PROJECT FOR STATISTICAL METHODS I

Exploratory Data Analysis

Students of B.Stat. First Year (2020-21)

Semester I

INDIAN STATISTICAL INSTITUTE

Online Classes and their effect on Students in 2020

Fall 2021

Contents

1 Introduction

The year 2020 has been a tough one for all. The pandemic had led to the postponement of all events that the year had in store. Needless to say, it was a tough one for the students too. From the customary chalk-blackboards to Zooms and Skypes, we faced all the transitions that could be. Since we, as students in this fateful year, faced a lot of ordeals and challenges while attending online classes, we were curious to know how our peers adapted to the same. Hence, we present here our project on "**The effect of Online Classes on the students in 2020**".

2 Procedure of data collection

"Without data, you are just another person with an opinion."

-Edwards Deming.

We have collected data through a Google form for this project. The link to the form is here

The form was intended to be filled by our friends from other colleges and schools. The population for the data collection involved students from class 12, freshman and sophomore year. Initially, there were 240 people who volunteered to provide us data, and after scrutinisation, the number reduced to 226, with 167 first year students, 37 second year students and 22 class 12 students.

The students were from different parts of the country, but mostly there were students from West Bengal. Our population consisted of the people mostly from urban areas, but there were a few respondents from rural areas as well. This showed that people from various localities are having online classes for their colleges and schools. Due to several restrictions on social gatherings and mobility, we could not collect data from an even wider demographic and thus, the inferences we get from the data analysis may be a little biased towards the data obtained from the students who lead an urban lifestyle.

3 Exploratory Data Analysis.

"Data by itself is useless. Data is only useful if you apply it."

-Todd Park.

The data that we have collected have pointed out several aspects of the online classes, that we wouldn't have known otherwise. For example, the maximum number of students in the data collected use a laptop/desktop to attend online classes, and most of the users of laptops/desktops also use mobile phones, possibly as a backup. Moreover the data collected from students from rural areas tell us that all the online class attendees use mobile phones and a laptop/desktop, and there are none who use a mobile phone alone. The cause of this maybe traced back to the recent explosion in accessibility of affordable technology and easy access to the internet, and since it is an uphill task to study, for both pedagogues and students from phones, a laptop/desktop is usually necessary.

The entire data was analysed by 6 groups. The following page contains the variables for the project, and then we have the conclusions for each of the groups.

The variables that were considered while analysing the data are as follows:-

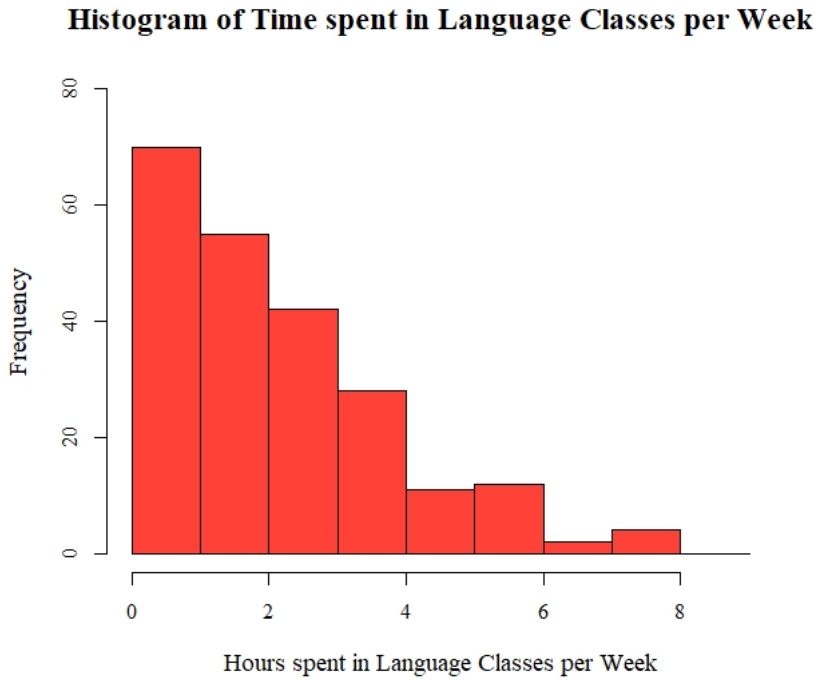
S. No.	Description	Variable	Type
1.	Name of Student	Name	Nominal
2.	Gender of student	Gender	Categorical
3.	Age of student	Age	Continuous
4.	Place of Residence	Place	Nominal
5.	Grade studying	Standard	Ordinal
6.	Institute of the Student	Institute	Nominal
7.	Devices used for class	Device	Categorical
8.	Average Internet Speed	Speed	Continuous
9.	Software platform used	Platform	Categorical
10.	Time spent on online classes in hours per week(Math)	M.Hours	Continuous
11.	Time spent in online classes in hours per week (Language)	L.Hours	Continuous
12.	% of marks in Math(online)	M.Mrks.on	Continuous
13.	% of marks in Language (online)	L.Mrks.on	Continuous
14.	% of marks in Math (offline)	M.Mrks.off	Continuous
15.	% of marks in Language (offline)	L.Mrks.off	Continuous
16.	Satisfaction Level Online Examination(Math)	M.Sat.on	Ordinal
17.	Satisfaction Level Teaching quality (Math)	M.Sat.teach	Ordinal
18.	Satisfaction Level Overall experience (Math)	M.Sat.total	Ordinal
19.	Satisfaction Level Online Examination (Language)	L.Sat.on	Ordinal
20.	Satisfaction Level Teaching quality (Language)	L.Sat.teach	Ordinal
21.	Satisfaction Level Overall experience (Language)	L.Sat.total	Ordinal

We present here the analysis done and the conclusions reached by each of the groups. Out of the 6 groups, 3 worked on the univariate plots, 2 on bivariate plots, and one group did the regression modelling of the data.

3.1 Univariate 1

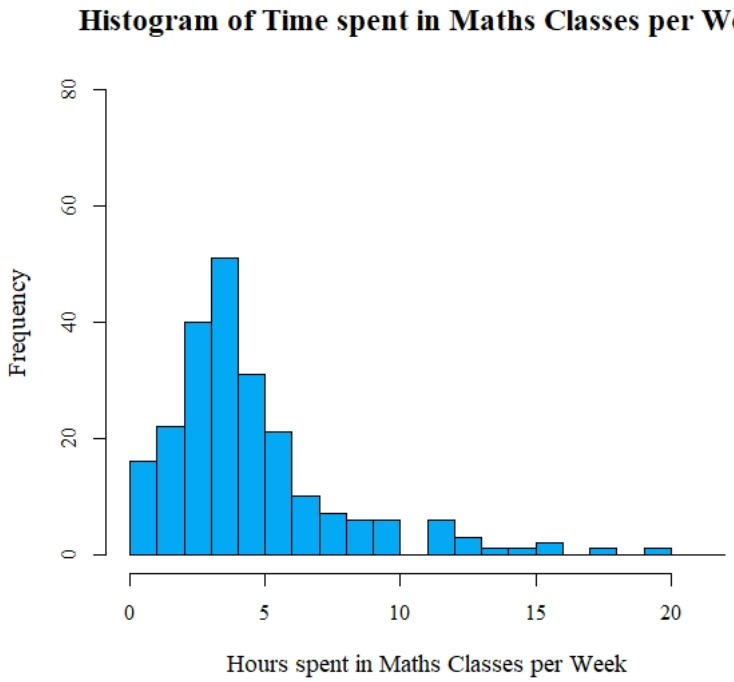
The students in this group were Samprit Chakraborty, Rahul Debnath, Abhishek Maiti, Arka Sinha, Arunsoumya Basu, Avinandan Roy, Sounak Das and Souvik Roy.

This group worked on univariate data. First, we have the histogram for the time spent in language classes per week.



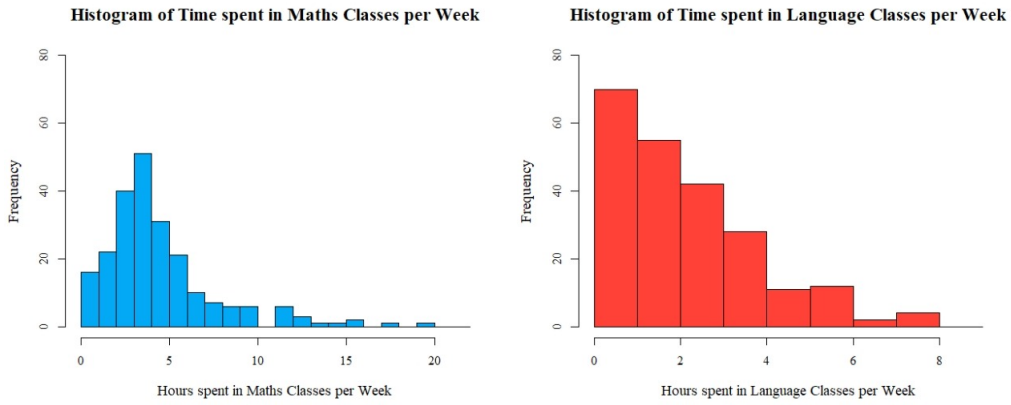
From the above histogram, we conclude that among the students who have any language as their subjects, the mode is attained in the interval 0 to 1. Thus, the class (0,1) is the modal class. This, means that most of the students who have language, have 0 to 1 hours of classes per week. The cause of this might be the population that we selected, as most of them were science students in their respective colleges, and so the number of weekly hours of Language classes is less.

If we similarly look at the time spent on Math classes per week, then the number of hours spent becomes substantially high.



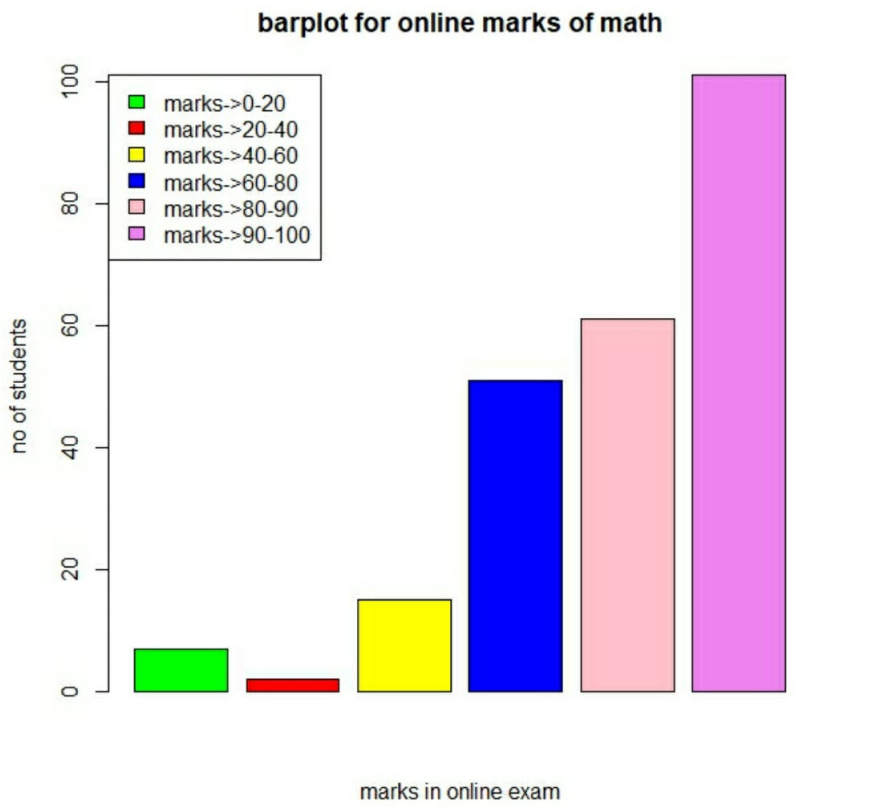
As it is evident from the histogram, the number of hours for Math classes has increased, which is obviously due to the population that was selected. Moreover, the class 3 to 4 has the highest number of datapoints, that is, most of the students have 3 to 4 hours of Math classes every week. There are also a few students who have more than 10 hours of Math classes per week. It is likely that they are pursuing a Major in Mathematics.

We now look at the variation of the time spent in Language and Math in a side by side histogram.

