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

1st Evaluation Report

Predictive Analytics on the Academic Record of NUCES

Dataset Version: [01]

Supervisor	Dr Jawwad Ahmed Shamsi	
Project Team	Obaid ur Rehman Areeka Aijaz Tooba Shahid	(17k-3848) (17k-3913) (17k-3731)
Submission Date	14th December 2020	

Introduction:

The aim is to answer numerous questions about how different factors affect students' academic performance and to make useful insights and find out the correlation between different attributes.
Since the dataset had only few main attributes (city , gender , CGPA) , following are the questions that we have answered in this evaluation:

1. Does there exist any correlation between the city and the CGPA?
2. What role does gender play in academic performance?

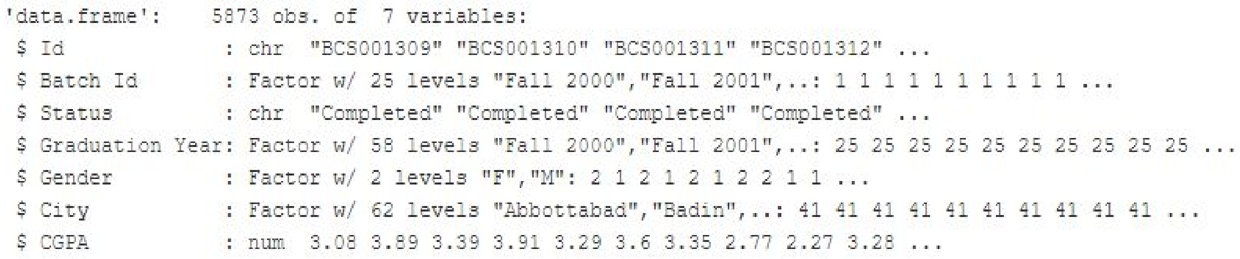
Implementation:

Initially we were provided with the dataset of Karachi campus and the data was given separately for three different departments CS , EE and BBA. After extracting the data, we performed pre processing techniques. Preliminary statistical analysis, through visualization have been performed to better understand the data. Bar charts and box plots are used to visualize the data. After pre-processing and exploratory data analysis we applied ANOVA testing to analyze the effect of gender and city on CGPA.

Individual Analysis (separately for each department)

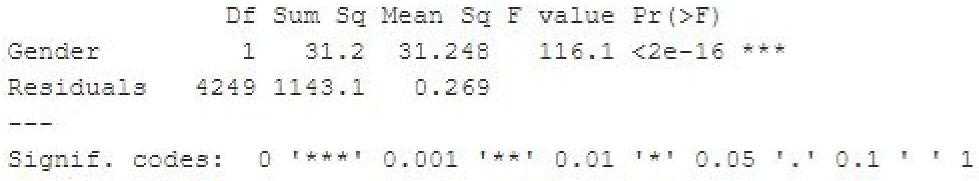
1. Computer Science department

- **Summary of the dataset:**



```
'data.frame':	5873 obs. of  7 variables:
 $ Id             : chr  "BCS001309" "BCS001310" "BCS001311" "BCS001312" ...
 $ Batch Id       : Factor w/ 25 levels "Fall 2000","Fall 2001",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Status         : chr  "Completed" "Completed" "Completed" "Completed" ...
 $ Graduation Year: Factor w/ 58 levels "Fall 2000","Fall 2001",..: 25 25 25 25 25 25 25 25 25 25 ...
 $ Gender         : Factor w/ 2 levels "F","M": 2 1 2 1 2 1 2 2 1 1 ...
 $ City           : Factor w/ 62 levels "Abbottabad","Badin",..: 41 41 41 41 41 41 41 41 41 41 ...
 $ CGPA           : num  3.08 3.89 3.39 3.91 3.29 3.6 3.35 2.77 2.27 3.28 ...
```

