Department of Computer Science
FAST National University of Computer and Emerging Sciences
Karachi Campus

Predictive Analytics on the Academic Record of NUCES

FYP-1 PROGRESS REPORT

Submitted by :

Obaid Ur Rehman
17k-3848

Areeka Ajaz
17k-3913

Tooba Shahid
17k-3731

Submitted On :

17th January 2021

Supervisor :

Dr Jawwad Ahmed Shamsi

Table of Contents

ABSTRACT

Predictive Analytics is the process of using past data to make future predictions . The past data is used to capture important trends with the help of a mathematical model and the model is then used to make predictions on current data . Our aim is to perform predictive analytics on the academic record of NUCES . The basis of any model is the features using which the model is built. In FYP-I we targeted those features which we will be using , in the second phase of our project , to build the predictive model. We used the past academic record of NUCES to make insights and find out correlations between different attributes. Finally a dashboard was built to display the insights and the analytics.

INTRODUCTION

Each year a number of students take admission in FAST NUCES . The students taking admission in FAST NUCES are from different educational backgrounds and different regions . Their academic performance throughout their university life is a reflection of different factors , not only but including their educational background , their previous academic records , the region/district from where they belong etc . The first phase of our project aimed to answer numerous questions about how these factors are related to the performance of students at FAST throughout their educational period.

LITERATURE REVIEW

Predictive analytics has become an influencing factor in improving educational experiences for students. The result of predictive analytics on academic record plays a big role in a way to achieve the highest level of quality of education. This analytics can not only be used to better understand student performance but also to boost graduation rates. Moreover , the predictive model may also help to identify the students who are subject to low performance at an early stage and do the necessary intervention. Hence , early

Student performance prediction can help universities to take appropriate actions on time to improve the success rates of students .

A standard predictive analytics process starts by integrating raw data – from different data sources. This data becomes the basis of the analytics , as this data can be utilized for discovering unknown patterns and trends as well as hidden relationships. However the data in its original form is usually not ready for analysis and modeling. Since the data is usually formed as a combination of different tables , the data contains duplications , missing values and inconsistencies. It is important to know how to handle them without compromising the quality of the prediction. Therefore the data has to go through an initial preparation (cleaning) , before it can be further utilized. All things considered, this cannot be done by a general procedure, and several methods need to be considered within the context of the problem. The main approaches of cleaning data involve listwise deletion and imputation.

Once the data is cleaned , preliminary statistical analysis, especially through visualization, is done which allows to better understand the data. This helps in identifying outliers and imbalance in the data which must be removed for better accuracy of the analysis.

After the preliminary statistical analysis the data preprocessing step begins. In this step , the data undergoes transformation of which the most commonly used methods are normalization and encoding. Then to remove imbalance from the data set either over sampling or under sampling is done. Now that the data has been cleaned and transformed it is ready to be used for finding patterns and trends.

To discover different patterns that can improve students' performance, many studies have been conducted. Especially during the last few years lots of research has been carried out to predict students' academic performance. The research begins with identifying the important factors (feature selection) that affect the students' academic performance. Feature selection, an important strategy to be followed , aims to choose a subset of attributes from the input data. Feature selection enables reduced computation time, improved prediction performance while allowing a better understanding of the data. For our problem , different researchers have identified different factors that affect academic performance .

Abeer Badr El Din Ahmed et. al. , in his study , used the course of the student, mid-term marks, Lab test grade, assignment, attendance, homework, student participation. Another research was carried by Fadhilah Ahmad and Azwa Abdul Aziz in which they used nine parameters like gender, race and hometown, GPA, family income, university entry mode, and grades in related courses. Mohammed M. Abu Tair and Alaa M. El-Halees in his study tried to extract some useful information from student's data of Science and Technology College – Khan Younis. They initially selected different attributes like Gender, date of Birth, Place of Birth, Speciality, Enrollment year, Graduation year, City, Location, Address, Telephone number, HSSC Marks, SSC school type, HSSC obtained the place, HSSC year, College CGPA for analysis. But after preprocessing the data they found that attributes like Gender, Speciality, City, HSSC Marks, SSC school type, College CGPA are most significant. . Jyoti Bansode for predicting student academics performance collected data from Shah and Anchor Kutchhi Polytechnic, Chembur, Mumbai. They considered student attributes like parent's education, parent' s occupation, category, SSC board, admission type, SSC medium, SSC class, first-semester result, second-semester , third-semester, fourth-semester, the fifth-semester and sixth-semester result as most important attributes. Maria Koutina and Katia Lida Kermanidis tried to find out the best techniques for predicting the final grade of the postgraduate students of Ionian University Informatics, Greece. On the basis of reviewed literature, they considered Gender, Age, Marital Status, Number of children, Occupation, Job associated with computers, Bachelor, Another master, Computer literacy, Bachelor in informatics. Mashael A. Al-Barrak and Mona S. AlRazgan collected a dataset of student's from the Information Technology department at King Saud University, Saudi Arabia for their analysis. They further used the different attributes for the prediction like student ID, student name, student grades in three different quizzes, midterm1, midterm2, project, tutorial, final exam, and total points obtained in the Data structure course of computer science department [8]. Edin Osmanbegović and Mirza Suljic collected data from surveys in the midst of first-year students and the data taken during the enrollment at the University of Tuzla. They further used the different attributes for the prediction like Gender, Family, Distance, High School, GPA, Entrance exam, Scholarships, Time, Materials, the Internet, Grade importance, Earnings. Raheela Asif and Mahmood K. Pathan in their study used data from four academic batches of Computer Science & Information Technology (CS & IT) department at NED University, Pakistan. They used HSSC marks , Maths marks in HSSC , and marks in programming courses like Logic design, OOP, DBMS and Data Structures .

To conclude , after the review of different research papers it was found that in most of the cases the factors which affect the student performance are gender, high school grade, student's parental education, financial background, living location, medium of teaching, student's family status, students' previous semester marks, class test grade, seminar performance, assignment performance, general proficiency, attendance in class and lab work, interest in particular course, admission type and previous schools marks .

PROPOSED WORK

Our proposed work for FYP-1 was divided into two parts, the first one being cleaning , transforming and EDA of the data followed by feature selection and the second part was integrating our EDA and feature selection with a dashboard.

Through our feature selection we aimed to answer the following :

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?

Feature selection was to be done using statistical methods such as Pearson correlation , ANOVA to find out correlation between cgpa and different attributes.

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

Student Data :

This data set contained all the relevant detail about a particular student i-e gender , batch , campus , program code , CGPA , first semester , last semester , city , SSC Board , SSC obtained , SSC Total , HSSC Board , HSSC obtained , HSSC Total, O Level Board , O Level Obtained , O Level Total , A Level Board , A Level Obtained , A Level Total , warnings , credits attempted

, credits completed. This all was given against a unique student id.

Semester Data:
This data set contained the 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 i-e for a single student there will multiple rows each for a particular semester.

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 students were split into two , one half was in Course Data 1 and the other in Course Data 2 . Each row showed data about a particular course of a particular student i-e data of each student was given in several rows to cover all his/her courses.

DATA PREPROCESSING

Whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Therefore the data has to go through pre-processing in which it is transformed (as needed) into a form which can be easily used for the analysis.

Student Data

Columns such as warnings , credits attempted , credits completed , SSC Total , HSSC total were dropped. A single column was made for Secondary Education from SSC Board and O Level Board . Also SSC Obtained and O Level Obtained were combined to form a column Secondary Grade. The columns SSC Board and O Level Board were combined to form a column School. Similarly , a single column was made for Higher Secondary Education from HSSC Board and A Level Board . Also HSSC Obtained and A Level Obtained were combined to form a column Secondary Grade. The columns HSSC Board and A Level Board were combined to form a column 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.

Finally above two data sets were joined on unique student ids to form a single data set.

Course Data 1 & Course Data 2

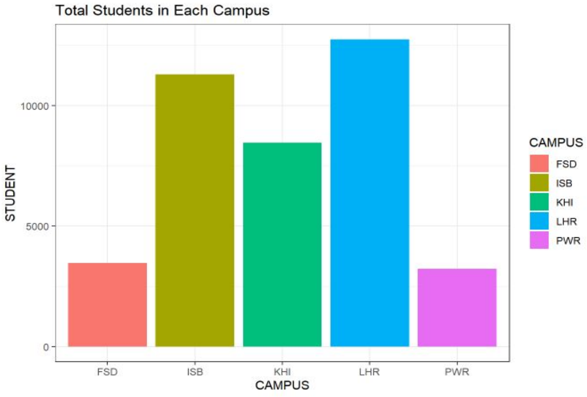
From course code which was given like SS123 the course domain i-e SS was extracted. From the relation id attribute only core courses were retained and the elective courses were dropped , as they were not a part of our FYP scope. Columns that weren't useful were dropped and only columns student id , title , 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 and separate sheets were maintained for each domain to find out relation between different courses of the same domain. Point to be noted is that we chose only those courses which were prerequisites of some other courses , because for finding relationships between courses we only needed the courses in chain.

EXPLORATORY DATA ANALYSIS

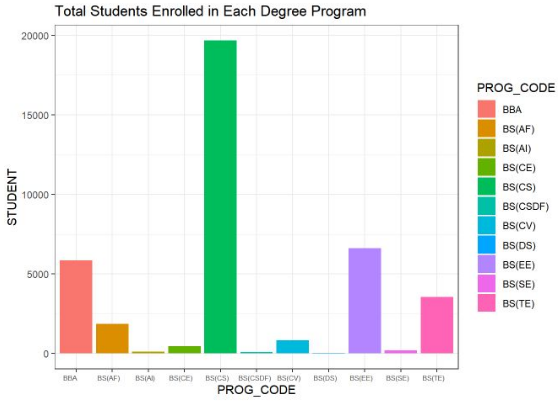
Exploratory Data Analysis is one of the crucial steps in data analytics. Before we jump to learning and modeling the data , EDA was to be performed. In our case , the EDA was first performed for all the data as a whole (all FIVE campuses together) and then separately for each campus .

Visualizing the whole data :

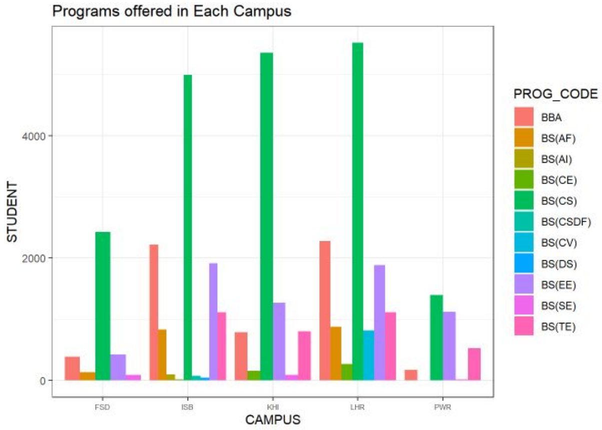
As it was already mentioned that the data provided was for FIVE different campuses , the first bar chart shows the number of students in each campus .



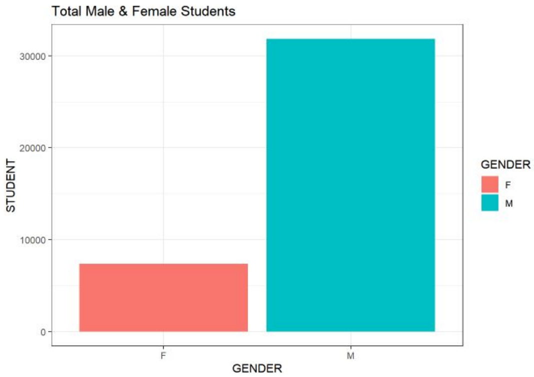
This is quite clear from the above bar chart that for the past 19 years , out of all the five campuses the greatest number of students were enrolled in Lahore Campus .

