CS-306: Data Analysis and Visualization

Project Report

Pratvi Shah
201801407@daiict.ac.in

Instructor of the course: Prof. Pankaj Kumar
Teaching Assistant: Umang Patel

1 Introduction

Marks of the students are given a lot of emphasis but the grading is done solely based on the pen-paper test or other forms of examinations. There is not much consideration of the background information or the supplements/facilities available to the students. Through this project, my objective would be to analyze how the marks of the students vary depending upon the external factors like family background, economic well-being and amenities available. In addition to this, there would be address to the notion that female students are generally weaker in mathematics in comparison to other subjects and other students.

2 Dataset Description

Standard dataset containing results of national assessments for secondary and university education in engineering students[1] is used for the project as it captured majority of the factors of daily life ranging from access to internet, computer, mobile, phone to the economic class of the family represented by SISBEN. Following attributes are considered to be important for the scope of this project:

1. Amenities: Internet, TV, computer, washing machine, microwave oven, car, DVD, phone, mobile
2. Subjects: MAT S11, CR S11, CC S11, BIO S11, ENG S11
3. Family background: Education levels of mother and father, occupation of mother and father, number of people in a house
4. Socio-Economic factors: Stratum (Social status ranking), SISBEN ranking(well-being measure w.r.t. health, education, housing, and vulnerability)[2]

3 Data cleaning and grouping

The dataset used for the project consists of 34 features but recurring features including percentile and decile values and the ones conveying the branch of engineering in which the student wished to apply or the Labels which could not be deciphered were dropped for a better and focused analysis of the categories mentioned in Section 2.

Category 1(amenities) consists of boolean data and hence one-hot encoding with access to the amenities being considered as a positive impact hence, Yes=1 and No=0 was used.
Category 2(scores in subjects) data was specifically analyzed for outliers and IQR methodology was used to clean the data followed by subsequent standardization to zero mean and one standard deviation.
Category 3 consists of categorical data which was either converted to numeric data based on the inherent meaning or was acted upon using label encoding
Category 4 was converted to corresponding numeric value which was evident from the description of the same.

To group the data within each category, statistical methods of averaging were used after the characteristic behaviour of the components was explored using visualizations.

4 Analysis using Visualizations

After obtaining a clean dataset we further move towards retrieving information that is readily available from the data.
NOTE: For all future reference the TOTAL mentioned in the figures and in the text suggest the average marks of students in 5 subjects. The codes for all the implementations are available in the colab link given below[3].

1. **Category 2/general comments**
   We can start analyzing the marks in different subjects using the following plots.

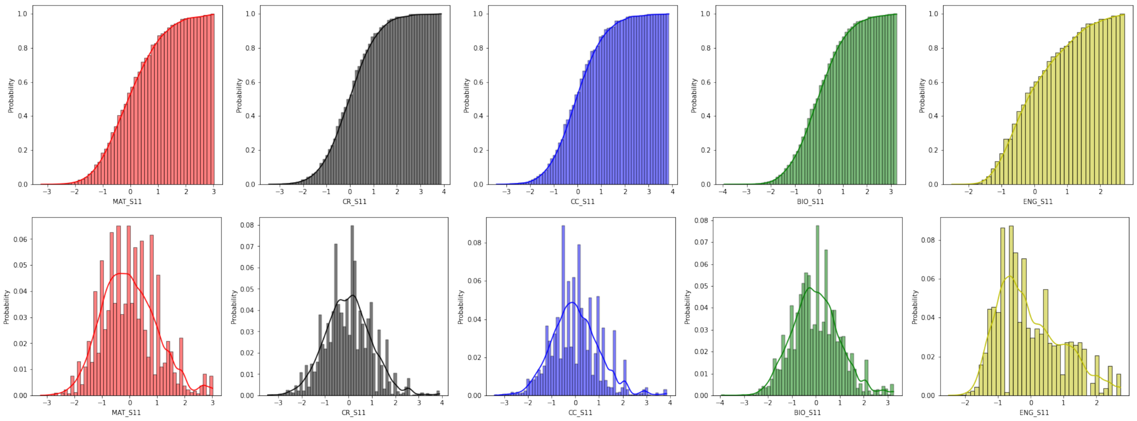


Figure 1: Distribution of marks in different subjects

Above distribution plots suggest that the **marks of students in ENG S11 is right skewed** whereas distribution of other subject marks are centered at their corresponding mean. This skewed nature suggests that the distribution of marks of students in that subject is not normally distributed hence, grading based on that assumption would not be justified. Whereas for the other plots we can not comment directly looking at the distribution whether the marks are normal distribution or not.

Further, to analyze the overall distribution of marks, average of all the five subjects was taken and then it was compared with the normal distribution having same mean and standard deviation. The blue curve represents the average marks distribution and it follows the emperical rule(68-95-99.7) for normal distribution.

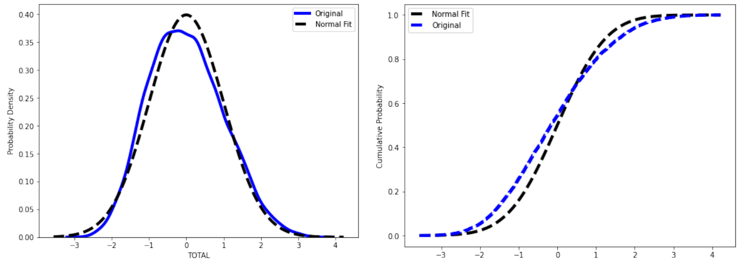


Figure 2: Average marks and the corresponding normal distribution fit

From Fig. 2 we can say that there is not much difference between the two datasets but it might be possible that the statistically significant difference exists but it is not visible to the naked eye.