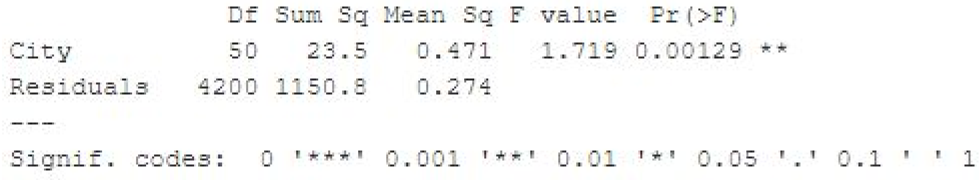
- **Relationship between CGPA and Gender:**



```
               Df Sum Sq Mean Sq F value Pr(>F)    
Gender          1   31.2  31.248   116.1 <2e-16 ***
Residuals    4249 1143.1   0.269                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test concludes that there doesn't exist any significant relationship between gender and CGPA for the CS department.

- **Relationship between CGPA and City:**

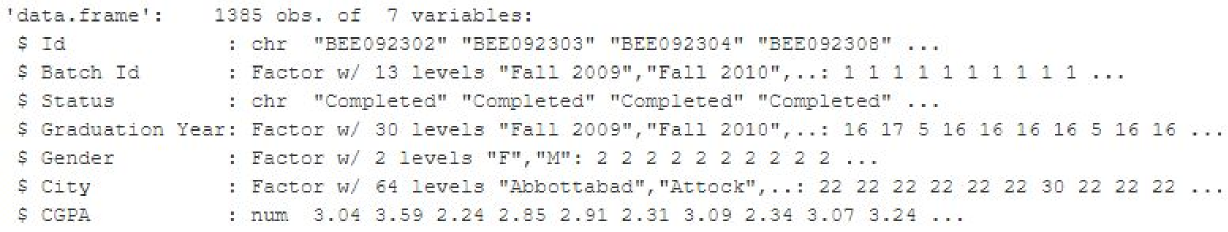


```
               Df Sum Sq Mean Sq F value  Pr(>F)   
City           50   23.5   0.471   1.719 0.00129 **
Residuals    4200 1150.8   0.274                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value is less than 0.05 for the CS department and hence indicates that the city of students doesn't affect the academic performance.

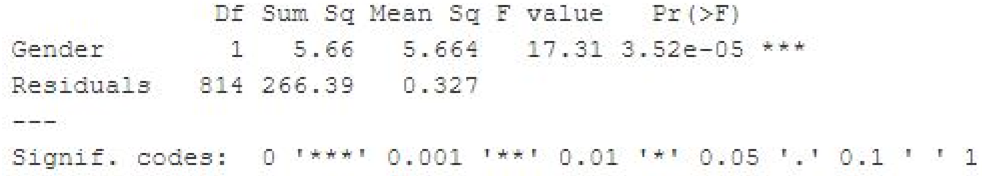
2. Electrical Engineering department

- **Summary of the dataset:**



```
'data.frame':	1385 obs. of  7 variables:
 $ Id             : chr  "BEE092302" "BEE092303" "BEE092304" "BEE092308" ...
 $ Batch Id       : Factor w/ 13 levels "Fall 2009","Fall 2010",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Status         : chr  "Completed" "Completed" "Completed" "Completed" ...
 $ Graduation Year: Factor w/ 30 levels "Fall 2009","Fall 2010",..: 16 17 5 16 16 16 16 5 16 16 ...
 $ Gender         : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ City           : Factor w/ 64 levels "Abbottabad","Attock",..: 22 22 22 22 22 22 30 22 22 22 ...
 $ CGPA           : num  3.04 3.59 2.24 2.85 2.91 2.31 3.09 2.34 3.07 3.24 ...
```

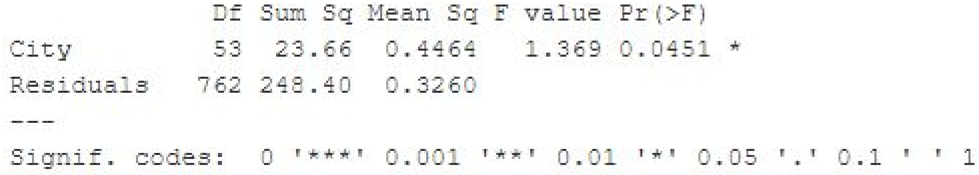
- **Relationship between CGPA and Gender:**



```
              Df Sum Sq Mean Sq F value   Pr(>F)    
Gender         1   5.66   5.664   17.31 3.52e-05 ***
Residuals    814 266.39   0.327                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the EE department P-value is too small that shows there is no correlation between CGPA and gender.

- **Relationship between CGPA and City:**

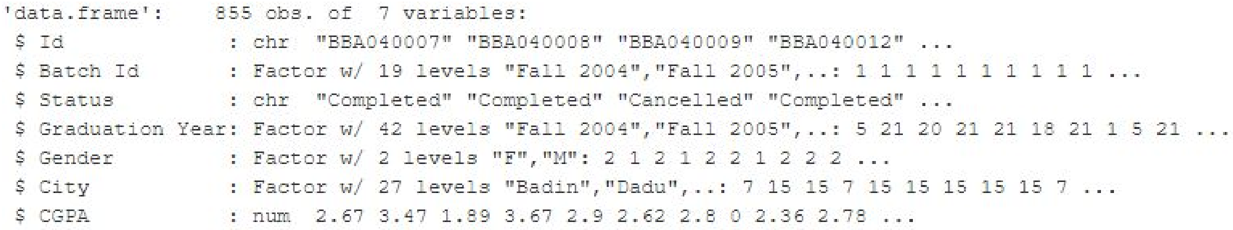