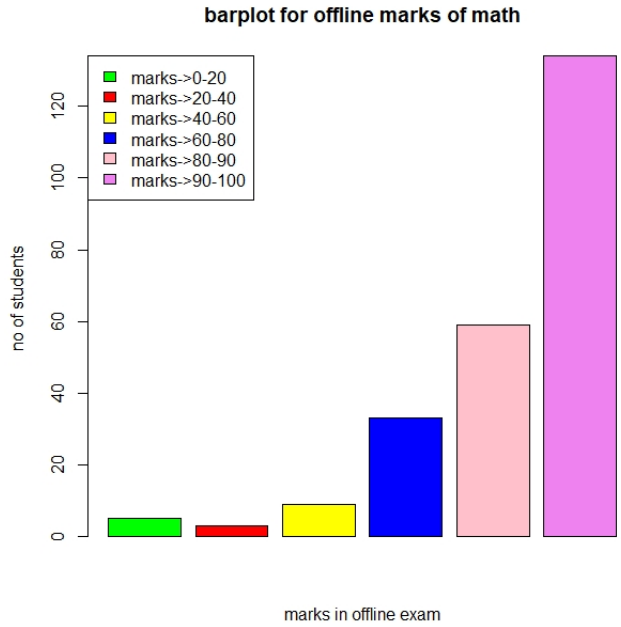


3.1.1 Marks in Math for Offline and Online examinations

The following histogram shows the marks in Math only for online examinations.

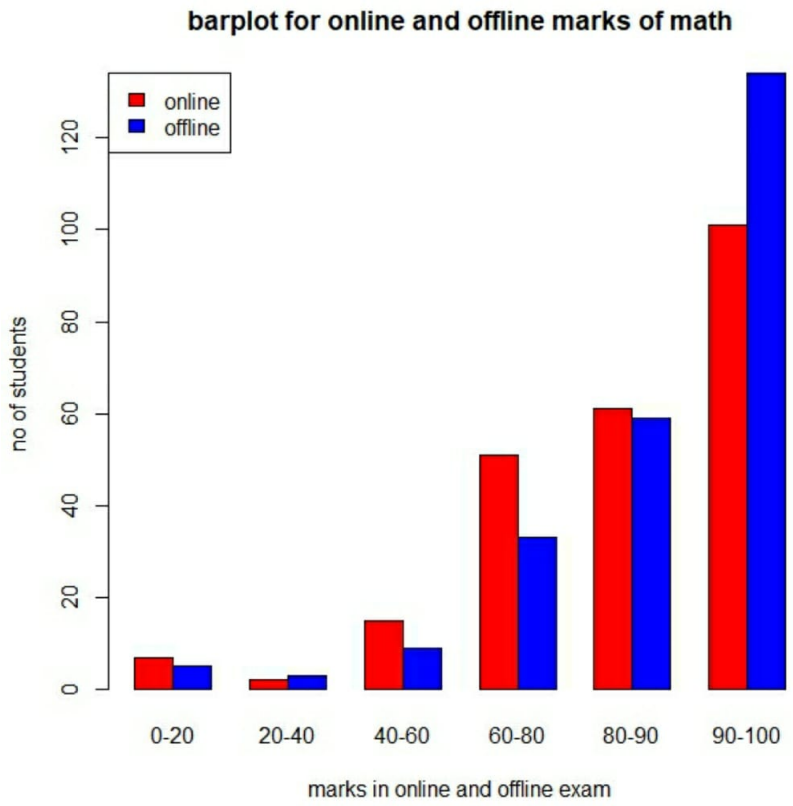

barplot for online marks of math
marks->0-20
marks->20-40
marks->40-60
marks->60-80
marks->80-90
marks->90-100
no of students
0
20
40
60
80
100
marks in online exam

The following one shows the histogram for the offline scores in Math.

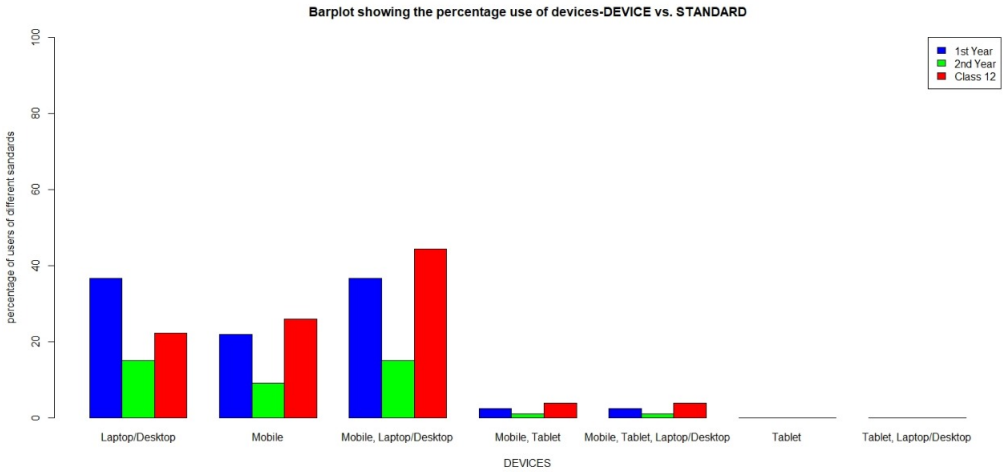


We observe that the number of students scoring more than 60 marks shows an increasing trend for both offline and online, and moreover, most of the students have marks in the range 90 to 100. This shows the trend in the education that has varied over the years, and the subsequent grading systems and changes in the patterns of the questions asked to students, that has resulted in most of the students getting high marks. Also, these being the marks in the online examinations, and most of the students having online examinations for the first time, we can assume that the grading procedure was lenient enough. Moreover, it also depends largely on the students in our population.

We now have the histogram comparing the offline and the online marks.



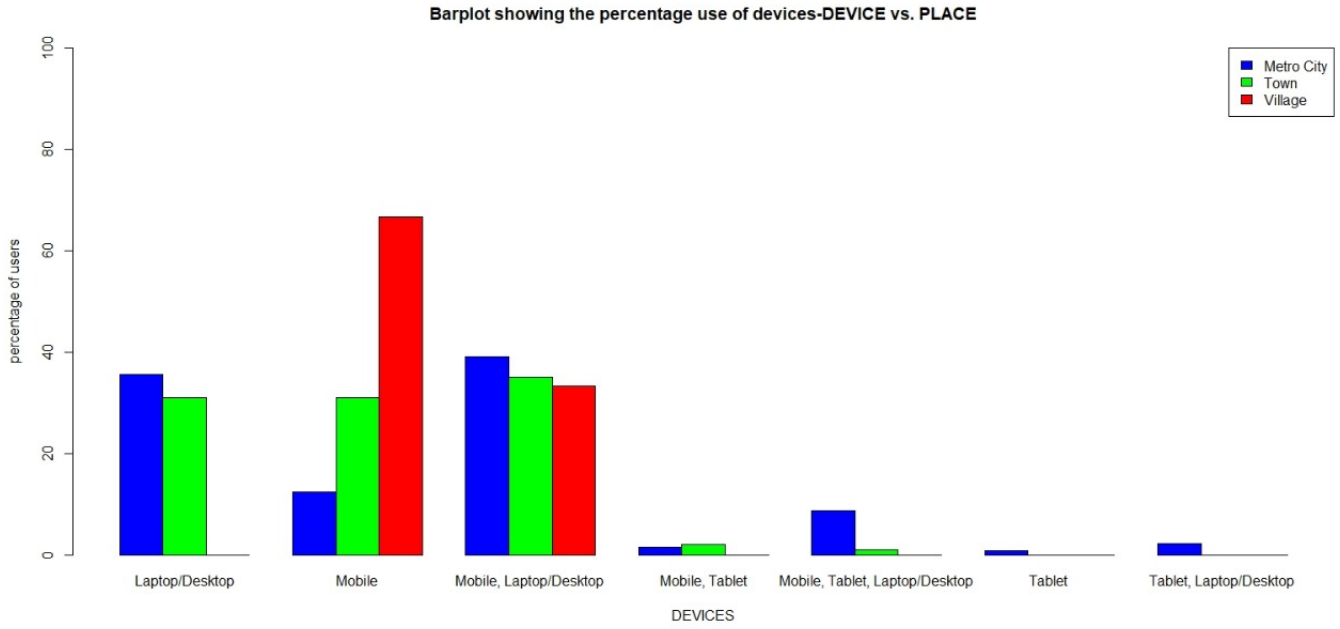
We now study the devices used by the students and compare them to the standard they study in. The following histogram shows the relevant comparison.



We observe that irrespective of the device, the students from class 12 are the ones who use them the most. This maybe because the class 12 students were the ones who began their classes before the first or the second year students during the lockdown that was imposed due to the pandemic. But the first year students are the immediate next ones, followed by second year students.

3.1.2 Devices vs Place

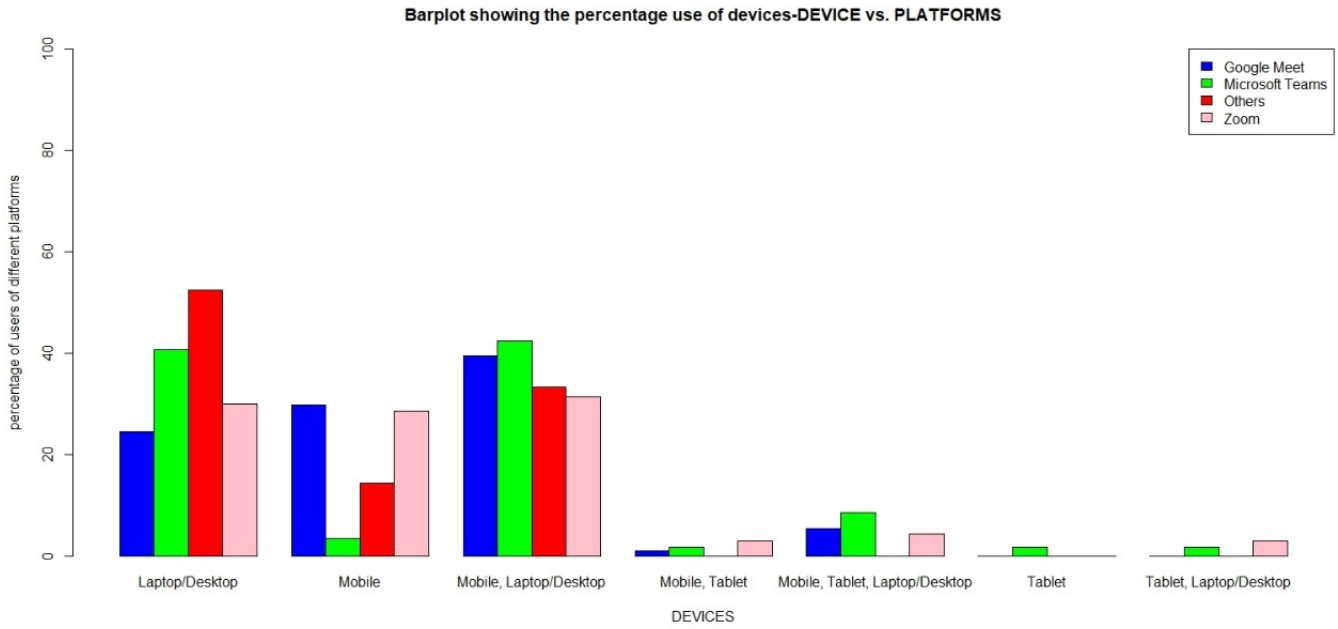
We expect the students from rural areas to have the minimum number of devices, and the following histogram solidfies the same.



It is also evident that most of the students from villages use a phone, as compared to laptop/desktop. It is also noticeable that all the students from villages use a phone and all the students from the villages who use the laptop/desktops, also use a phone.

