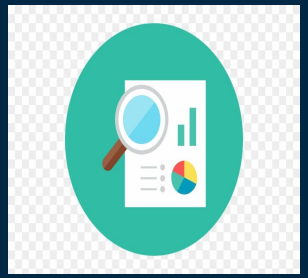
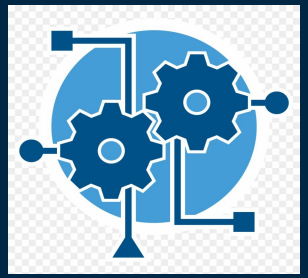
Predictive Analytics on the Academic Record of NUCES.

DATA SCIENCE



ANALYTICS



PREDICTION

VISUALIZATION

Project Team

Obaid Ur Rehman
K17-3848

Areeka Aijaz
K17-3913

Tooba Shahid
K17-3731

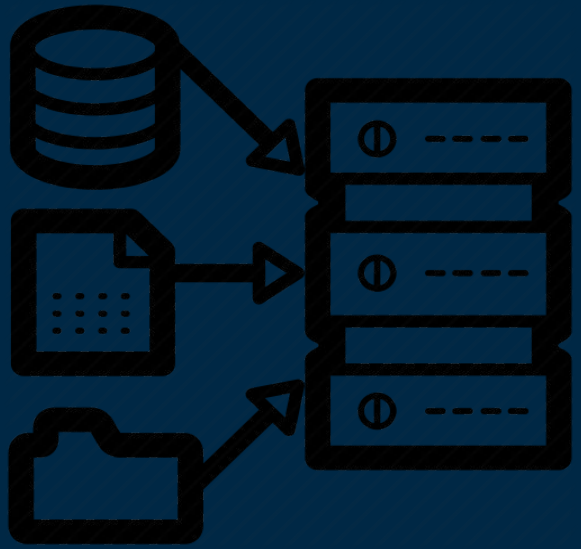
Introduction

- The first phase of our FYP aimed to answer several questions about factors that relate to student performance.

1. Does the previous educational background (Intermediate / A levels) affect the performance at FAST?
2. Does the previous educational background (Matriculation / O levels) affect the performance at FAST?
3. What is the correlation between matriculation / equivalence grade and the performance at FAST?
4. What is the correlation between intermediate / equivalence grade and the performance at FAST?
5. Does there exist any correlation between the city/district (a person belongs to) and their academic performance?
6. Does the performance in initial CS courses affect the performance in the later ones?
7. Does academic performance vary campus wise ?
8. What role does gender play in academic performance ? Do girls tend to perform better than boys or vice versa?
9. What is the correlation of a school with academic performance at FAST ?
10. What is the correlation of a college with academic performance at FAST ?
11. What is the correlation of year of admission with CGPA ?
12. What is the correlation of the year of graduation with CGPA ?
13. Does a degree program affect CGPA?

Phases of Achieving the Goals

1. Pre Processing , EDA and Data Cleaning
2. Feature Selection
3. Dashboard Development



Experimental Setup

Programming language

R language was used to perform all the work related to data analytics. R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

Software Tools

For the data cleaning , transformation and EDA , R studio was used . RStudio is an integrated development environment for R, a programming language for statistical computing and graphics.

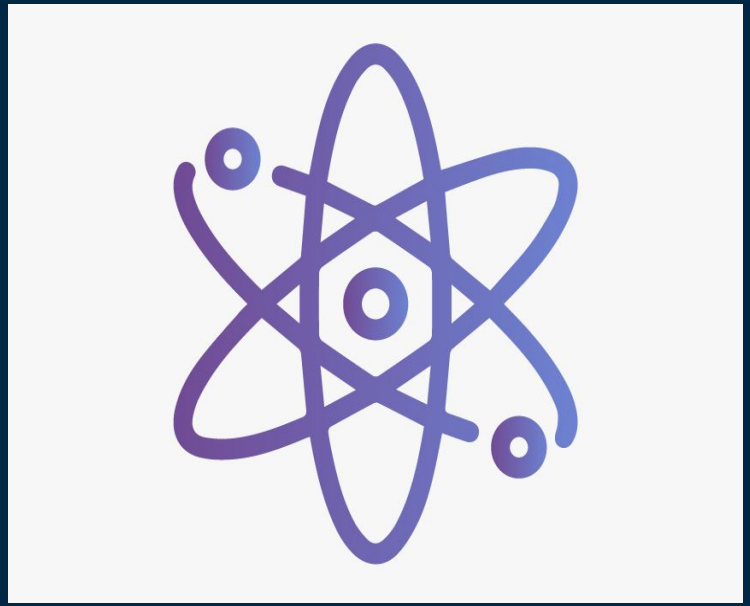
Dataset

- The data used for our project was provided by FAST NUCES for all the FIVE campuses of NUCES i-e Faisalabad , Islamabad , Karachi , Lahore and Peshawar.
- The data contained academic records of undergraduate level (Bachelors) students for the past 19 years from Fall 2001 to Summer 2019 .
- The dataset provided was given in four separate excel sheets Student Data , Semester Data , Course Data 1 , Course Data 2 .

Student Data : Contained details about student gender , batch , campus , program code , CGPA , first semester , last semester , city , SSC Board , SSC obtained etc.

Semester Data: Contained academic details of students for each semester throughout the graduation cycle. The attributes included semester , sgpa , cgpa , core course count , elective course count . Information about each semester of a particular student was given row wise.

Course Data 1 & Course Data 2: Both these datasets had the same columns: semester , student id , code , title , credit hours , course type , relation id , grade , grade point . The data of each student was given in several rows to cover all his/her courses.



Data Preprocessing

Student Data

- Columns such as warnings , credits attempted , credits completed , SSC Total , HSSC total were dropped.
- SSC Board and O Level Board transformed to Secondary Education
- SSC Obtained and O Level Obtained transformed to Secondary Grade
- SSC Board and O Level Board transformed to School
- HSSC Board and A Level Board transformed to Higher Secondary Education
- HSSC Obtained and A Level Obtained transformed to Higher Secondary Grade
- HSSC Board and A Level Board transformed to College