```
              Df Sum Sq Mean Sq F value Pr(>F)  
City          53  23.66  0.4464   1.369 0.0451 *
Residuals    762 248.40  0.3260                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value is nearer to 0.05 for the EE department, indicates the model is likely significant.

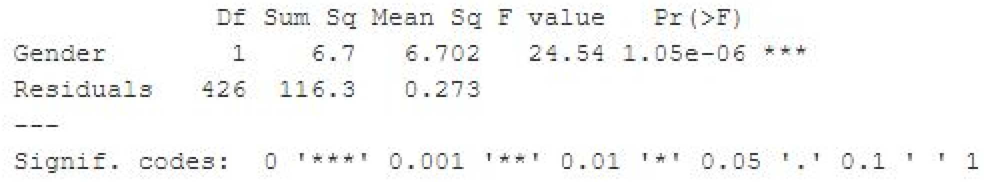
3. Business Administration department

- **Summary of the dataset:**



```
'data.frame':	855 obs. of  7 variables:
 $ Id             : chr  "BBA040007" "BBA040008" "BBA040009" "BBA040012" ...
 $ Batch Id       : Factor w/ 19 levels "Fall 2004","Fall 2005",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Status         : chr  "Completed" "Completed" "Cancelled" "Completed" ...
 $ Graduation Year: Factor w/ 42 levels "Fall 2004","Fall 2005",..: 5 21 20 21 21 18 21 1 5 21 ...
 $ Gender         : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 1 2 2 2 ...
 $ City           : Factor w/ 27 levels "Badin","Dadu",..: 7 15 15 7 15 15 15 15 15 7 ...
 $ CGPA           : num  2.67 3.47 1.89 3.67 2.9 2.62 2.8 0 2.36 2.78 ...
```

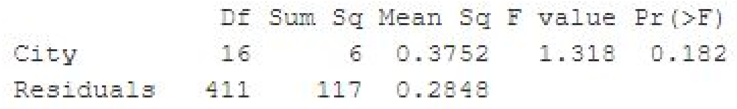
- **Relationship between CGPA and Gender:**



```
             Df Sum Sq Mean Sq F value   Pr(>F)    
Gender        1    6.7   6.702   24.54 1.05e-06 ***
Residuals   426  116.3   0.273                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model summarizes that there doesn't exist any significant relationship between CGPA and gender for the BBA department.

- **Relationship between CGPA and City:**

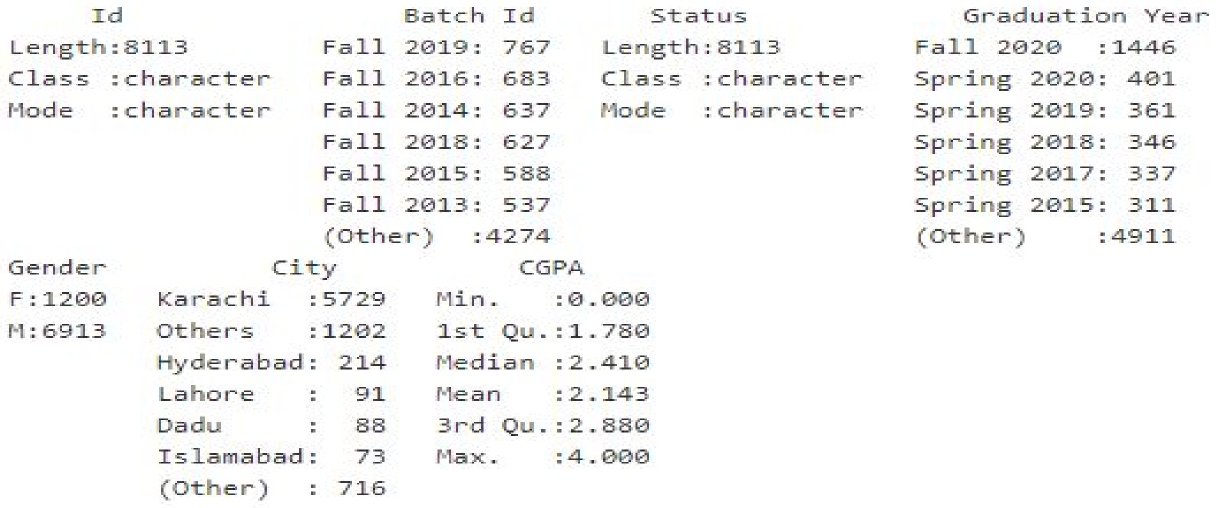