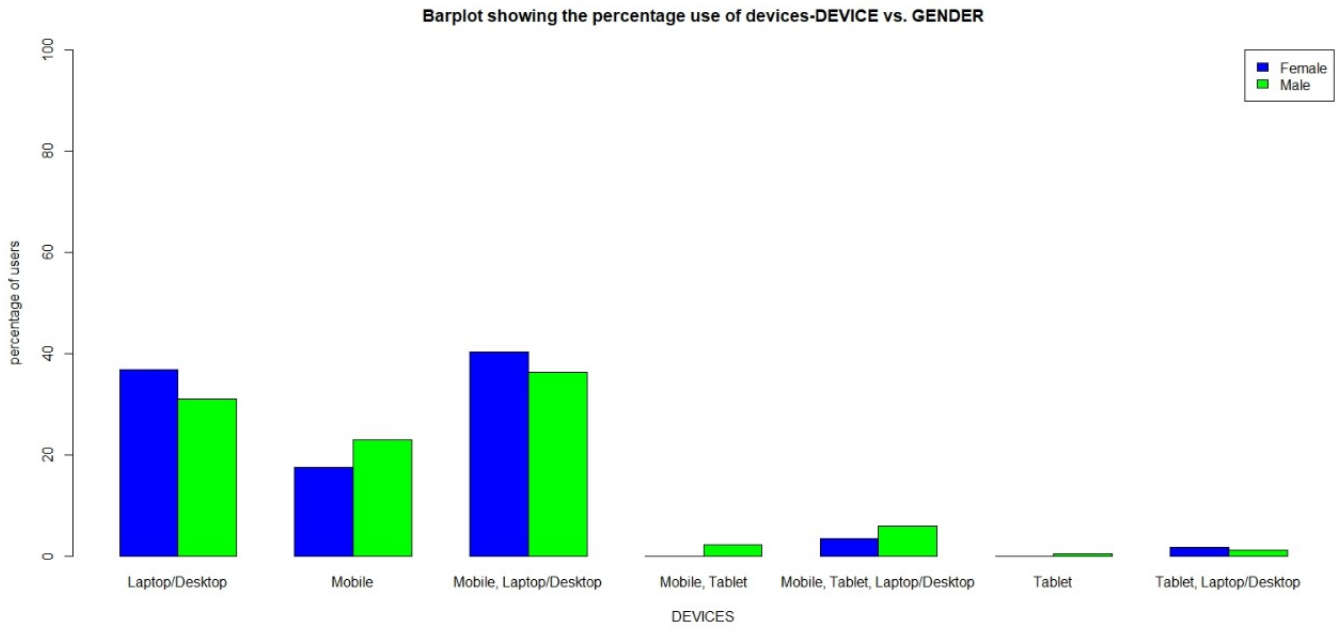
3.1.3 Devices vs Platforms

We now look at the histogram of the devices vs platform.



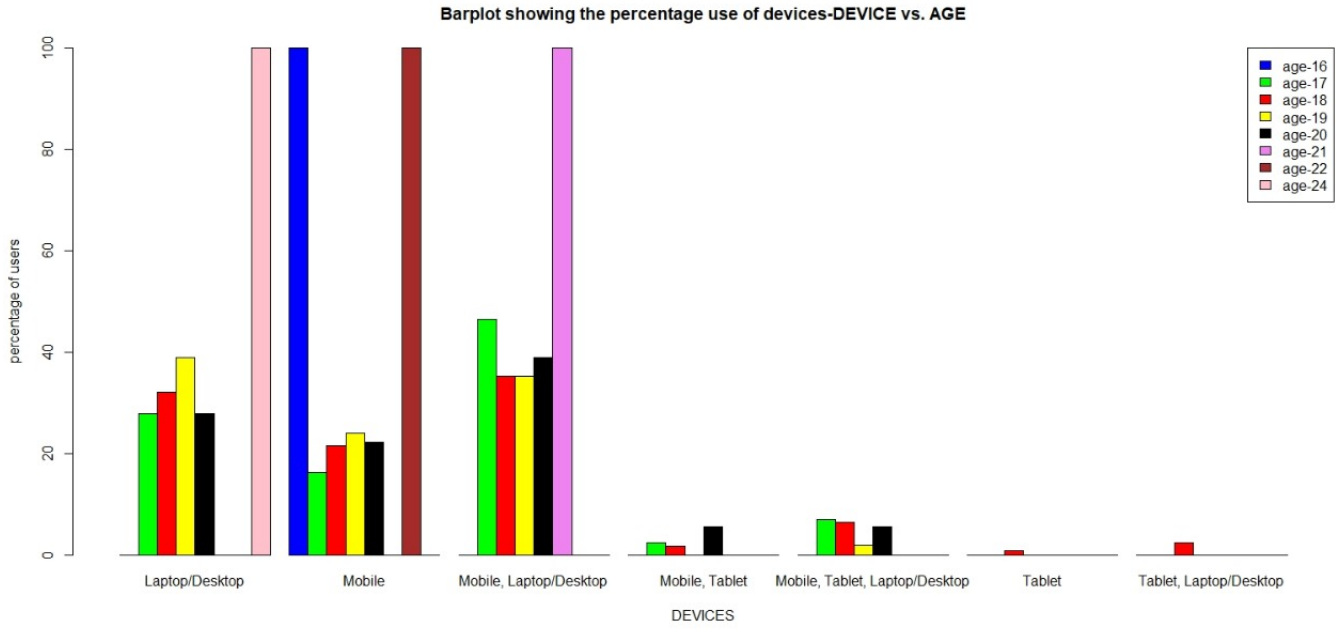
We deduce that most of the users use a laptop/desktop and/or a mobile phones. It is also to be noted that tablets are used by a very small proportion of the population, probably because it is not as convenient as a mobile phone or computer. Moreover, we have that Zoom, **contrary** to our expectation, is **NOT** the most used platform across all the students in our population.

3.1.4 Devices vs Gender



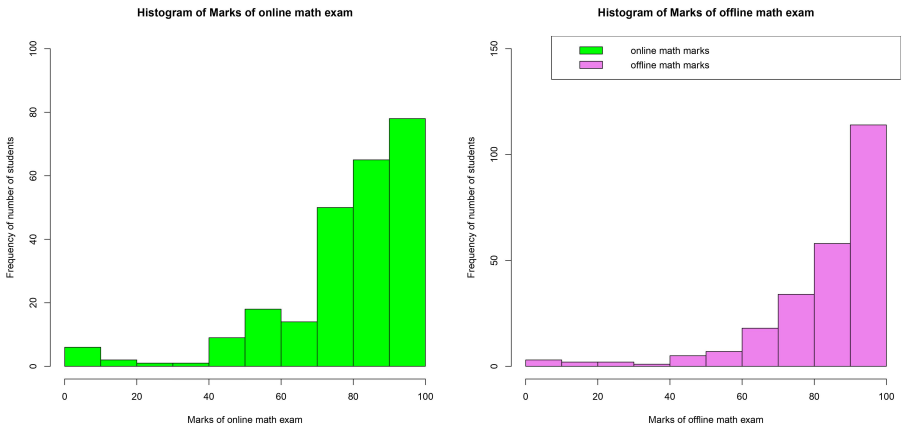
We observe that the percentage of users using any particular combination of devices is almost same for both males and females.

3.1.5 Devices vs Age



The above graph shows the histograms for the proportions of the ages of students who use the respective devices. It is observed that most of the students who are aged 20 or above use a mobile phone to attend online classes. On the other hand, most of the users using most of the devices are aged 18, which is the usual age of first year students and class 12 students.

3.1.6 Marks in Mathematics



The above histogram shows both the online and the offline marks obtained by the students in Mathematics. It is easy to see that most of the students, irrespective of the medium of the classes, have scored more than 90 in Mathematics. Moreover, it is also observable that, there is a sharp increase in the number of the students getting marks more than 70 for both offline and online classes. A steadier increase can be noticed for the offline classes, when the marks are above 40. Moreover, it is observable that, for online classes a substantial amount of the population has entered 0 as their marks in Mathematics, which may be due to the fact that they did not write any online examination in Mathematics before the survey was conducted.

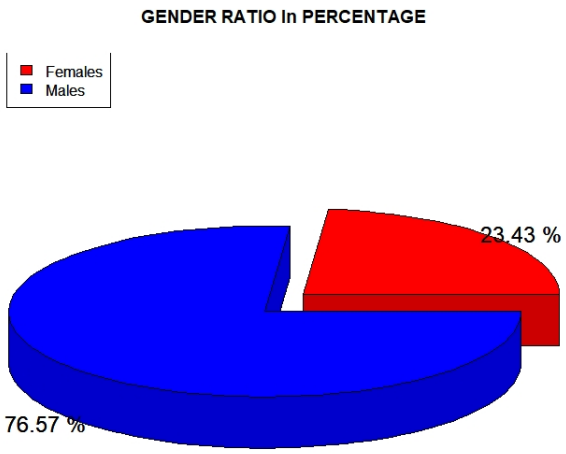
3.2 Univariate 2

The students in this group are Rajarshi Biswas, Pranay Samadder, Saptarshi Sinha, Ishan Paul, Tarun Agarwal, Debabrata Sarkar and Yash Maurya.

This group focused on the univariate analysis of:

- Gender of the students.
- Time spent in online language classes.
- Marks obtained in previous examinations on language.
- Satisfaction Level regarding online language course.

3.2.1 Gender

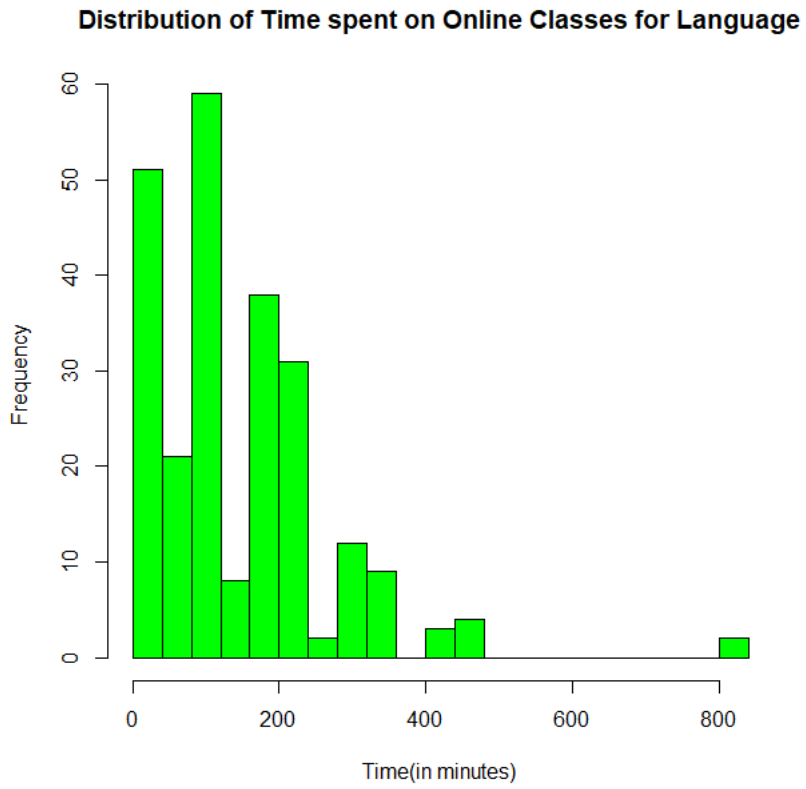


GENDER	NUMBER	PERCENTAGE(%)
MALE	187	76.57
FEMALE	57	23.43

It was initially observed that the male-female ratio in the population was 76.57: 23.43. This ratio was for the corrected dataset, which had 244 datapoints. The criteria adhered to by Group 3 while discarding the datapoints during scrutiny was the time spent and the marks correlation.

A datum was discarded if it was found that the time spent in Language online classes was 0, and the student scored positive marks. The software **R** was used for all kinds of computations.

3.2.2 Time



This histogram shows the time spent(over a week) by the respondents while attending online classes for Language. The time spent is in minutes. The five point summary and mean obtained for the above histogram is as follows,

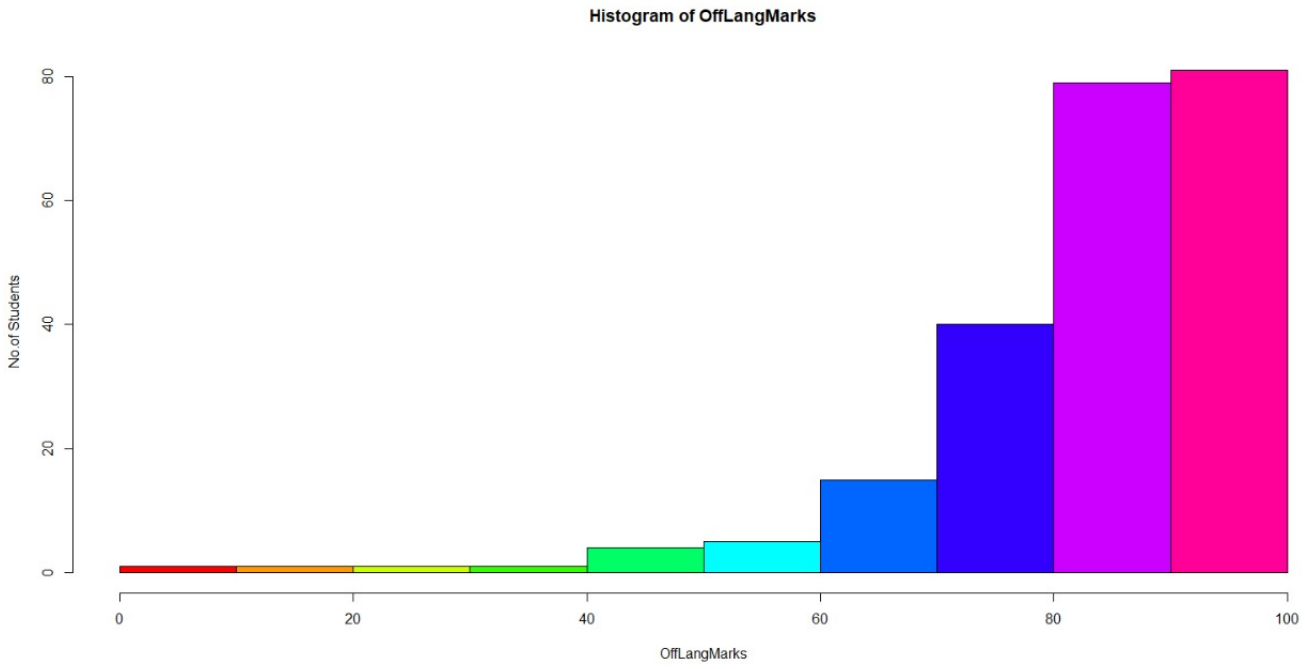
Min.	1st Quartile	Median	Mean	3rd Quartile	Max
0.0	60.0	120.0	148.6	217.6	840.0

3.2.3 Offline Marks in Language

Here we are analysing the performance of students in their last offline language examination. At first we have created the frequency distribution table of the offline marks by deleting the 0 marks as it is impossible to get 0(nowadays) in the last offline examination in language.

