :8.610
NA's   :46         NA's   :49           NA's   :49         NA's   :46
 Semester2         Exam.Type.1        Passing.Month.and.Year.1
Length:77          Length:77          Length:77
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character



Exam.Seat.Number.1 Total.Marks.Obtained.1 Total.Out.of.Marks.1     SGPI.1
Min.   : 6021567   Min.   :  6.37         Min.   : 10.0        Min.   :5.280
1st Qu.: 6021574   1st Qu.:425.25         1st Qu.:725.0        1st Qu.:6.665
Median : 6021581   Median :485.00         Median :725.0        Median :7.180
Mean   :13333965   Mean   :436.66         Mean   :683.4        Mean   :7.258
3rd Qu.: 6036395   3rd Qu.:504.00         3rd Qu.:725.0        3rd Qu.:7.695
Max.   :82200467   Max.   :615.00         Max.   :725.0        Max.   :9.550
NA's   :46         NA's   :49             NA's   :49           NA's   :46
 Semester.3        Exam.Type.2        Passing.Month.and.Year.2
Length:77          Length:77          Length:77
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character



Exam.Seat.Number.2 Total.Marks.Obtained.2 Total.Out.of.Marks.2     SGPI.2
Length:77          Length:77              Length:77            Min.   :6.250
Class :character   Class :character       Class :character     1st Qu.:8.170
Mode  :character   Mode  :character       Mode  :character     Median :8.450
                                                               Mean   :8.405
```

```
Mean   :8.405
3rd Qu.:8.780
Max.   :9.520
```

3. DATA PREPROCESSING

```
Class :character   Class :character   Class :character
```

```
# Data Cleaning
# Checking for NA values
colSums(is.na(df))
```

```
ResponseId: 0 Gender: 0 PRN.Number: 19 Admitted.Year: 0 Admission.Type:
Are.you.DSE..Direct.2nd.Year.Student.: 0 Semester1: 0 Exam.Type: 0 Passing.Month.
Exam.Seat.Number: 46 Total.Marks.Obtained: 49 Total.Out.of.Marks: 49 SGPI:
0 Exam.Type.1: 0 Passing.Month.and.Year.1: 0 Exam.Seat.Number.1: 46 Total.M
49 Total.Out.of.Marks.1: 49 SGPI.1: 46 Semester.3: 0 Exam.Type.2: 0
Passing.Month.and.Year.2: 0 Exam.Seat.Number.2: 0 Total.Marks.Obtained.2: 1 Total.
1 SGPI.2: 1 Semester.4: 0 Exam.Type.3: 0 Passing.Month.and.Year.3: 0
Exam.Seat.Number.3: 0 Total.Marks.Obtained.3: 1 Total.Out.of.Marks.3: 1 SGPI.3:
NA's :1
```

```
vars <- c("ResponseId", "SGPI.2", "SGPI.3", "SGPI.4")
newdf <- df[vars]
head(newdf)
```

A data.frame: 6 × 4

	ResponseId	SGPI.2	SGPI.3	SGPI.4
	<int>	<dbl>	<dbl>	<dbl>
1	1	8.45	8.57	8.83
2	2	8.45	8.57	8.83
3	3	7.52	7.61	7.30
4	4	8.57	8.74	8.52
5	5	8.52	9.30	8.83
6	6	8.83	8.78	8.96

```
#Preprocessing the NA values present in SGPI of all semesters by replacing them with mean.