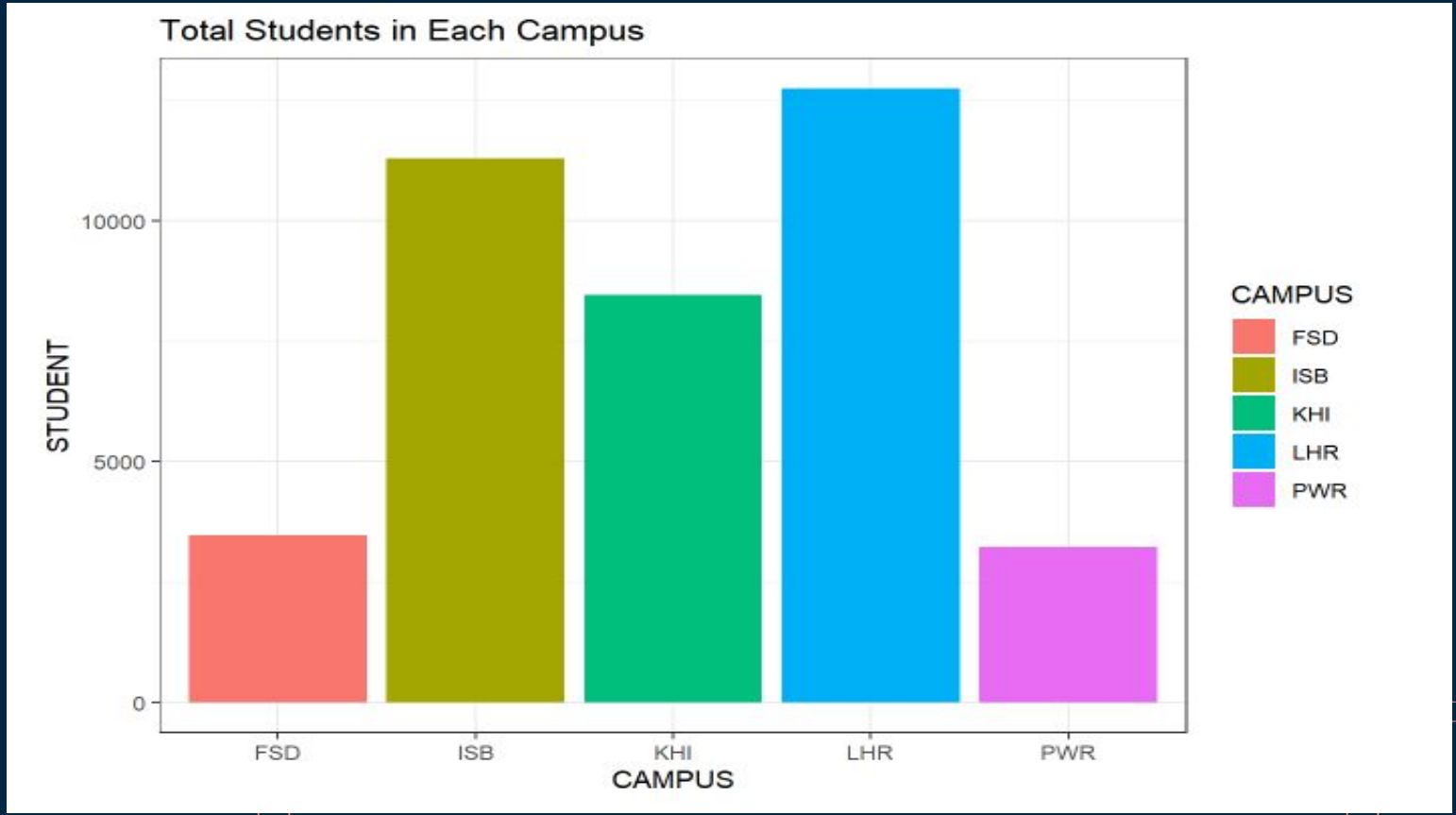
Semester Data

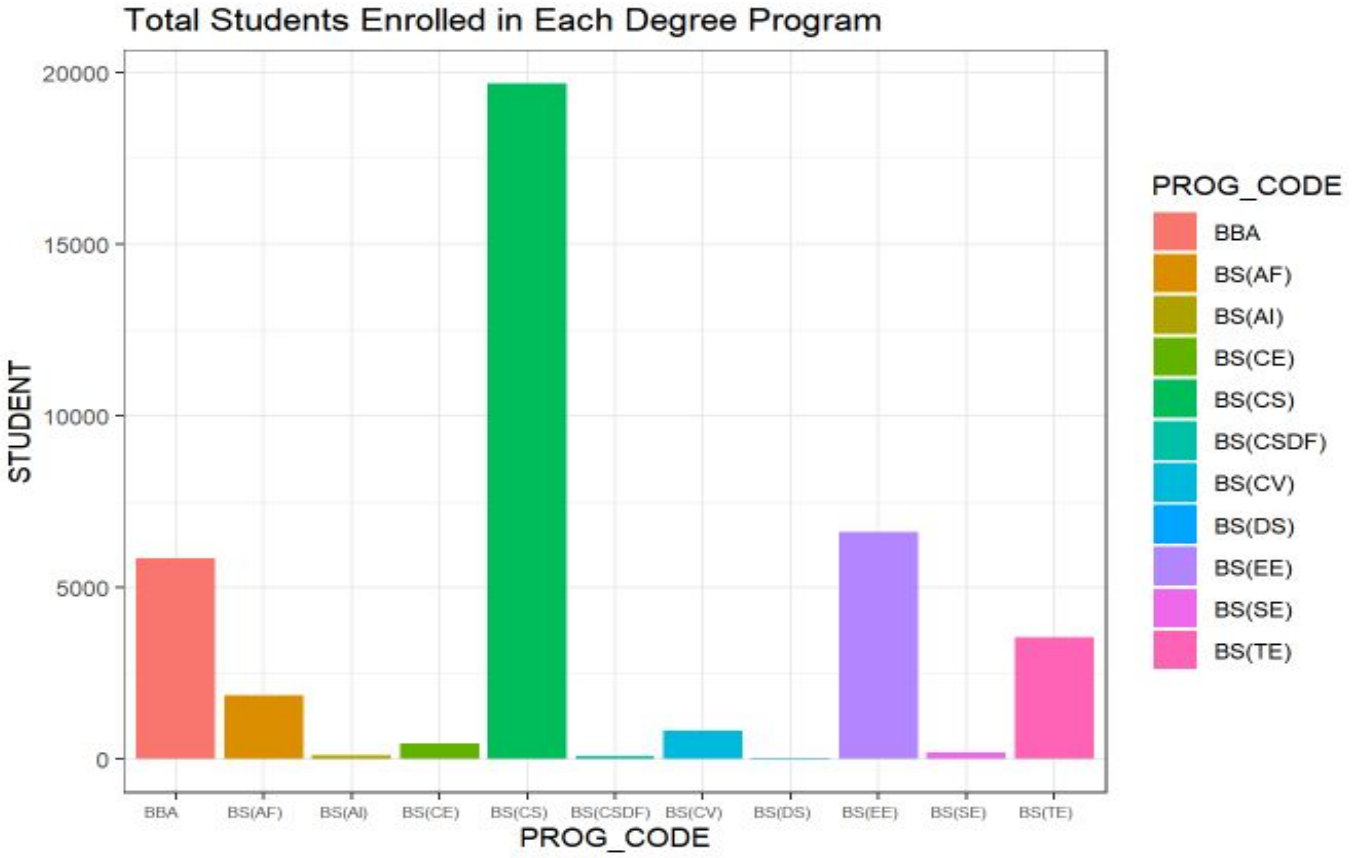
- For sorting the data the semester attribute was splitted into year and session.
- The data was transformed into a new dataframe in which each row had a unique student id against which there were columns for sgpa and cgpa from the first to the last semester.
- Elective Course Count and Core Course Count were dropped as they were not of any use for our analysis.

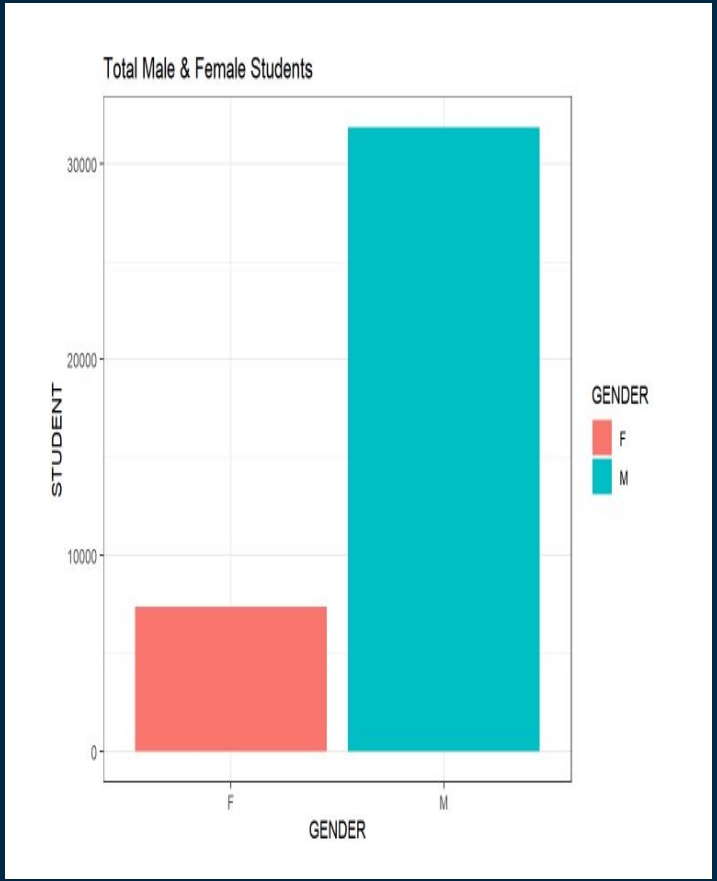
Course Data 1 & Course Data 2

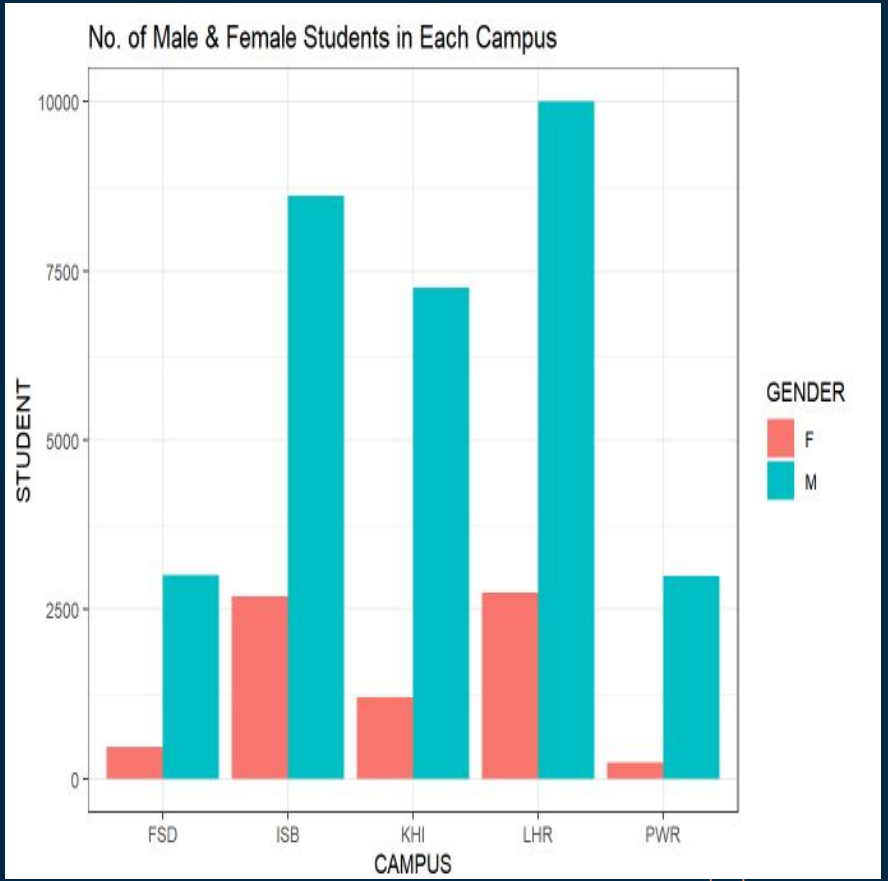
- From course code which was given like SS123 the course domain i-e SS was extracted.
- From the relation id attribute only core courses were retained and the elective courses were dropped
- Columns that weren't useful were dropped and only columns student id , title , domain and grade point were kept for further work .
- The courses were then splitted domain wise i-e CS , EE , SS , MG , CV , MT , EL , CL , VL, FYP.
- The dataset was transformed in a way that all courses were placed column wise
- Separate sheets were maintained for each domain to find out relation between different courses of the same domain.
- Only those courses which were prerequisites of some other courses were selected