In order to verify the claim that the two distributions are identical ks-test was performed and the following results were obtained:

```
x=np.linspace(xmin,xmax,size_l)
c = norm.cdf(x, data3['TOTAL'].mean(), data3['TOTAL'].std())
```

```
#get a close to emperical curve
temp= data3['TOTAL'].copy()
temp=np.sort(np.array(temp))
H,X1 = np.histogram(temp, bins =len(x), normed = True )
dx = X1[1] - X1[0]
F1 = np.cumsum(H)*dx

print(stats.ks_2samp(F1,c))
print(stats.kstest(F1,'norm'))
```

When *size_l* attribute was set to 100:
Ks_2sampResult (statistic=0.1, pvalue=0.70205)
KstestResult (statistic=0.5, pvalue=1.20332e-23)

When *size_l* attribute was set to $len(data3['TOTAL]) = 11207$:
Ks_2sampResult (statistic=0.096, pvalue=8.57132e-46)
KstestResult (statistic=0.5, pvalue=0.0)

From the kstest values we can see that if the sample size is small then the test gives us pvalue to be high suggesting that we cannot reject the null hypothesis but when the sample size is made equal to the population size then we can see that the pvalue obtained is very low suggesting that we can reject the null-hypothesis.

Further to notice the difference between the means, in other words to analyze how different the average values of the two samples are Cohen's d value was calculated and it came out to be $\approx$ **0.325** which suggests that the means of the two samples are 0.325 standard deviation away from each other. This according to Cohen comes under small effect size which implies that the difference between the means of two samples is very small to be visible but can have statistically significant difference depending upon the error tolerance allowed in the problem taken into consideration.

2. **CATEGORY 1(Amenities)**

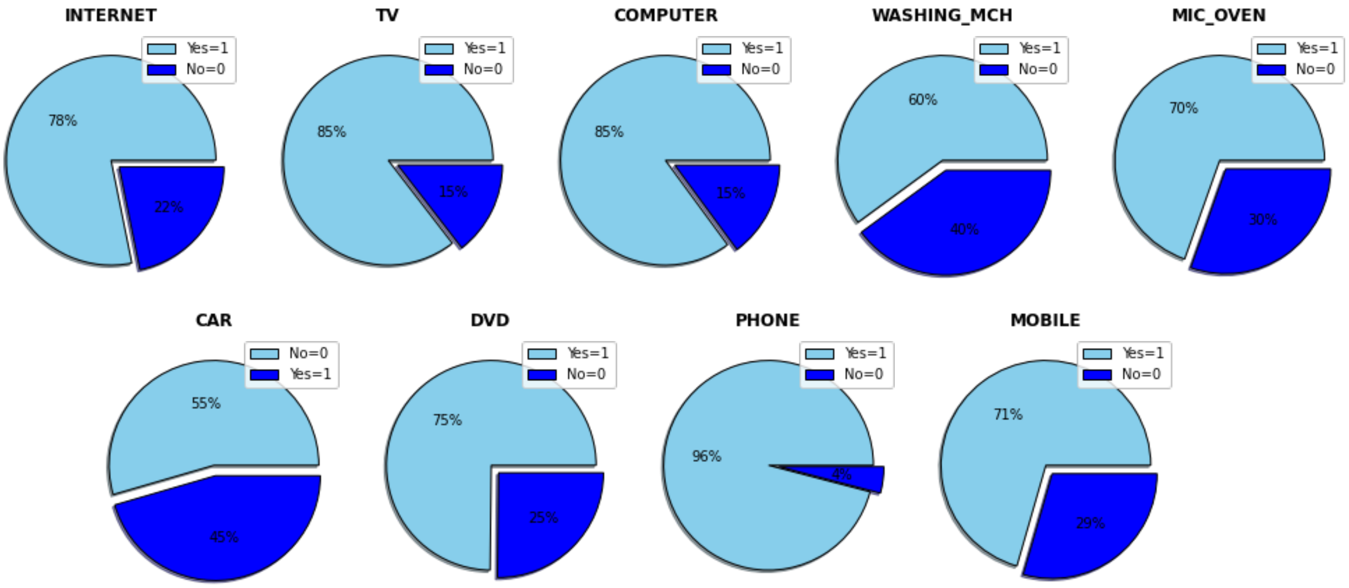


Figure 3: Distribution representing access to amenities

We can observe from the above distribution that substantial population($> 50\%$) has access to majority of the amenities except CAR.

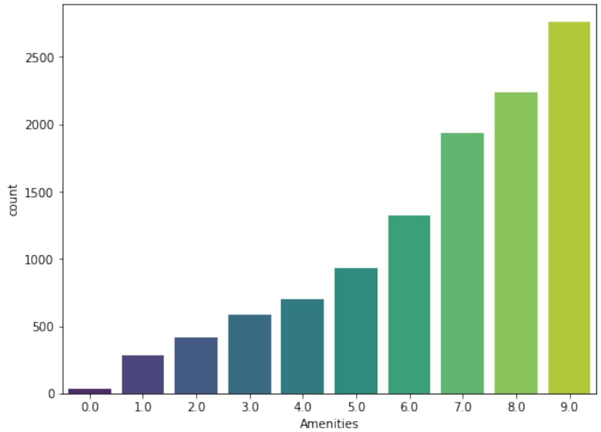


Figure 4: Frequency v/s Number of amenities available

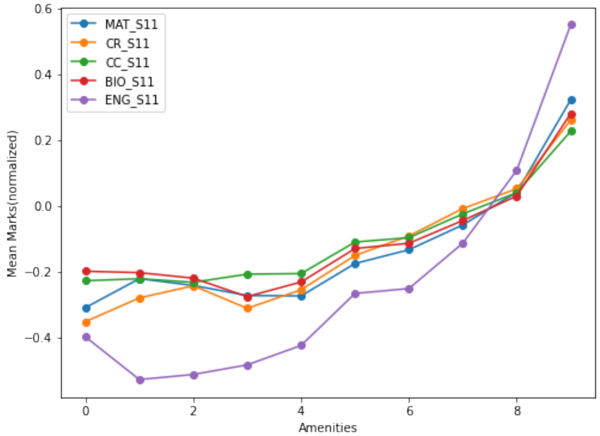


Figure 5: Average marks v/s Number of amenities available

Fig. 8 clearly shows that the population into consideration is a privileged population with multiple facilities available for them to make their daily routines easier.(While calculating this number of amenities there is an underlying assumption that all the amenities have the same impact on the individuals. For subsequent sections this total number of amenities is considered as a representative of all the amenities.)