newdf$SGPI.2[is.na(newdf$SGPI.2)] <- mean(newdf$SGPI.2, na.rm = TRUE)
newdf$SGPI.3[is.na(newdf$SGPI.3)] <- mean(newdf$SGPI.3, na.rm = TRUE)
newdf$SGPI.4[is.na(newdf$SGPI.4)] <- mean(newdf$SGPI.4, na.rm = TRUE)
newdf
```

A data.frame: 77 × 4

ResponseId	SGPI.2	SGPI.3	SGPI.4
<int>	<dbl>	<dbl>	<dbl>
1	8.45	8.57	8.83
2	8.45	8.57	8.83
3	7.52	7.61	7.30
4	8.57	8.74	8.52
5	8.52	9.30	8.83
6	8.83	8.78	8.96
7	9.17	8.96	9.26
8	8.70	9.35	8.83
9	8.17	8.61	8.48
10	8.35	8.87	9.13
11	6.29	5.00	8.65
12	8.43	8.61	8.61
13	8.22	9.09	8.30
14	8.22	8.04	8.52
15	8.39	8.65	8.48
16	7.96	7.91	7.61
17	8.65	9.17	8.96
18	8.13	9.04	8.43
19	8.91	8.39	8.39
20	8.22	8.87	8.61
21	8.78	9.00	8.30
22	8.52	8.87	9.09
23	8.83	9.17	8.57
24	8.43	8.61	8.52
25	9.22	9.70	9.30
26	9.22	9.74	9.13
27	6.91	7.75	7.83
28	7.70	6.71	6.74
29	8.65	8.66	8.74
30	9.17	9.30	8.65
⋮	⋮	⋮	⋮

48	7.390000	7.750000	7.830000
49	8.130000	8.650000	8.780000
50	8.480000	8.740000	8.130000
51	8.220000	8.390000	8.430000
52	8.610000	8.870000	8.260000
53	8.650000	9.300000	8.610000
54	8.960000	9.260000	9.400000
55	6.250000	4.820000	7.830000
56	7.650000	7.570000	7.520000
57	8.700000	8.870000	8.430000
58	8.170000	8.170000	7.430000
59	9.000000	9.000000	8.610000
60	9.300000	8.870000	8.520000
61	8.570000	9.830000	8.780000
62	7.520000	7.830000	9.090000
63	8.910000	9.610000	9.220000
64	8.960000	9.780000	8.650000
65	8.610000	9.430000	8.480000
66	8.300000	9.670000	8.830000
67	7.910000	8.780000	8.350000
68	8.520000	8.130000	8.780000
69	9.130000	9.130000	8.740000
70	8.870000	9.260000	8.830000
71	8.300000	8.520000	7.610000
72	8.520000	8.910000	8.960000

4. Asking 2-3 different analytical questions

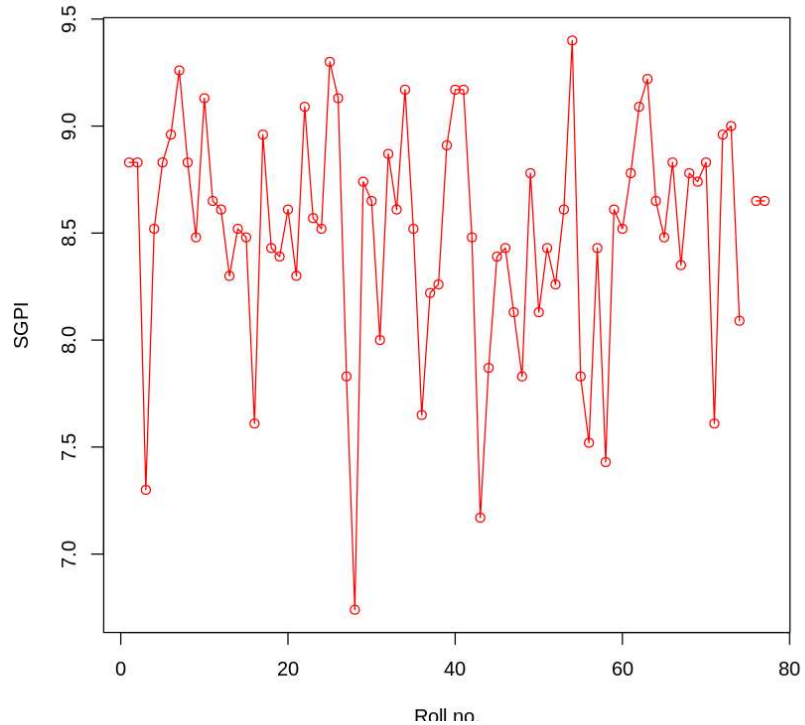
Q1) What are the SGPI of students in semester 3? What are maximum, minimum and average SGPI of the same semester?

76	8.780000	9.260000	8.650000

Here, we can see a graphical distribution semester 3 SGPI with highest being at 9.4, lowest being 6.74 and average being 8.49

```
plot(df$`SGPI.4`,type="o",col="red",xlab="Roll no.", ylab="SGPI")
summary(df$`SGPI.4`)
```

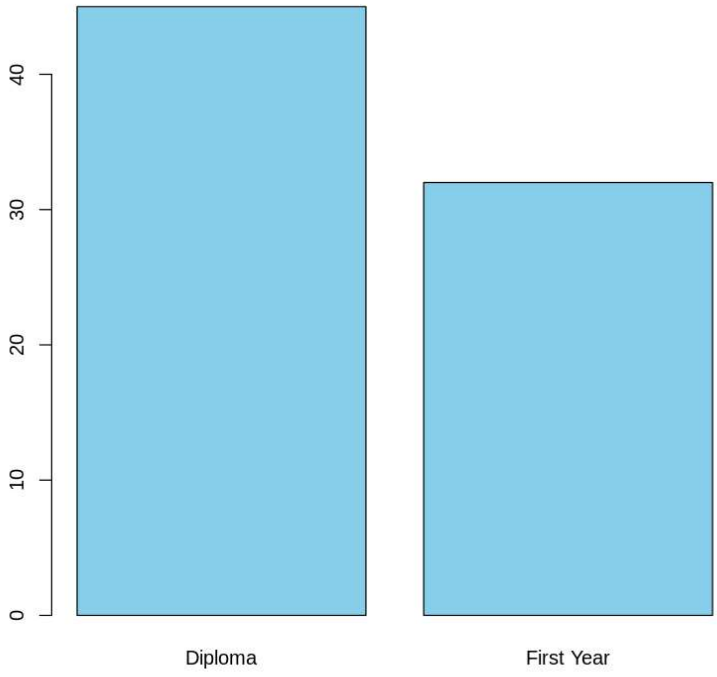
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  6.740   8.290   8.590   8.499   8.830   9.400       1
```



Q2) What group of students are more, Diploma or First Year?