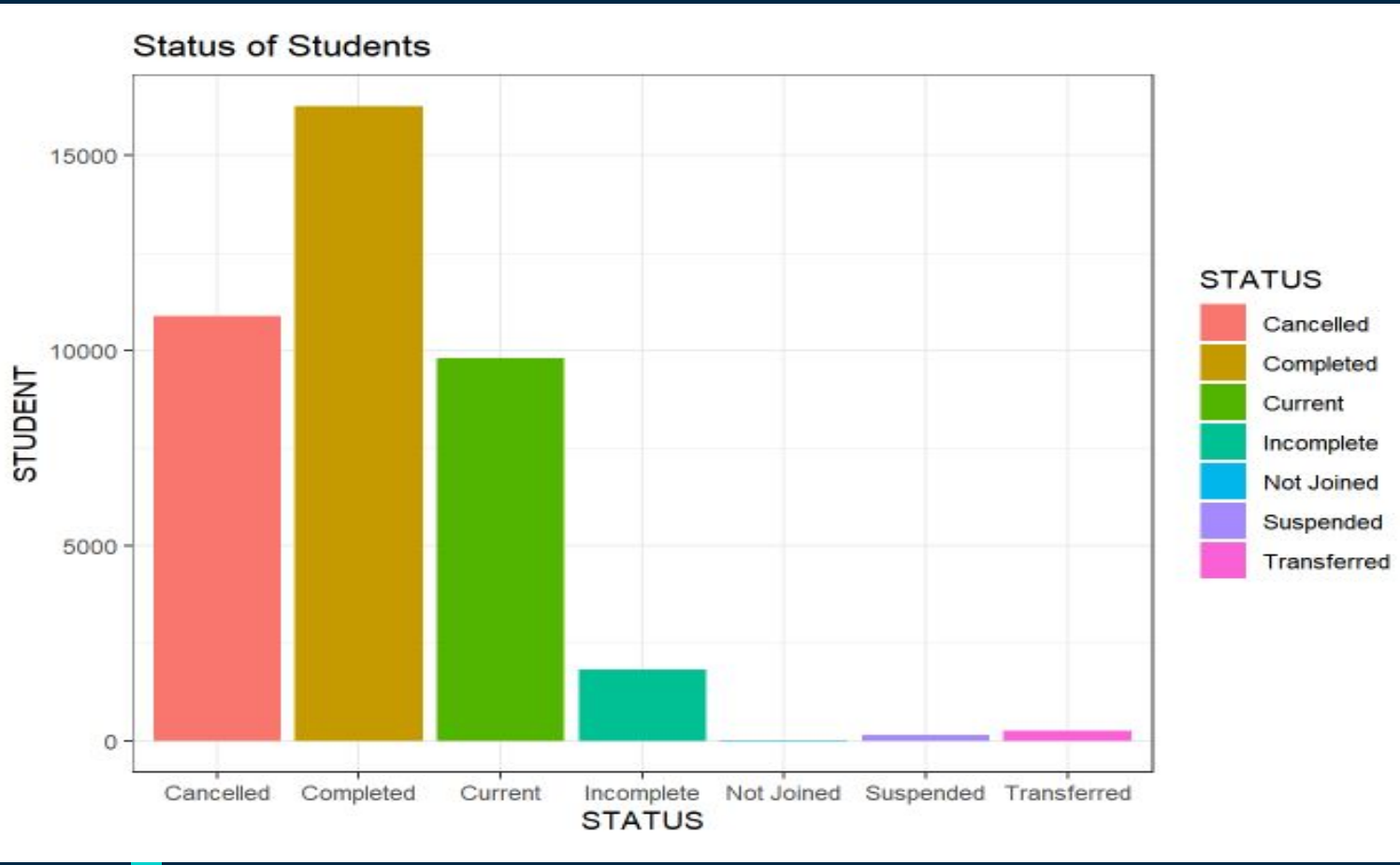
Exploratory Data Analysis

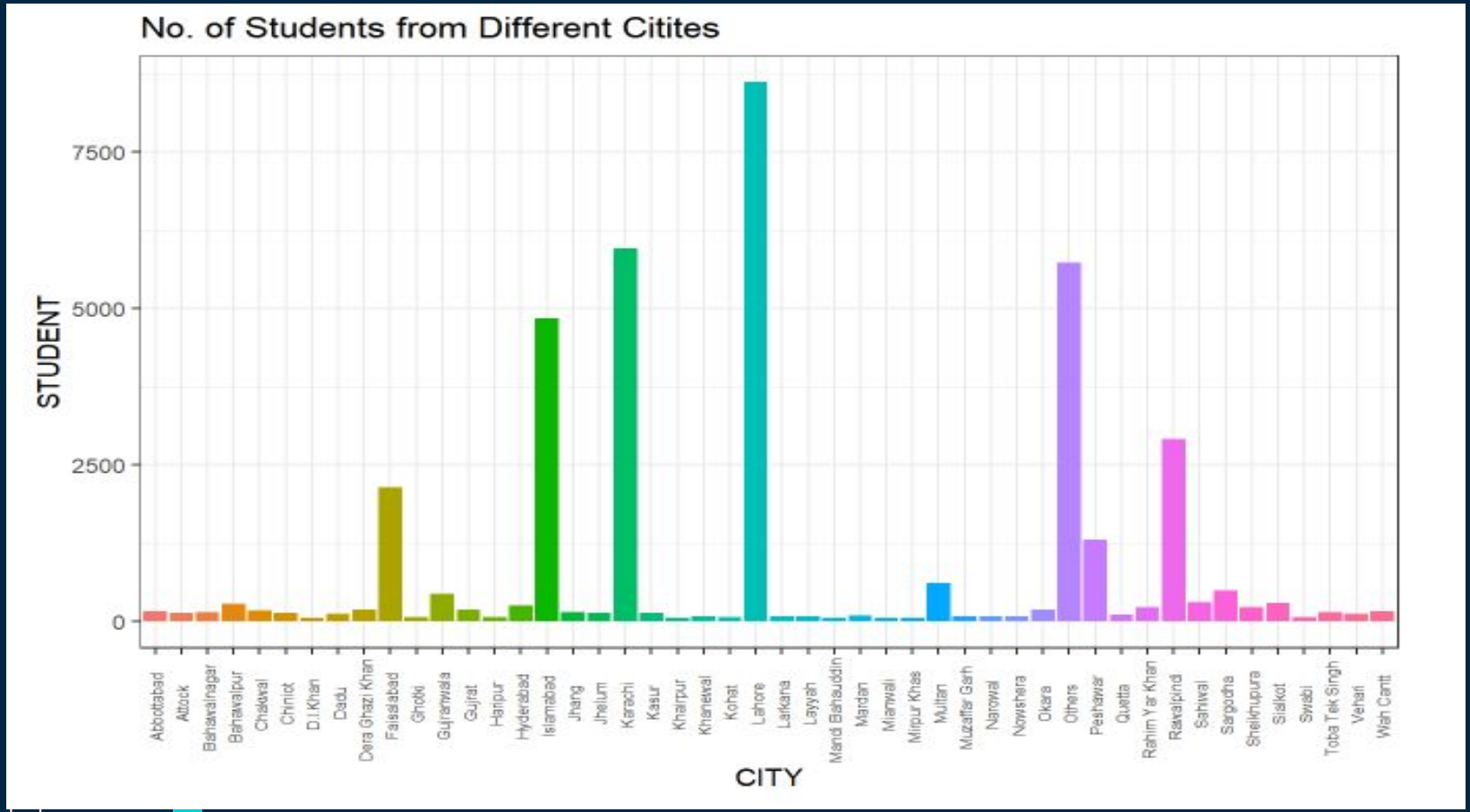

Total Students in Each Campus
STUDENT
10000
5000
0
FSD
ISB
KHI
LHR
PWR
CAMPUS
CAMPUS
FSD
ISB
KHI
LHR
PWR

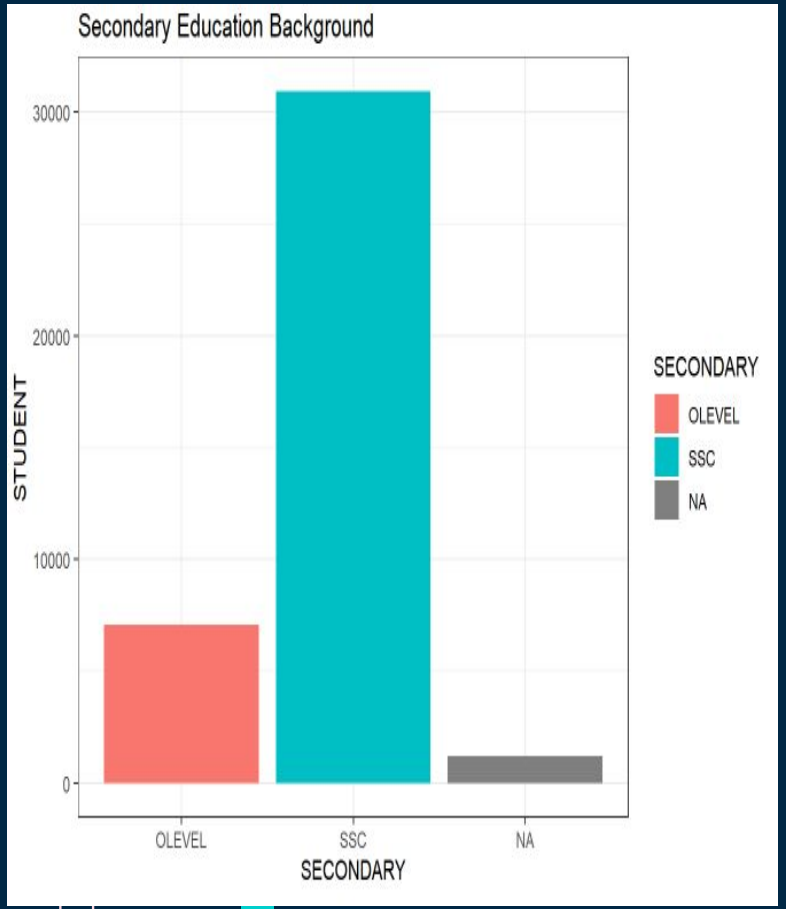

Total Students Enrolled in Each Degree Program
STUDENT
20000
15000
10000
5000
0
BBA
BS(AF)
BS(AI)
BS(CE)
BS(CS)
BS(CSDF)
BS(CV)
BS(DS)
BS(EE)
BS(SE)
BS(TE)
PROG_CODE
PROG_CODE
BBA
BS(AF)
BS(AI)
BS(CE)
BS(CS)
BS(CSDF)
BS(CV)
BS(DS)
BS(EE)
BS(SE)
BS(TE)

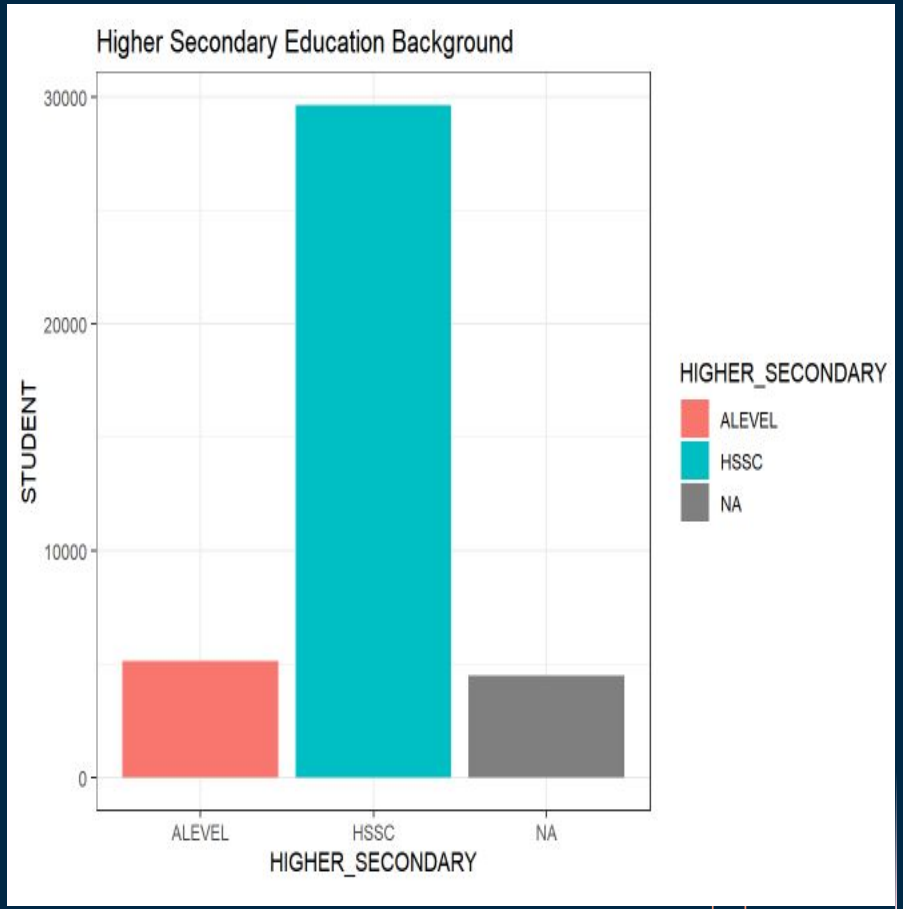

Total Male & Female Students
STUDENT
30000
20000
10000
0
F
M
GENDER
GENDER
F
M


No. of Male & Female Students in Each Campus
STUDENT
10000
7500
5000
2500
0
FSD
ISB
KHI
LHR
PWR
CAMPUS
GENDER
F
M