Next we consider how the amenities affect the average marks of the students in each subject(Fig. 9). Here, we group each class of students having the same number of amenities and then calculate the average marks of those students in each of the 5 subjects. This plots clearly shows how with increase in the number of amenities the average marks of students increase.

3. **CATEGORY 3(Family background)**

For this case we shall first see the range of average total marks for each category of education level of the parent and then the occupation oof the parent.

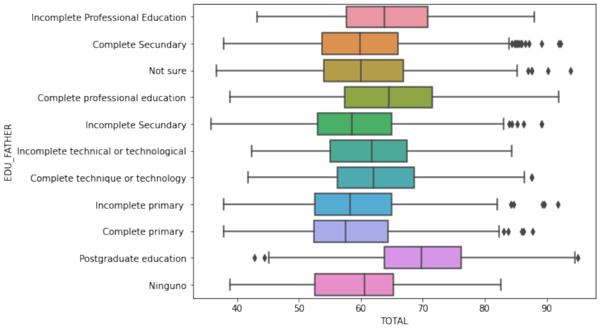


Figure 6: Education Level of Father v/s Average Marks(normalized)

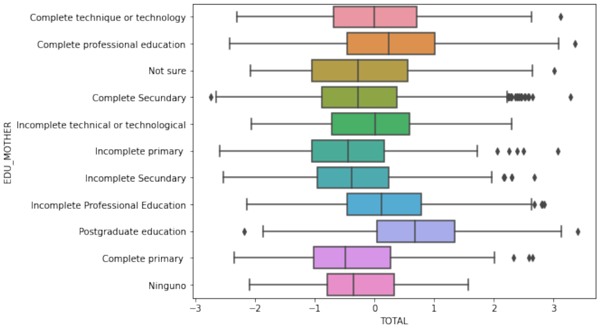


Figure 7: Education Level of Mother v/s Average Marks(normalized)

We can observe from the above plot that the students whose parents are postgraduates have higher marks than the other students whereas no such clear trend is observed in case of occupation. In order to make a

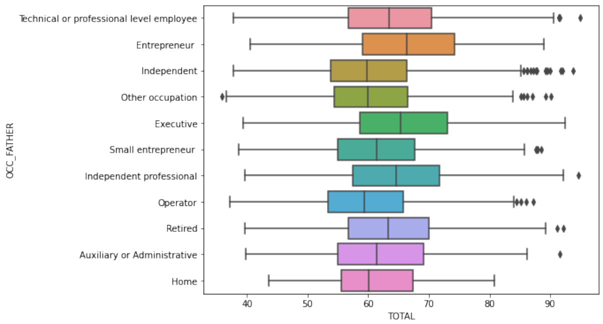


Figure 8: Occupation of Father v/s Average Marks

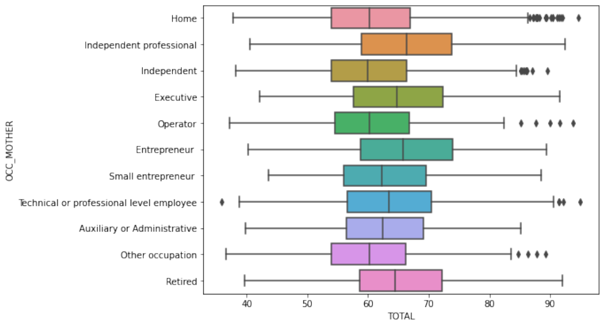


Figure 9: Occupation of Mother v/s Average Marks

comment about the same, grouping of students based on their parents' occupation and thereby the average marks obtained were calculated.

Following mapping was used for education level and occupation of parents:

```
#set manually according to the prevalent hierarchy of education level
edu_cols={'Ninguno':0,'Not sure':0,'Incomplete primary ':1,'Complete primary ':2,
'Incomplete Secundary':3,'Complete Secundary':4,'Incomplete Professional Education':5,
'Complete professional education':6,'Incomplete technical or technological':7,
'Complete technique or technology':8,'Postgraduate education':9}


occ_cols={'Auxiliary or Administrative':0,'Entrepreneur ':1,'Executive':2, 'Home':3,
'Independent':4,'Independent professional':5,'Operator ':6, 'Other occupation':7,
'Retired':8,'Small entrepreneur ':9, 'Technical or professional level employee':10}



#from sklearn.preprocessing import LabelEncoder
#le = LabelEncoder()
#data3['OCC_FATHER'] =le.fit_transform(data3['OCC_FATHER'])
#data3['OCC_MOTHER'] =le.fit_transform(data3['OCC_MOTHER'])
#this would obtain us a sequence which is given in occ_cols
```

