CAUSES OF THE DROPOUTS -PREDICTION:

-Charanya Devi P S

2023-02-19

DESCRIPTION:

This dataset provides an information of students enrolled in various courses in an institution. It includes social-economic factors and academic performance information that can be used to analyze the dropout rate and causes. This dataset contains multiple disjoint databases consisting of relevant information available at the time of enrollment, such as application mode, marital status, course chosen and more. Additionally, this data contains the student performance at the end of the semester by assessing curricular units credited/enrolled/evaluated/approved as well as their respective grades. Finally, we have unemployment rate, inflation rate and GDP from the region to predict student dropout rates.

```
library(readxl)
df=read_excel("dropouts.xlsx")
head(df)

## # A tibble: 6 × 28
##   Marit…¹ Appli…² Appli…³ Course Dayti…⁴ Previ…⁵ Natio…⁶ Mothe…⁷ Fathe…⁸
Mothe…⁹
##     <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
## 1       1       8       5      2       1       1       1      13      10
6
## 2       1       6       1     11       1       1       1       1       3
4
## 3       1       1       5      5       1       1       1      22      27
10
## 4       1       8       2     15       1       1       1      23      27
6
## 5       2      12       1      3       0       1       1      22      28
10
## 6       2      12       1     17       0      12       1      22      27
10
```

```
## # … with 18 more variables: Father_occupation <dbl>, Displaced <dbl>,
## #   Educational_special_needs <dbl>, Debtor <dbl>, Tuition_fees <dbl>,
## #   Gender <dbl>, Scholarship_holder <dbl>, Age_at_enrollment <dbl>,
## #   International <dbl>, one_sem_credited <dbl>, one_sem_enrolled <dbl>,
## #   one_sem_evaluations <dbl>, one_sem_approved <dbl>, one_sem_grade
<dbl>,
## #   Unemployment_rate <dbl>, Inflation_rate <dbl>, GDP <dbl>, Target
<chr>, and
## #   abbreviated variable names ¹Marital_status, ²Application_mode, …
```

```
tail(df)
```

```
## # A tibble: 6 × 28
##   Marit…¹ Appli…² Appli…³ Course Dayti…⁴ Previ…⁵ Natio…⁶ Mothe…⁷ Fathe…⁸
Mothe…⁹
##     <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
## 1       1       1       1      5       1       1       1      13       1
4
## 2       1       1       1     14       1       1       1      22      27
6
## 3       1       1       6     12       1       1       1       1       3
4
## 4       1       1       1      6       1       1       1      22      27
10
## 5       1       1       1     16       1       1       1       3       1
3
## 6       1       4       3      8       1       3       1      13      27
12
## # … with 18 more variables: Father_occupation <dbl>, Displaced <dbl>,
## #   Educational_special_needs <dbl>, Debtor <dbl>, Tuition_fees <dbl>,
## #   Gender <dbl>, Scholarship_holder <dbl>, Age_at_enrollment <dbl>,
## #   International <dbl>, one_sem_credited <dbl>, one_sem_enrolled <dbl>,
## #   one_sem_evaluations <dbl>, one_sem_approved <dbl>, one_sem_grade
<dbl>,
## #   Unemployment_rate <dbl>, Inflation_rate <dbl>, GDP <dbl>, Target
<chr>, and
## #   abbreviated variable names ¹Marital_status, ²Application_mode, …
```

```
str(df)
```

```
## tibble [496 × 28] (S3: tbl_df/tbl/data.frame)
##  $ Marital_status            : num [1:496] 1 1 1 1 2 2 1 1 1 1 ...
##  $ Application_mode          : num [1:496] 8 6 1 8 12 12 1 9 1 1 ...
##  $ Application_order         : num [1:496] 5 1 5 2 1 1 1 4 3 1 ...
##  $ Course                    : num [1:496] 2 11 5 15 3 17 12 11 10 10 ...
##  $ Daytime/evening_attendance: num [1:496] 1 1 1 1 0 0 1 1 1 1 ...
##  $ Previous qualification    : num [1:496] 1 1 1 1 1 12 1 1 1 1 ...
##  $ Nationality               : num [1:496] 1 1 1 1 1 1 1 1 15 1 ...
##  $ Mother_qualification      : num [1:496] 13 1 22 23 22 22 13 22 1 1 ...
##  $ Father_qualification      : num [1:496] 10 3 27 27 28 27 28 27 1 14 ...
##  $ Mother_occupation         : num [1:496] 6 4 10 6 10 10 8 10 10 5 ...
##  $ Father_occupation         : num [1:496] 10 4 10 4 10 8 11 10 10 8 ...
##  $ Displaced                 : num [1:496] 1 1 1 1 0 0 1 1 0 1 ...
##  $ Educational_special_needs : num [1:496] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Debtor                    : num [1:496] 0 0 0 0 0 1 0 0 0 1 ...
##  $ Tuition_fees              : num [1:496] 1 0 0 1 1 1 1 0 1 0 ...
##  $ Gender                    : num [1:496] 1 1 1 0 0 1 0 1 0 0 ...
##  $ Scholarship_holder        : num [1:496] 0 0 0 0 0 0 1 0 1 0 ...
##  $ Age_at_enrollment         : num [1:496] 20 19 19 20 45 50 18 22 21 18
...
##  $ International             : num [1:496] 0 0 0 0 0 0 0 0 1 0 ...
##  $ one_sem_credited          : num [1:496] 0 0 0 0 0 0 0 0 0 0 ...
##  $ one_sem_enrolled          : num [1:496] 0 6 6 6 6 5 7 5 6 6 ...
##  $ one_sem_evaluations       : num [1:496] 0 6 0 8 9 10 9 5 8 9 ...
##  $ one_sem_approved          : num [1:496] 0 6 0 6 5 5 7 0 6 5 ...
##  $ one_sem_grade             : num [1:496] 0 14 0 13.4 12.3 ...
##  $ Unemployment_rate         : num [1:496] 10.8 13.9 10.8 9.4 13.9 16.2
15.5 15.5 16.2 8.9 ...
##  $ Inflation_rate            : num [1:496] 1.4 -0.3 1.4 -0.8 -0.3 0.3 2.8
2.8 0.3 1.4 ...
##  $ GDP                       : num [1:496] 1.74 0.79 1.74 -3.12 0.79 -0.92
-4.06 -4.06 -0.92 3.51 ...
##  $ Target                    : chr [1:496] "Dropout" "Graduate" "Dropout"
"Graduate" ...
```

ASSUMPTION:

From the dataset, I assume that dropout rate of the students are higher than the success rate of the students.