Language Marks in offline exams (in form (a,b])	Class Frequency (f_i)
0	0
0-10	1
10-20	1
20-30	1
30-40	1
40-50	4
50-60	5
60-70	15
70-80	40
80-90	79
90-100	81
Total:	$\sum f_i$=228

The marks distribution in the form of a histogram is as follows:



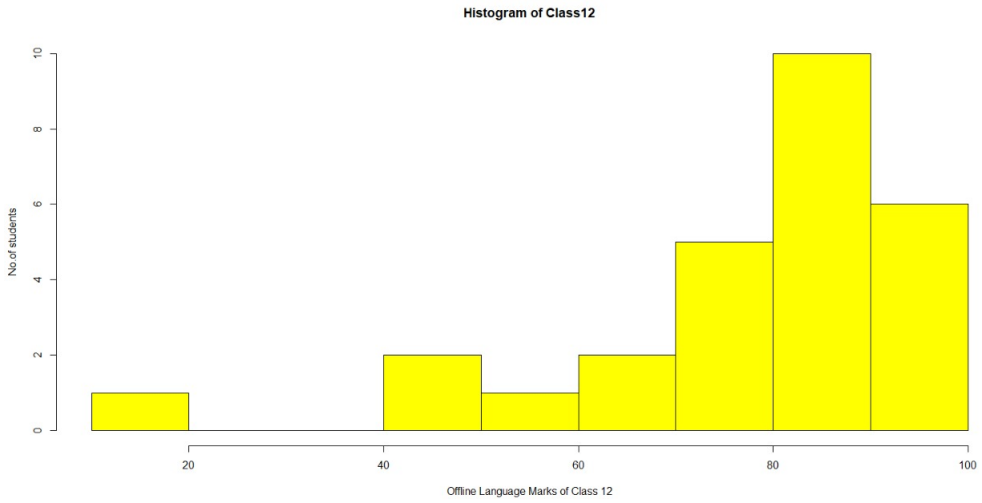
Min.	1st Quartile	Median	3rd Quartile	Max.
6.00	80.00	90.00	95.00	100.00

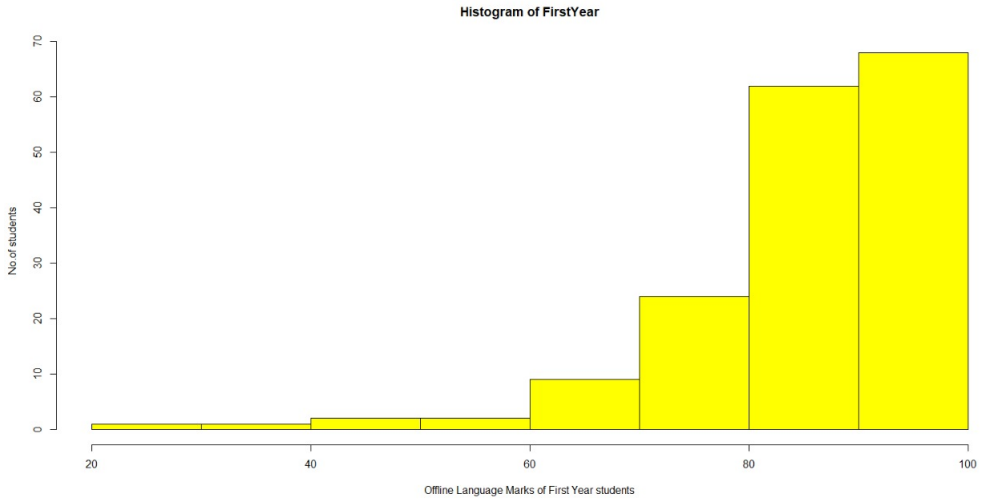
Note: The above marks are reported in percentage scores.

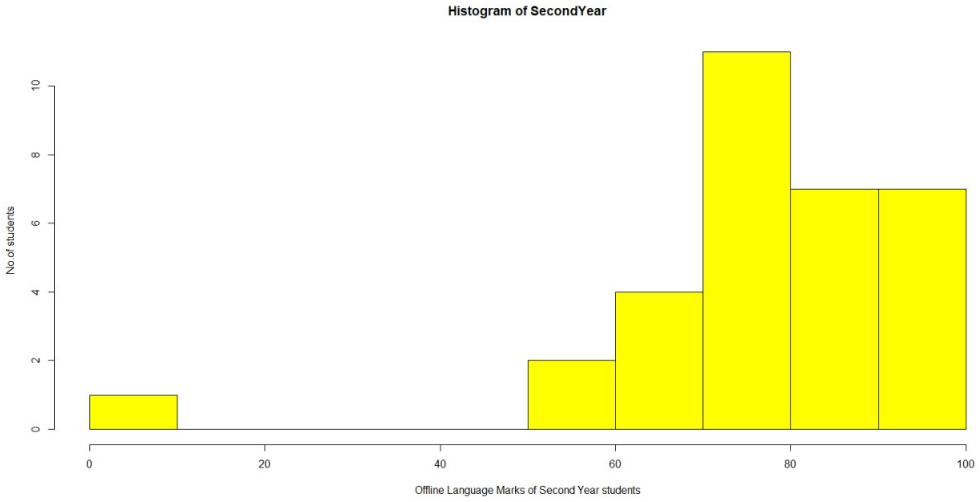
- The mean score in language came out to be: 85.13

- About 75% of the student population scored 80 or above and about 25% of the total students got 95 or above.
- The IQR in offline language marks is **15.00**
- We know that the online semesters started during the time when the students are promoted to a higher class so the last offline exam that the students gave were their final examinations so they had to pass. One can see that there are very few datapoints below 40 marks and the graph goes up after the 40 mark. The reason for this can be that the passing criteria is about 40% at most of the institutions.

The remaining histograms follow.







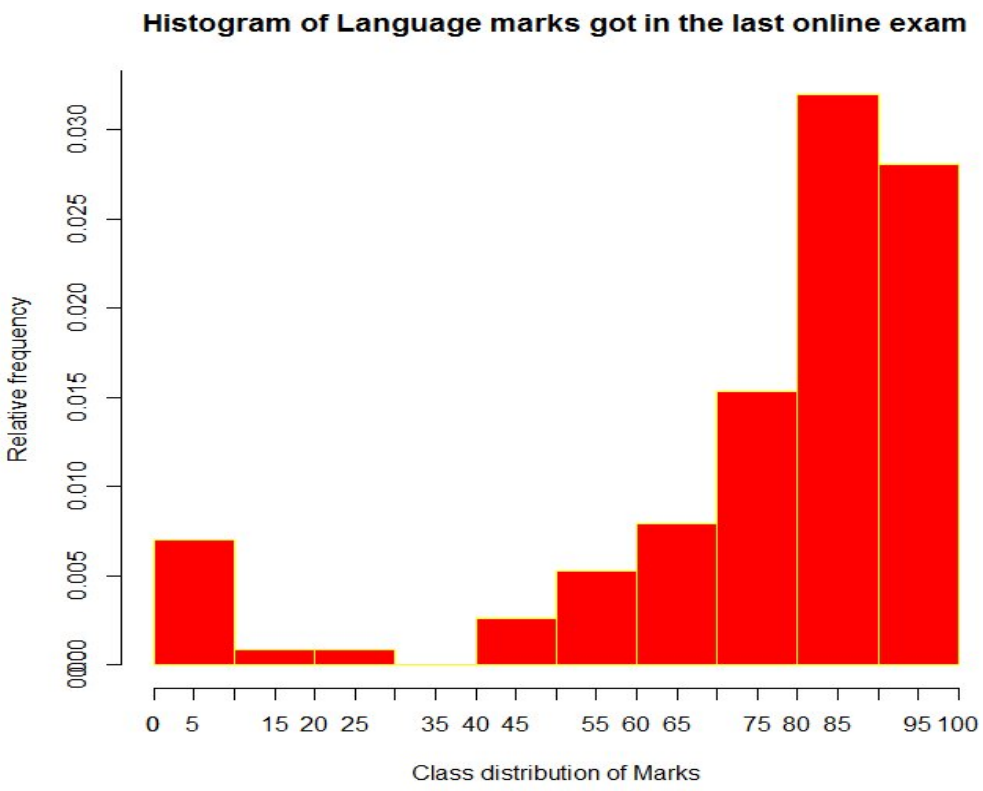
From these histograms we observed that the region below 40% is still scarce and majority of people have above 70% marks for all three groups.

3.2.4 Online Classes in Language

We are going to analyze the marks obtained by the students (who cooperated with us in this survey) in the last online exam they have given on subjects like language and literature. Obviously there are students who have not given any test in online mode on these subjects. They have entered 0 in the marks section. The frequency distribution table looks like this:

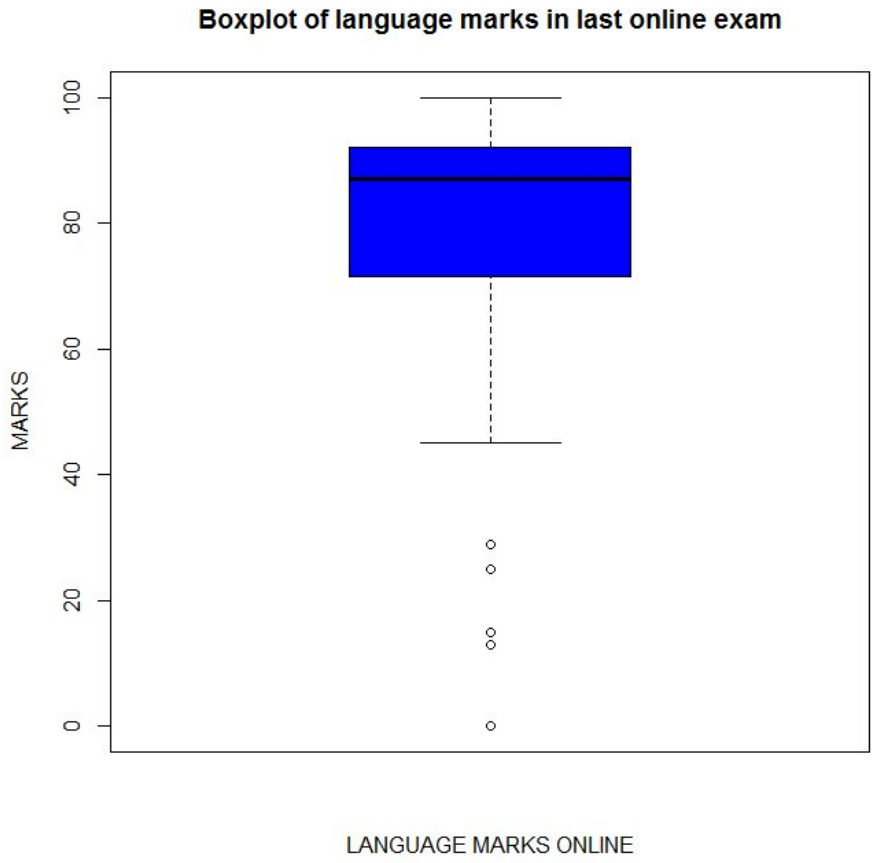
Language Marks in online exams (in form (a,b])	Class Frequency (f_i)
0	16
0-10	0
10-20	2
20-30	2
30-40	0
40-50	6
50-60	12
60-70	18
70-80	35
80-90	73
90-100	64
Total:	$\sum f_i$=228

We also draw the histogram of the online marks with the marks(X) in percentage(out of 100) Vs. their relative frequency(Y). The histogram looks like this:



We plotted the histogram to observe something from the distribution of the number of students along the marks classes. From the histogram, we see that the height of the bars in the region from marks=0 to marks=40 excluding 0 is extremely small. This means that almost all of the students scored more than the passing marks (generally taken as 40%). The students who got 0 in this plot probably had not given any online exams on this field. Majority of the students are in the region 80%-90% making this the modal class.