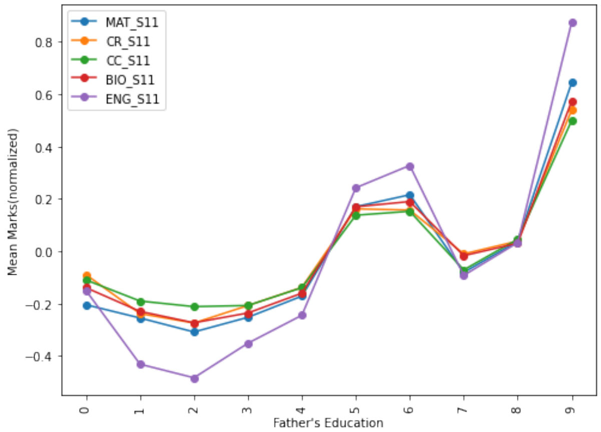


Figure 10: Education Level of Father v/s Average Marks(normalized)

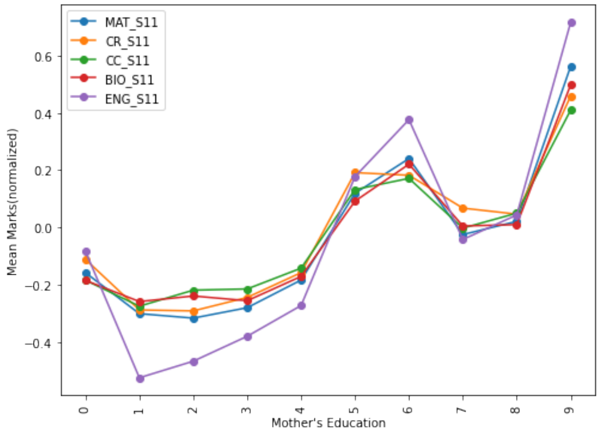


Figure 11: Education Level of Mother v/s Average Marks(normalized)

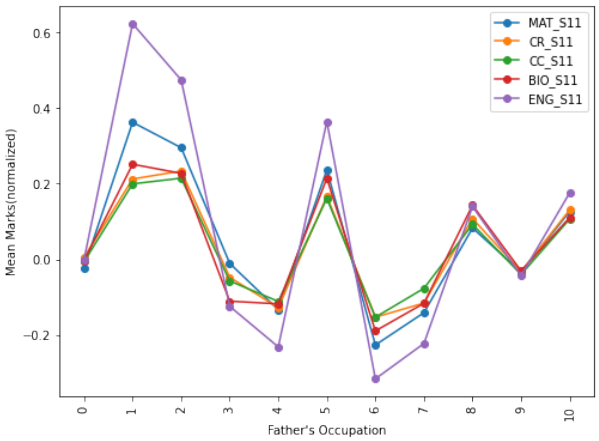


Figure 12: Occupation of Father v/s Average Marks

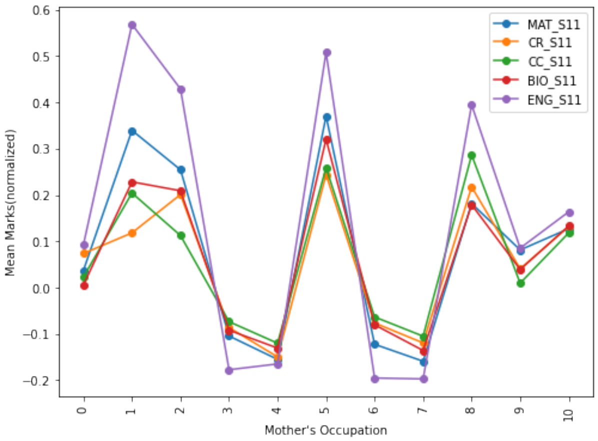


Figure 13: Occupation of Mother v/s Average Marks

Fig 10 to 13 shows how the average marks(normalized) differ based on the education and occupation of the parents. Higher education of parents relates to higher average score of the students and when the parents are Entrepreneurs, Executive, Independent Professional or retired then in those cases the marks of students are observed to be higher than the average. The positive effect of retired parents is more significant when the mother is retired. We can also observe that working mother in roles of independent professional or entrepreneur correspond to higher positive effect on the average marks of the students.

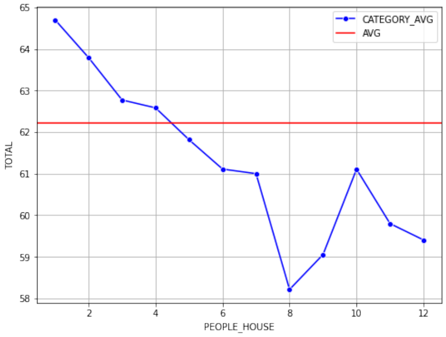


Figure 14: Number of people in house v/s Average total marks

Finally, we see the variation of the number of people in the house and the average total marks. We can observe from Fig 14 that when the number of people increase beyond five the average marks of that category (and higher) of students decrease. This shows that as the number of people increase in the house the average total marks decreases.

4. **CATEGORY 4(Socio-economic factors)**

In this part we mainly focus on the Status and the vulnerability index along with the factors like the revenue(in terms of legal monthly minimum wage LMMV) and whether the student has a job as an additional source of income or not. In addition to this we also have factors like which school the student belongs i.e., private or public.

Following mapping was used for this section:

```
col2num={
    'REVENUE':{'less than 1 LMMW':0,'Between 1 and less than 2 LMMW':1,
    'Between 2 and less than 3 LMMW':2, 'Between 3 and less than 5 LMMW':3,
    'Between 5 and less than 7 LMMW':4, 'Between 7 and less than 10 LMMW':5, '10 or more LMMW':6},

    'JOB':{'No':0,'Yes, 20 hours or more per week':1,'Yes, less than 20 hours per week':2},

    'SISBEN':{'It is not classified by the SISBEN':4, 'Level 2':2,
    'Level 1':1,'Esta clasificada en otro Level del SISBEN':4, 'Level 3':3}
}

#For Stratum
data[['Trash','STRATUM']] = data['STRATUM'].str.split(expand=True)
data=data.drop(['Trash'],axis= 1)
```

Above replacements help us to convert categorical to numerically equivalent data. The numeric equivalent are chosen in accordance with the meaning of the variables and thereby maintains the underlying meaning of the feature.