The dropout rate is mainly caused by student’s personal issue (such as marital status, Age of the student, Gender); Academic issues (such as semester-grade, Displaced, Unemployment of the course taken, Course, previous qualification); Financial issues (such as lack of Scholarship, debt).These factors affects the dropout rate of the students, which makes them greater than success rate of the students.

```
summary(df)
```

```
##  Marital_status  Application_mode Application_order     Course
##  Min.   :1.000   Min.   : 1.000   Min.   :1.000     Min.   : 1.00
##  1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:1.000     1st Qu.: 8.00
##  Median :1.000   Median : 8.000   Median :1.000     Median :11.00
##  Mean   :1.113   Mean   : 6.306   Mean   :1.798     Mean   :10.36
##  3rd Qu.:1.000   3rd Qu.:12.000   3rd Qu.:2.000     3rd Qu.:13.00
##  Max.   :4.000   Max.   :17.000   Max.   :6.000     Max.   :17.00
##  Daytime/evening_attendance Previous qualification  Nationality
##  Min.   :0.0000             Min.   : 1.000         Min.   : 1.000
##  1st Qu.:1.0000             1st Qu.: 1.000         1st Qu.: 1.000
##  Median :1.0000             Median : 1.000         Median : 1.000
##  Mean   :0.9052             Mean   : 2.427         Mean   : 1.107
##  3rd Qu.:1.0000             3rd Qu.: 1.000         3rd Qu.: 1.000
##  Max.   :1.0000             Max.   :17.000         Max.   :15.000
##  Mother_qualification Father_qualification Mother_occupation
Father_occupation
##  Min.   : 1.00        Min.   : 1.00        Min.   : 1.000    Min.   :
1.000
##  1st Qu.: 2.00        1st Qu.: 3.00        1st Qu.: 5.000    1st Qu.:
5.000
##  Median :13.00        Median :14.00        Median : 6.000    Median :
8.000
##  Mean   :12.02        Mean   :16.74        Mean   : 7.137    Mean   :
7.597
##  3rd Qu.:22.00        3rd Qu.:27.00        3rd Qu.:10.000    3rd
Qu.:10.000
##  Max.   :27.00        Max.   :29.00        Max.   :29.000    Max.
:46.000
##    Displaced      Educational_special_needs     Debtor
Tuition_fees
##  Min.   :0.0000   Min.   :0.00000           Min.   :0.00000   Min.
:0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000           1st Qu.:0.00000   1st
Qu.:1.0000
##  Median :1.0000   Median :0.00000           Median :0.00000   Median
:1.0000
##  Mean   :0.5484   Mean   :0.01411           Mean   :0.09476   Mean
:0.9254
##  3rd Qu.:1.0000   3rd Qu.:0.00000           3rd Qu.:0.00000   3rd
Qu.:1.0000
##  Max.   :1.0000   Max.   :1.00000           Max.   :1.00000   Max.
:1.0000
##      Gender        Scholarship_holder Age_at_enrollment International
##  Min.   :0.00000   Min.   :0.0000     Min.   :18.00     Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000     1st Qu.:18.00     1st Qu.:0.00000
##  Median :0.00000   Median :0.0000     Median :20.00     Median :0.00000
##  Mean   :0.07863   Mean   :0.2863     Mean   :21.94     Mean   :0.01008
##  3rd Qu.:0.00000   3rd Qu.:1.0000     3rd Qu.:22.00     3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000     Max.   :55.00     Max.   :1.00000
##  one_sem_credited  one_sem_enrolled one_sem_evaluations one_sem_approved
##  Min.   : 0.0000   Min.   : 0.00    Min.   : 0.000      Min.   : 0.000
##  1st Qu.: 0.0000   1st Qu.: 6.00    1st Qu.: 6.000      1st Qu.: 4.000
##  Median : 0.0000   Median : 6.00    Median : 8.000      Median : 5.000
##  Mean   : 0.5141   Mean   : 6.24    Mean   : 8.077      Mean   : 4.964
##  3rd Qu.: 0.0000   3rd Qu.: 7.00    3rd Qu.:10.000      3rd Qu.: 6.000
##  Max.   :19.0000   Max.   :21.00    Max.   :24.000      Max.   :21.000  
##  one_sem_grade   Unemployment_rate Inflation_rate        GDP          
##  Min.   : 0.00   Min.   : 7.60     Min.   :-0.800   Min.   :-4.06000  
##  1st Qu.:11.32   1st Qu.: 9.40     1st Qu.: 0.300   1st Qu.:-1.70000  
##  Median :12.33   Median :11.10     Median : 1.400   Median : 0.32000  
##  Mean   :11.13   Mean   :11.59     Mean   : 1.218   Mean   : 0.01762  
##  3rd Qu.:13.38   3rd Qu.:13.90     3rd Qu.: 2.600   3rd Qu.: 1.79000  
##  Max.   :17.12   Max.   :16.20     Max.   : 3.700   Max.   : 3.51000  
##     Target         
##  Length:496        
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