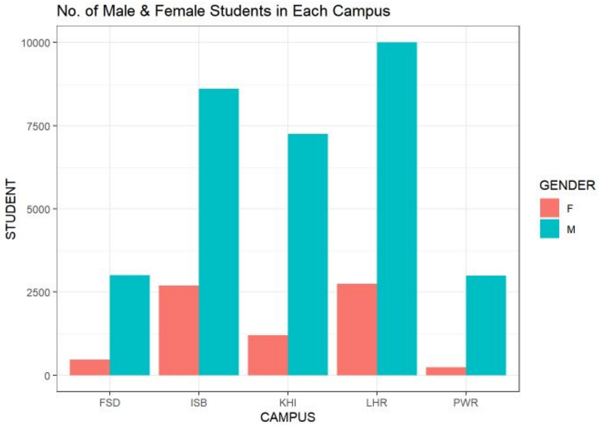


For the past 19 years , 11 different degree programs have been offered . Being the best Computer Science university in Pakistan , the majority of the students at FAST are enrolled in BS(CS) - Bachelors in Computer Science.

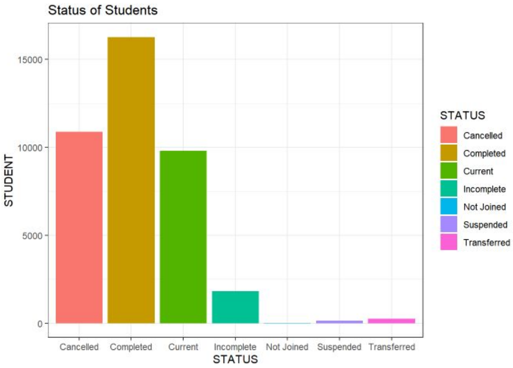


The graphs below are a clear evidence of an imbalance of male and female students. Around 81.22% students enrolled in five campuses of FAST NUCES around the country are found to be males.

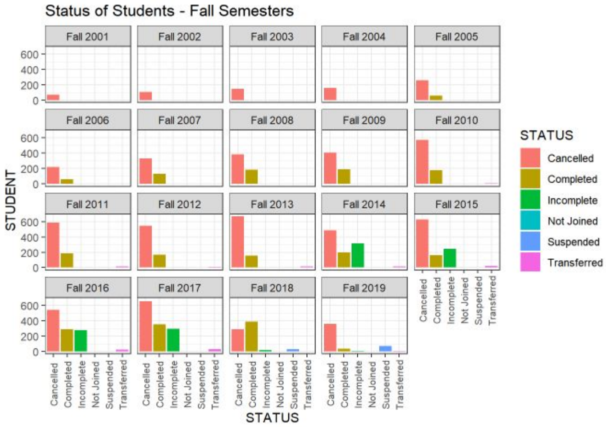


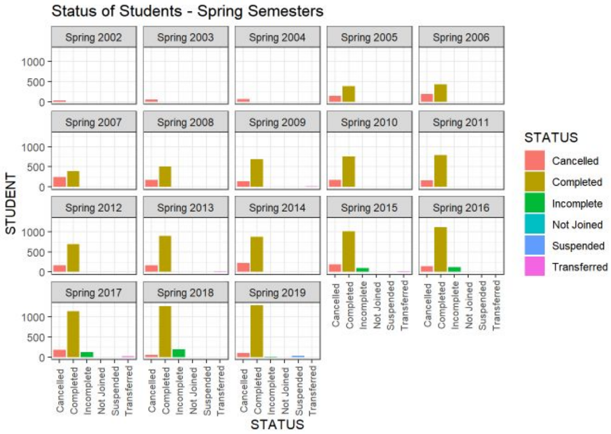


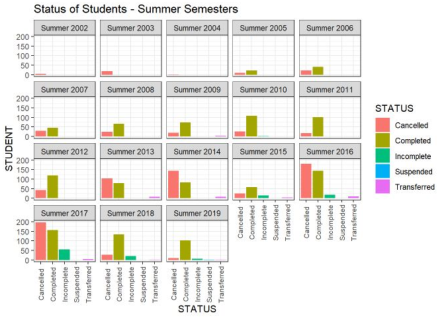
The dataset was provided with an attribute of student status which had 7 different categories.



The count of students with incomplete , not joined , suspended and transferred is negligible as compared to the others therefore these weren't used for the purpose of analysis. Also the students with cancelled status , didn't complete their graduation and not enough data was available about them they were also excluded from the analytics .



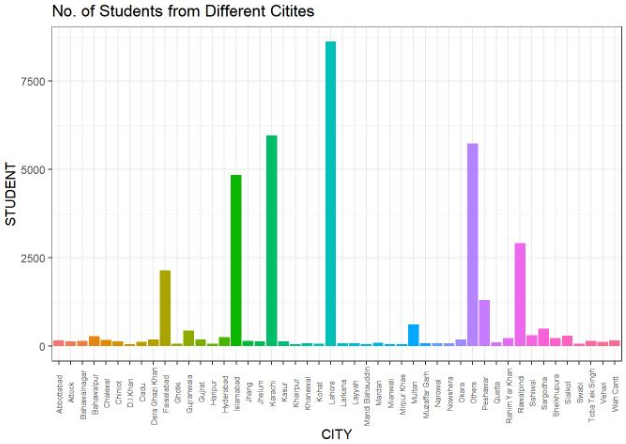


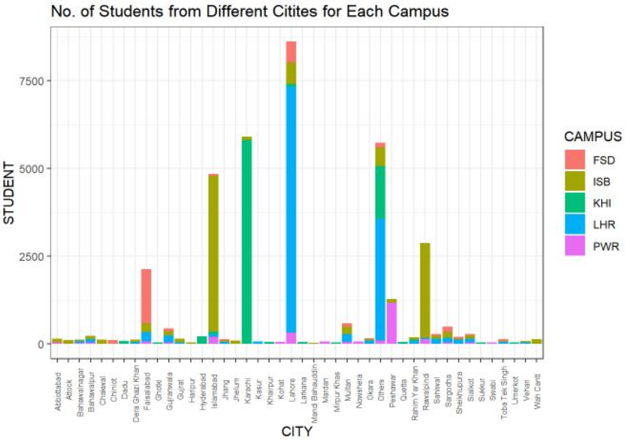


The above graphs are the visualization of students status for different semesters of Fall , Spring and Summer for the past 19 years.

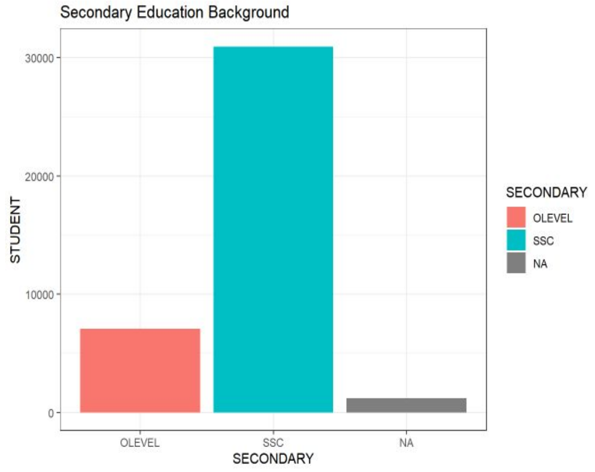
```
## [1] "Total Cities: 151"
```

The students enrolled in different campuses of FAST NUCES are from 151 different cities around the country . Some of the cities have a great majority of students but there are some who have quite negligible student count .

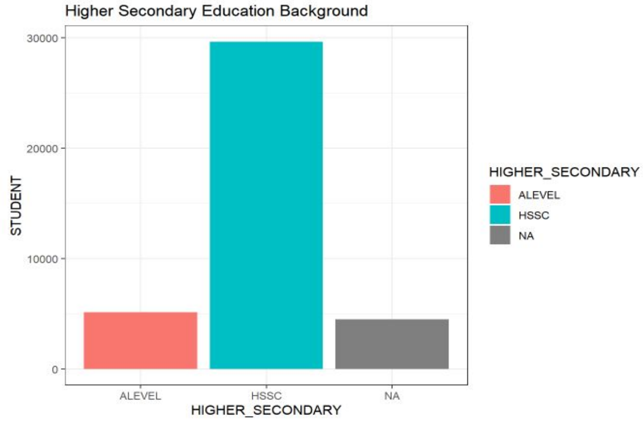




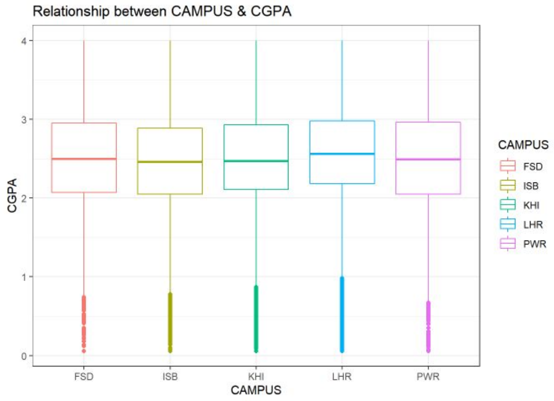
For better visualizations after a threshold of min 50 students , top 48 cities are displayed in both the graphs above.



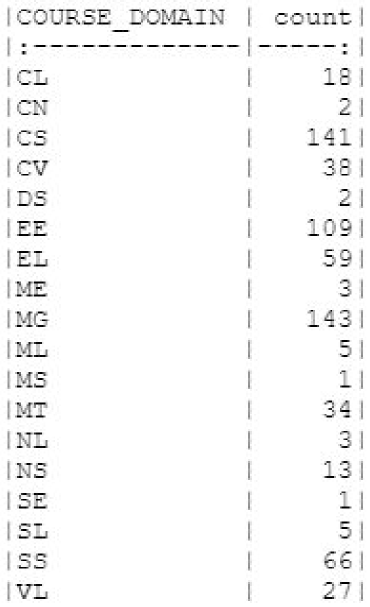
Majority of students , around 78.92% , enrolled in FAST NUCES across the country are from SSC educational background.



Similarly , since most of the students' secondary education is from SSC background , the higher secondary education of most students' , around 75.5% , is from HSSC background.



The boxplot visualizes the campus wise CGPAs. It can be seen that in each campus there are outliers . However , the mean of each campus is quite the same .

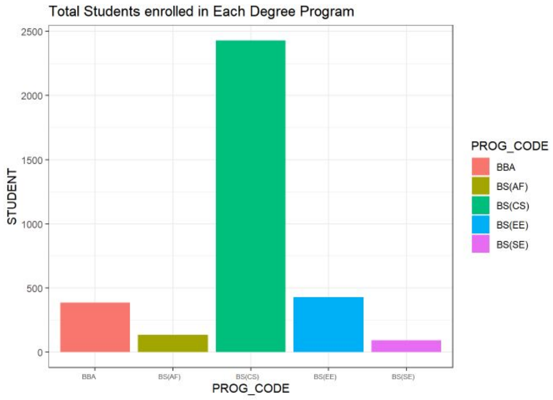


```
|COURSE_DOMAIN | count|
|:-------------|-----:|
|CL            |    18|
|CN            |     2|
|CS            |   141|
|CV            |    38|
|DS            |     2|
|EE            |   109|
|EL            |    59|
|ME            |     3|
|MG            |   143|
|ML            |     5|
|MS            |     1|
|MT            |    34|
|NL            |     3|
|NS            |    13|
|SE            |     1|
|SL            |     5|
|SS            |    66|
|VL            |    27|
```

The total number of courses offered in 5 different campuses in different degree programs for the past 19 years were found to be 641. These courses were from different domains . The above below table shows total courses in each domain.

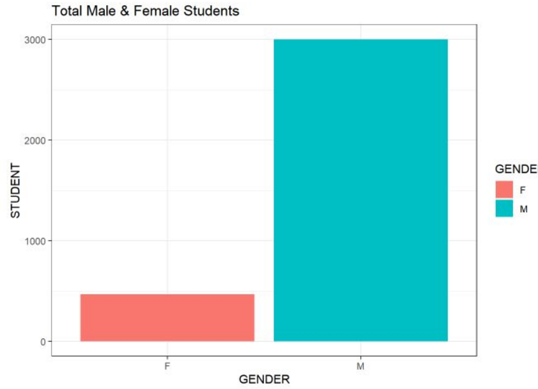
Visualizing the data of FAISALABAD Campus :

At Faisalabad Campus , 5 degrees were offered .