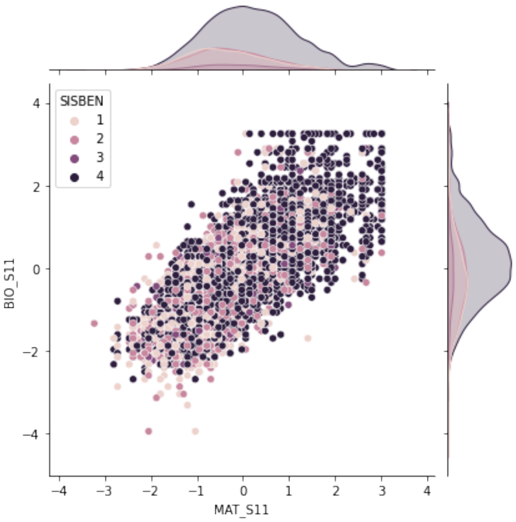


Figure 15: Marks in MATH and BIO according to SISBEN Level

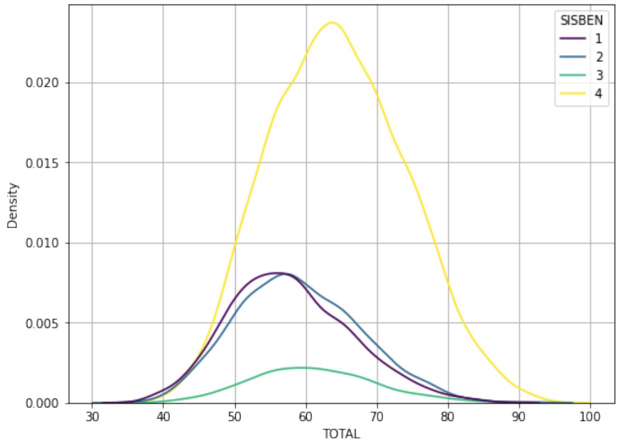


Figure 16: Variation of Total marks density with respect to SISBEN Level

Above figures show the relationship between the vulnerability index mainly highlighting the economic level of the family. This index gives every family a score between 0-100 and then classifies them within 4 levels the criteria for which varies depending on whether the family is located in urban or rural areas. Lower levels 1 & 2 suggest poor families and increases thereafter.

The plot show clear indication that higher SISBEN ranking is related to students with higher marks as evident from the darker shades accumulated on the top right while ligher shades representing lower SISBEN level accumulated around the lower left in Fig. 15.
Fig 16 shows the variation of the total marks of students belonging to different SISBEN levels. This plot again verifies our assumption that the population taken into consideration is that of a privileged section of the society as majority of the students either belong to higher levels of SISBEN or belong to class not falling under SISBEN. The delayed peaks of higher SISBEN levels suggest that the mean score obtained by students belonging to higher SISBEN levels is higher.

Next we look for the correlation existing between the economic factors/levels present in the data i.e., Revenue, STRATUM, SISBEN and their corresponding correlation with the total marks.