```
#libraries
library(lattice)
library(rmarkdown)
library(dplyr)
```

```
## 
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```

```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
glimpse(df)
```

```
## Rows: 496
## Columns: 28
## $ Marital_status                 <dbl> 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1,
1, 1,…
## $ Application_mode               <dbl> 8, 6, 1, 8, 12, 12, 1, 9, 1, 1, 1, 1,
1, …
## $ Application_order              <dbl> 5, 1, 5, 2, 1, 1, 1, 4, 3, 1, 1, 1,
2, 1,…
## $ Course                         <dbl> 2, 11, 5, 15, 3, 17, 12, 11, 10, 10,
14, …
## $ `Daytime/evening_attendance`   <dbl> 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1,
1, 1,…
## $ `Previous qualification`       <dbl> 1, 1, 1, 1, 1, 12, 1, 1, 1, 1, 1, 1,
1, 1…
## $ Nationality                    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 15, 1, 1, 1,
1, 1…
## $ Mother_qualification           <dbl> 13, 1, 22, 23, 22, 22, 13, 22, 1, 1,
23, …
## $ Father_qualification         <dbl> 10, 3, 27, 27, 28, 27, 28, 27, 1, 14,
14,…
## $ Mother_occupation            <dbl> 6, 4, 10, 6, 10, 10, 8, 10, 10, 5, 6,
10,…
## $ Father_occupation            <dbl> 10, 4, 10, 4, 10, 8, 11, 10, 10, 8,
8, 10…
## $ Displaced                    <dbl> 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1,
1, 1,…
## $ Educational_special_needs    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ Debtor                       <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0,
0, 0,…
## $ Tuition_fees                 <dbl> 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1,
1, 1,…
## $ Gender                       <dbl> 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0,
0, 0,…
## $ Scholarship_holder           <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1,
0, 1,…
## $ Age_at_enrollment            <dbl> 20, 19, 19, 20, 45, 50, 18, 22, 21,
18, 1…
## $ International                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0,…
## $ one_sem_credited             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ one_sem_enrolled             <dbl> 0, 6, 6, 6, 6, 5, 7, 5, 6, 6, 6, 8,
6, 6,…
## $ one_sem_evaluations          <dbl> 0, 6, 0, 8, 9, 10, 9, 5, 8, 9, 6, 8,
6, 7…
## $ one_sem_approved             <dbl> 0, 6, 0, 6, 5, 5, 7, 0, 6, 5, 6, 7,
0, 6,…
## $ one_sem_grade                <dbl> 0.00000, 14.00000, 0.00000, 13.42857,
12.…
## $ Unemployment_rate            <dbl> 10.8, 13.9, 10.8, 9.4, 13.9, 16.2,
15.5, …
## $ Inflation_rate               <dbl> 1.4, -0.3, 1.4, -0.8, -0.3, 0.3, 2.8,
2.8…
## $ GDP                          <dbl> 1.74, 0.79, 1.74, -3.12, 0.79, -0.92,
-4.…
## $ Target                       <chr> "Dropout", "Graduate", "Dropout",
"Gradua…
```

```
#dimension
dim(df)
```

```
## [1] 496  28
```

```
#histogram

histogram(~Marital_status|Target,main="histogram_of_dropouts",xlab ="Target",
ylab="Marital_status", breaks = 50,col='red',df)
```

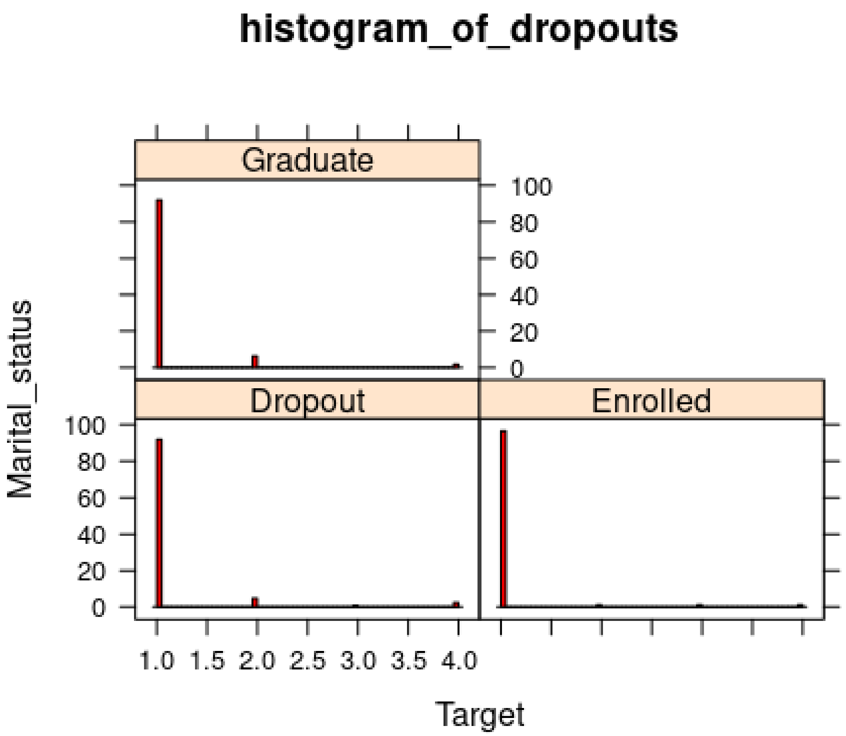

FIG-1

```
histogram(~Scholarship_holder|Gender,main="Schlorship_distribution",xlab
="Gender",ylab ="Scholarship_holder",breaks = 20,col='pink',df)
```