Status of Students
STUDENT
15000
10000
5000
0
Cancelled
Completed
Current
Incomplete
Not Joined
Suspended
Transferred
STATUS
STATUS
Cancelled
Completed
Current
Incomplete
Not Joined
Suspended
Transferred


No. of Students from Different Citites
STUDENT
7500
5000
2500
0
Abbottabad
Attock
Bahawalnagar
Bahawalpur
Chakwal
Chiniot
D.I.Khan
Dadu
Dera Ghazi Khan
Faisalabad
Ghotki
Gujranwala
Gujrat
Haripur
Hyderabad
Islamabad
Jhang
Jhelum
Karachi
Kasur
Khairpur
Khanewal
Kohat
Lahore
Larkana
Layyah
Mandi Bahauddin
Mardan
Mianwali
Mirpur Khas
Multan
Muzaffar Garh
Narowal
Nowshera
Okara
Others
Peshawar
Quetta
Rahim Yar Khan
Rawalpindi
Sahiwal
Sargodha
Sheikhupura
Sialkot
Swabi
Toba Tek Singh
Vehari
Wah Cantt
CITY


Secondary Education Background
30000
20000
10000
0
STUDENT
OLEVEL
SSC
NA
SECONDARY
SECONDARY
OLEVEL
SSC
NA


Higher Secondary Education Background
30000
20000
10000
0
STUDENT
ALEVEL
HSSC
NA
HIGHER_SECONDARY
HIGHER_SECONDARY
ALEVEL
HSSC
NA

Conclusions from EDA

- Overall 11 different degree programs were offered out of which Islamabad offered 9 , Lahore offered 7 , Karachi offered 6 , Peshawar also offered 6 and Faisalabad offered 5.
- With all the campuses have majority students enrolled in BS(CS) , Peshawar and Lahore have a good number of students enrolled in BS(TE) also.
- Peshawar also has nearly same number of students enrolled both in BS(CS) and BS(EE) whereas Faisalabad and Lahore also have significant number of students enrolled in BS(EE) and BBA.
- Where in Karachi and Lahore the maximum students enrolled are from the same city , but Islamabad also has a good number of students from Rawalpindi , Faisalabad has a good number of students from Lahore too and Peshawar has good number of students from both Lahore and Hyderabad.
- With male student count superseding female count in nearly every campus , Peshawar campus has the highest ratio of males which is more than 90% of the total students.
- In Faisalabad , the number of students who cancelled their degree are greater than the ones who completed.

Data Cleaning

Missing Values in Dataset

```
##       STUDENT_ID       SEM_1_SGPA       SEM_1_CGPA       SEM_2_SGPA
##                0             5783             5786            10812
##       SEM_2_CGPA       SEM_3_SGPA       SEM_3_CGPA       SEM_4_SGPA
##             7628            12977            12081            15550
##       SEM_4_CGPA       SEM_5_SGPA       SEM_5_CGPA       SEM_6_SGPA
##            13571            16788            15994            18509
##       SEM_6_CGPA       SEM_7_SGPA       SEM_7_CGPA       SEM_8_SGPA
##            16859            19377            18690            20947
##       SEM_8_CGPA       SEM_9_SGPA       SEM_9_CGPA      SEM_10_SGPA
##            19469            28122            27433            32003
##      SEM_10_CGPA      SEM_11_SGPA      SEM_11_CGPA      SEM_12_SGPA
##            31549            34667            34389            36599
##      SEM_12_CGPA      SEM_13_SGPA      SEM_13_CGPA      SEM_14_SGPA
##            36398            37748            37598            38473
##      SEM_14_CGPA      SEM_15_SGPA      SEM_15_CGPA        TOTAL_SEM
##            38372            38807            38751                0
##           GENDER            BATCH           CAMPUS        PROG_CODE
##                0                0                0                0
##             CGPA        FIRST_SEM         LAST_SEM           STATUS
##             5572                0                0                0
##             CITY        SECONDARY           SCHOOL        SEC_GRADE
##                0             1195             5587             1195
## HIGHER_SECONDARY          COLLEGE    HIG_SEC_GRADE
##             4481            10506             4481
```

- The duplicate student ids within rows were removed from the dataset
- To cater inconsistencies in school name and college name , upper casing was done and extra spaces were removed.
- To cater null values in categorical variables such as school name , college name , secondary , higher secondary row removal was done.
- For numerical attributes such as the sgpa , secondary grade , higher secondary grade mean imputation was done .
- For cgpa , to fill null values , the proper cgpa calculation was done using spga.
- Columns for sgpa and cgpa of semester above 8 were dropped , since most of the values in the column were null.
- Mean imputation was also done to fill missing values of grade points of courses.

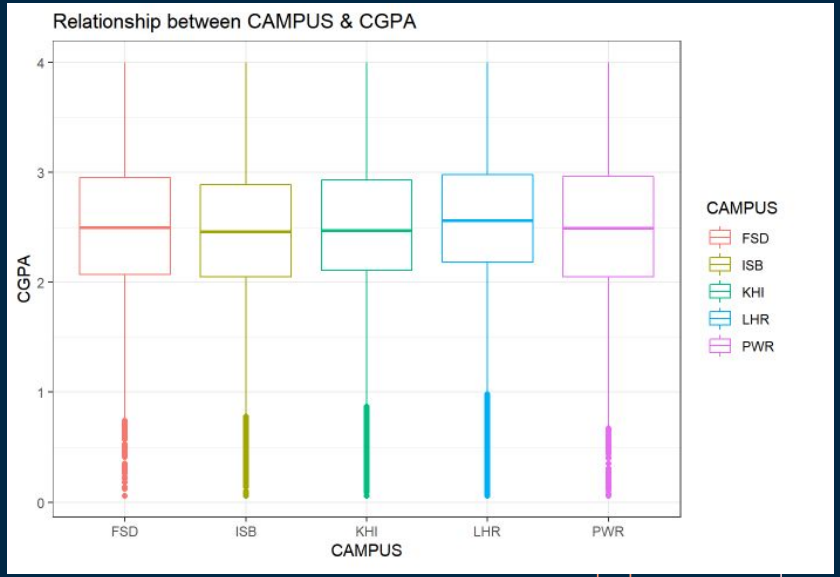
Results and Discussions

Working on the whole data

Null Hypothesis 1: Campus affect CGPA
Alternative Hypothesis 1: Campus doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## CAMPUS         4    5.9  1.4819   7.505 5.25e-06 ***
## Residuals   2495  492.7  0.1975
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Rejected !!
P value < Significance Value



Null Hypothesis 2: Degree Program affect CGPA
Alternative Hypothesis 2: Degree Program doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## PROG_CODE      5   5.89  1.1774   5.826 2.52e-05 ***
## Residuals   1194 241.29  0.2021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Rejected !!
P value < Significance Value

Null Hypothesis 3: Gender affect CGPA
Alternative Hypothesis 3: Gender doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## GENDER         1  14.57  14.571   73.23 <2e-16 ***
## Residuals   1198 238.36   0.199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Rejected !!
P value < Significance Value

Null Hypothesis 4: City affect CGPA
Alternative Hypothesis 4: City doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## CITY          16   7.34  0.4590   2.461 0.00113 **
## Residuals   1003 187.12  0.1866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 5: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 5: Secondary Education (SSC/O Level) doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## SECONDARY      1   2.93  2.9344    14.2 0.000174 ***
## Residuals    998 206.19  0.2066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value == Significance Value

Null Hypothesis 6: School affect CGPA
Alternative Hypothesis 6: School doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## SCHOOL       32   8.73  0.2728   1.507 0.0378 *
## Residuals   627 113.54  0.1811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted !!
P value > Significance Value

Null Hypothesis 7: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 7: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## HIGHER_SECONDARY   1   4.96   4.959   24.49 8.75e-07 ***
## Residuals        998 202.07   0.202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected !!
P value < Significance Value

Null Hypothesis 8: College affect CGPA
Alternative Hypothesis 8: College doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## COLLEGE      51  13.45  0.2638   1.379 0.0428 *
## Residuals   988 189.07  0.1914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 9: Admission Year affect CGPA
Alternative Hypothesis 9: Admission Year doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## FIRST_SEM    12    3.6  0.3000   1.419  0.152
## Residuals   637  134.7  0.2114
```

Hypothesis Accepted!!

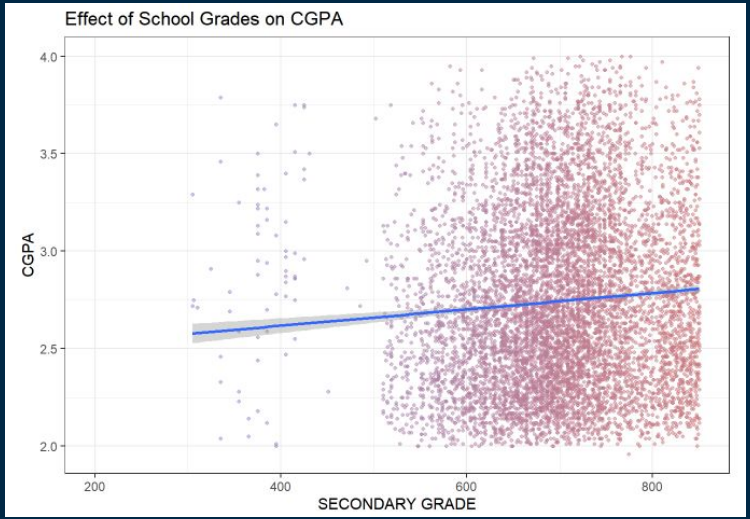
Null Hypothesis 10: Graduation Year affect CGPA
Alternative Hypothesis 10: Graduation Year doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## LAST_SEM      30  66.68  2.2228   16.55 <2e-16 ***
## Residuals   1519 204.07  0.1343
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected !!
P value < Significance Value

For the remaining attributes we calculated correlation coefficient .

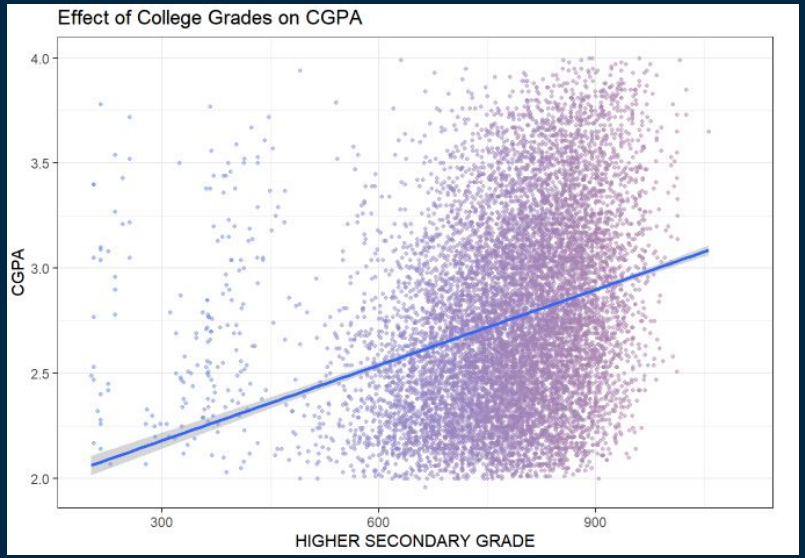
School Grade vs CGPA