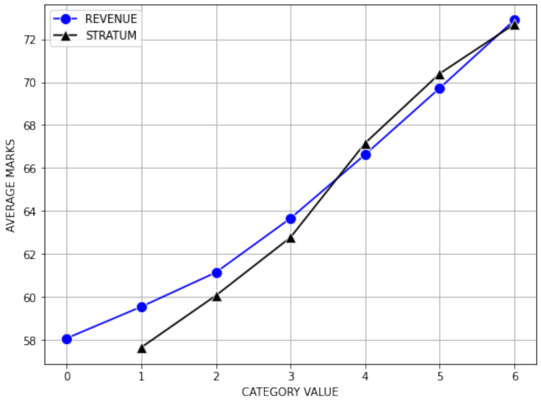


Figure 17: Average total Marks per category v/s Value of that category level

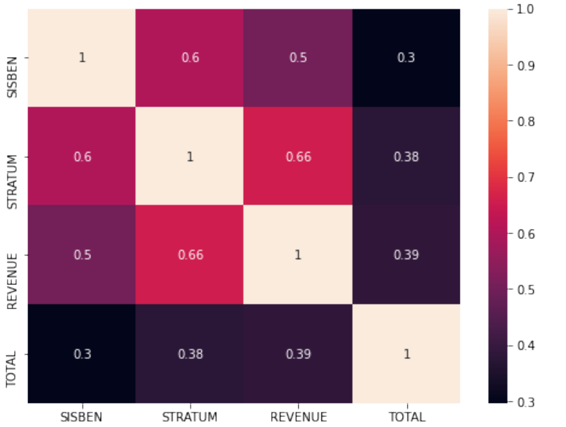


Figure 18: Correlation between different economic factors

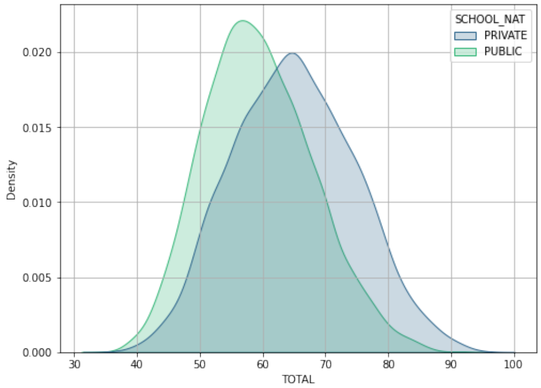


Figure 19: Category of the school and the corresponding marks distribution

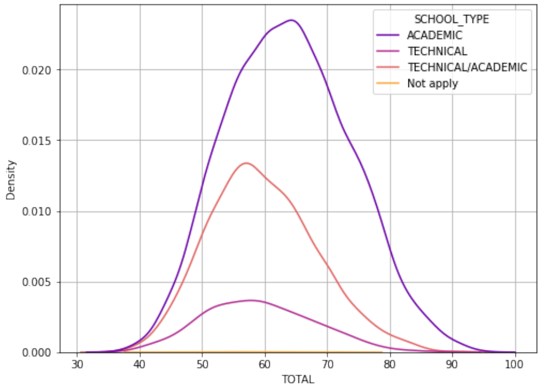


Figure 20: School type and the corresponding marks distribution

We can observe from Fig 17 and 18 that there is significant correlation between STATUS(STRATUM), Revenue and the SISBEN. In addition to this we observe that the correlation of these factors is almost similar accounting to ≈ 0.35. This suggest that there is some(weak) correlation of these factors with the total marks of the students.

Fig 19 and 20 shows the type of school and the affiliation of the school(private or public) and the corresponding distribution of the total average marks. The plots show that on an average the private school students perform better than the public school students while the peak is higher in case of public schools suggesting greater number of students having a higher score(high score of public category). Other plot shows that majority of the students appearing for the test belong to academic schools and the one belong to school classified as TECHNICAL/ACADEMIC occupy an intermediate position between the academic and technical school as expected.

5. **Relation with gender**

The notion that women perform worse than men when it comes to mathematics is widespread and to address the same notion let us consider the marks distribution distribution when grouped with respect to gender.

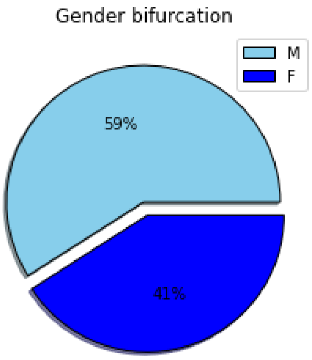


Figure 21: Population distribution according to gender

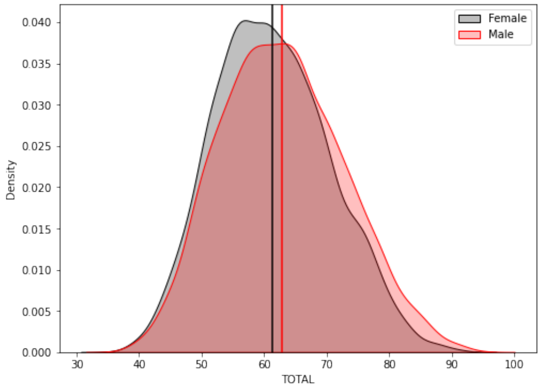


Figure 22: Total average marks distribution according to gender

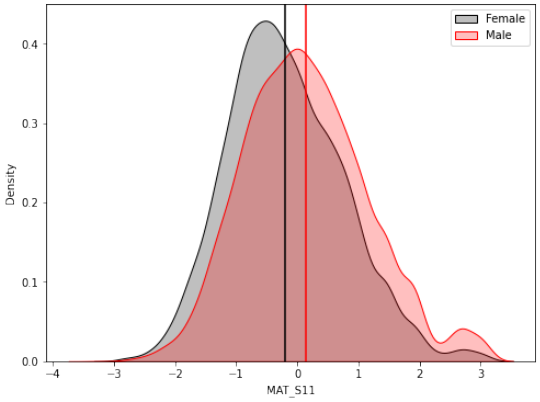


Figure 23: MAT-S11 (normalized) marks distribution according to gender

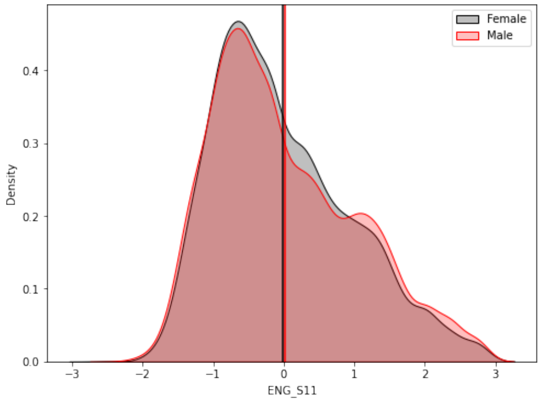


Figure 24: BIO-S11 (normalized) marks distribution according to gender

Fig 23 shows that the mean of the marks of Male candidates is higher than the mean of the marks of the Female candidates suggesting that this dataset is in accordance with the prevalent view of the society. We can also observe that on an average female students perform at par with male students in ENG_S11.
On calculating the confidence interval for the difference in the means of marks in MAT_S11 of female and male students when $\alpha = 0.05$, the interval came out to be **(0.3314 , 0.3329)** which suggests that the marks of male students are higher than female students and the average value of difference lies in the obtained interval with 95% accuracy.

6. **Grading**
NOTE: Grading policy followed is in intervals of 10 starting from F as $< 30, 30 < DD < 40$ and so on with $90 <= AA <= 100$ Finally we consider the grading policy that we generally use. We can come up with two types of grading either based on the total average marks obtained by the students or based on the percentile of the student.