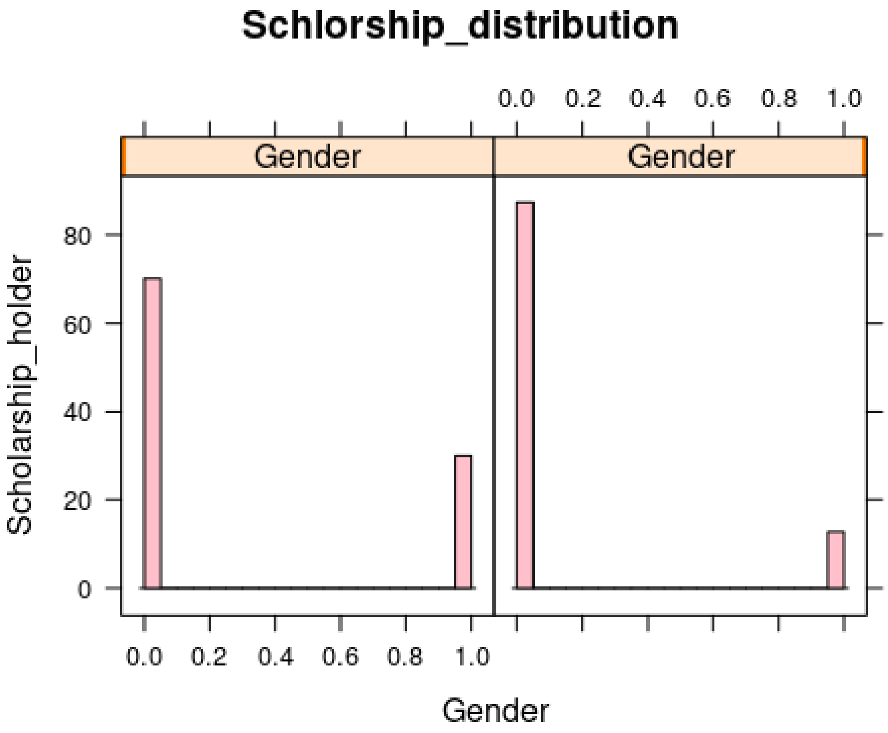

FIG-2

```
histogram(~Unemployment_rate,main="histogram_of_unemployment",breaks =
50,col='skyblue',df)
```

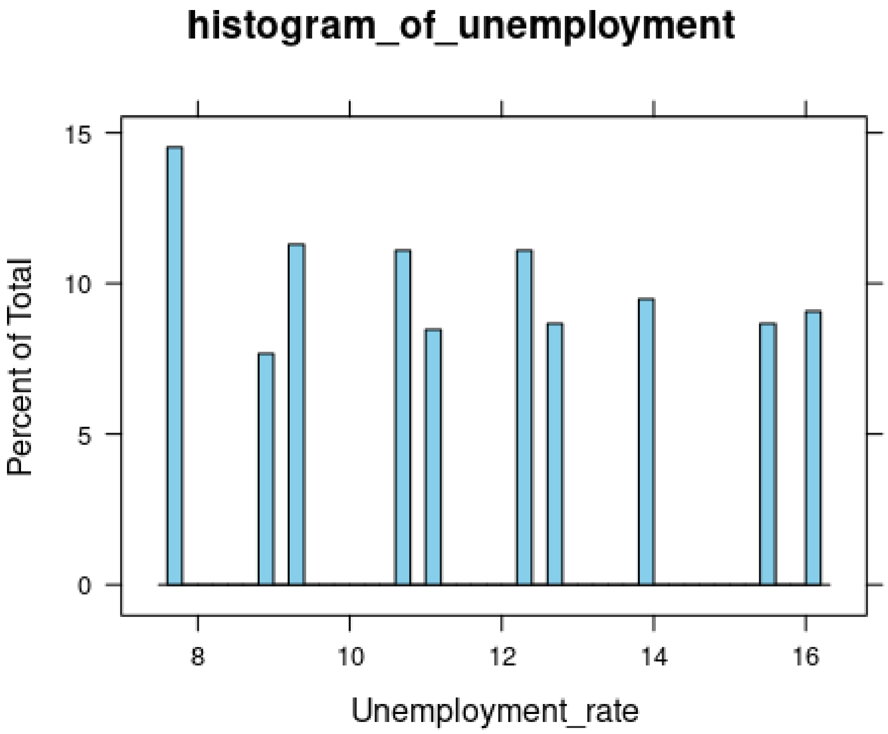

FIG-3

```
histogram(~Scholarship_holder|Debtor,main="histogram_of_debtors",xlab
="Scholarship_holder",ylab ="Debtor",breaks =60 ,col='blue',df)
```

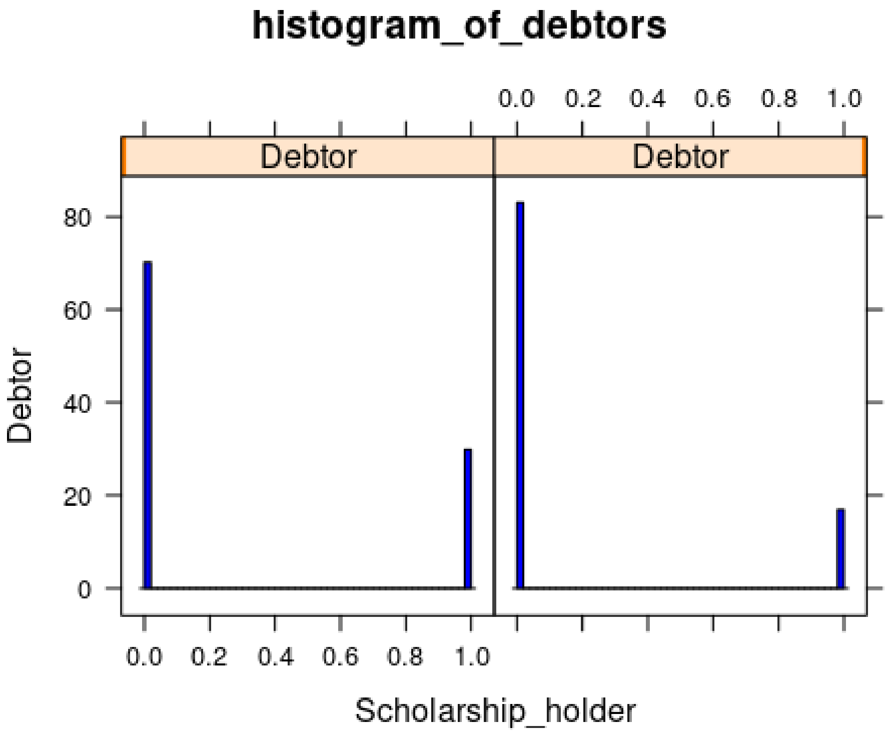

FIG-4

```
histogram(~Age_at_enrollment,main="histogram_of_age",breaks =40
,col='grey',df)
```