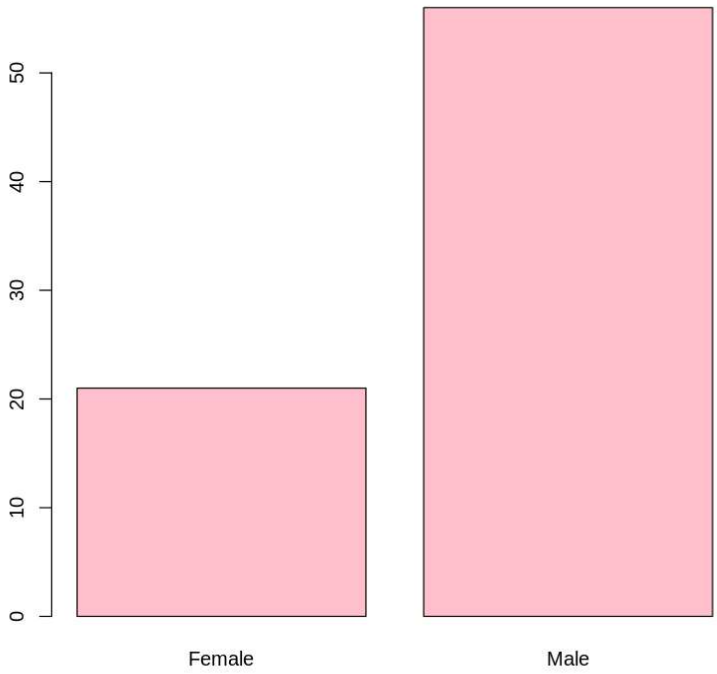
Here, we observe that number of Diploma students are more than First Year students

```
barplot(table(df$Admission.Type), col="skyblue")
```



Q3) What is the gender of majority of students? By how much does it lead?

```
barplot(table(df$Gender), col = "pink")
```



Here we observe that number of male students are more than twice the number of female students

5. Machine Learning Problem

Defining problem: Creating a Machine Learning model which predicts the SGPI of next semester by processing SGPIs of previous semesters.

Model used: Linear Regression

```
set.seed(30)
# Train Test Split
train_index <- sample(1:nrow(newdf), 0.8 * nrow(newdf))
test_index <- setdiff(1:nrow(newdf), train_index)
X_train <- newdf[train_index, -15]
y_train <- newdf[train_index, "SGPI.4"]
X_test <- newdf[test_index, -15]
y_test <- newdf[test_index, "SGPI.4"]
head(X_train)
```

A data.frame: 6 × 4

	ResponseId	SGPI.2	SGPI.3	SGPI.4
	<int>	<dbl>	<dbl>	<dbl>
74	74	7.70	9.04	8.09
50	50	8.48	8.74	8.13
46	46	7.91	8.39	8.43
13	13	8.22	9.09	8.30
10	10	8.35	8.87	9.13

```
head(X_test)
head(y_test)
```

A data.frame: 6 × 4

	ResponseId	SGPI.2	SGPI.3	SGPI.4
	<int>	<dbl>	<dbl>	<dbl>
7	7	9.17	8.96	9.26
22	22	8.52	8.87	9.09
24	24	8.43	8.61	8.52
30	30	9.17	9.30	8.65
31	31	8.39	8.65	8.00
36	36	7.70	7.74	7.65

9.26 · 9.09 · 8.52 · 8.65 · 8 · 7.65

```
library(readxl)
Sem5 = lm(SGPI.4 ~ SGPI.2+SGPI.3 , data = newdf)
summary(Sem5)
AIC(Sem5)
BIC(Sem5)
```

```
Call:
lm(formula = SGPI.4 ~ SGPI.2 + SGPI.3, data = newdf)
```

```
distPred <- predict(Sem5, X_test)
actuals_preds <- data.frame(cbind(actuals=X_test$SGPI.4, predicteds=distPred))
min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
print(min_max_accuracy)
mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
print(mape)
```

```
[1] 0.9437557
[1] 0.05909289
Signif. codes:  0       0.001      0.01     0.05  .  0.1     1
```

We get 94.3% min-max accuracy in Linear Regression and Mean Absolute Percentage Error of 0.04%

```
95.15064053784 95
```

```
predict(Sem5, data.frame(SGPI.2 = 8.0, SGPI.3 = 8.0))
```

1: 8.27305363077389

Explaination of output: If a student scores 8.0 SGPIs in two semesters, he/she will score 8.27 SGPI in next semester. This prediction is 94.3% accurate.