To get a clear idea about the median marks and outliers, we also plotted a boxplot below



From the boxplot, we can see that the median is around 90% and half of the data points lie between around 70%-90%. The IQR is **20.25** There are outliers which are in the low-scoring regions, which is under 40%.

Summaries

After plotting, we computed the following summaries (Using R):

- The mean of the online language marks is 77.52, pulled down due to the 16 0s in our data.
- The five point summaries are

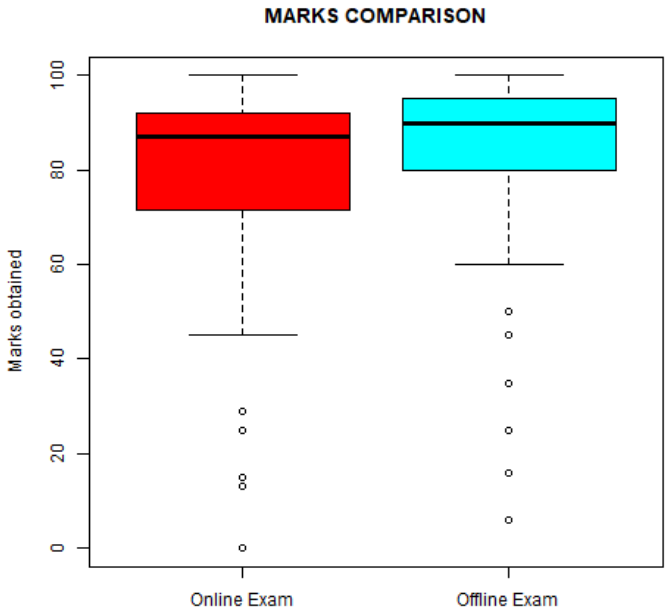
Min.	1st Quartile	Median	3rd Quartile	Max.
0.00	71.75	87.00	92.00	100.00

- The Standard deviation is 25.86614. The variance is due to the presence of many outliers.

We are considering the online marks in language for 228 students out of 244 students.

3.2.5 Offline vs Online marks Comparison

Following is the boxplot for marks scored in online exams(left) and offline exams(right).



From these plots we observe that:

- Median of offline marks is higher as compared to online marks.
- Variation in marks is lower in offline exams as compared to online exams, which implies that performance of the students is more consistent in the former.
- The IQR for online language exam marks is **20.25** and the IQR for Offline language examination marks is **15.00**

Clearly, offline classes are more beneficial for the students.

We may speculate reasons for this drop in performance. Possible reasons are:

- Less time spent in classes.
- Difficulty in following the classes.

3.2.6 Satisfaction Level

We have analyzed the satisfaction and thoughts of students of class 12, first and second year of college for the quality of online exams, teaching quality of language subjects like English or any

other regional language over online platforms and the overall satisfaction regarding this new medium.
We could see in the bar plot of satisfaction level of online exam quality, teaching quality of language subjects and overall satisfaction, that students are categorized in five different categories i.e., **Great** , **Satisfied** , **Neutral** , **Not so good**, and **Very bad**.

Firstly let us see frequency of all categories : -
Total population= 244

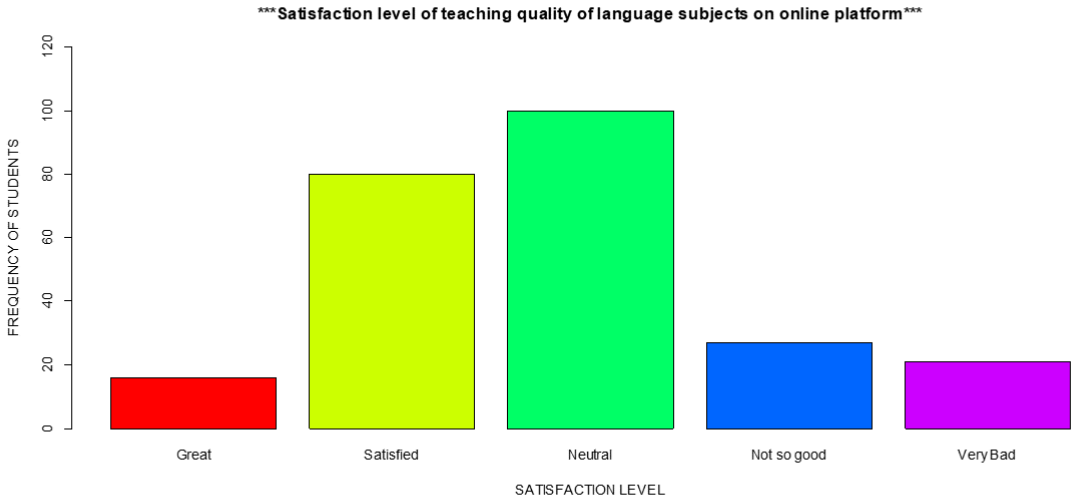
NOTE: 244 data points have been used instead of 285 data points as some of the population haven't chosen the language subjects.

Online Teaching Quality

Great	Satisfied	Neutral	Not So Good	Very Bad
16 (6.55 %)	80 (32.78 %)	100 (40.98 %)	27 (11.06%)	21 (8.60%)

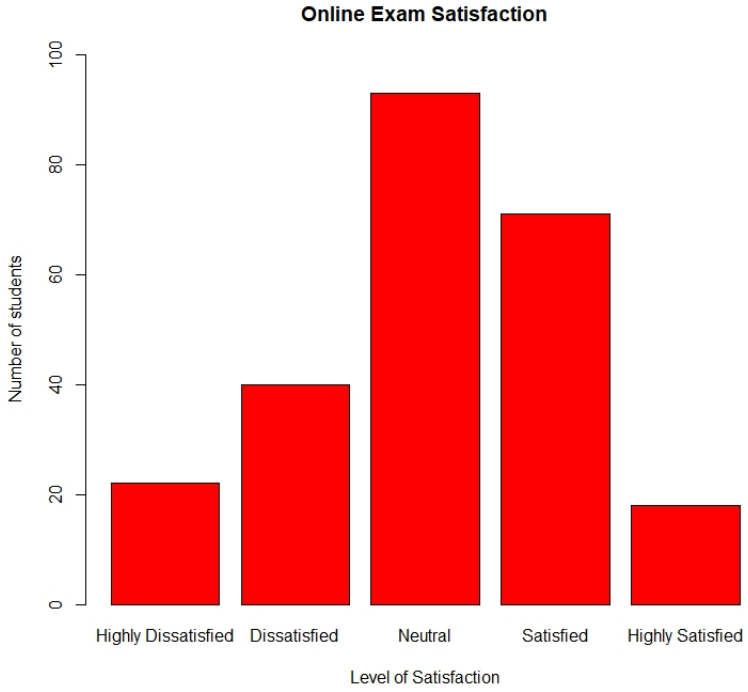
We can see that approximately 41% of the population are having Neutral opinion over teaching of language subjects through online mode. This can be because this dataset has been collected mostly from students studying science primarily and it is seen that generally, scoring in language subjects is not in their priority. Nearly 33% of the population is satisfied with online teaching of language subjects. Then approx. 11% population is not satisfied and 8.6% population has very bad experience with language being taught online. At last 6.55% of the population has had great experience.

The histogram depicting the above follows.



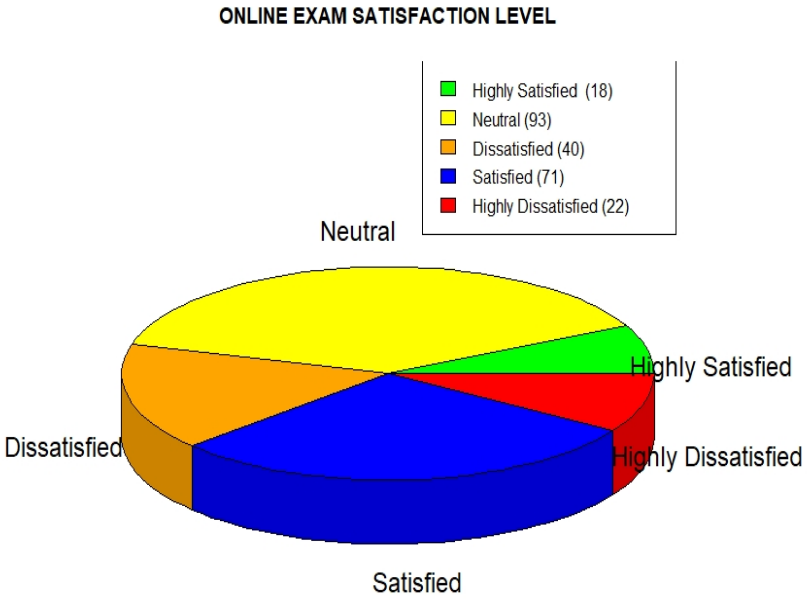
ONLINE EXAM QUALITY

Great	Satisfied	Neutral	Not So Good	Very Bad
18 (7.37 %)	71 (29.09 %)	93 (32.13 %)	40 (16.39 %)	22 (9.02%)



It is observed that the online exam of language is found to be neutrally likely i.e. neither satisfying nor dissatisfying by most people . A little amount of people did not like this mode of examination

as per the graph. As majority are inclined towards central and not towards extreme poles, so it is more inclined towards being symmetric.

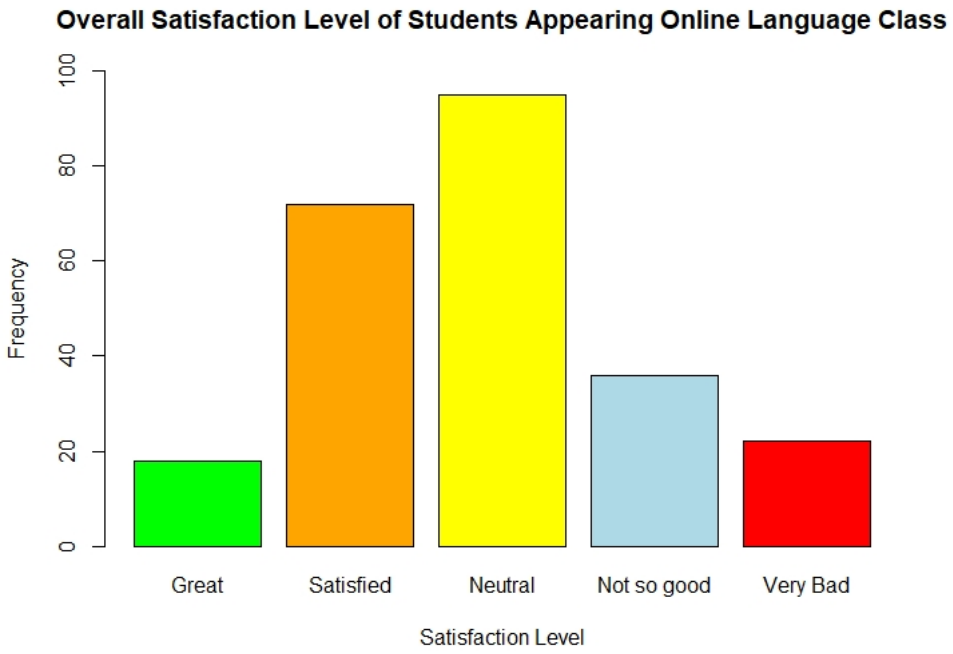


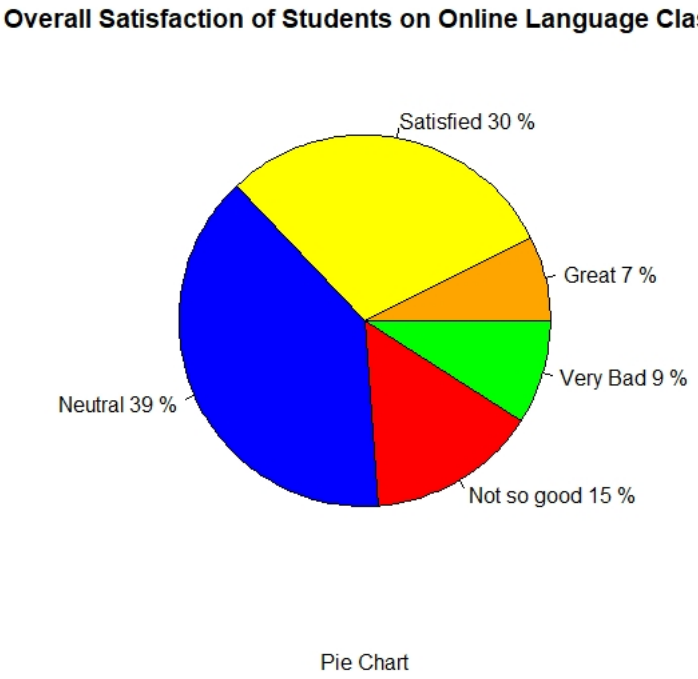
Most people do not have a negative outlook towards online exams that have started due to the pandemic . So overall it seems that online exams in language have not affected students' views at all .