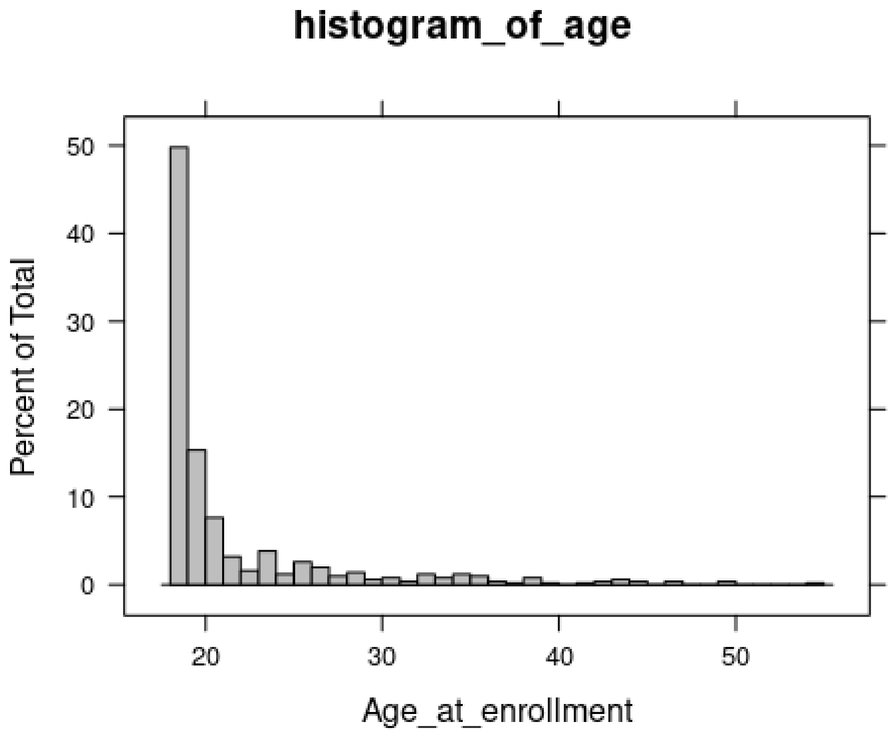

FIG-5

```
histogram(~one_sem_grade|Target,main="distribution of target",xlab
="one_sem_grade",ylab ="Target",breaks =40 ,col='yellow',df)
```