A great many students from Faisalabad are enrolled in BS(CS) whereas BS(EE) and

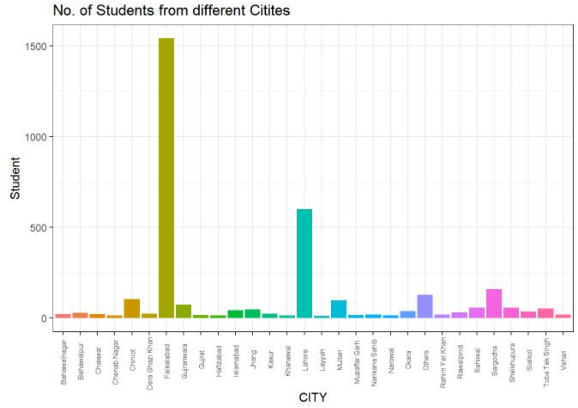
BBA have almost the same student count over the past 19 years .



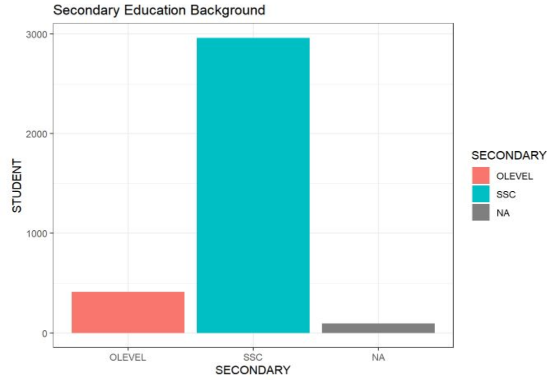
From a total of 3,468 students , around 86.5% students are male .

```
## [1] "Total Cities:  69"
```

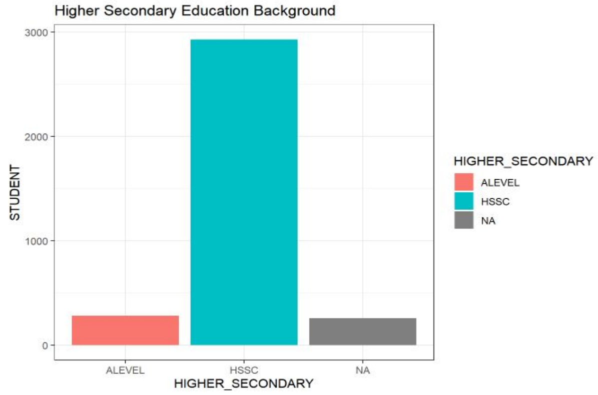
Students enrolled in Faisalabad campus are from 49 different cities. For better visualization , after a threshold of minimum 10 students , 30 cities are displayed in the graph below.



Majority of the students in Faisalabad Campus are from Faisalabad. However from Lahore even there are a representative number of students.



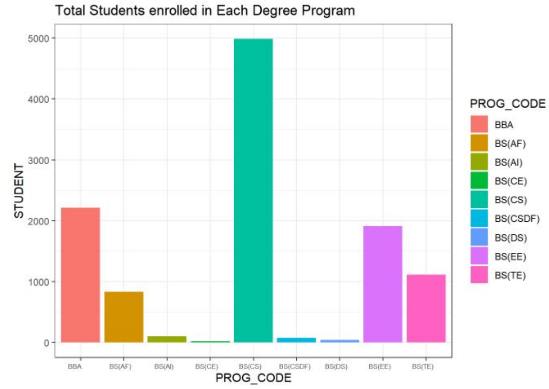
A great majority i-e around 85.38% students had SSC in their Secondary Education.



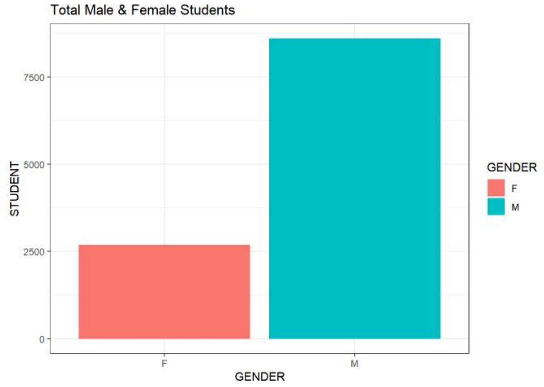
Around 84.48% students had HSSC in their Higher Secondary Education.

Visualizing the data of ISLAMABAD Campus :

Islamabad Campus offered 9 different degree programs.



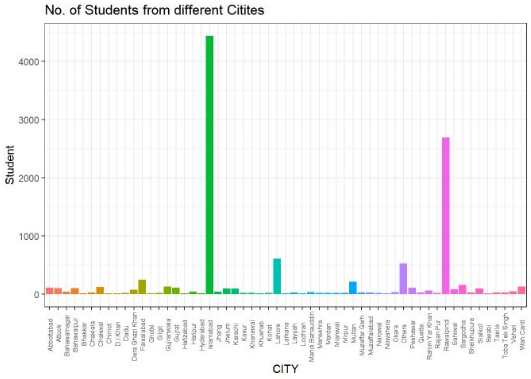
The students enrolled in BS(CS) clearly exceed the other degree programs , but still a significant proportion of students have also enrolled in BBA and BS(EE).



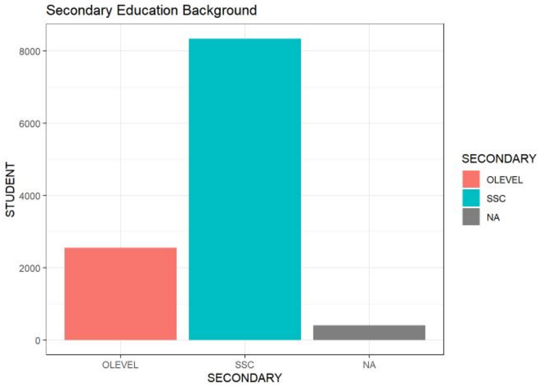
Out of 11,305 students , 76.16% students are male at Islamabad Campus.

```
## [1] "Total Cities:  119"
```

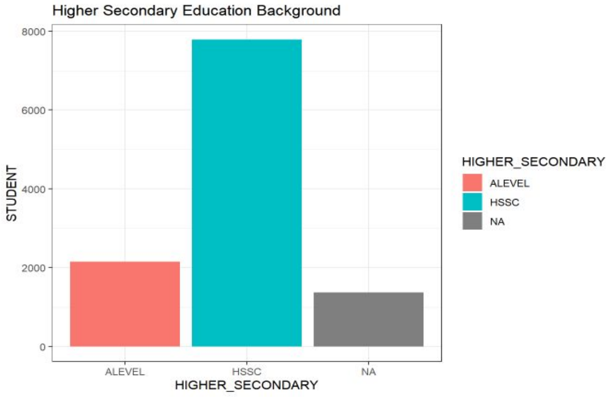
Students from 119 different cities around the country are enrolled at Islamabad Campus. For better visualizations , after a minimum threshold of 10 students , 57 different cities are displayed in the bar chart below .



A significant number of students enrolled in Islamabad Campus are from Islamabad and Rawalpindi.



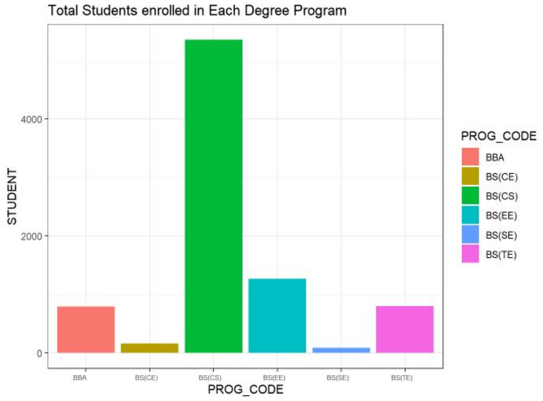
From 11,305 students at Islamabad 73.82% students are from SSC background.



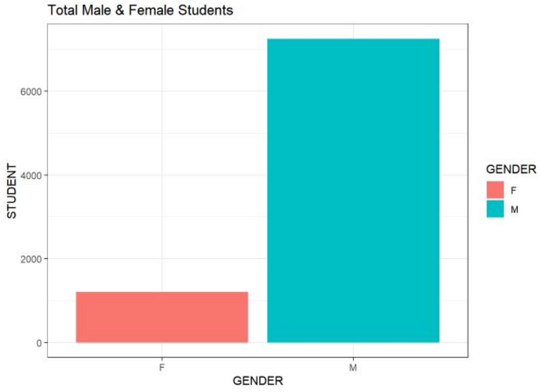
However , in Islamabad , 68.93% students are from HSSC background.

Visualizing the data of KARACHI Campus :

Karachi Campus offered 6 different degree programs .



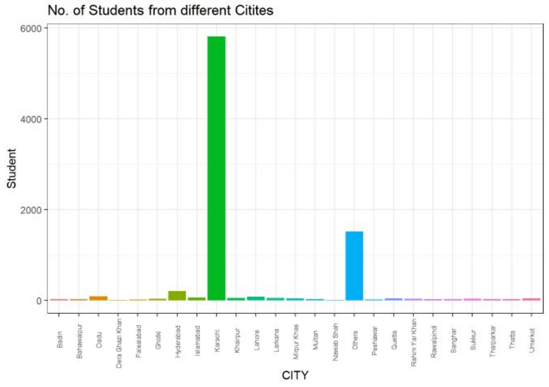
Out of which majority are enrolled in BS(CS) whereas students count in BBA and BS(EE) is nearly the same.



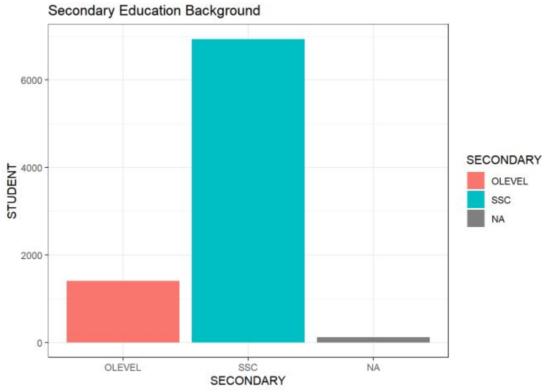
Out of 8,459 students , 85.77% students are male at Karachi Campus.

```
## [1] "Total Cities:  78"
```

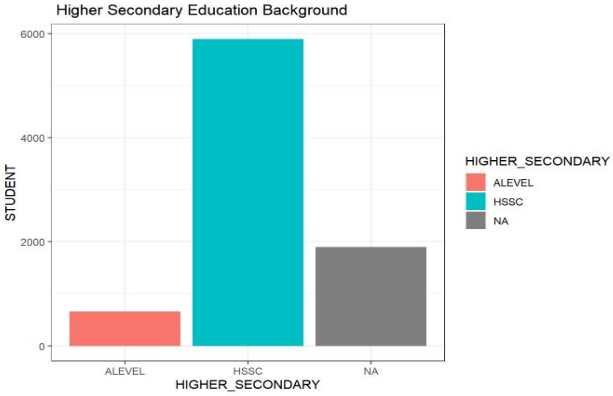
Students enrolled in Karachi Campus are from 78 different cities . However after the threshold of minimum 10 students , 25 cities are displayed in the bar chart below.



The majority of students in Karachi Campus are from Karachi .



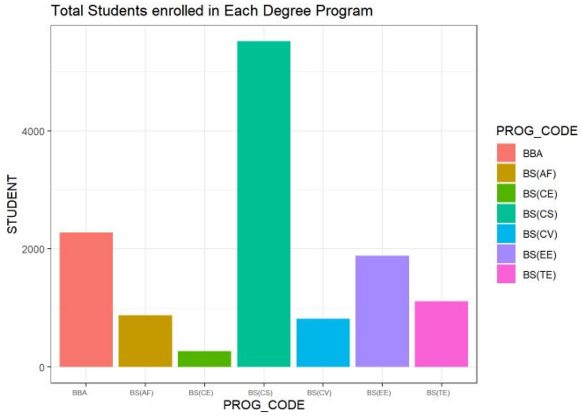
At Karachi Campus , 81.99% students from 8,459 students are from SSC background.