Overall Level of Satisfaction with Online classes.

Great	Satisfied	Neutral	Not So Good	Very Bad
18 (7%)	72 (30%)	95 (39%)	36 (15%)	22 (9%)

We have also studied the **overall** satisfaction levels among students and observed that 37% of the students are comfortable with online classes. The plots look like this:





From the Bar graph, as well as the Pie chart, we saw that there is about 37% positive response(satisfied and great) and 24% negative response(not so good and very bad) excluding the neutral ones. This means that there were more happy people than unhappy ones in the sample.

3.3 Univariate 3

The members in this group were Manas Patnayakuni, Valay Ashish Shah, Sanku Bhaskara Rahul, Nithya Dev Allamneni, Shivendra Singh, Pilli Srikar Babu and Amarjeet Kumar Bhanu.

3.3.1 Devices used for online classes

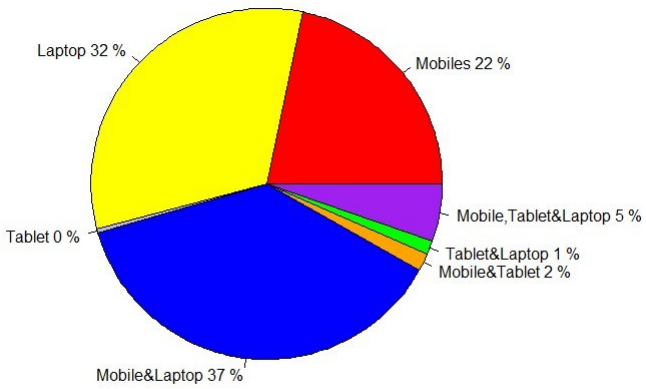
The devices usd by the students for attending the online classes are:

- Mobile
- Laptop/Desktop
- Tablet

From the survey we got,

Devices	Frequency
Mobile	53
Laptop	79
Tablet	1
Mobile and Laptop	91
Mobile and Tablet	4
Tablet and Laptop	3
Mobile, Tablet and Laptop	13

The pie chart denoting the above proportions is as follows:



Summary

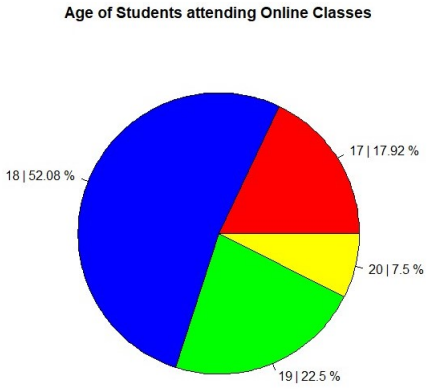
The pie chart provides the proportion of the students using various devices for the online classes. We can say that most of the students use multiple devices (mostly laptop and mobile phones). If we consider the students using single devices then most of them use laptops. We also observe that tablet is the least used device to attend the online classes.

3.3.2 Age of the students attending online classes

Frequency Table

Age	16	17	18	19	20	21	22	24
Frequency	1	43	125	54	18	1	1	1

Pie chart



Summary

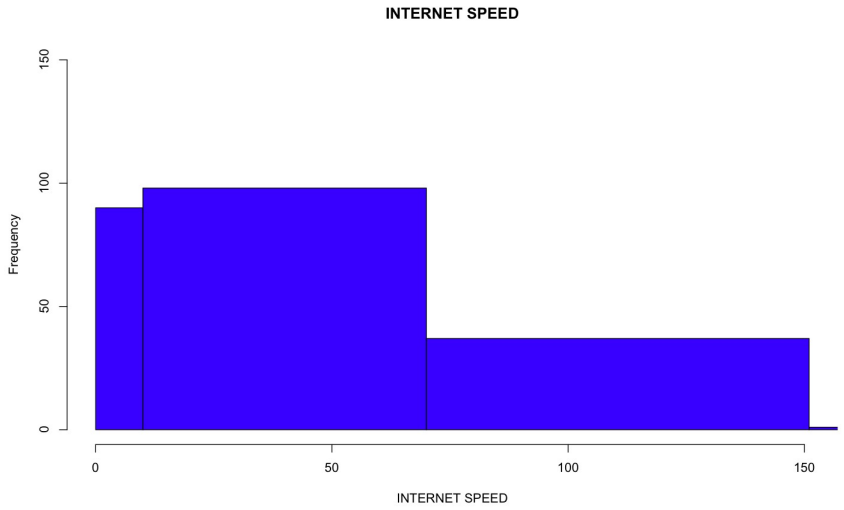
Here the ages of 16, 21, 22, 24 are outliers, so we have excluded them in our pie chart. Our data is collected from +2 students, 1st year students and 2nd year students.

In our data set, 2nd year students comprise of a small population, hence the ages of 21, 22 and 24 are outliers. From the pie chart, we see that the students of age 18 have the highest population. This shows that our data comprises mostly of students from 1st year of college.

3.3.3 Internet Speed

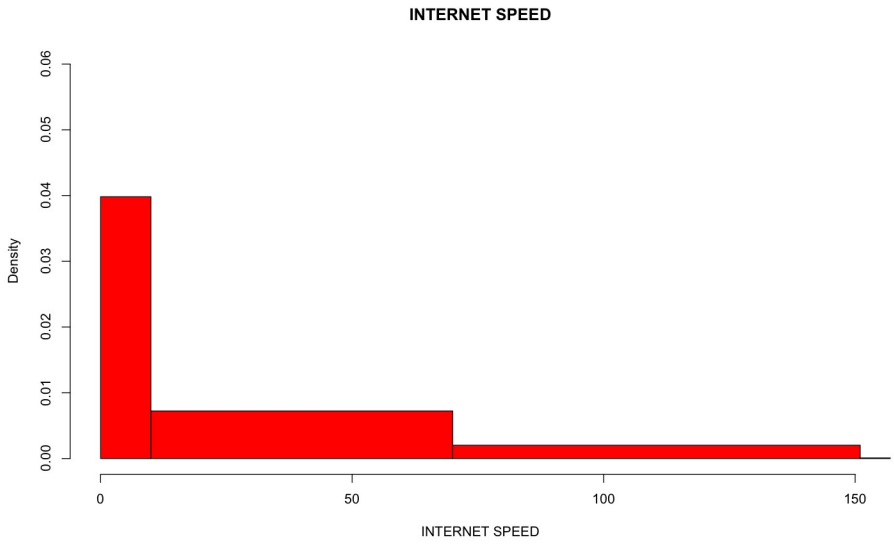
This section covers the plots and analysis of the data collected on the internet speed. The number of points was 226.

Histograms



The above frequency histogram is divided into 3 intervals, namely,: low speed (0 to 10), medium speed (10 to 70) and high speed (70 to 150), where internet speed is measured in mbps.

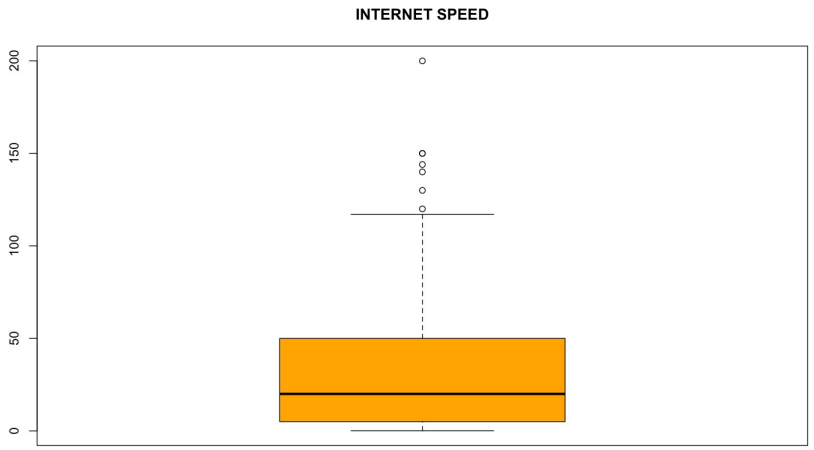
We see that majority of the data lies in the lower half (0 to 70) of the range which says that this data is highly skewed on the right side, also called positively skewed.
But this does not reveal the true picture. Look at the following Density-Histogram.



The data which looked to be packed in the first half, now looks highly concentrated in the range (0 to 10). As the widths are unequal, the frequencies are proportional to the area of the rectangles. For example, Freq(0 to 10)=$0.04 \times 10 \times 226 = 90$, i.e, (density $\times$ width $\times$ total points).

On calculating the other frequencies, we notice that still majority of the people are deprived of high speed internet, which may be due to the fact that they rely on mobile data internet.

Boxplots

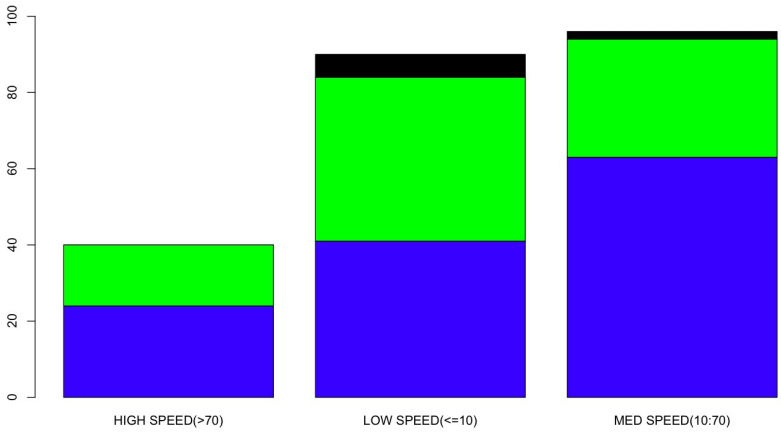


Min	Max	Q1	Median	Q3	IQR	1.5 IQR
0.1	200	5	20	50	45	67.5

Hence, the students with internet speed above $1.5IQR + Q3 = 117.5$ are outliers. Hence, the points: (120, 130,140,144,150,200) are the outliers. Note:

- IQR = 28.58 $<<$ Range = 199.9 (Min - Max).
- This suggests that kurtosis of this given data set is very high.
- Median. Being closer to the lower hinge also suggests that data is positively skewed.

Stacked Bar-Charts



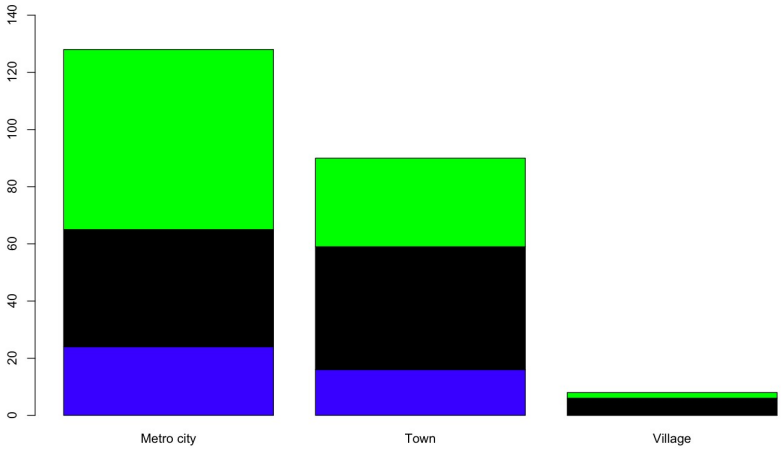
- Blue represents Metro City
- Green represents Town
- Black represents Village

We see that village comprises majority of the low speed internet and Metro city and towns constitute major portions of the high and medium speed internet.

As we go from low to medium and high speed internet, we speed that proportion of blue increases and that of green decreases. Hence speed in:

Metro City $>$ Town $>$ Villages.

But then, what is missing? Let us find out.



- Blue for High speed
- Green for Medium speed
- Black for Low speed

We see that in villages there is total scarcity of high speed internet and the frequency of high speed increases from village to town to metro city.

- But, the important thing to note here is that even in metro cities, the proportion of black dominates over blue, which says that low speed internet is still very common in the metro cities.