```
##
##  Pearson's product-moment correlation
##
## data:  data$SEC_GRADE and data$CGPA
## t = 9.459, df = 10741, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07210533 0.10961323
## sample estimates:
##       cor
## 0.09089151
```

The value of 0.09 shows that there is no correlation between Secondary Grade and CGPA.

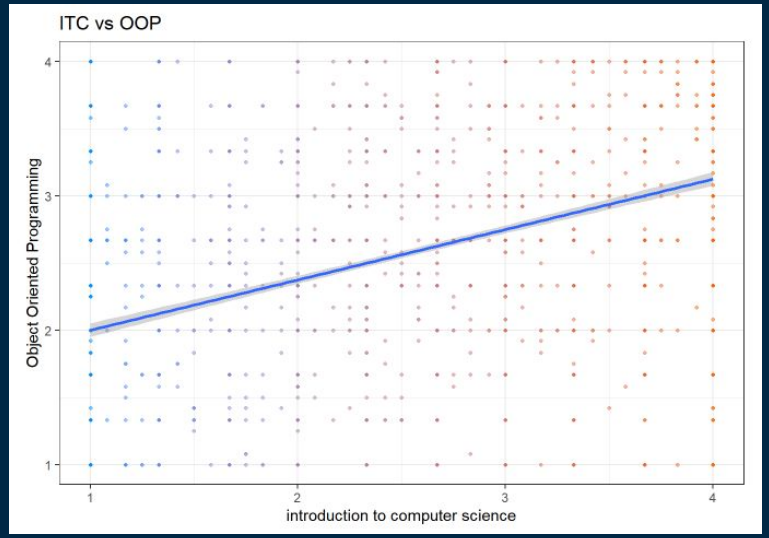
College Grade vs CGPA




```
##
##  Pearson's product-moment correlation
##
## data:  data$HIG_SEC_GRADE and data$CGPA
## t = 27.917, df = 10741, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2423769 0.2776395
## sample estimates:
##       cor
## 0.2600949
```

Higher Secondary Grade and CGPA show a very weak correlation.

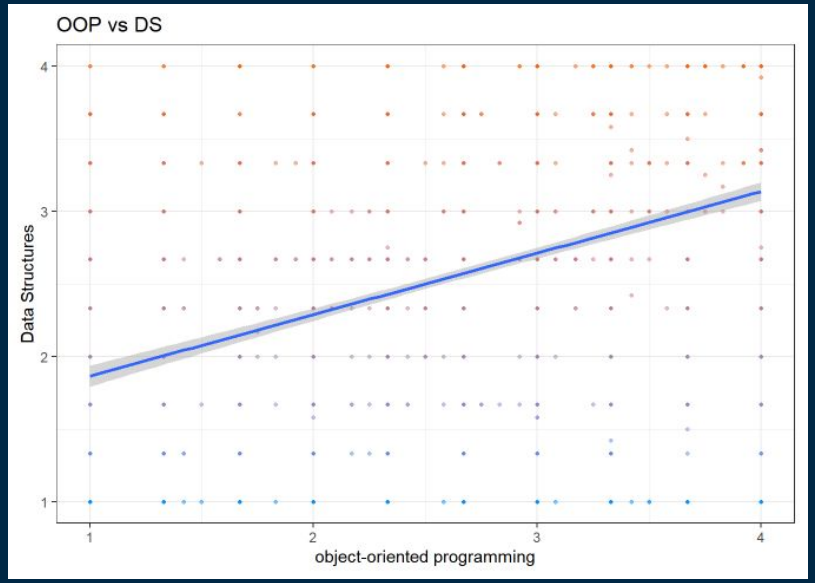
ITC vs OOP




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`object-oriented programming` and CS_courses$`introduction to computer science`
## t = 25.652, df = 3679, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3617646 0.4165830
## sample estimates:
##       cor
## 0.3895187
```

The value 0.38 of correlation coefficient shows that somehow performance of Introduction to Computing and Object Oriented Programming is related.

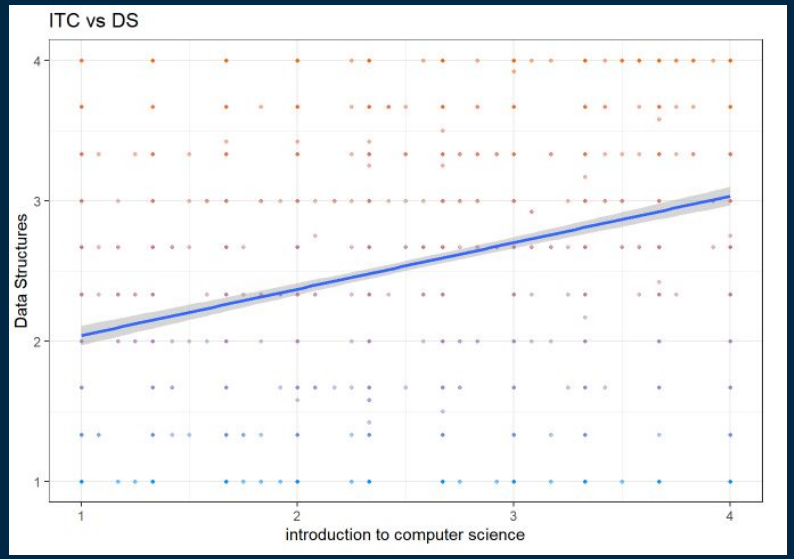
OOP vs DS




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`object-oriented programming` and CS_courses$`data structures`
## t = 21.18, df = 2114, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3826000 0.4529283
## sample estimates:
##       cor
## 0.4183911
```

The courses Object Oriented Programming and Data Structures also have a positive relationship.

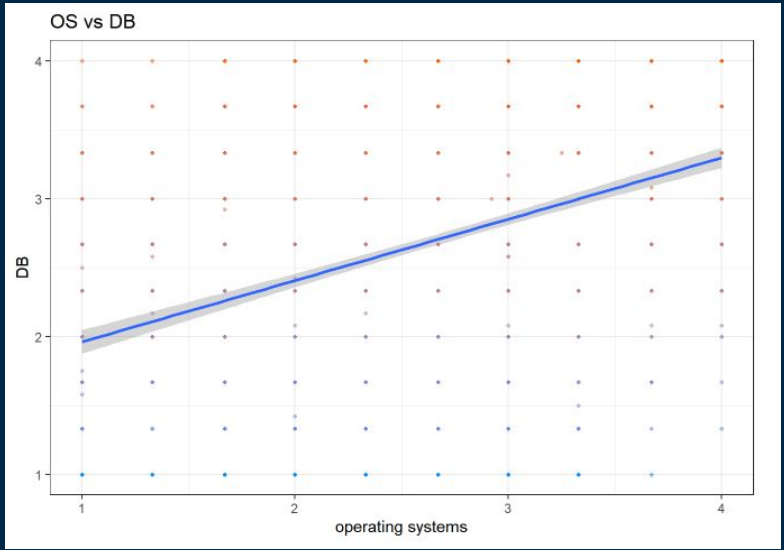
ITC vs DS




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`introduction to computer science` and CS_courses$`data structures`
## t = 16.653, df = 2114, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3023191 0.3776760
## sample estimates:
##       cor
## 0.3405443
```

As compared to OOP , Introduction to Computing affects the grade in Data Structures less than it does in OOP.