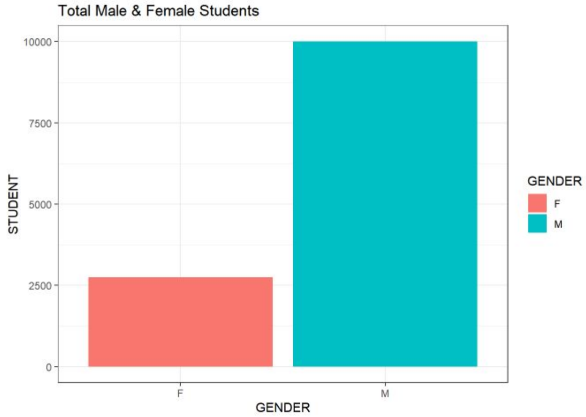
However , 69.74% students from 8,459 students are from HSSC background.

Visualizing the data of LAHORE Campus :

At Lahore Campus 7 different degree programs are offered.



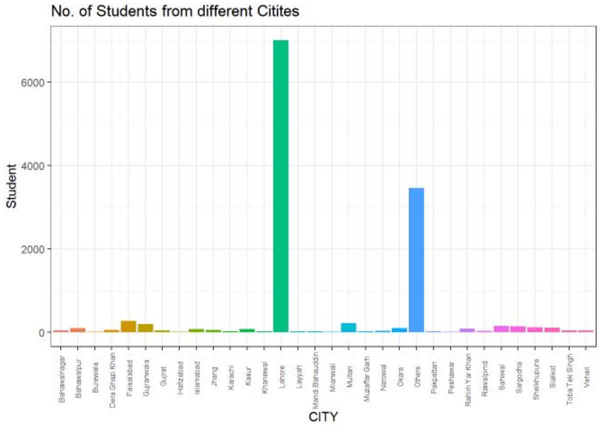
Compared to other students Lahore campus has a number of students enrolled in BS(TE) and also a good number of students in BBA and BS(EE) .



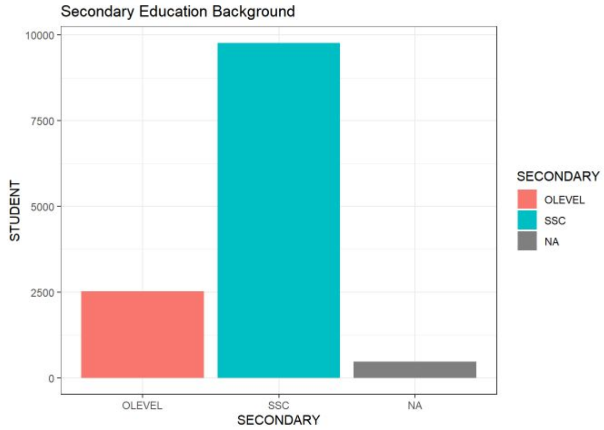
Out of 12,576 students at Lahore , male students are in a majority of 78.48% .

```
## [1] "Total Cities:  80"
```

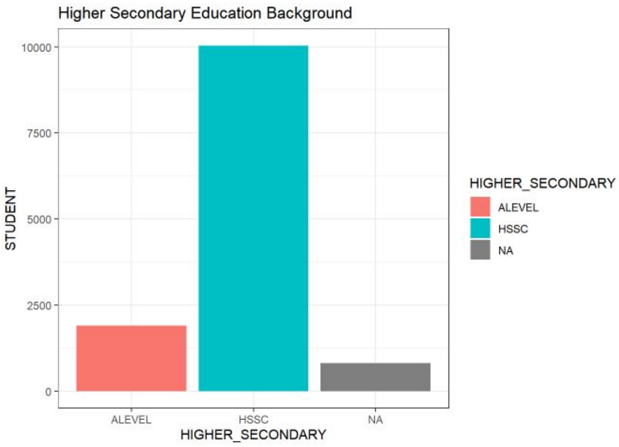
In Lahore Campus , students from 80 different cities are enrolled. After a minimum threshold of 10 students , 32 cities are displayed in the bar chart .



As obvious , the majority of students at Lahore Campus are from Lahore .



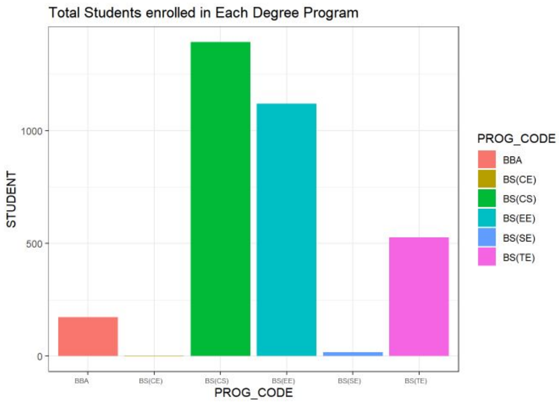
At Lahore Campus , 76.58% students are from SSC background.



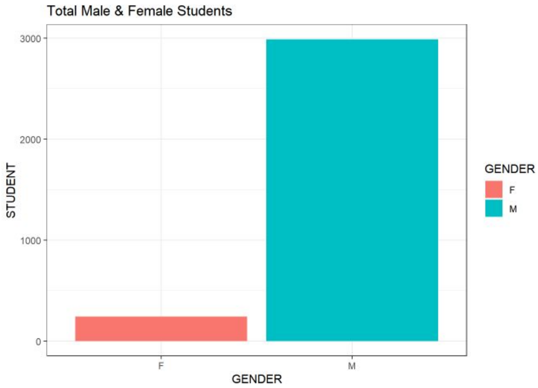
At Lahore Campus , 78.71% students are from HSSC background.

Visualizing the data of PESHAWAR Campus :

Peshawar Campus offered 6 different degree programs .



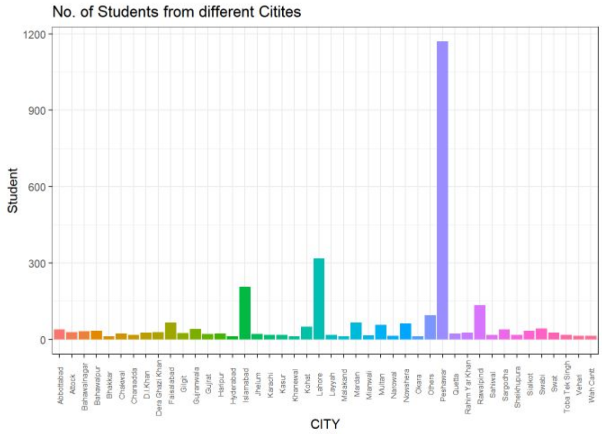
At Peshawar Campus , a significant number of students are from BS(CS) and BS(EE). Also the count of students enrolled in BS(TE) is more than other cities.



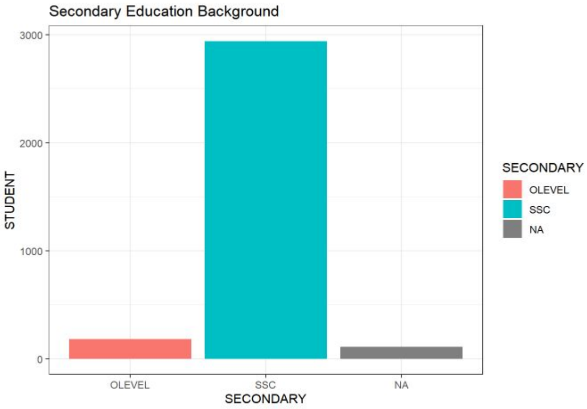
Out of 3,229 students , Peshawar Campus has male students around 92.53% .

```
## [1] "Total Cities:  109"
```

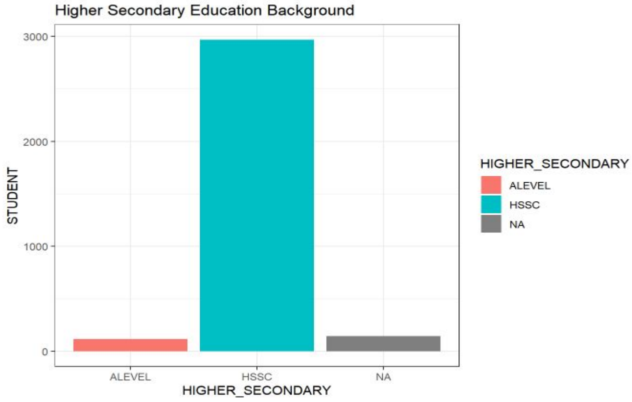
Students from 109 different cities took admission at Peshawar Campus. To visualize at a better level , only 44 cities are displayed after a minimum threshold of 10 students .



In descending order students from Peshawar , Lahore and Hyderabad have majority representation.



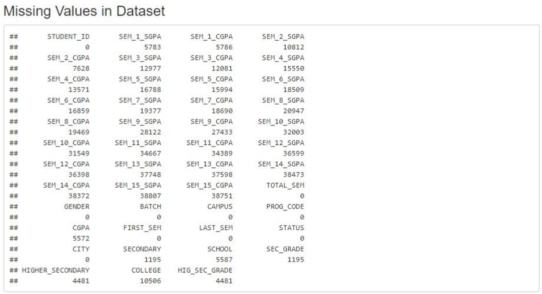
At Peshawar Campus , 91.04% students are from SSC background.



At Peshawar Campus , 91.97% students are from HSSC background.

DATA CLEANING

Before finding out patterns in data , it has to be cleaned against inconsistencies and missing values. In our data set we found a lot of missing data and a few inconsistencies.



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.

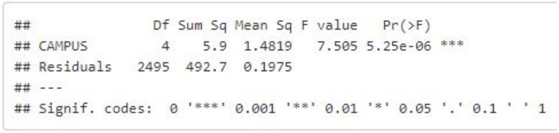
RESULTS AND DISCUSSIONS

Now when the data is pre processed , transformed and cleaned , we move forward to the feature selection . The results here are also explored in two ways: first for the whole data and then separately for each campus.

For attributes where the data set was imbalanced down sampling is done to balance the dataset before finding out correlations.

Working on the whole data :

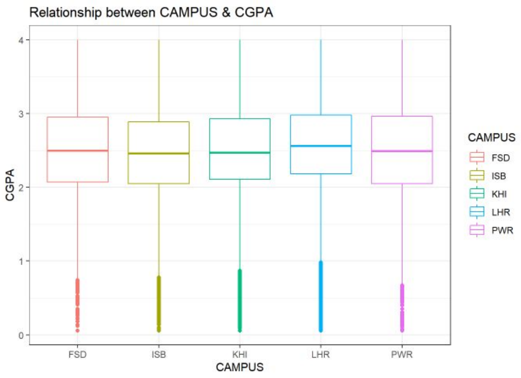
Null Hypothesis 1: Campus affect CGPA
Alternative Hypothesis 1: Campus doesn't affect CGPA .



```
##              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
```

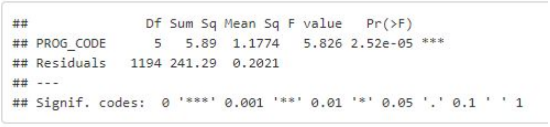
With p value less than significance value it can be seen that the null hypothesis is clearly rejected.

This can also be shown from the box plot below.



It can clearly be seen that the CGPA for each campus is almost the same , so campus doesn't really matter in terms of student performance.

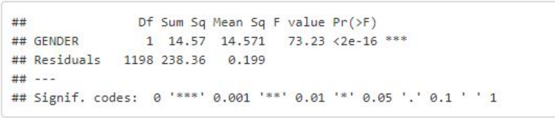
Null Hypothesis 2: Degree Program affect CGPA
Alternative Hypothesis 2: Degree Program doesn't affect CGPA .



```
##              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
```

This null hypothesis is also rejected with p value less than significance value.