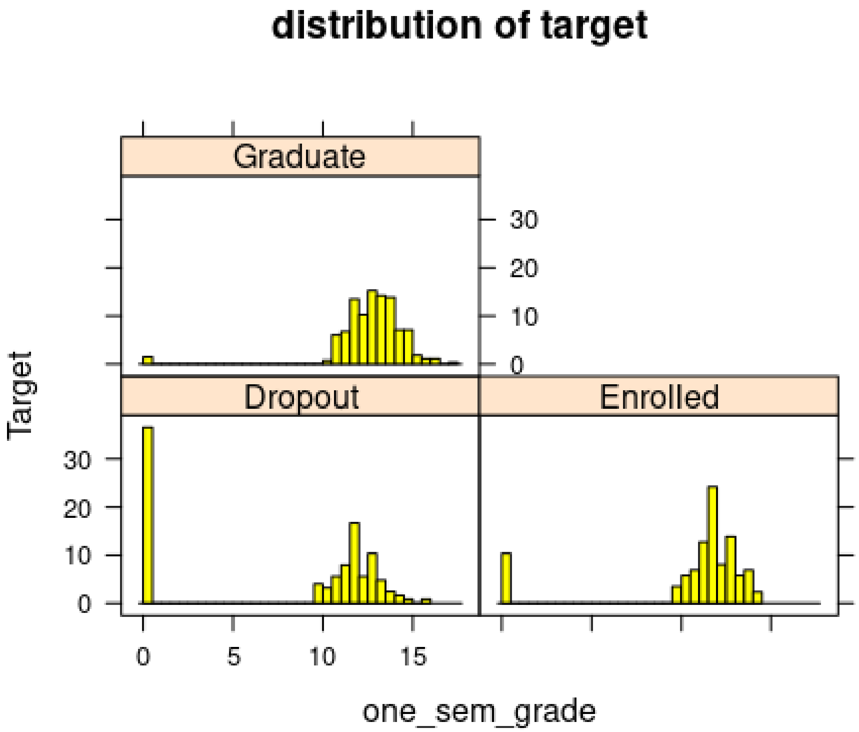

FIG-6

```
#Boxplot

bwplot(df$Marital_status)
```

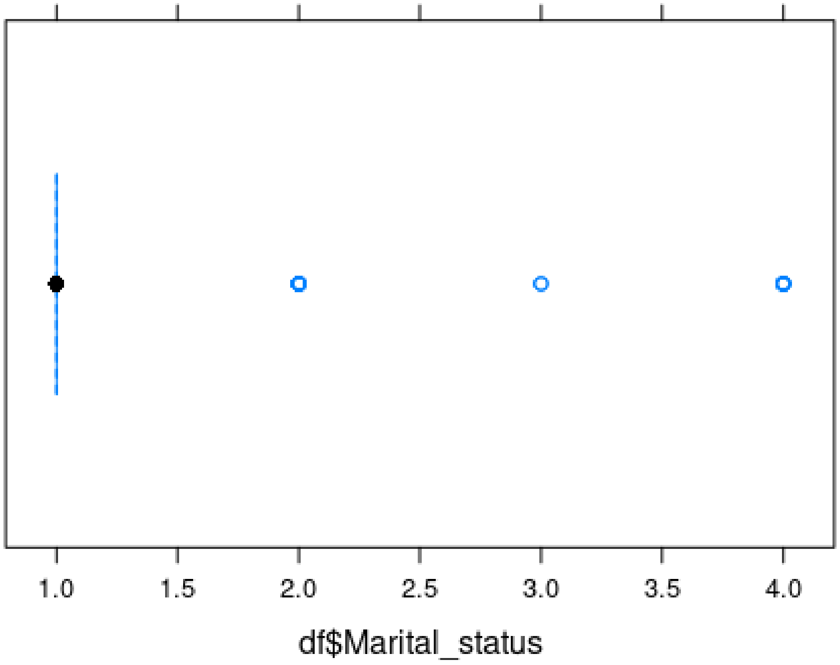

FIG-7

```
bwplot(df$Age_at_enrollment,box.width=0.5)
```

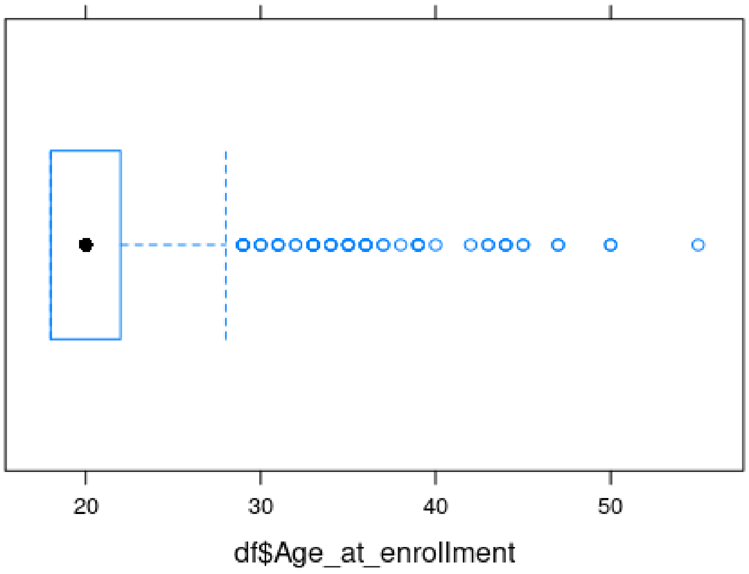

FIG-8

```
bwplot(df$one_sem_grade,box.width=1)
```

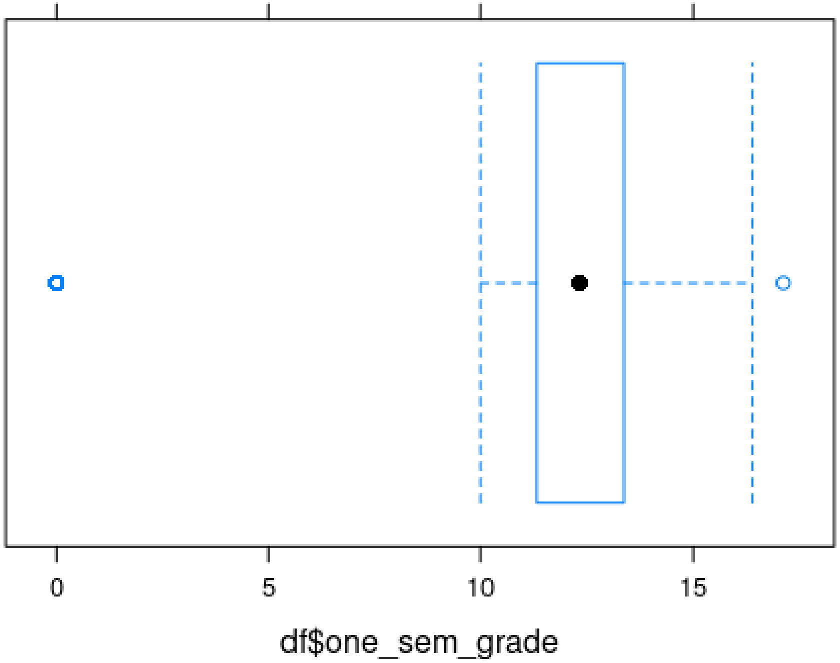

FIG-9

```
bwplot(df$Unemployment_rate,box.width=0.5)
```

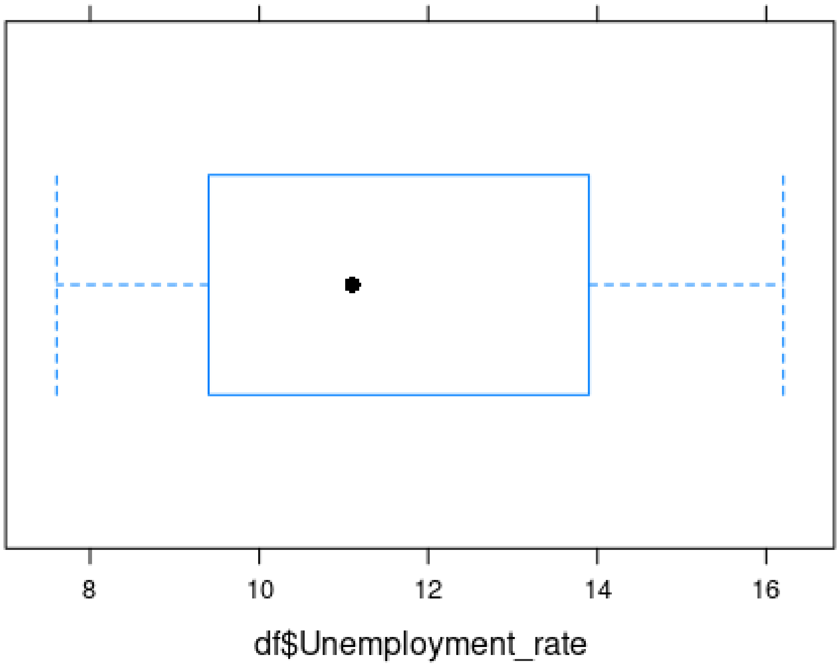

FIG-10

```
bwplot(df$Debtor,box.width=0.5)
```

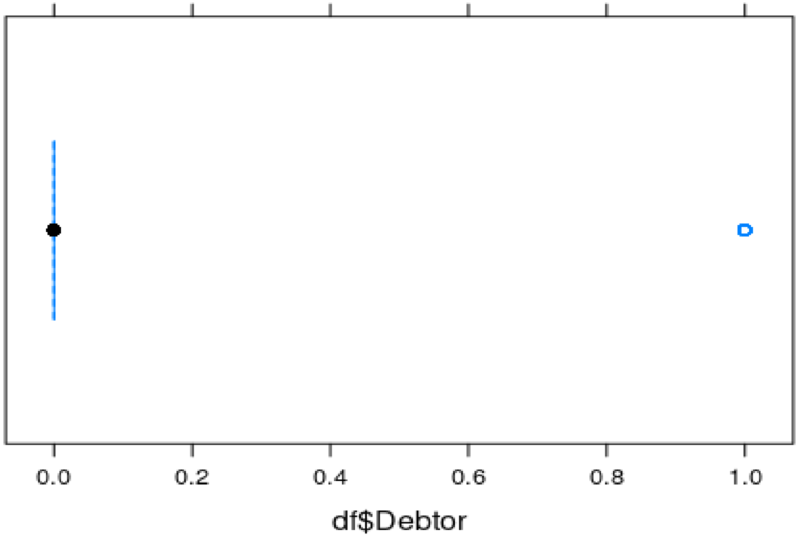

FIG-11

```
#scatter plot
```

```
xyplot(Unemployment_rate~Course, main = "distribution of course&
unemployment", xlab = "Unemployment_rate", ylab = "Course",df)
```