OS vs DB




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`operating systems` and CS_courses$`database systems`
## t = 18.952, df = 1424, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4063716 0.4893245
## sample estimates:
##       cor
## 0.4488144
```

With the correlation coefficient of 0.44 Operating Systems and Database Systems show a relation.

Summary of Feature Selection

Features	Evaluation Metric					
	Overall	Faislabad	Islamabad	Karachi	Lahore	Peshawar
Degree	No	Yes	Yes	No	Yes	Yes
Gender	No	Yes	No	No	No	Yes
City	Yes	Yes	Yes	Yes	Yes	Yes
Secondary Education	Yes	Yes	No	Yes	No	Yes
School	Yes	Yes	Yes	Yes	Yes	Yes
Higher Secondary	No	No	No	Yes	No	Yes
College	Yes	Yes	Yes	Yes	Yes	Yes
Admission Year	Yes	Yes	Yes	Yes	Yes	No
Graduation Year	No	No	No	No	No	No
School Grade	No	No	No	No	No	No
College Grade	No	No	No	No	No	No

Faisalabad Campus

Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## PROG_CODE     1  0.301  0.3014   1.563  0.214
## Residuals    98 18.903  0.1929
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## GENDER        1   1.54  1.5383   8.737 0.0035 **
## Residuals   198  34.86  0.1761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## CITY          2  1.441  0.7204   3.871 0.0245 *
## Residuals    87 16.192  0.1861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 4: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 4: Secondary Education (SSC/O Level) doesn't affect CGPA .

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## SECONDARY    1  1.336  1.3363   6.153 0.0148 *
## Residuals   98 21.284  0.2172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn't affect CGPA .

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## SCHOOL      12  5.244  0.4370   2.582 0.00448 **
## Residuals  117 19.802  0.1693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## HIGHER_SECONDARY   1  3.833   3.833   25.13 3.27e-06 ***
## Residuals         78 11.894   0.152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected!!
P value << Significance Value

Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## COLLEGE      11  3.702  0.3366   1.719 0.0786 .
## Residuals   108 21.147  0.1958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn't affect CGPA .

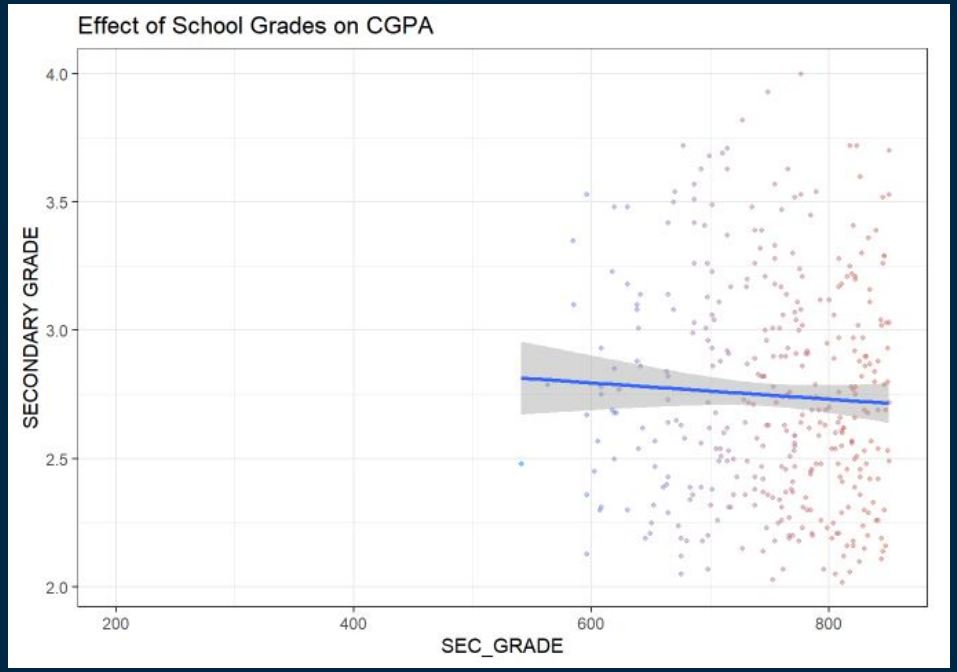
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## FIRST_SEM     3  1.612  0.5375   3.296  0.023 *
## Residuals   116 18.916  0.1631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn't affect CGPA .

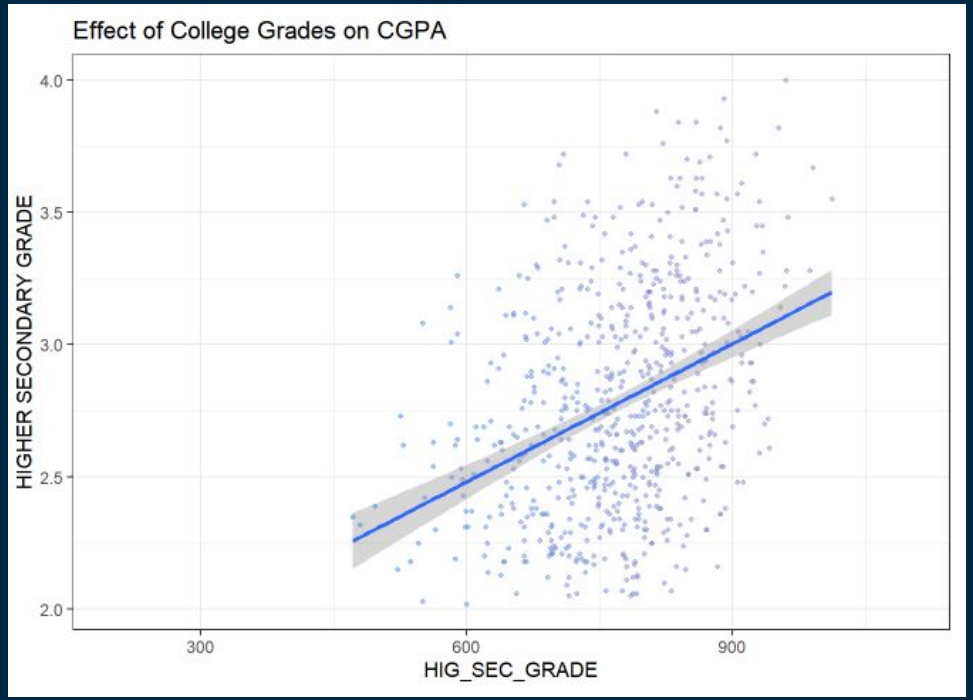
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## LAST_SEM      6  4.491  0.7485   5.247 4.79e-05 ***
## Residuals   203 28.961  0.1427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected!!
P value << Significance Value




```
##
##  Pearson's product-moment correlation
##
## data:  FAISALABAD_data$SEC_GRADE and FAISALABAD_data$CGPA
## t = 1.8394, df = 659, p-value = 0.0663
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.004814656  0.146928797
## sample estimates:
##       cor
## 0.07147059
```

A really weak correlation between school grades and CGPA




```
##
##  Pearson's product-moment correlation
##
## data:  FAISALABAD_data$HIG_SEC_GRADE and FAISALABAD_data$CGPA
## t = 7.6245, df = 659, p-value = 8.578e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2130837 0.3533042
## sample estimates:
##       cor
## 0.2847162
```

Higher Secondary Grade and CGPA show a very weak correlation.

Islamabad Campus

Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## PROG_CODE     4   1.54  0.3856   2.088 0.0803 .
## Residuals   995 183.72  0.1846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## GENDER         1  14.14  14.136   69.14 2.29e-16 ***
## Residuals   1298 265.36   0.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Rejected!!
P value < Significance Value

Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## CITY           6   1.87  0.3120   1.637  0.137
## Residuals    273  52.02  0.1905
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 4: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 4: Secondary Education (SSC/O Level) doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## SECONDARY     1   5.02   5.024   25.13 6.35e-07 ***
## Residuals   998 199.54   0.200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Rejected!!
P value < Significance Value

Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## SCHOOL       18   7.10  0.3942   1.986 0.00993 **
## Residuals   361  71.66  0.1985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## HIGHER_SECONDARY   1   3.26   3.262   16.03 6.7e-05 ***
## Residuals        998 203.07   0.203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected!!
P value << Significance Value

Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## COLLEGE      22   6.81  0.3095   1.507 0.0666 .
## Residuals   437  89.76  0.2054
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn't affect CGPA .

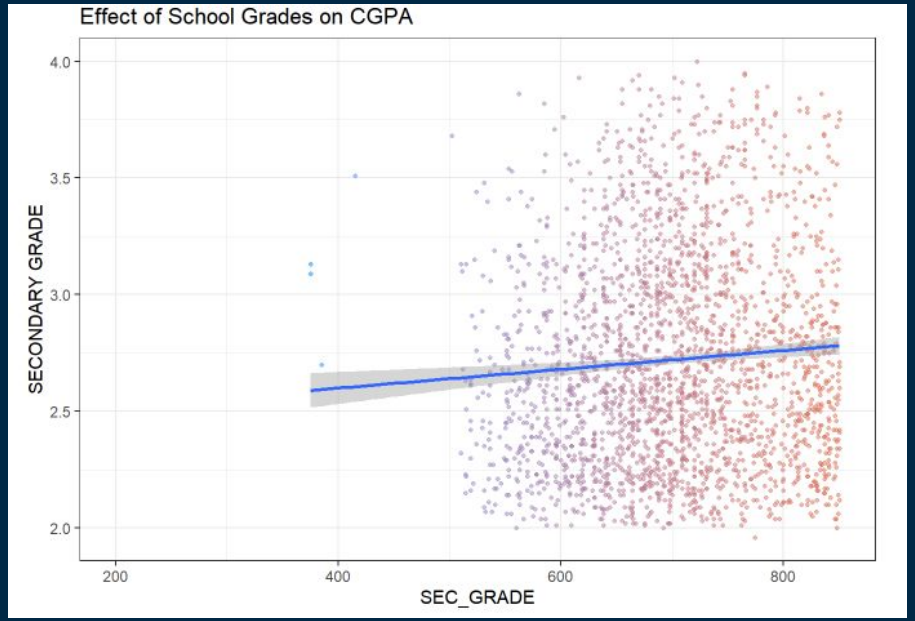
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## FIRST_SEM    10   5.02  0.5018   2.436 0.0076 **
## Residuals   539 111.03  0.2060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn't affect CGPA .

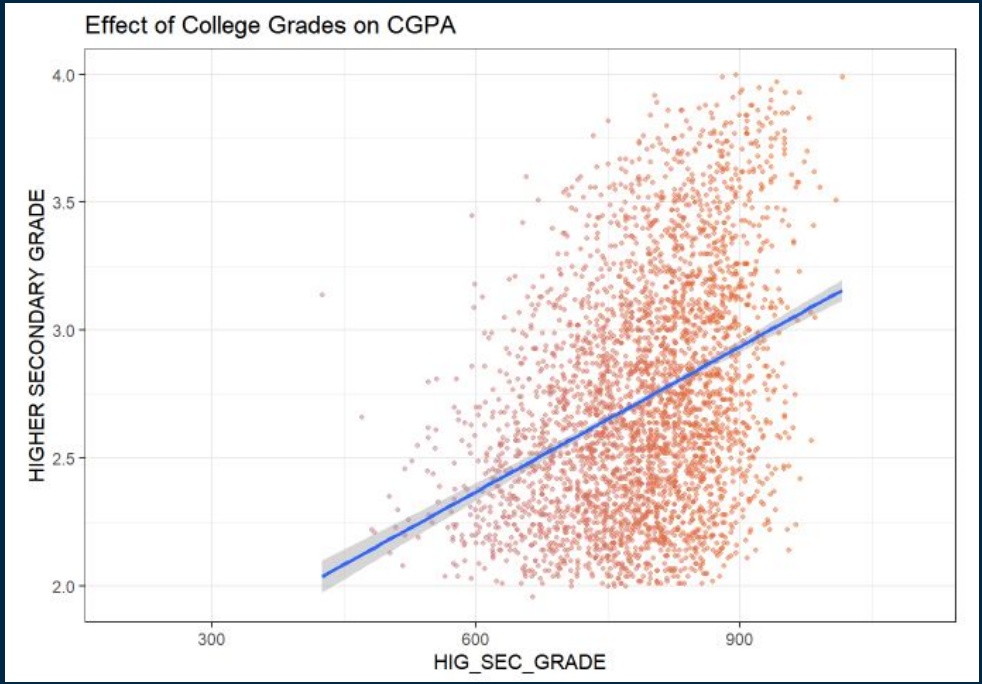
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## LAST_SEM     15  33.36  2.2238   13.93 <2e-16 ***
## Residuals   784 125.13  0.1596
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected!!
P value << Significance Value




```
##
##  Pearson's product-moment correlation
##
## data:  ISLAMABAD_data$SEC_GRADE and ISLAMABAD_data$CGPA
## t = 3.6636, df = 3344, p-value = 0.0002526
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02940458 0.09690491
## sample estimates:
##       cor
## 0.06322705
```

A really weak correlation between school grades and CGPA




```
##
##  Pearson's product-moment correlation
##
## data:  ISLAMABAD_data$HIG_SEC_GRADE and ISLAMABAD_data$CGPA
## t = 21.783, df = 3344, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3224721 0.3818303
## sample estimates:
##       cor
## 0.3525057
```

College grades and CGPA are not highly correlated.

KARACHI Campus

Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## PROG_CODE     3   8.47  2.8218    12.7 4.08e-08 ***
## Residuals   796 176.82  0.2221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis has strong rejection

Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## GENDER        1   4.42   4.416   20.66 6.83e-06 ***
## Residuals   518 110.72   0.214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected !!
P value < Significance Value

Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## CITY          2  1.508  0.7540   4.432 0.0139 *
## Residuals   117 19.903  0.1701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted!!

Null Hypothesis 4: Secondary Education (SSC / O-LEVEL) affect CGPA
Alternative Hypothesis 4: Secondary Education (SSC / O-LEVEL) doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## SECONDARY     1   0.89  0.8930   3.749 0.0534 .
## Residuals   498 118.63  0.2382
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!

Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## SCHOOL        7   2.01  0.2878   1.216  0.297
## Residuals   152  35.97  0.2367
```

Clear Evidence for Hypothesis to be Accepted!