Hence, it is one of the big drawbacks of online classes.

3.3.4 Place of Residence

The areas where the students surveyed live are categorised as:

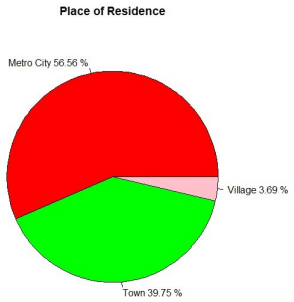
- Metro City
- Town
- Village

From the surveyed result:

- Metro City: 138
- Town: 97

- Village: 9

Pie chart



The above pie diagram provides us the proportion of students living in different areas and attending the online classes. From this, we can say that most of the students attending the online classes are from urban areas (Metro City and Town) and a lesser portion of students are from rural or village areas.

Mode

The mode of the data is the metro city i.e, most of the students attending the online classes are from Metro Cities.

3.4 Bivariate 1

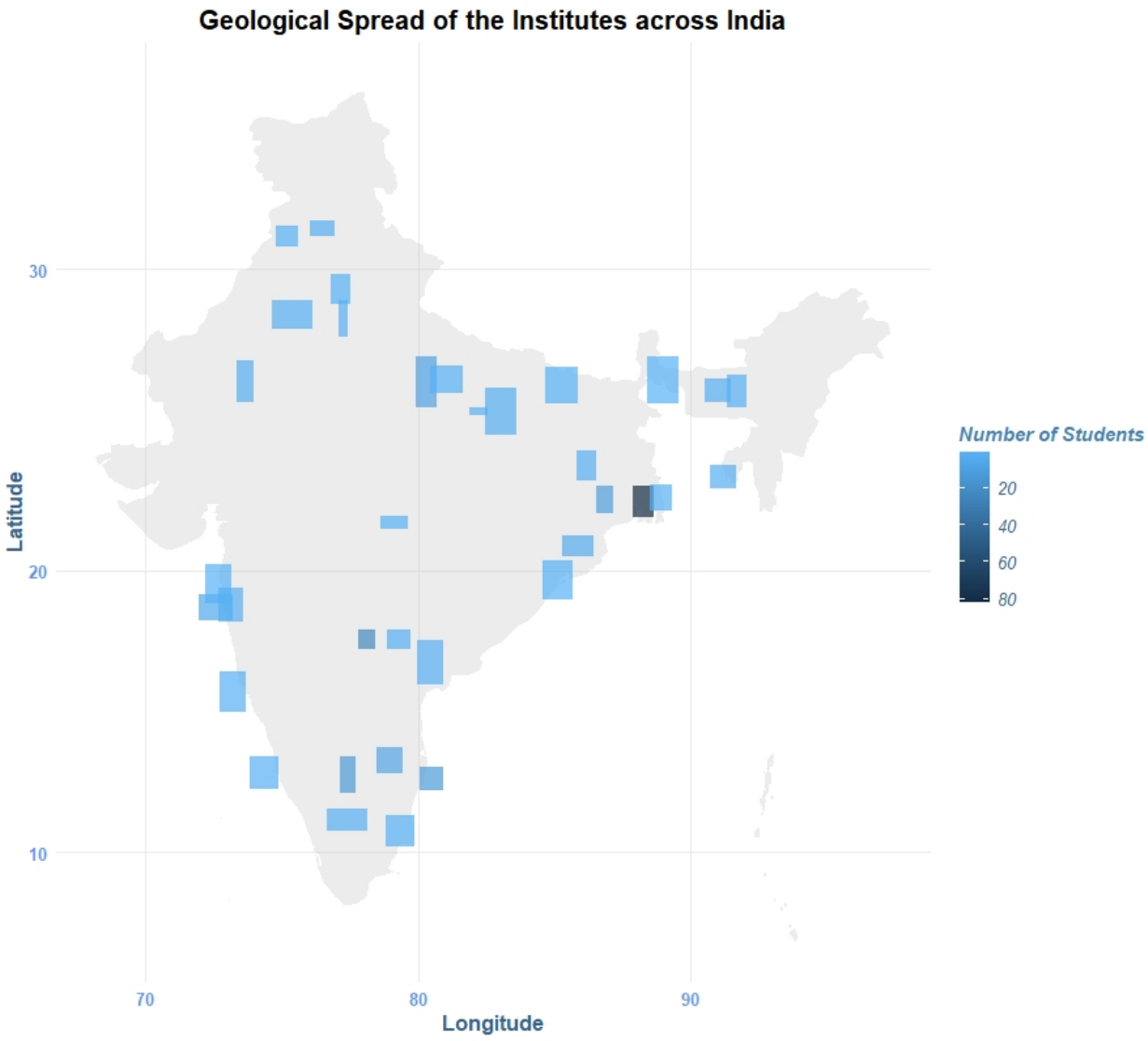
The students in this group were Arnab Modak, Sattick Das, Progyan Sarkar, Sirsha Dey, Mrinmoy Banik, Reetanjan Kali Roychowdhury, Rishi Dey Chowdhury.

This group worked on bivariate data.First let us have a look at the geographical spread of the data.

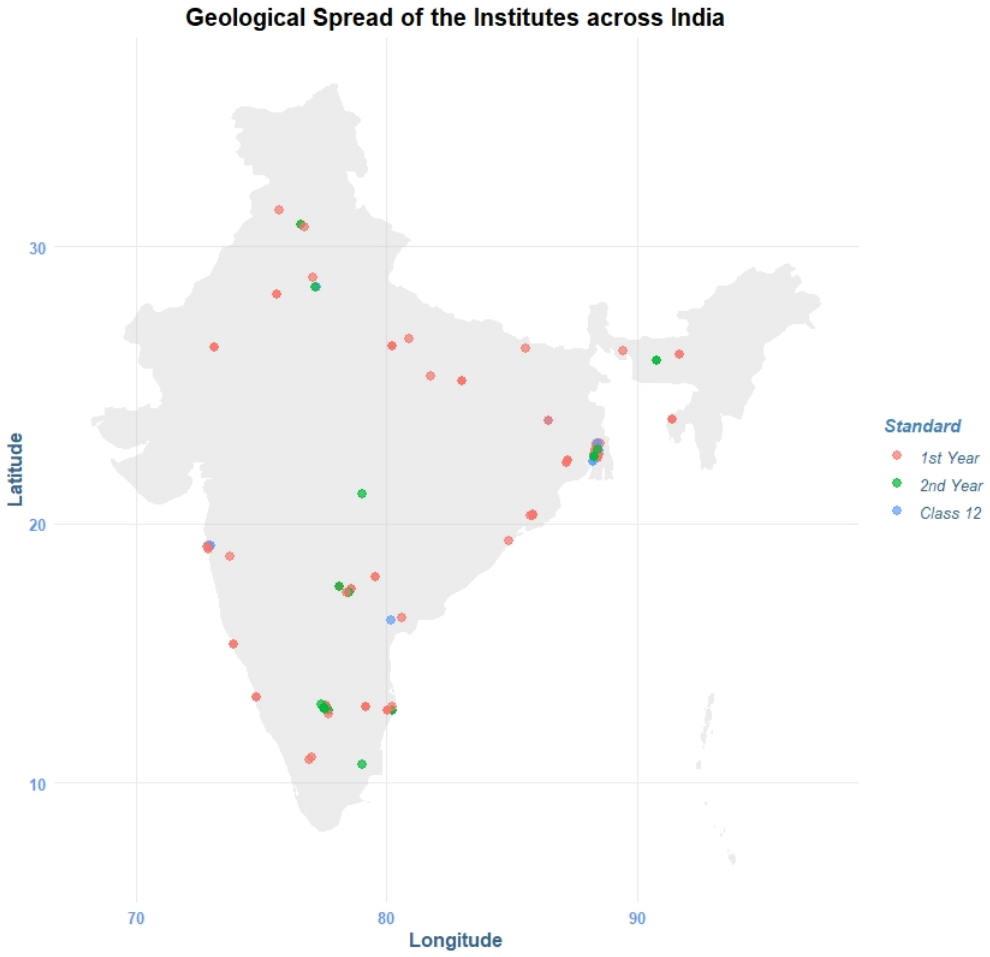
3.4.1 Geographical Spread of Sample Data

In order to analyse Data and draw suitable insights from them , we need to ensure that there is variability in the Data Collection Procedure . That is , we should not focus excessively on a particular category of the concerned Data Variable but try to diversify their types as much as possible to get a much clearer picture of the situation (or else several lurking factors may emerge to give us a Biased Perspective from the Collected Data !) .

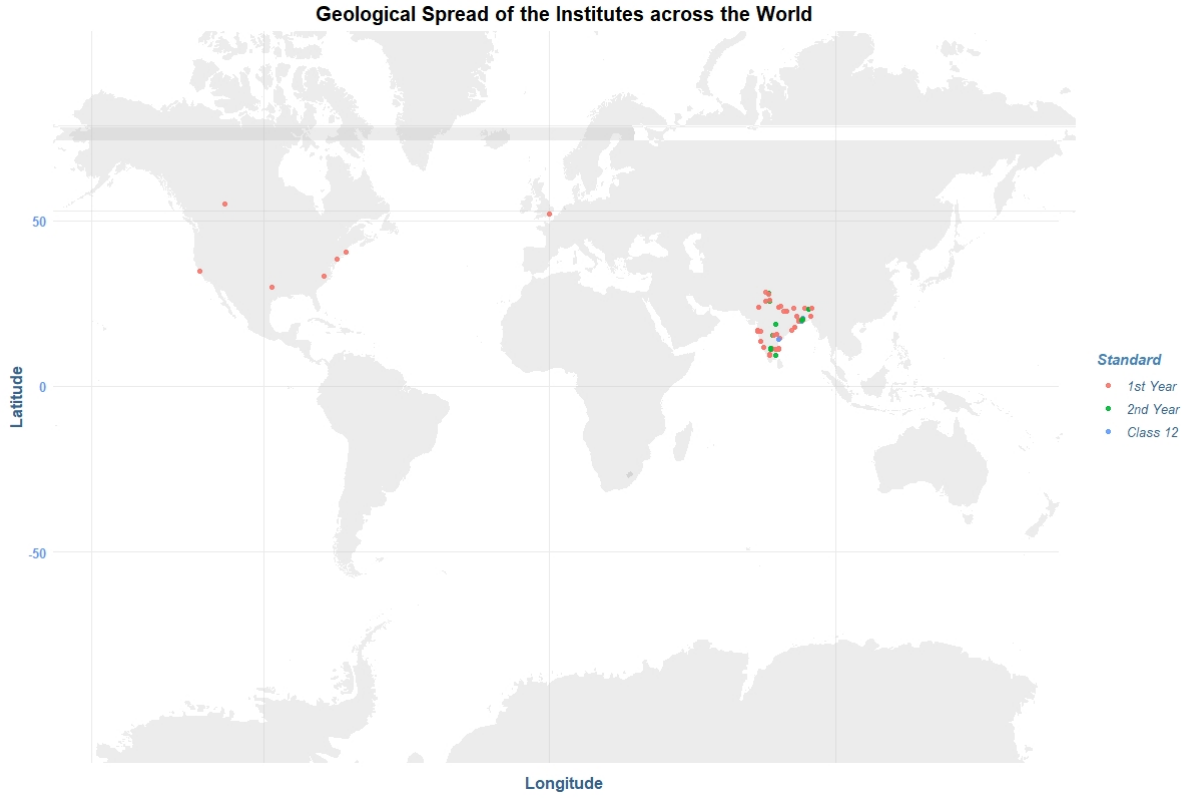
Thus , now we include some distribution plots of some of our collected data set variables depicted on a regional basis :-



The above figure shows us the **Distribution of the Institutions** (which can also be mapped to the distribution of Students) on the map of India . We can clearly see , that there is quite a variation in the distribution of students in different parts of the country . (This implies that our collected data is diverse and is not biased towards any particular Region in India .)



The above figure gives the **Distribution of Institutes** (correspondingly students according to their standard i.e. 1st year , 2nd year and Class 12) across the map of India .The density of students ,who responded to this survey, is greater in institutes belonging to Karnataka ,Tamil Nadu ,Telengana and West Bengal.Due to the difference in the population of students in different Standards , we need to further analyse them according to their Standards.



The above figure gives the Distribution of Institutes (correspondingly students according to their standards i.e. 1st year , 2nd year and Class 12) across the map of the entire World . Here the data is scattered non uniformly across the Map . (Mostly concentrated near India and few across North America) . So, whatever inference we draw from the Data is mostly concerned with India .