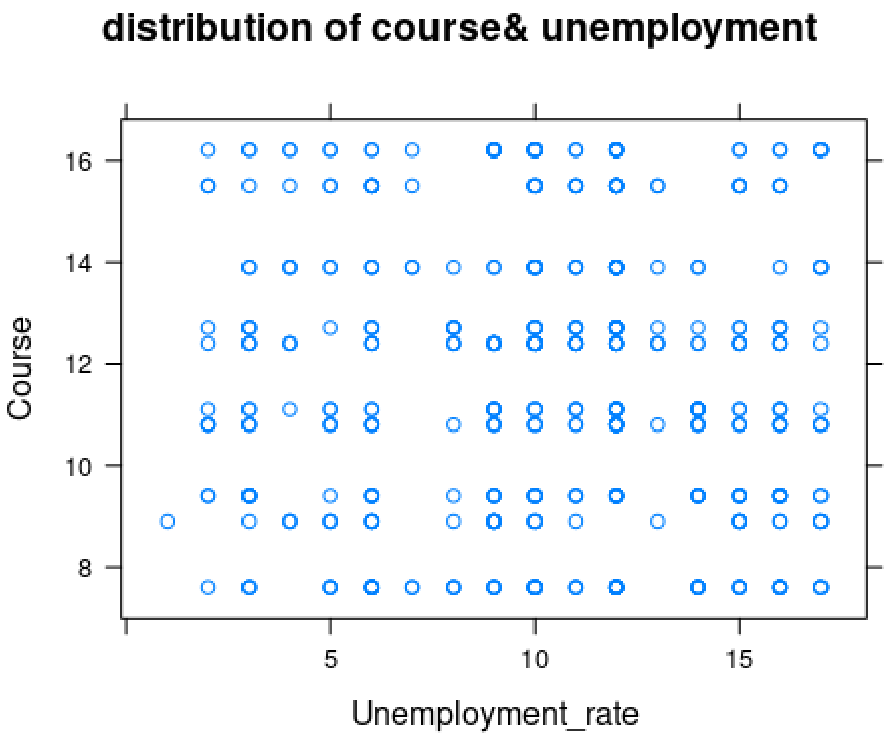

FIG-12

```
xyplot(Course~Age_at_enrollment, main = "distribution of enrollment
age&course", xlab = "Course", ylab = "Age_at_enrollment",df)
```

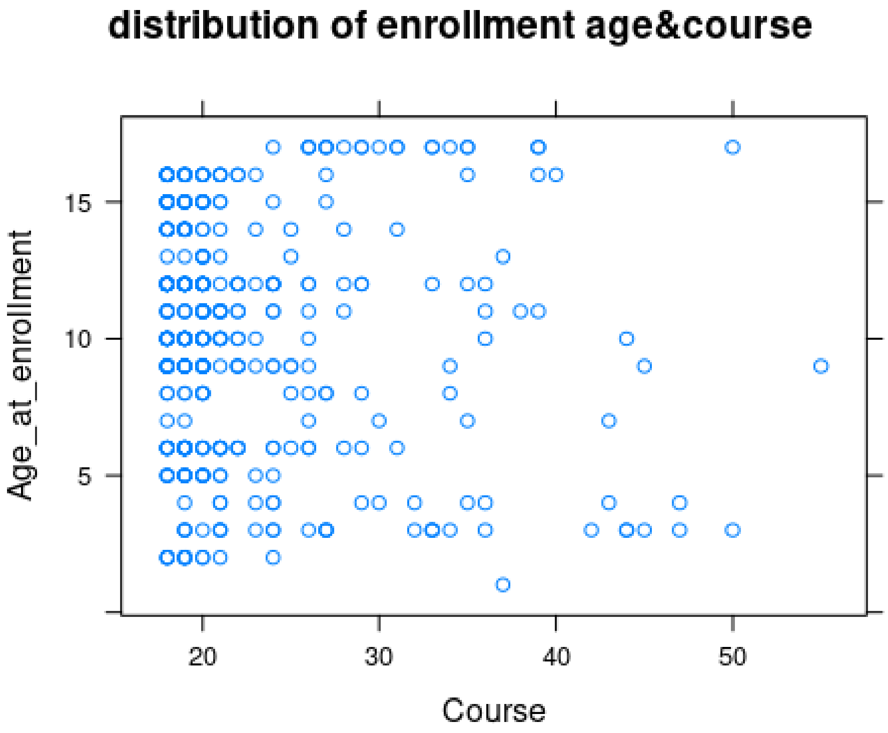

FIG-13

```
#ggplot

library(ggplot2)
ggplot(df,aes(x = one_sem_grade, y = Unemployment_rate)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("one_sem_grade") +
  ylab("Unemployment_rate") +
  ggtitle("grade~Unemployment")

## `geom_smooth()` using formula = 'y ~ x'
```