Null Hypothesis 6: Higher Secondary Education (HSSC / A-LEVEL) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A-LEVEL) doesn't affect CGPA .

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## HIGHER_SECONDARY   1   1.42  1.4236   6.351  0.012 *
## Residuals        498 111.62  0.2241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!
P-value == Significance Value

Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## COLLEGE      10   4.13  0.4127   1.786 0.0647 .
## Residuals   209  48.29  0.2311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!

Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn't affect CGPA .

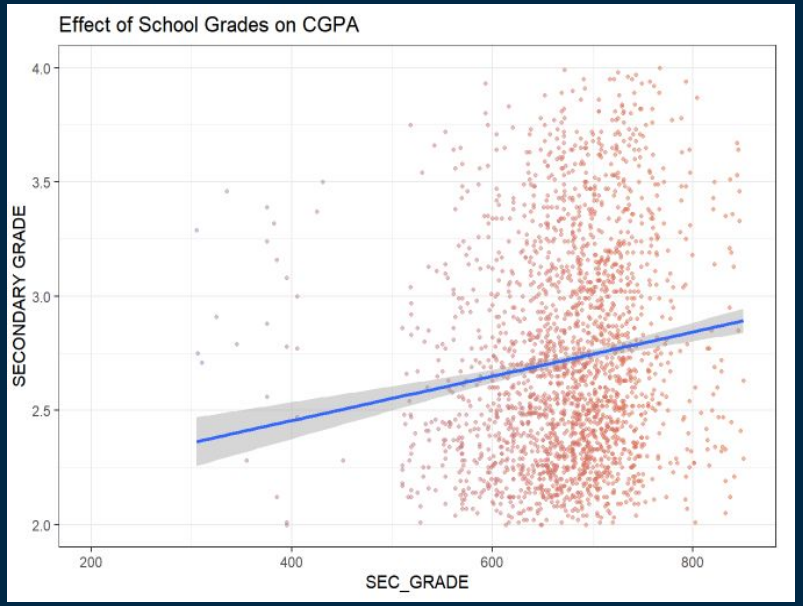
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## FIRST_SEM    12   2.17  0.1806   0.826  0.623
## Residuals   377  82.42  0.2186
```

Null Hypothesis Accepted!

Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn't affect CGPA .

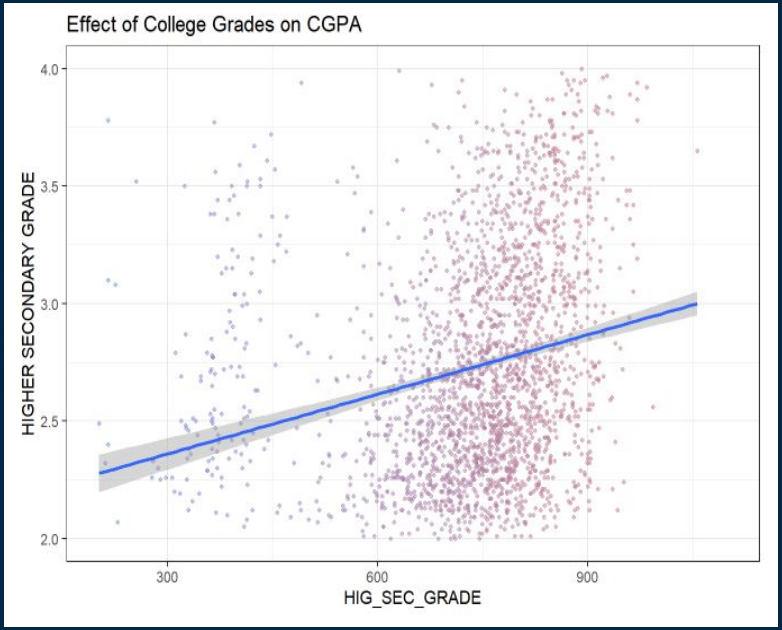
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## LAST_SEM     13  12.21  0.9395   4.388 5.83e-07 ***
## Residuals   406  86.92  0.2141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Rejected!
P-value is much small.




```
##
##  Pearson's product-moment correlation
##
## data:  KARACHI_data$SEC_GRADE and KARACHI_data$CGPA
## t = 8.654, df = 2367, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1358136 0.2138917
## sample estimates:
##      cor
## 0.175128
```

There is no correlation between Secondary Grades and CGPA.




```
##
##  Pearson's product-moment correlation
##
## data:  KARACHI_data$HIG_SEC_GRADE and KARACHI_data$CGPA
## t = 9.8485, df = 2367, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1594056 0.2367845
## sample estimates:
##       cor
## 0.1984042
```

There exist weak relationship between College Grades and CGPA.

Lahore Campus

Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## PROG_CODE     4   3.19  0.7982   4.357 0.00169 **
## Residuals   995 182.28  0.1832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## GENDER         1  17.84  17.836   99.94 <2e-16 ***
## Residuals   1098 195.96   0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Rejected!!
P value < Significance Value

Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn't affect CGPA .

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## CITY           5  1.415  0.2831    1.99 0.0823 .
## Residuals    174 24.744  0.1422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 4: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 4: Secondary Education (SSC/O Level) doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## SECONDARY     1   4.03   4.030    21.6 3.94e-06 ***
## Residuals   798 148.91   0.187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Rejected!!
P value < Significance Value

Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## SCHOOL       18  6.998  0.3888   2.128 0.00678 **
## Residuals   171 31.247  0.1827
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## HIGHER_SECONDARY   1   5.32   5.320   27.72 1.8e-07 ***
## Residuals        798 153.15   0.192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected!!
P value << Significance Value

Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## COLLEGE      23   7.04  0.3063   1.747  0.022 *
## Residuals   216  37.87  0.1753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn't affect CGPA .

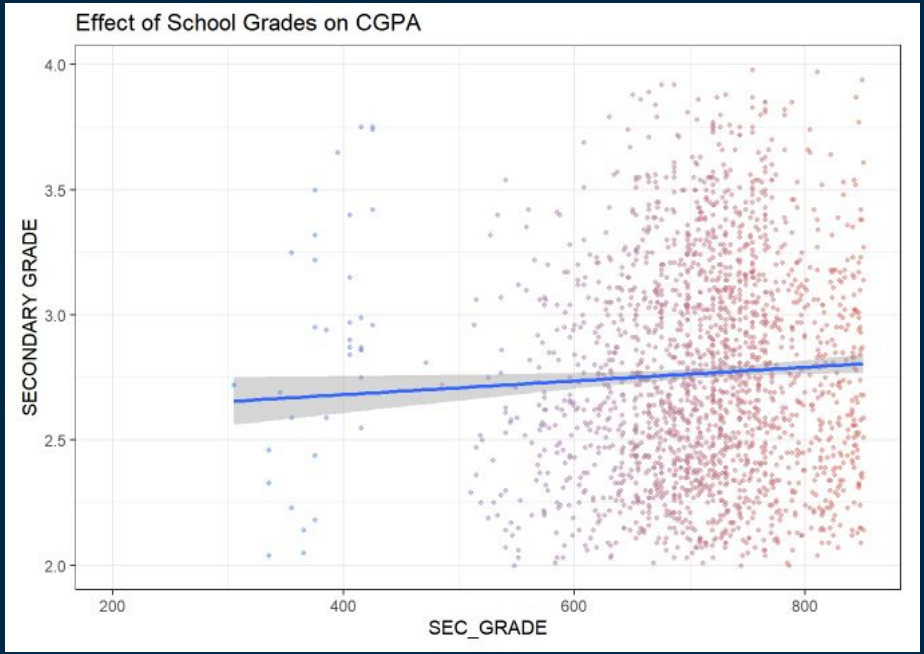
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## FIRST_SEM     7   1.73  0.2467   1.511  0.164
## Residuals   232  37.89  0.1633
```

Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn't affect CGPA .

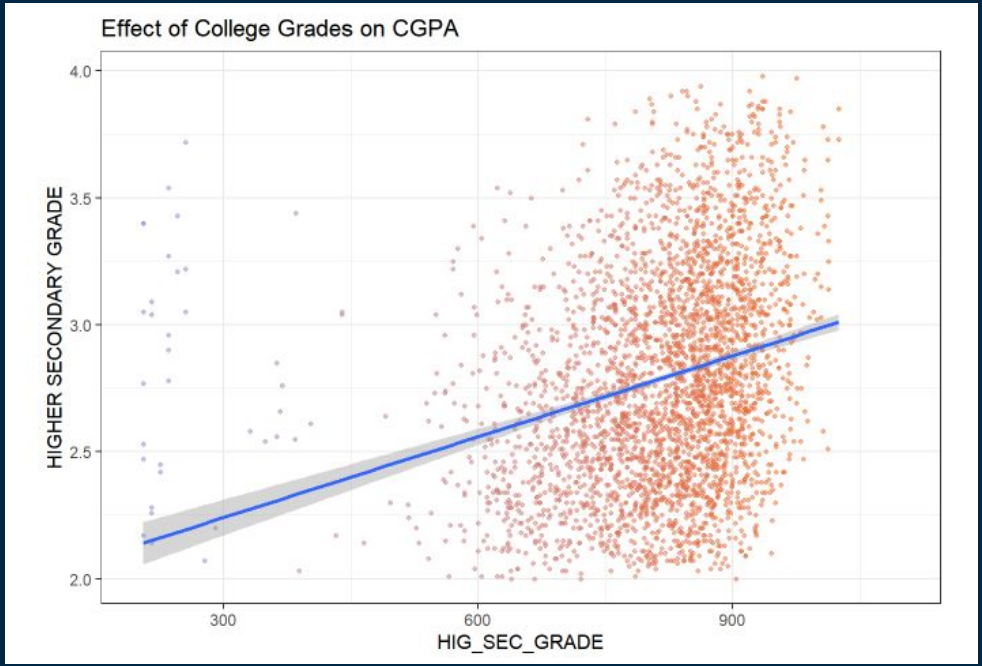
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## LAST_SEM     20  31.07  1.5534   12.69 <2e-16 ***
## Residuals   609  74.57  0.1225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected!!
P value << Significance Value




```
##
##  Pearson's product-moment correlation
##
## data:  LAHORE_data$SEC_GRADE and LAHORE_data$CGPA
## t = 5.1042, df = 3184, p-value = 3.516e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05553715 0.12442619
## sample estimates:
##       cor
## 0.09008943
```

A really weak correlation between school grades and CGPA




```
##
##  Pearson's product-moment correlation
##
## data:  LAHORE_data$HIG_SEC_GRADE and LAHORE_data$CGPA
## t = 14.453, df = 3184, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2152479 0.2804291
## sample estimates:
##       cor
## 0.2481194
```

College grades and CGPA do not show strong correlation.