```
             Df Sum Sq Mean Sq F value Pr(>F)
City         16      6  0.3752   1.318  0.182
Residuals   411    117  0.2848               
```

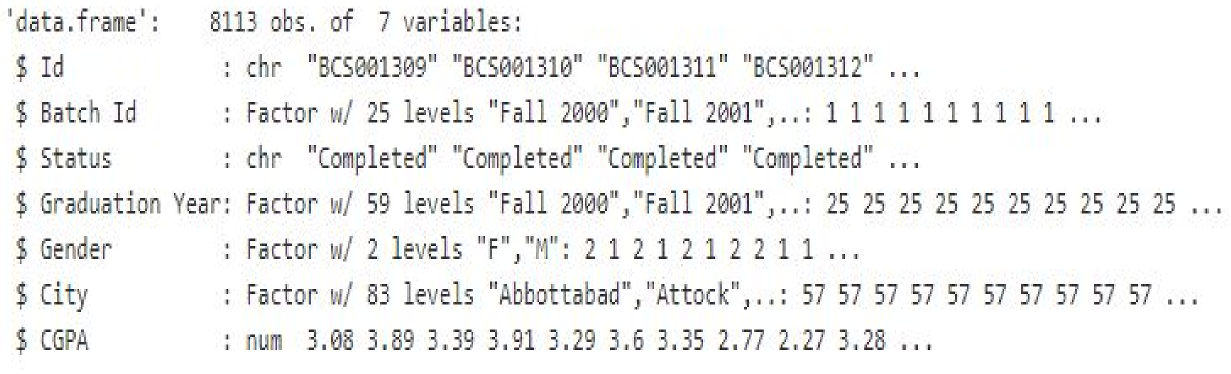
For the BBA department the ANOVA test summarizes that the CGPA of students is dependent on their City .

Combined Analysis (grouped all the departments)

- **Summary of the dataset:**



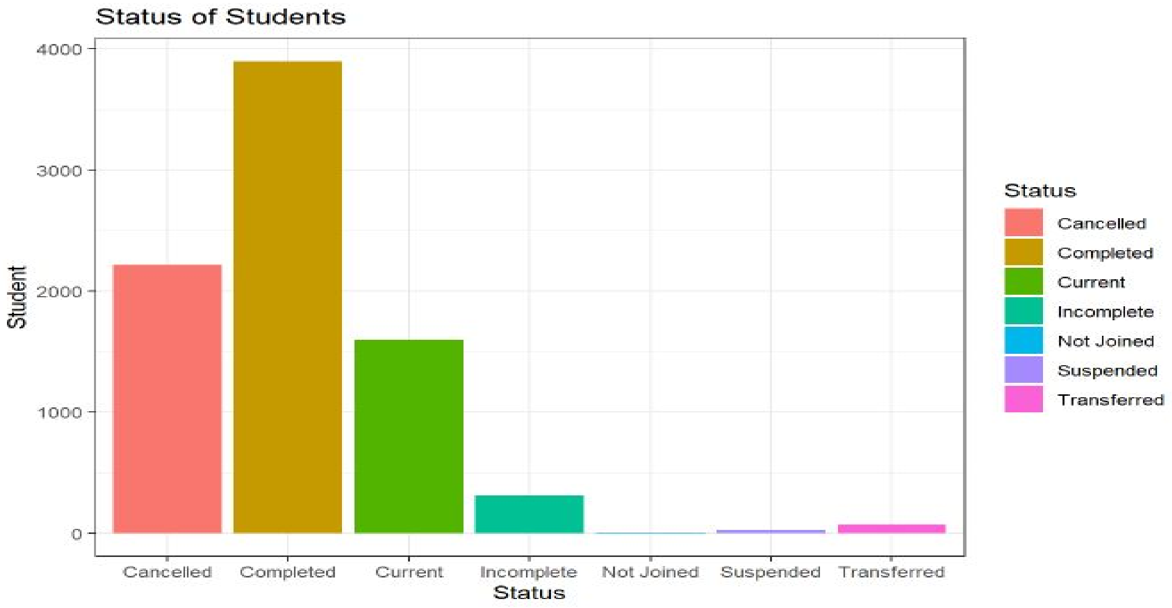
```
      Id                  Batch Id        Status              Graduation Year
 Length:8113        Fall 2019: 767   Length:8113         Fall 2020  :1446   
 Class :character   Fall 2016: 683   Class :character    Spring 2020: 401   
 Mode  :character   Fall 2014: 637   Mode  :character    Spring 2019: 361   
                    Fall 2018: 627                       Spring 2018: 346   
                    Fall 2015: 588                       Spring 2017: 337   
                    Fall 2013: 537                       Spring 2015: 311   
                    (Other)  :4274                       (Other)    :4911   
 Gender          City           CGPA      
 F:1200   Karachi  :5729   Min.   :0.000  
 M:6913   Others   :1202   1st Qu.:1.780  
          Hyderabad: 214   Median :2.410  
          Lahore   :  91   Mean   :2.143  
          Dadu     :  88   3rd Qu.:2.880  
          Islamabad:  73   Max.   :4.000  
          (Other)  : 716                  
```