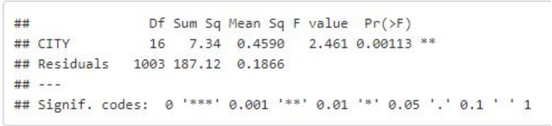
Null Hypothesis 3: Gender affect CGPA
Alternative Hypothesis 3: Gender doesn't affect CGPA .



```
##               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
```

And clearly CGPA doesn't relate to gender.

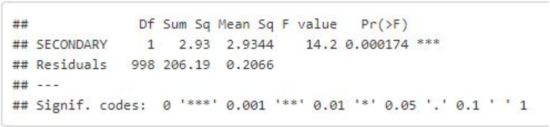
Null Hypothesis 4: City affect CGPA
Alternative Hypothesis 4: City doesn't affect CGPA .



```
##               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
```

Comparing the p value and significance value , we accept the null hypothesis .

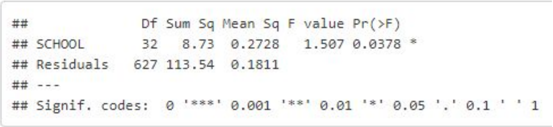
Null Hypothesis 5: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 5: Secondary Education (SSC/O Level) doesn't affect CGPA .



```
##              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
```

With p level almost equal to significance value we accept the null hypothesis.

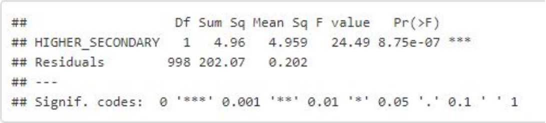
Null Hypothesis 6: School affect CGPA
Alternative Hypothesis 6: School doesn't affect CGPA .



```
##              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
```

As can be seen that p value is greater than significance level , clearly accept the null hypothesis.

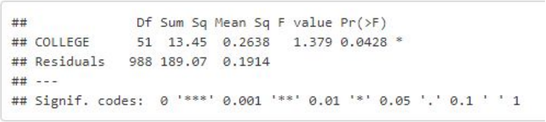
Null Hypothesis 7: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 7: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .



```
##                   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
```

With a p value way too small than the significance value , we reject the null hypothesis.

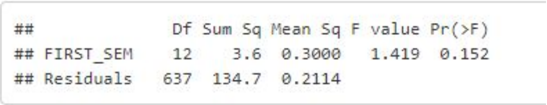
Null Hypothesis 8: College affect CGPA
Alternative Hypothesis 8: College doesn't affect CGPA .



```
##              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
```

Since p value is greater than significance value , the null hypothesis is accepted.

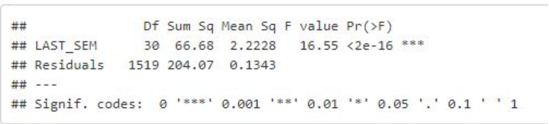
Null Hypothesis 9: Admission Year affect CGPA
Alternative Hypothesis 9: Admission Year doesn't affect CGPA .



```
##              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               
```

With a greater p value , there is no evidence against the null hypothesis hence accepted.

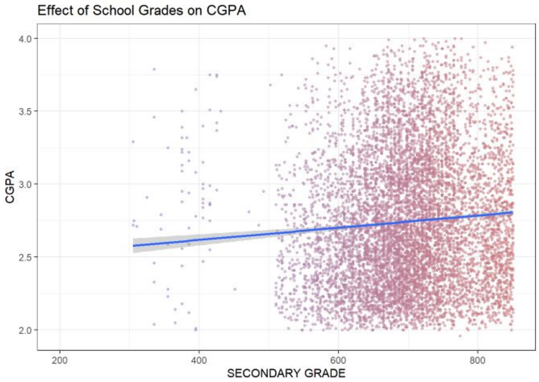
Null Hypothesis 10: Graduation Year affect CGPA
Alternative Hypothesis 10: Graduation Year doesn't affect CGPA .

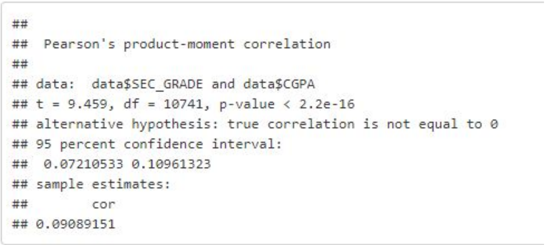


```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## LAST_SEM      30  66.68  2.2228   16.55 <2e-16 ***
## Residuals   1519 204.07  0.1343                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p value is less than the significance value , we will go with the alternative hypothesis.

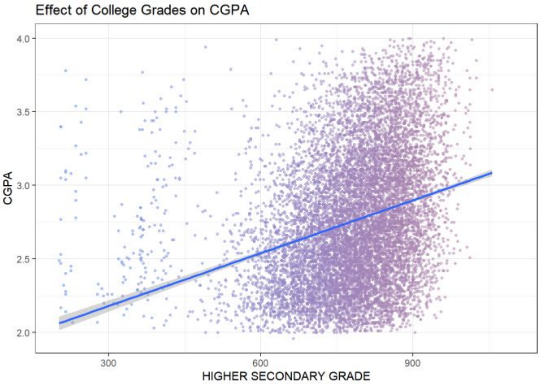
For the remaining attributes we calculate correlation coefficient .

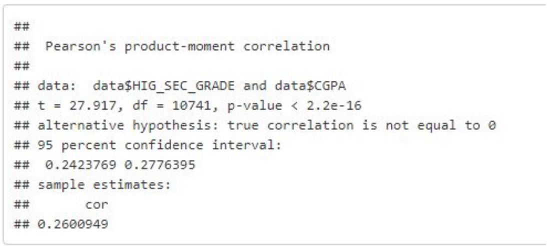




```
##
##  Pearson's product-moment correlation
##
## data:  data$SEC_GRADE and data$CGPA
## t = 9.459, df = 10741, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07210533 0.10961323
## sample estimates:
##       cor
## 0.09089151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

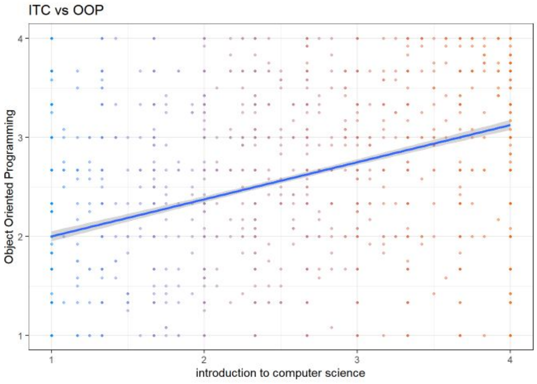
The value of 0.09 shows that there is no correlation between Secondary Grade and CGPA.

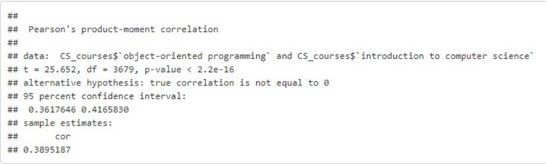




```
##
##  Pearson's product-moment correlation
##
## data:  data$HIG_SEC_GRADE and data$CGPA
## t = 27.917, df = 10741, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2423769 0.2776395
## sample estimates:
##       cor
## 0.2600949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

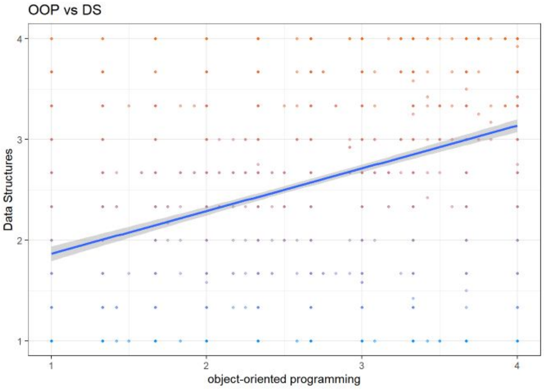
Higher Secondary Grade and CGPA show a very weak correlation.

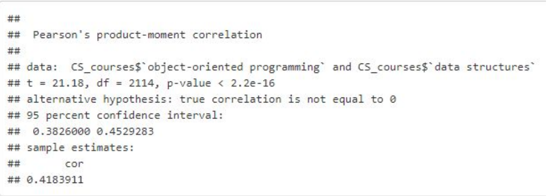




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`object-oriented programming` and CS_courses$`introduction to computer science`
## t = 25.652, df = 3679, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3617646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

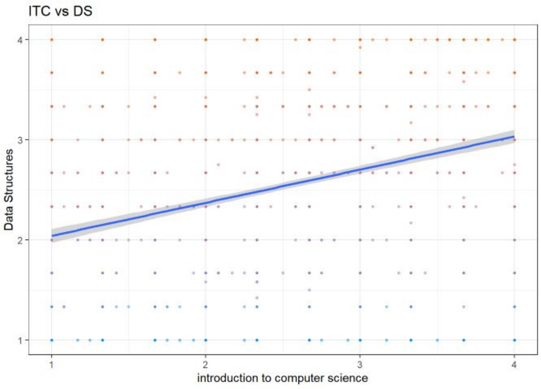
The value 0.38 of correlation coefficient shows that somehow performance of Introduction to Computing and Object Oriented Programming is related.

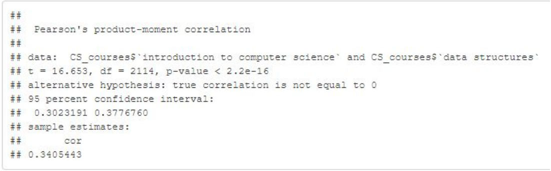




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`object-oriented programming` and CS_courses$`data structures`
## t = 21.18, df = 2114, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3826000 0.4529283
##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

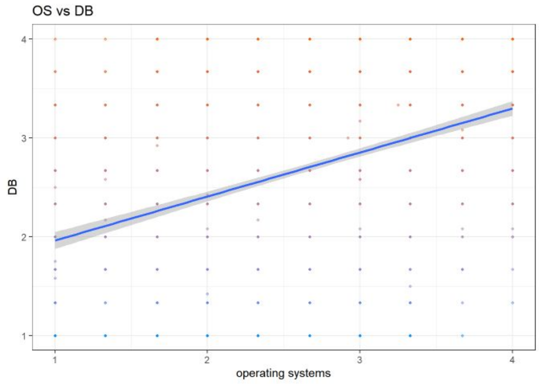
The courses Object Oriented Programming and Data Structures also have a positive relationship.

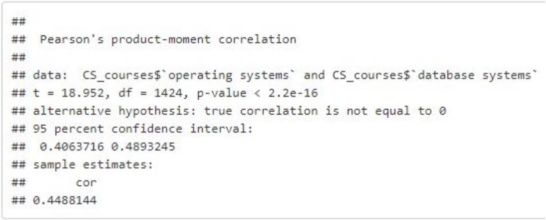




```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`introduction to computer science` and CS_courses$`data structures`
## t = 16.653, df = 2114, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3023191 0.3776760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

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



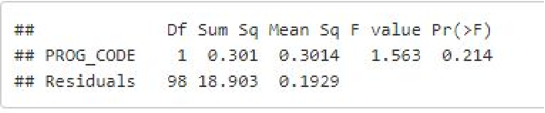


```
##
##  Pearson's product-moment correlation
##
## data:  CS_courses$`operating systems` and CS_courses$`database systems`
## t = 18.952, df = 1424, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4063716 0.4893245
## sample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

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

Faisalabad Campus :

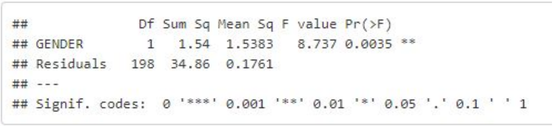
Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn't affect CGPA .



```
##              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
```

With a large p value , the null hypothesis is accepted.

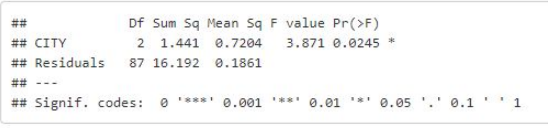
Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn't affect CGPA .



```
##              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
```

With p value slightly equal to significance value we can go with the null hypothesis.

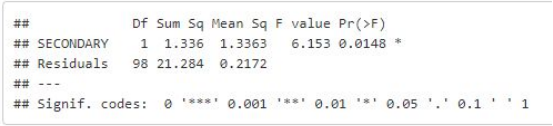
Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn't affect CGPA .