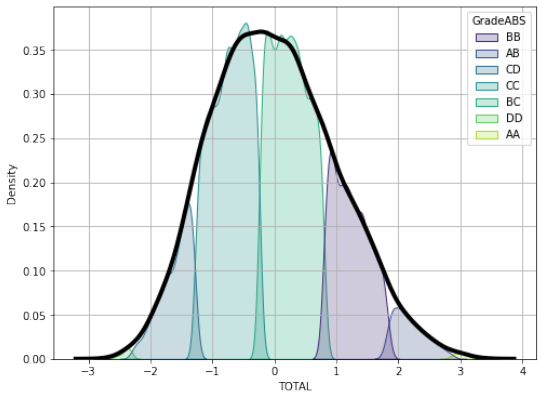


Figure 25: Conversion to 2 letter Grade from total average score

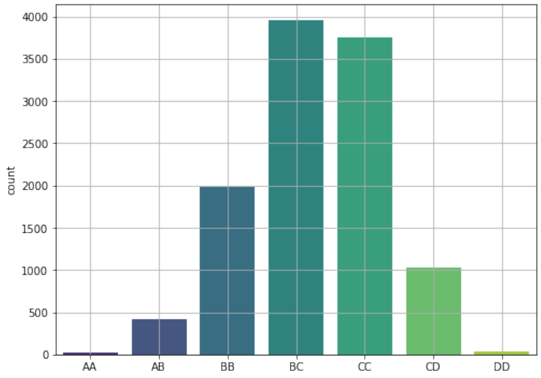


Figure 26: Frequency of 2 letter grades

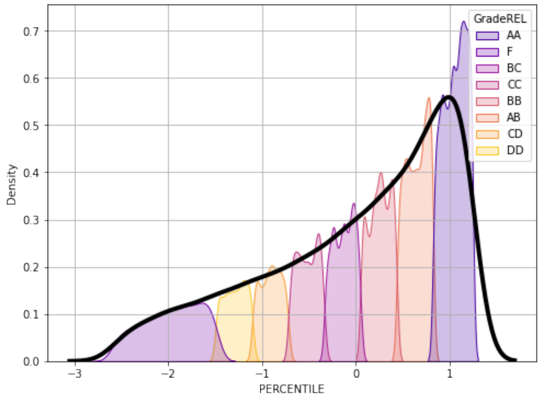


Figure 27: Conversion to 2 letter Grade from percentile

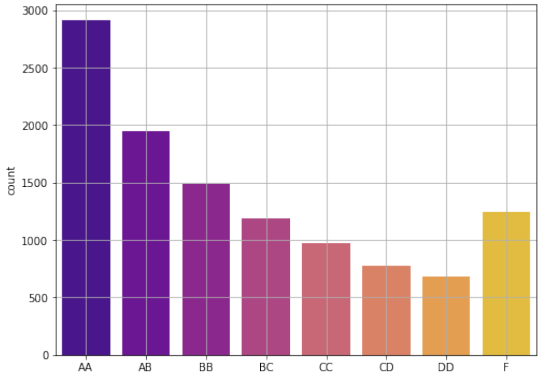
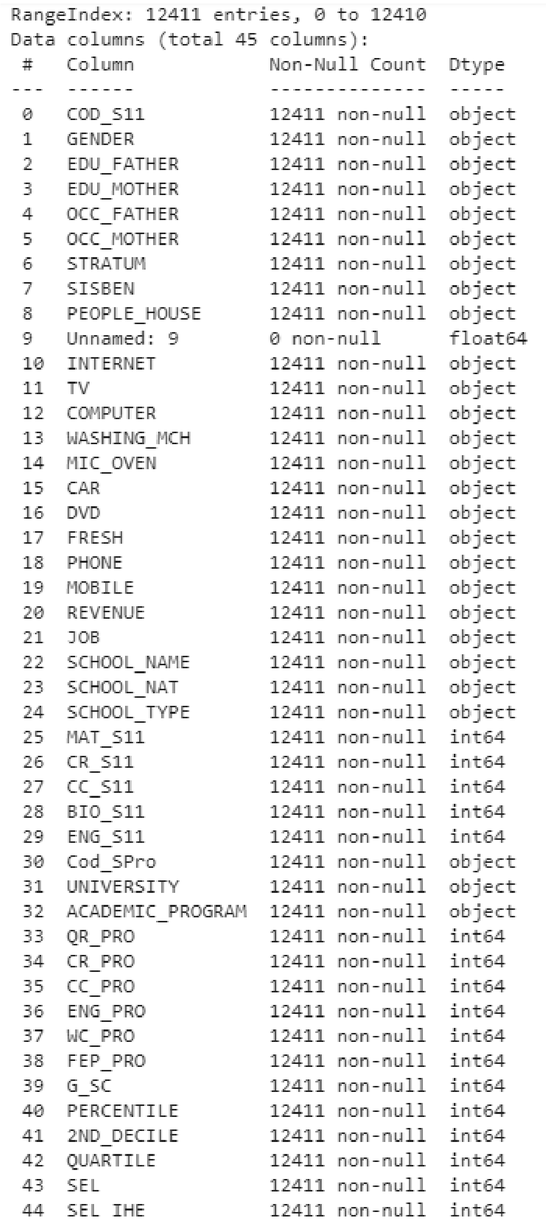


Figure 28: Frequency of 2 letter grades

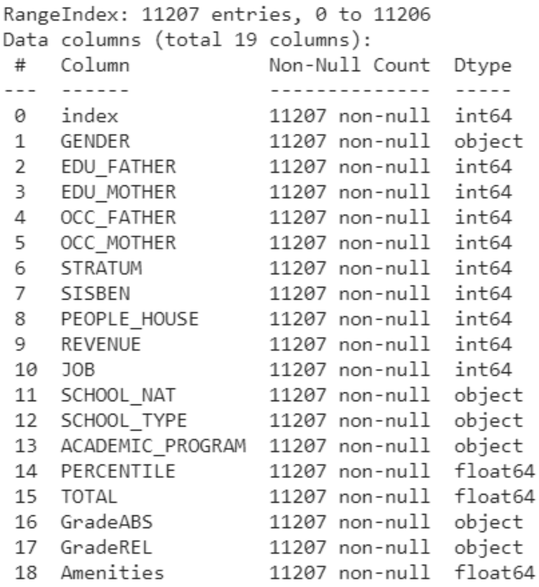
We can see that grading based on the average of all the subjects comes out to be a curve that approaches normal distribution whereas the grading done based on percentile does not resemble normal distribution. We can also observe from the Fig 27 and **??** that more students would receive higher grade(AA) as well as large number of students would fail(F grade) when grading is done based on the percentile instead of absolute score of the students. So, both grading policies in place have their own drawbacks and benefits.