```
'data.frame':	8113 obs. of  7 variables:
 $ Id             : chr  "BCS001309" "BCS001310" "BCS001311" "BCS001312" ...
 $ Batch Id       : Factor w/ 25 levels "Fall 2000","Fall 2001",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Status         : chr  "Completed" "Completed" "Completed" "Completed" ...
 $ Graduation Year: Factor w/ 59 levels "Fall 2000","Fall 2001",..: 25 25 25 25 25 25 25 25 25 25 ...
 $ Gender         : Factor w/ 2 levels "F","M": 2 1 2 1 2 1 2 2 1 1 ...
 $ City           : Factor w/ 83 levels "Abbottabad","Attock",..: 57 57 57 57 57 57 57 57 57 57 ...
 $ CGPA           : num  3.08 3.89 3.39 3.91 3.29 3.6 3.35 2.77 2.27 3.28 ...
```

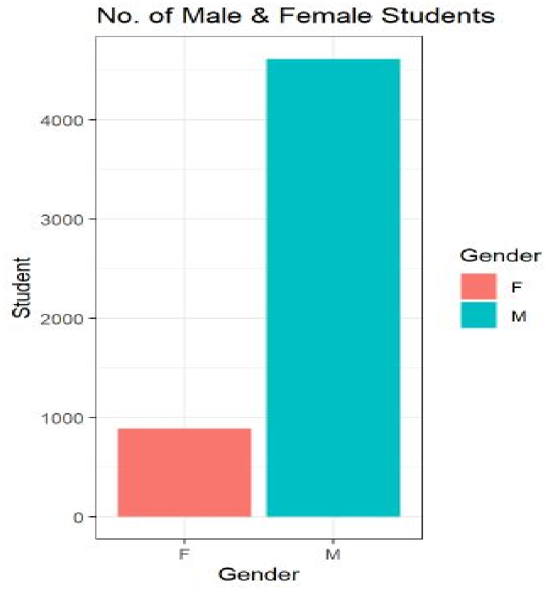
- **Visualized status of Students:**

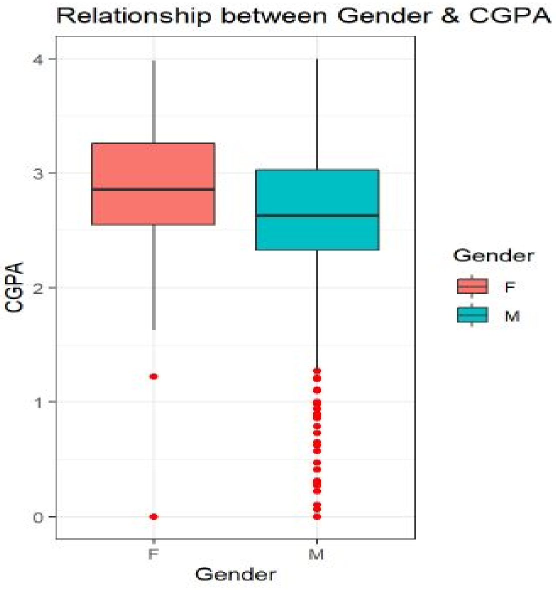
Total unique status: 7

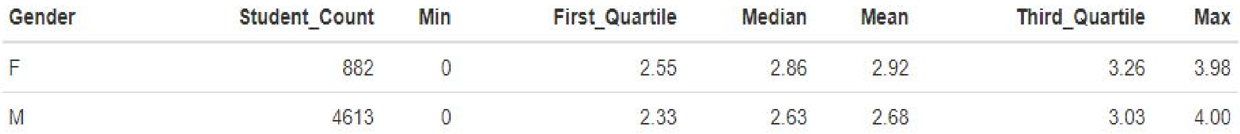


As we can see “Not Joined”, “Suspended” and “Transferred” statuses are negligible, hence we have only used the data of students whose statuses are either “Current” or “Completed”. Students with “Cancelled” status are of no use as they haven’t studied from the NUCES.

- **Relationship between CGPA and Gender:**


No. of Male & Female Students

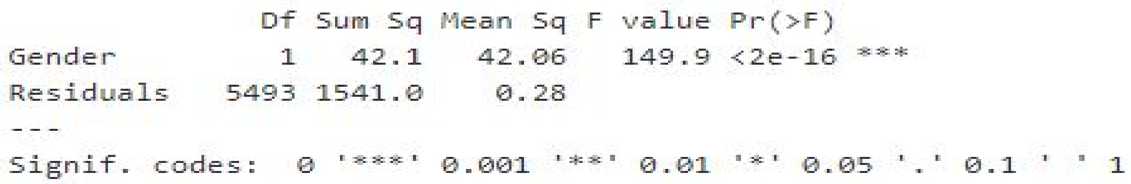

Relationship between Gender & CGPA



Gender	Student_Count	Min	First_Quartile	Median	Mean	Third_Quartile	Max
F	882	0	2.55	2.86	2.92	3.26	3.98
M	4613	0	2.33	2.63	2.68	3.03	4.00

1. In our data set there is a great difference in the frequency of male and female.
2. There are less female students as compared to male students, indicating imbalance. Therefore, the average of female students is greater than the average of male students.
3. A point to be noted is that there doesn't exist a significant difference in the performance based on gender.

Now, applying ANOVA to analyze the effect of gender on CGPA