```
##             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
```

A p value greater than significance value shows the clear acceptance of null hypothesis.

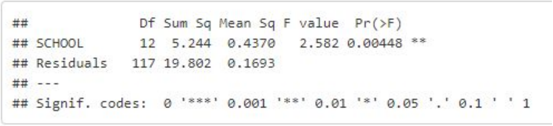
Null Hypothesis 4: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 4: Secondary Education (SSC/O Level) doesn't affect CGPA .



```
##             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
```

Here also the null hypothesis is accepted.

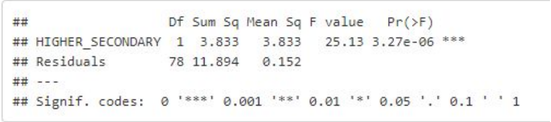
Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn't affect CGPA .



```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## SCHOOL       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 clearly accepted as p value is greater than significance value.

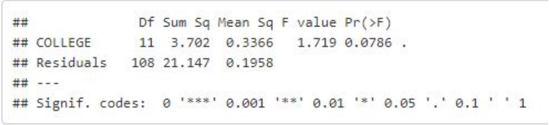
Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .



```
##                  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
```

Reject null hypothesis , p value is very small as compared to significance value.

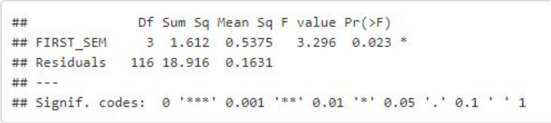
Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn't affect CGPA .



```
##            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
```

Here also we accept the null hypothesis because p value is greater than significance value.

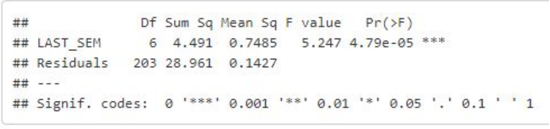
Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn't affect CGPA .



```
##            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
```

The null hypothesis is accepted here too.

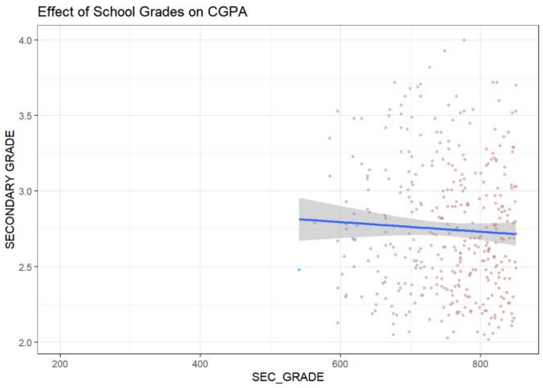
Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn't affect CGPA .

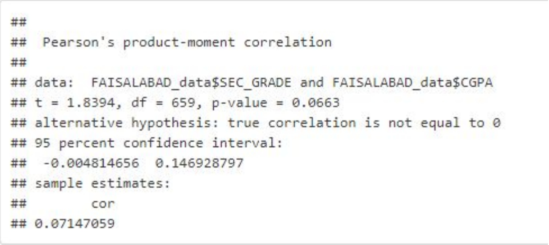


```
##            Df Sum Sq Mean Sq F value   Pr(>F)    
## LAST_SEM     6  4.491  0.7485   5.247 4.79e-05 ***
## Residuals  203 28.961  0.1427                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This hypothesis is rejected because p value is less than significance value.

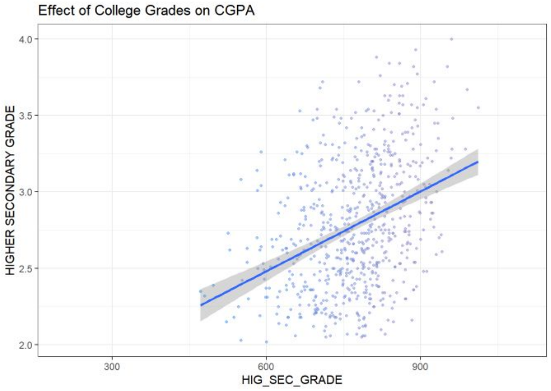
For the remaining attributes we calculate correlation coefficient .

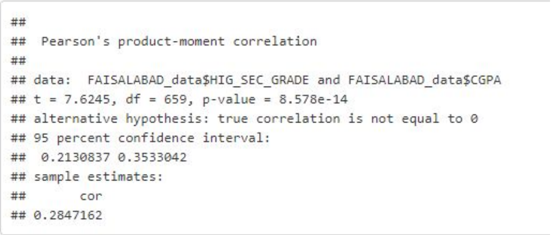




```
## 
##  Pearson's product-moment correlation
## 
## data:  FAISALABAD_data$SEC_GRADE and FAISALABAD_data$CGPA
## t = 1.8394, df = 659, p-value = 0.0663
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.004814656  0.146928797
## sample estimates: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

A really weak correlation between school grades and CGPA.



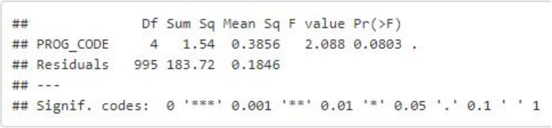


```
## 
##  Pearson's product-moment correlation
## 
## data:  FAISALABAD_data$HIG_SEC_GRADE and FAISALABAD_data$CGPA
## t = 7.6245, df = 659, p-value = 8.578e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2130837 0.3533042
## sample estimates: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

College grade is not very correlated with CGPA.

Islamabad Campus :

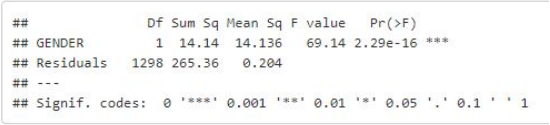
Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn’t affect CGPA .



```
##              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
```

The null hypothesis has to be accepted since p value is greater than significance value.

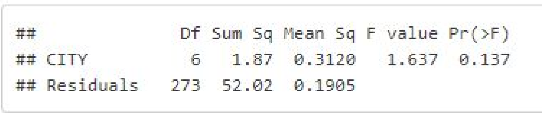
Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn’t affect CGPA .



```
##               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
```

The p value , being less than significance value , rejects the null hypothesis.

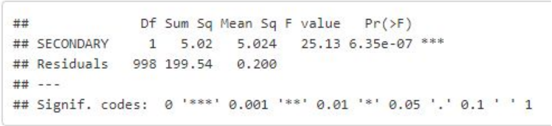
Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn’t affect CGPA .



```
##              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               
```

Evident enough to accept the null hypothesis.

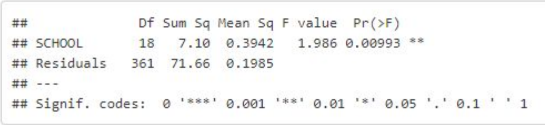
Null Hypothesis 4: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 4: Secondary Education (SSC/O Level) doesn’t affect CGPA .



```
##              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
```

The null hypothesis is rejected.

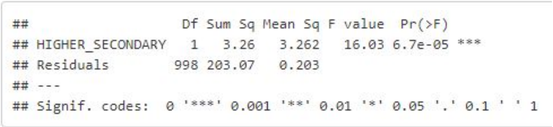
Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn’t affect CGPA .



```
##              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
```

Since here p value is greater , we accept the null hypothesis.

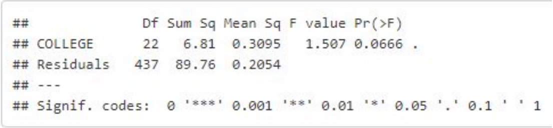
Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn’t affect CGPA .



```
##                   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
```

Null hypothesis clearly rejected.

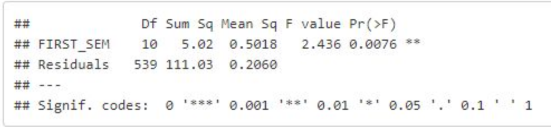
Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn’t affect CGPA .



```
##              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
```

With p value greater than significance level there is no point to reject the null hypothesis.

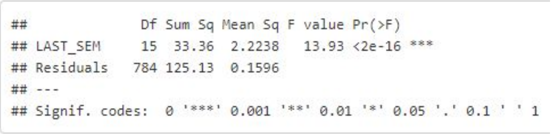
Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn’t affect CGPA .



```
##              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
```

It is evident that the null hypothesis has to be accepted.

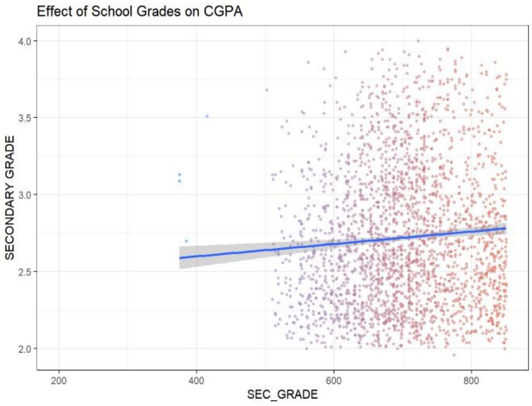
Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn’t affect CGPA .

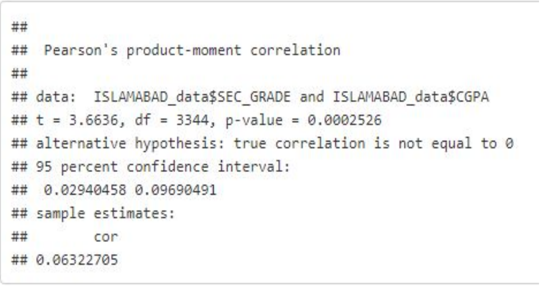


```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## LAST_SEM     15  33.36  2.2238   13.93 <2e-16 ***
## Residuals   784 125.13  0.1596                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A clear rejection of the null hypothesis.

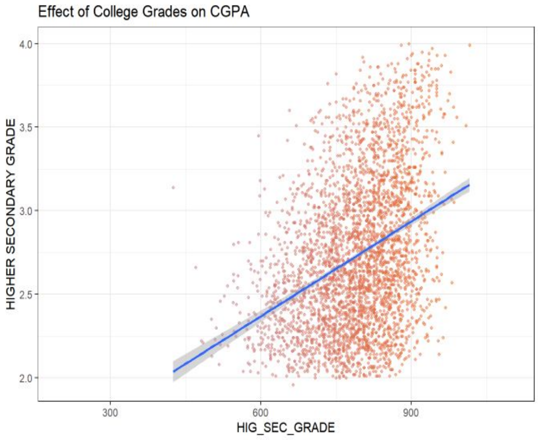
For the remaining attributes we calculate correlation coefficient .

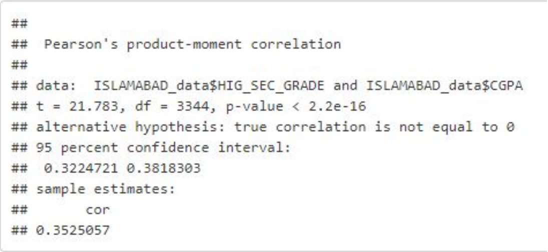




```
## 
##  Pearson's product-moment correlation
## 
## data:  ISLAMABAD_data$SEC_GRADE and ISLAMABAD_data$CGPA
## t = 3.6636, df = 3344, p-value = 0.0002526
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02940458 0.09690491
## sample estimates: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

The correlation between school grade and CGPA is negligible.



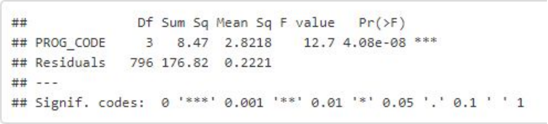


```
## 
##  Pearson's product-moment correlation
## 
## data:  ISLAMABAD_data$HIG_SEC_GRADE and ISLAMABAD_data$CGPA
## t = 21.783, df = 3344, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3224721 0.3818303
## sample estimates: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

College grades and CGPA are not highly correlated.