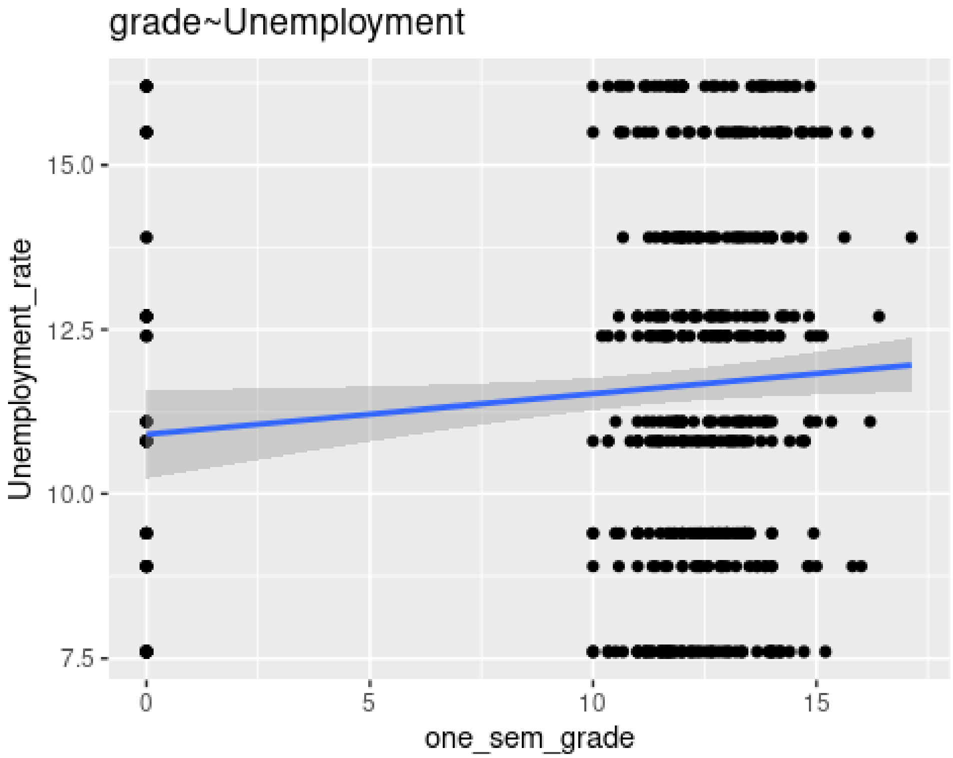

FIG-14

```
ggplot(df,aes(x =GDP , y =Displaced )) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("GDP") +
  ylab("Displaced") +
  ggtitle("GDP~Displaced")

## `geom_smooth()` using formula = 'y ~ x'
```

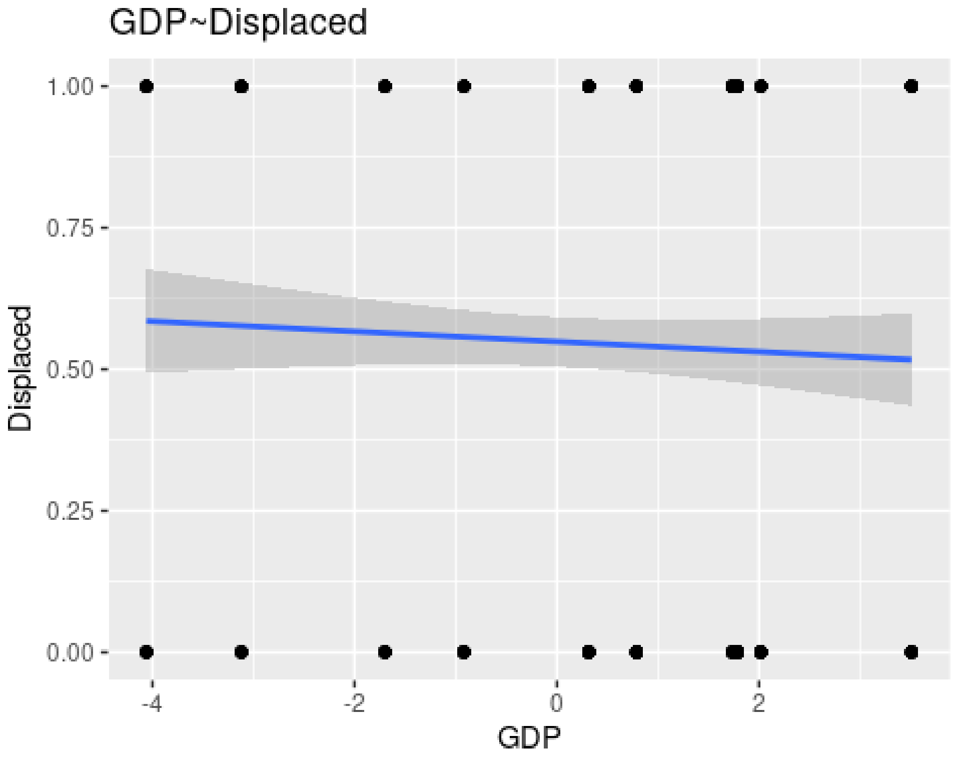

FIG-15

INFERENCE:

The average age of the students enrolled in various courses is 22(21.94). Most of the age lies between 17 to 28(FIG-8)

(In Fig-7)more than 70% of the students unmarried . mostly,

Grade of the students lies between 11.32 and 17.38 in the end of the semester which covers 70% of student population. The average grade is 13.38 (FIG-9)

Average of debtors is 0 which shows that maximum students have no debts (FIG-11)

Most of the time the unemployment rate varies between the range 9.40 to 13.90(FIG-10). The average of students with unemployment is 11.59.

Outliers are present in the attributes such as Age, Semester Grade and Marital status of the students.

Distribution of attributes like marital status of drop outs(fig-1), scholarship holding students(fig-2), Debt of students(fig-4), Age of enrollment(fig-5), Grade of the dropouts(fig-6) are positively skewed.

INSIGHTS:

(From FIG-1), The dropout and graduate students influenced by marital status are same. Marital status of students doesn't affect the rates of dropout.

Many Male students hold scholarship comparing with female students (FIG-2) whereas the female students having debt are higher than the male students (FIG_4). Dropout rate of students (female) is affected by lack of scholarships.

(From FIG-6) Academic or semester grade of the students is quite low for dropout students, which is the major cause for the increase in dropout rates.

Unemployment of students over a time decreases (FIG-3).so, dropout rate was not affected by lack of job opportunity. Also, majority of the students are younger i.e., less than 40. Age doesn't affect the rate of the dropouts.

Concluding that the rate of dropouts are higher which is due to academic (Grades) and Financial (scholarships) issues.