All the observations and the systematic analysis of the data enabled us to draw significant conclusions and also helped to reduce the dimensionality of the data to a large extent as we clubbed the data of Category 1 and 2



```
RangeIndex: 12411 entries, 0 to 12410
Data columns (total 45 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   COD_S11           12411 non-null  object
 1   GENDER            12411 non-null  object
 2   EDU_FATHER        12411 non-null  object
 3   EDU_MOTHER        12411 non-null  object
 4   OCC_FATHER        12411 non-null  object
 5   OCC_MOTHER        12411 non-null  object
 6   STRATUM           12411 non-null  object
 7   SISBEN            12411 non-null  object
 8   PEOPLE_HOUSE      12411 non-null  object
 9   Unnamed: 9        0 non-null      float64
 10  INTERNET          12411 non-null  object
 11  TV                12411 non-null  object
 12  COMPUTER          12411 non-null  object
 13  WASHING_MCH       12411 non-null  object
 14  MIC_OVEN          12411 non-null  object
 15  CAR               12411 non-null  object
 16  DVD               12411 non-null  object
 17  FRESH             12411 non-null  object
 18  PHONE             12411 non-null  object
 19  MOBILE            12411 non-null  object
 20  REVENUE           12411 non-null  object
 21  JOB               12411 non-null  object
 22  SCHOOL_NAME       12411 non-null  object
 23  SCHOOL_NAT        12411 non-null  object
 24  SCHOOL_TYPE       12411 non-null  object
 25  MAT_S11           12411 non-null  int64
 26  CR_S11            12411 non-null  int64
 27  CC_S11            12411 non-null  int64
 28  BIO_S11           12411 non-null  int64
 29  ENG_S11           12411 non-null  int64
 30  Cod_SPro          12411 non-null  object
 31  UNIVERSITY        12411 non-null  object
 32  ACADEMIC_PROGRAM  12411 non-null  object
 33  QR_PRO            12411 non-null  int64
 34  CR_PRO            12411 non-null  int64
 35  CC_PRO            12411 non-null  int64
 36  ENG_PRO           12411 non-null  int64
 37  WC_PRO            12411 non-null  int64
 38  FEP_PRO           12411 non-null  int64
 39  G_SC              12411 non-null  int64
 40  PERCENTILE        12411 non-null  int64
 41  2ND_DECILE        12411 non-null  int64
 42  QUARTILE          12411 non-null  int64
 43  SEL               12411 non-null  int64
 44  SEL IHE           12411 non-null  int64
```

Figure 29: Initial data



```
RangeIndex: 11207 entries, 0 to 11206
Data columns (total 19 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   index             11207 non-null  int64
 1   GENDER            11207 non-null  object
 2   EDU_FATHER        11207 non-null  int64
 3   EDU_MOTHER        11207 non-null  int64
 4   OCC_FATHER        11207 non-null  int64
 5   OCC_MOTHER        11207 non-null  int64
 6   STRATUM           11207 non-null  int64
 7   SISBEN            11207 non-null  int64
 8   PEOPLE_HOUSE      11207 non-null  int64
 9   REVENUE           11207 non-null  int64
 10  JOB               11207 non-null  int64
 11  SCHOOL_NAT        11207 non-null  object
 12  SCHOOL_TYPE       11207 non-null  object
 13  ACADEMIC_PROGRAM  11207 non-null  object
 14  PERCENTILE        11207 non-null  float64
 15  TOTAL             11207 non-null  float64
 16  GradeABS          11207 non-null  object
 17  GradeREL          11207 non-null  object
 18  Amenities         11207 non-null  float64
```

Figure 30: Final data

5 Conclusions

From all the above inferences we could conclude the following:

1. Not all distributions that appear to be normal are actually normal this was the case for the average marks calculated for the five subjects. The ks-test results are accurate for larger problem size and show us the difference which is not even visible to the naked eye.

2. The data used for this project has majority of the population that is privileged. This is evident from the fact that majority of the population had access to more than 5 amenities that were present in the data. This notion was also verified by the SISBEN index and the STRATUM and revenue values.

3. The analysis based on gender suggested that the gender ratio of the population into consideration is skewed with male being in majority. Furthermore, this population had female students who performed at par with the male students in subjects like ENG whereas for the other subjects the performance was a little worse than their male counterpart.

4. The analysis of the grading based on absolute marks and the percentile showed that when the data is dense in a certain range then the percentile grading will have distribution such that more number of students will get extreme grades in comparison to the absolute grading wherein for this data extreme values were not that prevalent.

5. Finally, we can conclude that students having greater access to amenities along with a great family background wherein parents are well educated and belonging to a decent sized family which is well to do in terms of economic and social status tend to perform better than the other section of students who do not have the listed conditions.

The analysis done in this project, with respect to the data available, show that the family background, socio-economic level and the access to amenities have an effect on the results of the students in their examination. So, these conditions should be factored in for evaluating in a holistic manner.

6 Bibliography

1. De La Hoz, Enrique (2020), "Data of Academic Performance evolution for Engineering Students", Mendeley Data, V1, `http://dx.doi.org/10.17632/83tcx8psxv.1`

2. `http://documents1.worldbank.org/curated/en/364521468019731045/pdf/32759.pdf`

3. CODE file: `https://colab.research.google.com/drive/1e8gF6OlTpj864mNN0Yb6CHrR_J1AS-Z1?usp=sharing`