Karachi Campus :

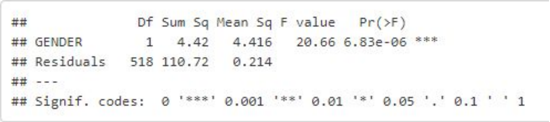
Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn’t affect CGPA .



```
##              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.

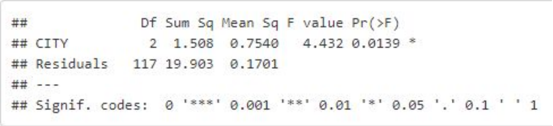
Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn’t affect CGPA .



```
##              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
```

The p value , being smaller than significance value , rejects null hypothesis.

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

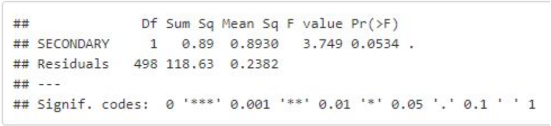


```
##              Df Sum Sq Mean Sq F value Pr(>F)  
## CITY          2  1.508  0.7540   4.432 0.0139 *
## Residuals   117 19.903  0.1701                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is accepted .

Null Hypothesis 4: Secondary Education (SSC/O Level) affect CGPA

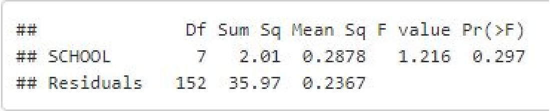
Alternative Hypothesis 4: Secondary Education (SSC/O Level) doesn’t affect CGPA .



```
##                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
```

Acception for the null hypothesis.

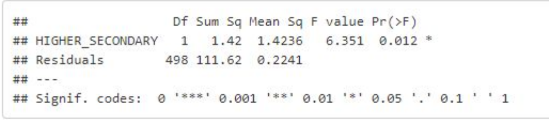
Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn’t affect CGPA .



```
##              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 of null hypothesis to be accepted.

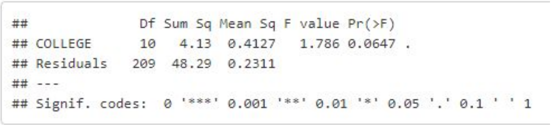
Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn’t affect CGPA .



```
##                    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
```

Here with p value equal to significance value , the null hypothesis is accepted.

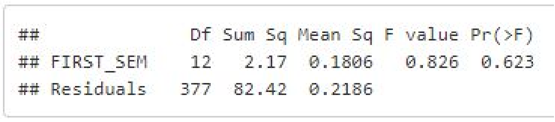
Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn’t affect CGPA .



```
##              Df Sum Sq Mean Sq F value Pr(>F)
## COLLEGE      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
```

An indication to accept null hypothesis.

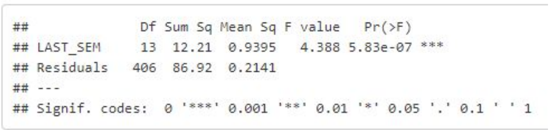
Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn’t affect CGPA .



```
##              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
```

The null hypothesis is accepted.

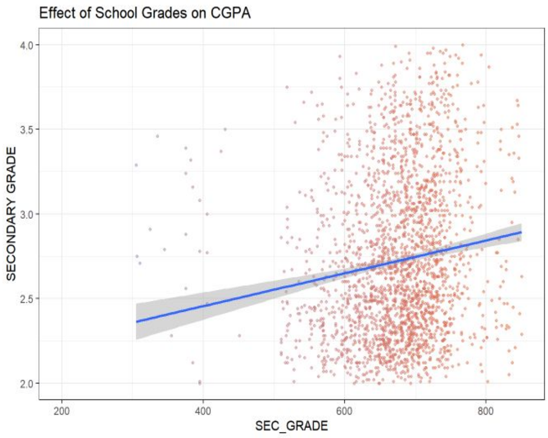
Null Hypothesis 9: Graduation Year affects CGPA.
Alternative Hypothesis 9: Graduation Year doesn’t affect CGPA .

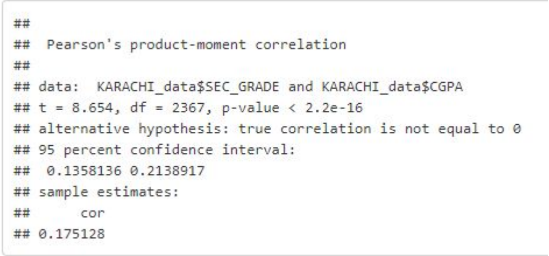


```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## LAST_SEM     13  12.21  0.9395   4.388 5.83e-07 ***
## Residuals   406  86.92  0.2141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a very small p value , the null hypothesis is rejected.

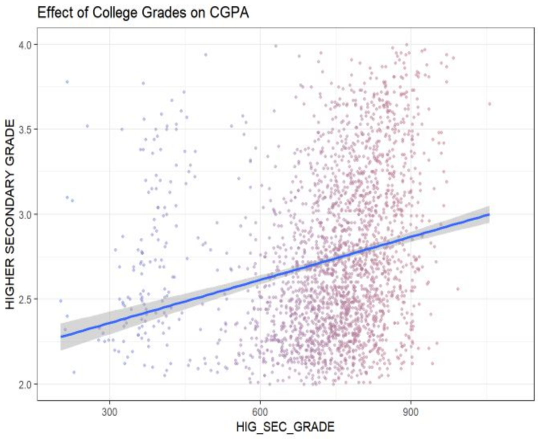
For the remaining attributes we calculate correlation coefficient .

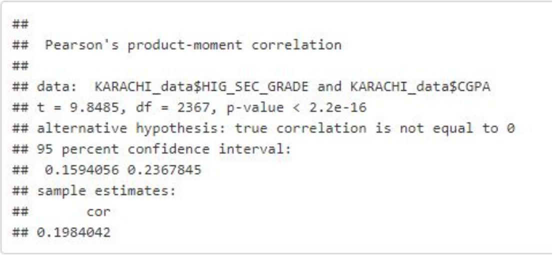




```
##
##  Pearson's product-moment correlation
##
## data:  KARACHI_data$SEC_GRADE and KARACHI_data$CGPA
## t = 8.654, df = 2367, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1358136 0.2138917
## sample estimates:
##      cor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

The correlation is not strong between school grade and CGPA.



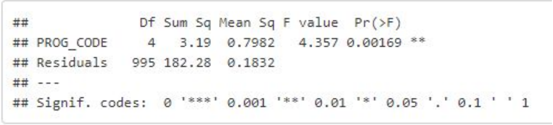


```
##
##  Pearson's product-moment correlation
##
## data:  KARACHI_data$HIG_SEC_GRADE and KARACHI_data$CGPA
## t = 9.8485, df = 2367, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1594056 0.2367845
## sample estimates:
##  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

Indicates a weak relationship between college grade and CGPA.

Lahore Campus :

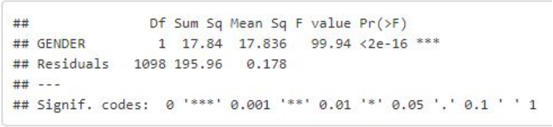
Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn’t affect CGPA .



```
##              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
```

This shows that the null hypothesis is accepted.

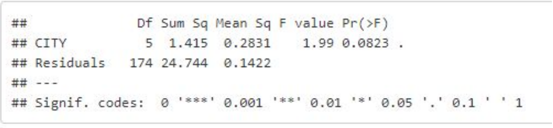
Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn’t affect CGPA .



```
##               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
```

Clear evidence of rejecting null hypothesis.

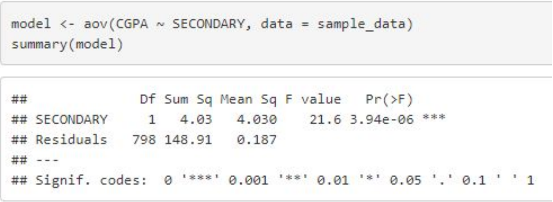
Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn’t affect CGPA .



```
##              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 is accepted.

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

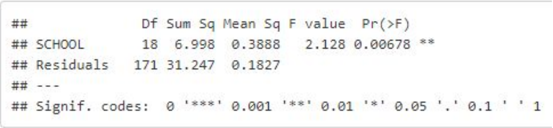


```
model <- aov(CGPA ~ SECONDARY, data = sample_data)
summary(model)
```

```
##              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
```

The p value , being quite smaller than significance value rejects the null hypothesis.

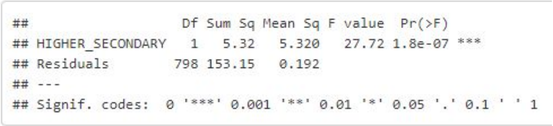
Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn’t affect CGPA .



```
##              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
```

The null hypothesis is accepted.

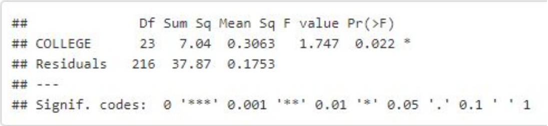
Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn’t affect CGPA .



```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## HIGHER_SECONDARY   1   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
```

Here the null hypothesis needs to be rejected.

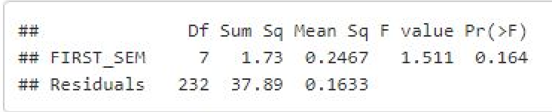
Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn’t affect CGPA .



```
##              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
```

Here , we accept the null hypothesis.

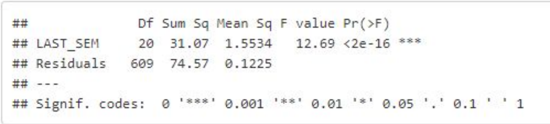
Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn’t affect CGPA .



```
##              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
```

Evident enough to be accepted.

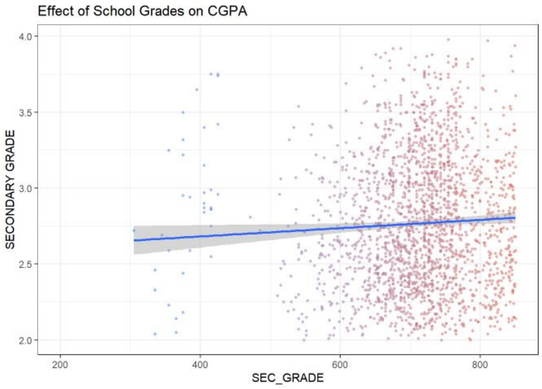
Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn’t affect CGPA .

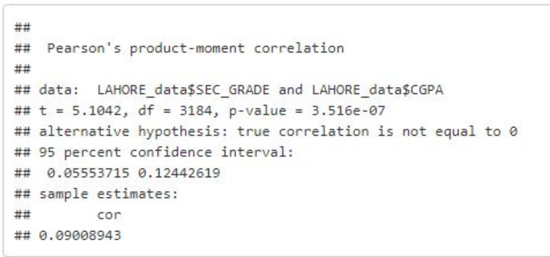


```
##             Df Sum Sq Mean Sq F value Pr(>F)
## LAST_SEM    20  31.07  1.5534   12.69 <2e-16 ***
## Residuals  609  74.57  0.1225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P value being too small than significance value rejects the null hypothesis.

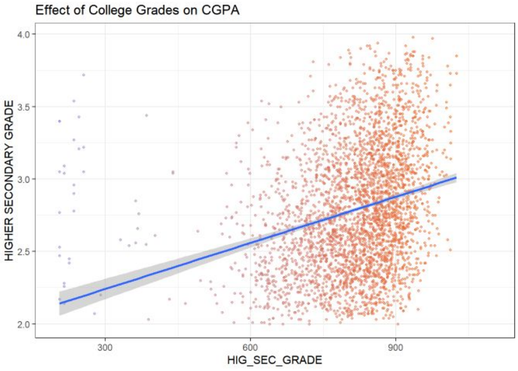
For the remaining attributes we calculate correlation coefficient .