PESHAWAR Campus

Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## PROG_CODE     3   2.46  0.8207   3.501  0.017 *
## Residuals   156  36.57  0.2345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## GENDER        1   1.48  1.4809    5.66 0.0183 *
## Residuals   198  51.80  0.2616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!

Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## CITY          3  1.824  0.6081   2.963 0.0351 *
## Residuals   116 23.808  0.2052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 4: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 4: Secondary Education (SSC/O Level) doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## SECONDARY     1  0.206  0.2059   1.224  0.276
## Residuals    38  6.394  0.1683
```

Null Hypothesis Accepted!!

Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn't affect CGPA .

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## SCHOOL       10  5.635  0.5635   3.395 0.000756 ***
## Residuals    99 16.430  0.1660
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## HIGHER_SECONDARY  1  0.986  0.9860   4.414 0.0423 *
## Residuals        38  8.489  0.2234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Accepted!!

Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn't affect CGPA .

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## COLLEGE        13  1.685  0.1296   0.681  0.779
## Residuals     126 23.976  0.1903
```

Hypothesis Accepted!!
P value > Significance Value

Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn't affect CGPA .

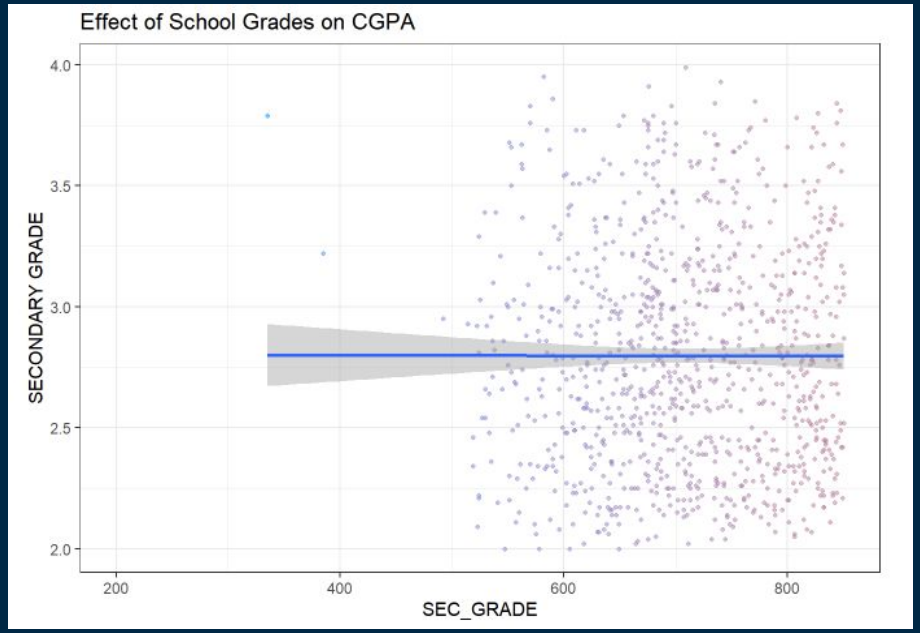
```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## FIRST_SEM    11   7.95  0.7227   3.583 8.4e-05 ***
## Residuals   348  70.18  0.2017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected!!
P value << Significance Value

Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn't affect CGPA .

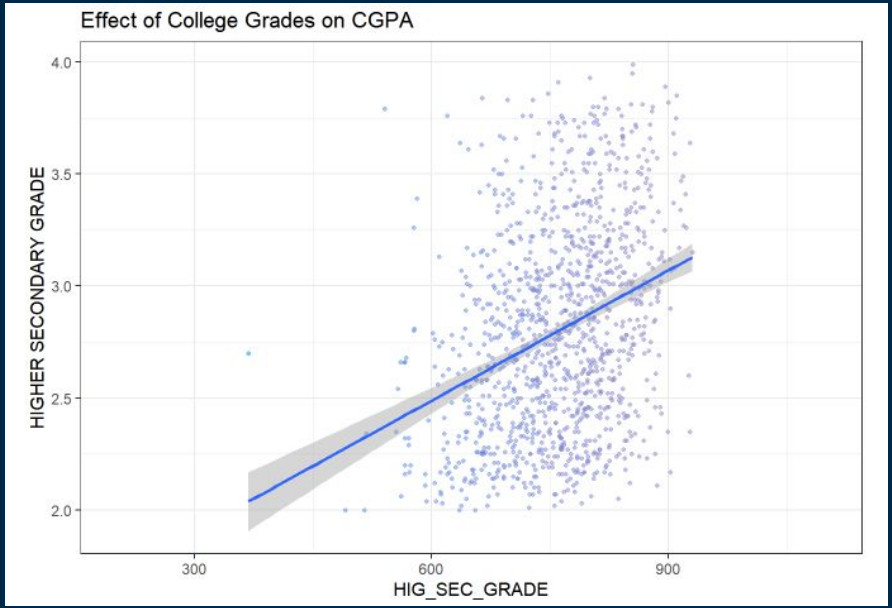
```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## LAST_SEM     12  20.07  1.6723   9.157 1.58e-15 ***
## Residuals   377  68.85  0.1826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Rejected!!
P value << Significance Value




```
##
##  Pearson's product-moment correlation
##
## data:  PESHAWAR_data$SEC_GRADE and PESHAWAR_data$CGPA
## t = 0.25193, df = 1179, p-value = 0.8011
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04972716  0.06435308
## sample estimates:
##         cor
## 0.007336832
```

A really weak correlation between school grades and CGPA




```
##
##  Pearson's product-moment correlation
##
## data:  PESHAWAR_data$HIG_SEC_GRADE and PESHAWAR_data$CGPA
## t = 10.495, df = 1179, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2392471 0.3436149
## sample estimates:
##       cor
## 0.2923011
```

College grades and CGPA do not show strong correlation.

Dashboard Development

- All the exploratory data analysis along with the work of finding correlations (feature selection) is added in the dashboard.

- Programming Language
  Used REACT for the purpose of dashboard development

- Software Tools
  For the purpose of building the dashboard VISUAL STUDIO CODE was used.

Home

Import Data

View Data

Data Visualization

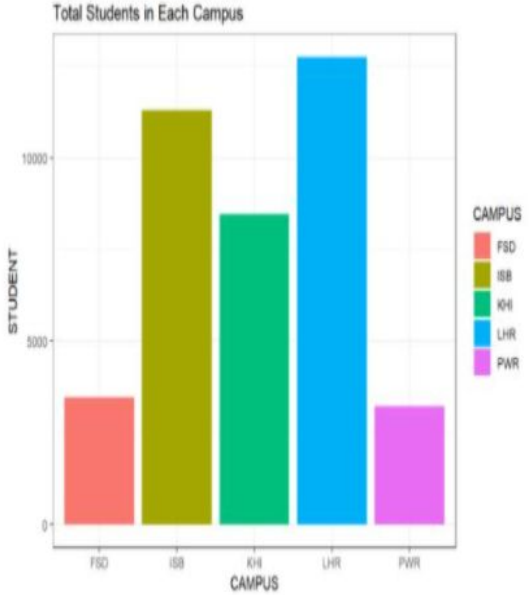
Data Analysis

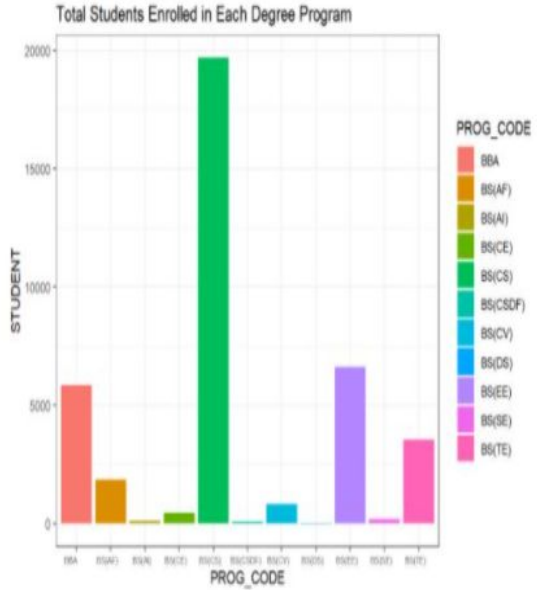
Prediction

Help

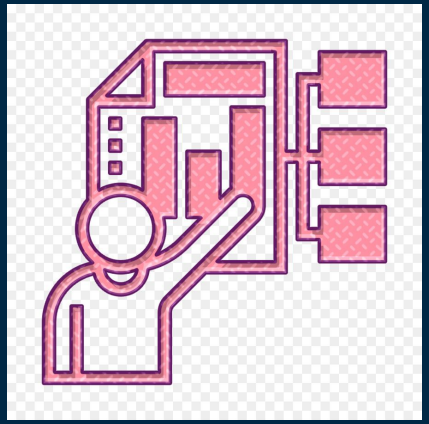
About Us

WHOLE DATA VISUALIZATION





Conclusion and Future Work



- Different factors affect a student's performance and their correlation with student's performance is quite dependent on the data .

- We have selected features

    1. Degree
    2. City
    3. Secondary education
    4. School
    5. College
    6. Admission year

- These features will be further used for building prediction models.
- Model Building : Linear Regression and Logistic Regression , Decision Trees , Random Forest , Support Vector Machine , kNN and k-means clustering .
- Which of the models implemented : linear regression , logistic regression, decision tree, random forest , or support vector machines , KNN and k-means clustering provide the best result .
- Fully functional website showing all the analysis and prediction along with the feature of query processing on the data set

Thank you for listening us!

We will , now , take your queries !