```
              Df Sum Sq Mean Sq F value Pr(>F)    
Gender         1   42.1   42.06   149.9 <2e-16 ***
Residuals   5493 1541.0    0.28                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To check whether or not there exists a dependency between gender and cgpa , we conduct ANOVA test . We can see that that p-value is way too small (mention the value) and hence indicates that there is no correlation between gender and CGPA .

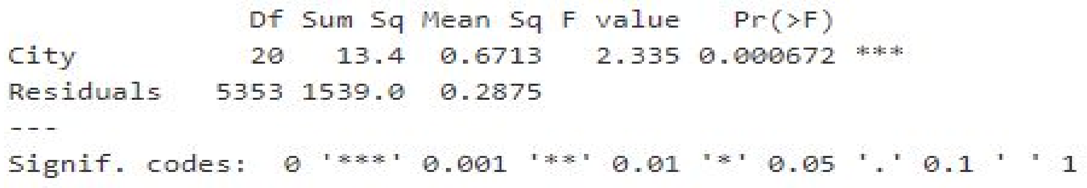
- **Relationship between CGPA and City:**

  We observed that students from a great number of cities take admission in NUCES Karachi . As the campus is located in the heart of Pakistan , Karachi , it is obvious that the majority of students admitted are from Karachi. However students from cities like Hyderabad , Lahore , Dadu , Islamabad , Larkana , Mirpur Khas , Quetta , Umerkot and Rahim Yar Khan also take admission in the Karachi Campus and are in better numbers as compared to the rest of the cities.

  Finding Correlation between City & CGPA

  Since in our cases (City vs CGPA) we have one categorical and one numerical variable , so for this type we typically perform one way ANOVA test .

  Now, applying ANOVA test to analyze the effect of City on CGPA



```
             Df Sum Sq Mean Sq F value   Pr(>F)    
City         20   13.4  0.6713   2.335 0.000672 ***
Residuals  5353 1539.0  0.2875                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

  Here also we conducted ANOVA test and here too the p value shows the absence of correlation between the attributes. There is no significant relationship between the city and CGPA. This can be easily established from the signif. codes from the summary of the model above .