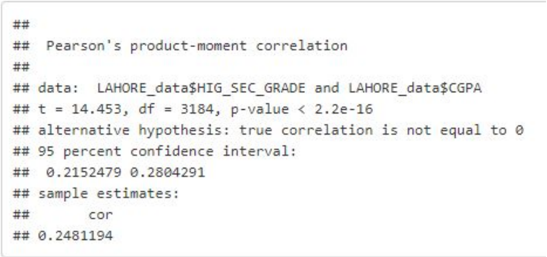




```
##
##  Pearson's product-moment correlation
##
## data:  LAHORE_data$SEC_GRADE and LAHORE_data$CGPA
## t = 5.1042, df = 3184, p-value = 3.516e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05553715 0.12442619
## sample estimates:
##    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

Correlation is weak between school grade and CGPA.



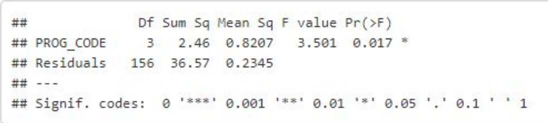


```
##
##  Pearson's product-moment correlation
##
## data:  LAHORE_data$HIG_SEC_GRADE and LAHORE_data$CGPA
## t = 14.453, df = 3184, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2152479 0.2804291
## sample estimates:
##    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

College grades and CGPA do not show strong correlation.

Peshawar Campus :

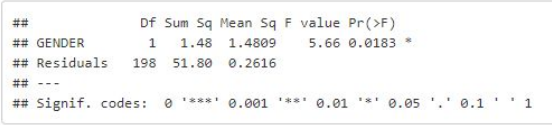
Null Hypothesis 1: Degree Program affect CGPA
Alternative Hypothesis 1: Degree Program doesn’t affect CGPA .



```
##            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
```

Here the null hypothesis is accepted.

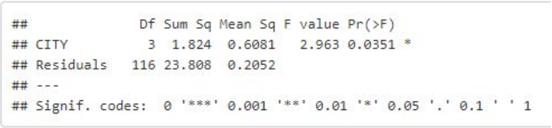
Null Hypothesis 2: Gender affect CGPA
Alternative Hypothesis 2: Gender doesn’t affect CGPA .



```
##             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
```

Accepting the null hypothesis.

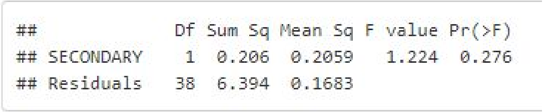
Null Hypothesis 3: City affect CGPA
Alternative Hypothesis 3: City doesn’t affect CGPA .



```
##              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
```

As p value is greater than significance value we accept the null hypothesis.

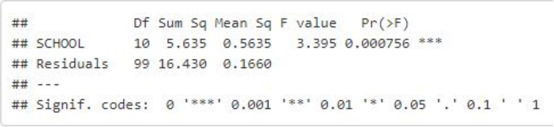
Null Hypothesis 4: Secondary Education (SSC/O Level) affect CGPA
Alternative Hypothesis 4: Secondary Education (SSC/O Level) doesn't affect CGPA .



```
##             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
```

Clearly , accept the null hypothesis.

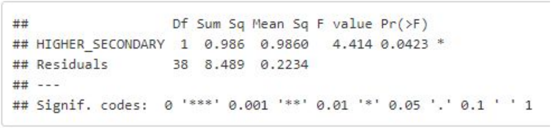
Null Hypothesis 5: School affect CGPA
Alternative Hypothesis 5: School doesn't affect CGPA .



```
##             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
```

The null hypothesis is acceptable.

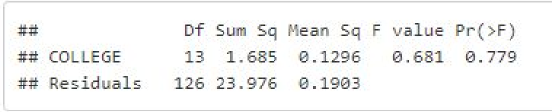
Null Hypothesis 6: Higher Secondary Education (HSSC / A Level) affect CGPA
Alternative Hypothesis 6: Higher Secondary Education (HSSC / A Level) doesn't affect CGPA .



```
##                  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
```

The null hypothesis is accepted.

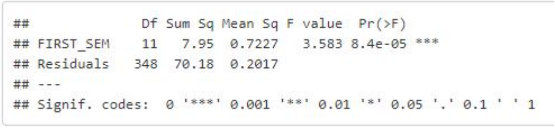
Null Hypothesis 7: College affect CGPA
Alternative Hypothesis 7: College doesn't affect CGPA .



```
##              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
```

The acceptance of the null hypothesis is evident.

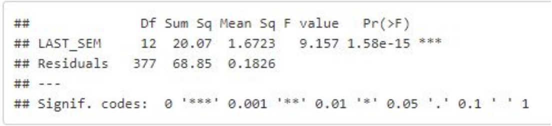
Null Hypothesis 8: Admission Year affect CGPA
Alternative Hypothesis 8: Admission Year doesn't affect CGPA .



```
##              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
```

A clear rejection of the null hypothesis.

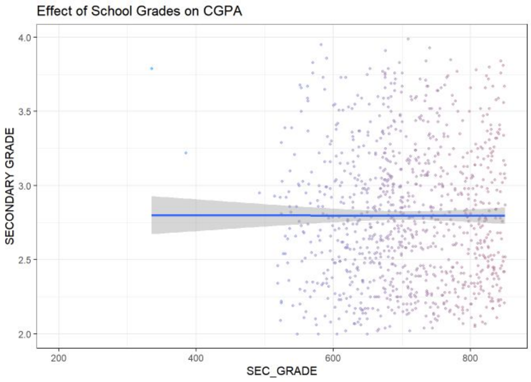
Null Hypothesis 9: Graduation Year affect CGPA
Alternative Hypothesis 9: Graduation Year doesn't affect CGPA .

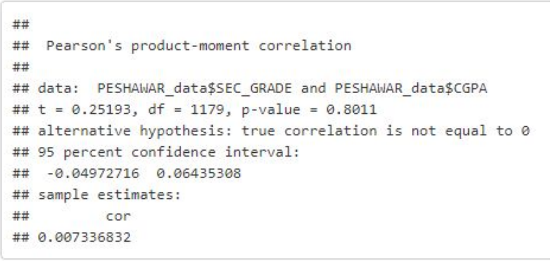


```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## LAST_SEM     12  20.07  1.6723   9.157 1.58e-15 ***
## Residuals   377  68.85  0.1826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rejected with p value being less than significance value.

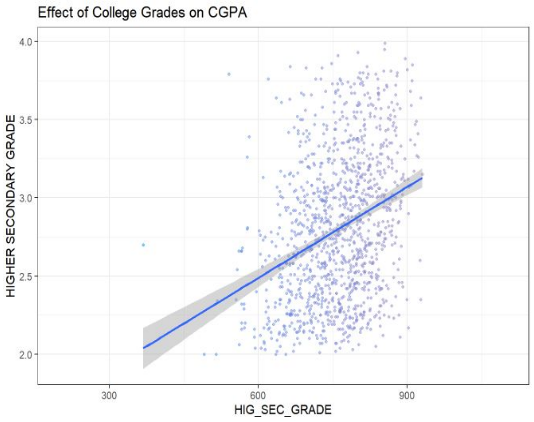
For the remaining attributes we calculate correlation coefficient .

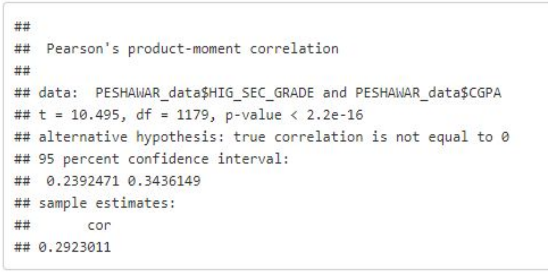




```
##
##  Pearson's product-moment correlation
##
## data:  PESHAWAR_data$SEC_GRADE and PESHAWAR_data$CGPA
## t = 0.25193, df = 1179, p-value = 0.8011
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04972716  0.06435308
## sample estimates:
##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

Very poor correlation between school grade 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
```

Not a strong relationship between college grade and CGPA.

About the statistical techniques used:

Our data had both numerical and categorical variables. Therefore , while finding correlations we applied appropriate techniques for numerical vs numerical and numerical vs categorical respectively.

For numerical vs numerical we have used Pearson Correlation Method which not only provides us the correlation coefficient but also the significance value , which tells us how reliable the result is. Value of correlation coefficient nearer to 1 shows highly positive correlation , nearer to -1 shows highly negative correlation whereas 0 shows no correlation.

For numerical vs categorical we have used One Way ANOVA. In this we take a null hypothesis and an alternate hypothesis. By applying ANOVA we get a p-value that is used to accept or reject the null hypothesis. If the p value is less than significance value we reject the null hypothesis.

DASHBOARD DEVELOPMENT

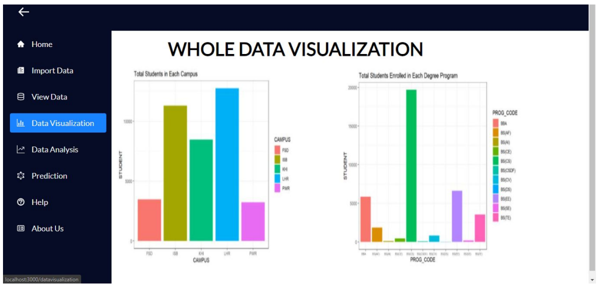
All the Exploratory Data Analysis along with the work of finding correlations (feature selection) is added in the dashboard.

Programming Language

Since our FYP-1 was only focused on building the dashboard , so we have used React for the purpose of dashboard development.

Software Tools

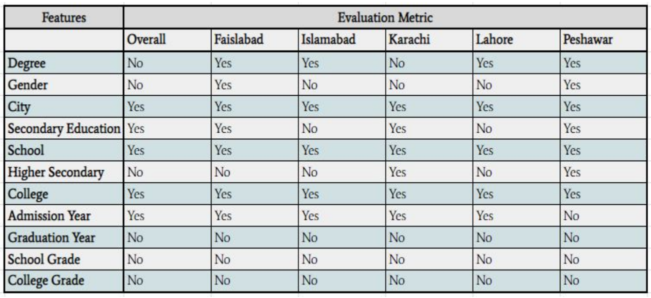
For the purpose of building the dashboard Visual Studio Code was used. Visual Studio Code is a code editor redefined and optimized for building and debugging modern web and cloud applications.



CONCLUSION AND FUTURE WORK

Student performance is dependent upon different factors and they may slightly vary for different data of students , but more or less there are few factors which stay the same.

The summary of our feature selection is shown in the table below :



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

So we selected degree , city , secondary education , school , college and admission year as important factors that affect student performance.

Since our FYP limited our scope to finding correlation between CS courses only , future work can be done on finding correlations between courses of other domains also for example EE , SS , MT etc.

The work will be further carried to build a predictive model using the selected features . Different models will be implemented including Logistic regression , Decision trees , Random Forest , Support Vector Machine , KNN etc. The performance for each model will be compared in terms of accuracy and other performance metrics. Finally , the whole work will be integrated with a fully functional website which will support data visualization and query processing features.

REFERENCES

1. Hajra Waheed , Saeed-Ul Hassan , Julie Hardman , Salem Alelyani , Raheel Nawaz , '"Predicting academic performance of students from VLE big data using deep learning models" , Volume 104, March 2020
2. Eyman Alyahyan , Dilek , "Predicting academic success in higher education: literature review and best practices" , International Journal of Educational Technology in Higher Education , 10 February 2020
3. Muhammad Yunus Iqbal Basheer, Sofianita Mutalib, Nurzeatul Hamimah Abdul Hamid, Shuzlina Abdul-Rahman, Ariff Md Ab Malik , "Predictive analytics of university student intake using supervised methods" , IAES International Journal of Artificial Intelligence , Vol. 8, No. 4, December 2019, pp. 367~374
4. Sonali Rawat , "Predictive Analytics for Placement of Student- A Comparative Study" , International Research Journal of Engineering and Technology (IRJET) , Volume: 06 Issue: 06 , June 2019
5. Mukesh Kumar , Prof A.J.Singh , Dr Disha Handa , "Literature Survey on Student's Performance Prediction in Education using Data Mining Techniques" , International Journal of Education and Management Engineering , November 2017
6. Ashish Dutt , Maizatul Akmar Ismail , Tutut Herawan , "A Systematic Review on Educational Data Mining" , IEEE Access , Volume: 5 , January 2017 .