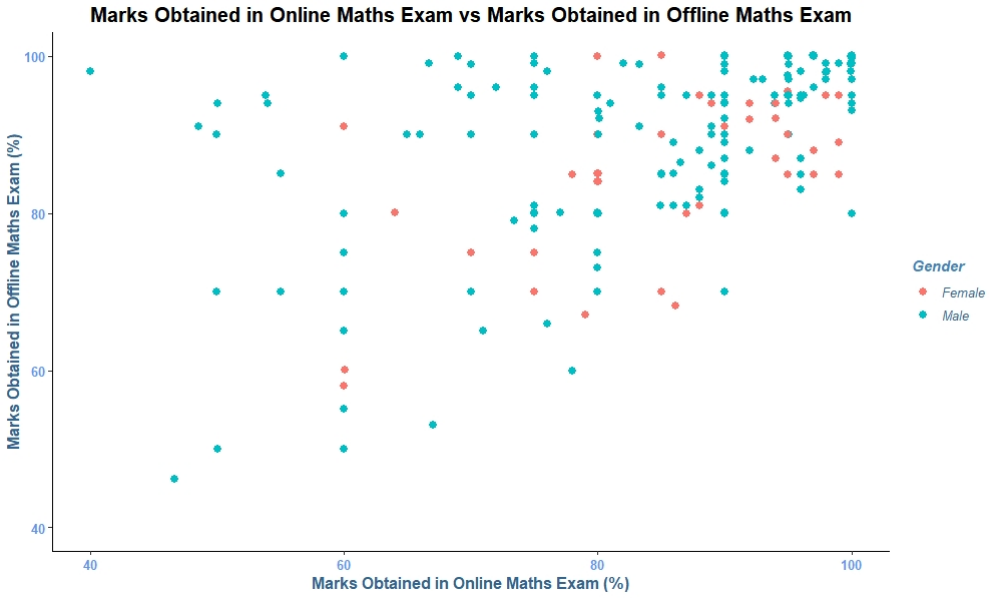
3.4.2 Marks Obtained in Online and Offline Classes

Standard	Gender	Online Mean Math Marks (%)	Offline Mean Math Marks (%)	Online Mean Language Marks (%)	Offline Mean Language Marks (%)	Marks Gain in Maths Mean Marks (%)	Marks Gain in Language Mean Marks (%)
1st Year	Female	79.44828	83.87931	78.51724	88.03448	↓ -4.43103448275862	↓ -9.51724137931033
1st Year	Male	82.05978	88.04058	77.77920	83.36051	↓ -5.98079710144927	↓ -5.58130434782609
2nd Year	Female	85.24385	84.24615	67.19231	78.60000	↑ 0.997692307692304	↓ -11.4076923076923
2nd Year	Male	82.55833	80.72917	45.70833	52.66667	↑ 1.82916666666667	↓ -6.95833333333333
Class 12	Female	81.71429	78.85714	84.28571	80.42857	↑ 2.85714285714285	↑ 3.85714285714286
Class 12	Male	82.03333	86.37333	76.40000	82.01333	↓ -4.34	↓ -5.61333333333333

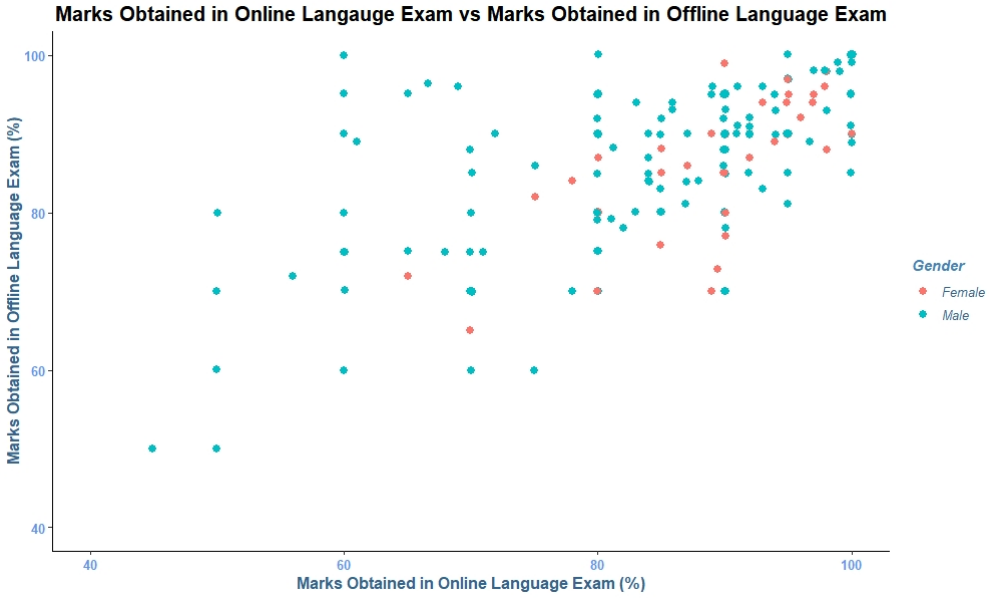
We try to find the variation in the data set to see and find any trends if they exist. As can be seen in the table above ,which briefly jot down the means of marks scored. We find a few indicators:-

- There is a strong decrease in average marks of the online phase compared to offline for language.(5-11% decrease compared to 3% increase)

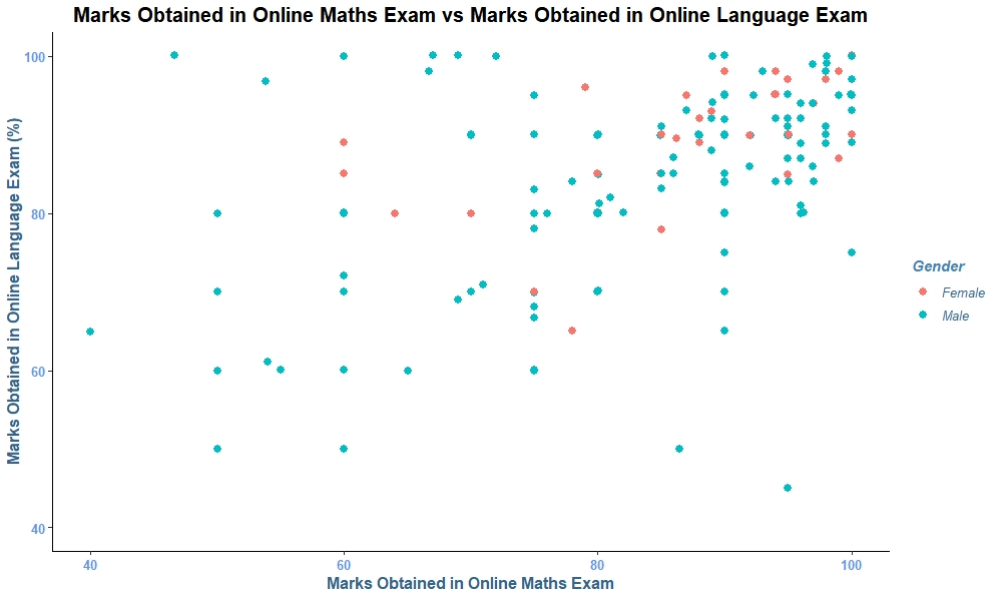
- While the data for Math shows an increase in marks for certain categories of students, the increment is quite small compared to the decrements in the other categories(0-2% increase compared to 4-5% decrease).
- The average marks scored by females in language is significantly better than their male counterparts for most categories of students.
- The average marks scored by males in Math is slightly higher than their female counterparts for most categories of students.



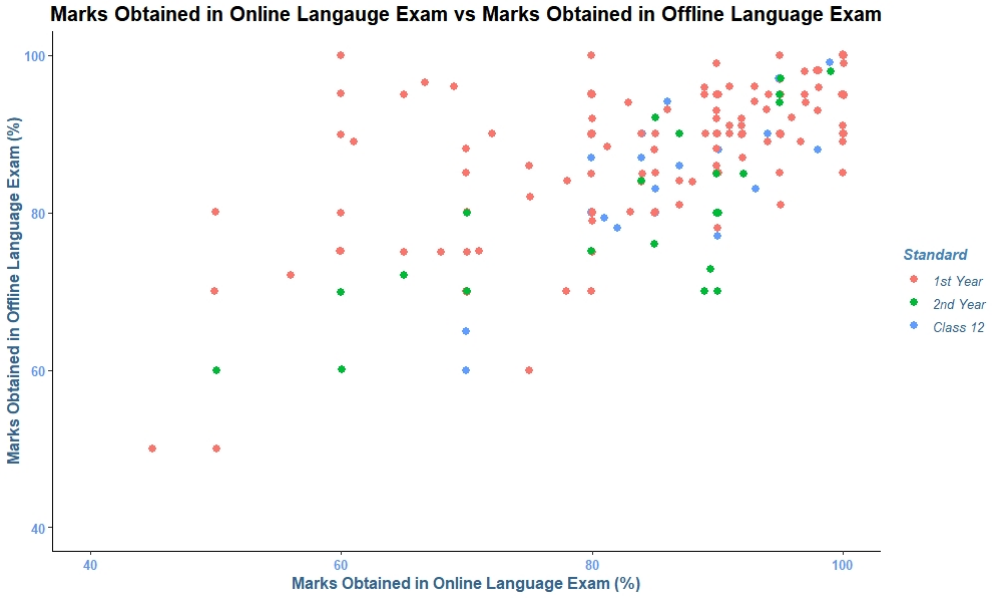
The above plot comprises all data points of the form (x,y), where 'x' represents the online Math marks and 'y' represents the offline Math marks obtained for an individual.(The female and male students are represented by orange and blue dots respectively).



The above plot comprises all data points of the form (x,y), where 'x' represents the online language marks and 'y' represents the offline language marks obtained for an individual.(The female and male students are represented by orange and blue dots respectively).



The above plot comprises all data points of the form (x,y), where 'x' represents the online Math marks and 'y' represents the offline Math marks obtained for an individual .(The 1st year ,2nd year and class 12 students are represented by orange ,green and blue dots respectively).



The above plot comprises all data points of the form (x,y), where 'x' represents the online language marks and 'y' represents the offline language marks obtained for an individual .(The 1st year ,2nd year and class 12 students are represented by orange ,green and blue dots respectively).

3.4.3 Marks obtained and time spent for Online classes:-

Standard	Online Mean Math Marks (%)	Avg. Time Spent in Online Maths Classes (Hours)	Online Mean Language Marks (%)	Avg. Time Spent in Online Langaugage Classes (Hours)
1st Year	81.60629	4.699102	77.90737	2.560878
2nd Year	83.50189	7.171171	53.25676	1.943694
Class 12	81.93182	3.814394	78.90909	2.352273

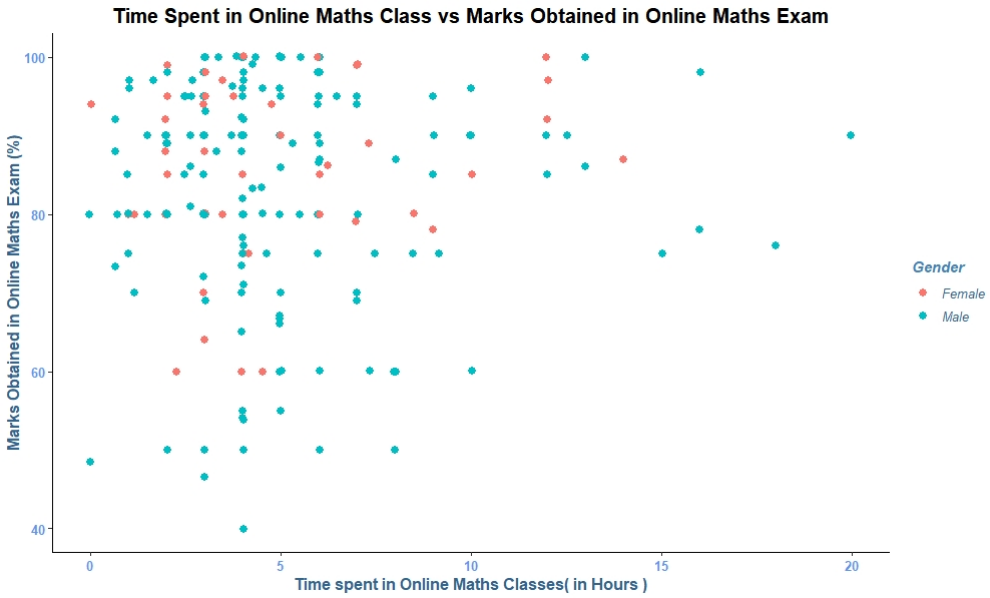
The above table gives us the Average marks and Time spent on both subjects (Math and Language) individually for the 3 sets of students (Class 12,1st Year,2nd Year) .

Now from the table , we can observe the following :-

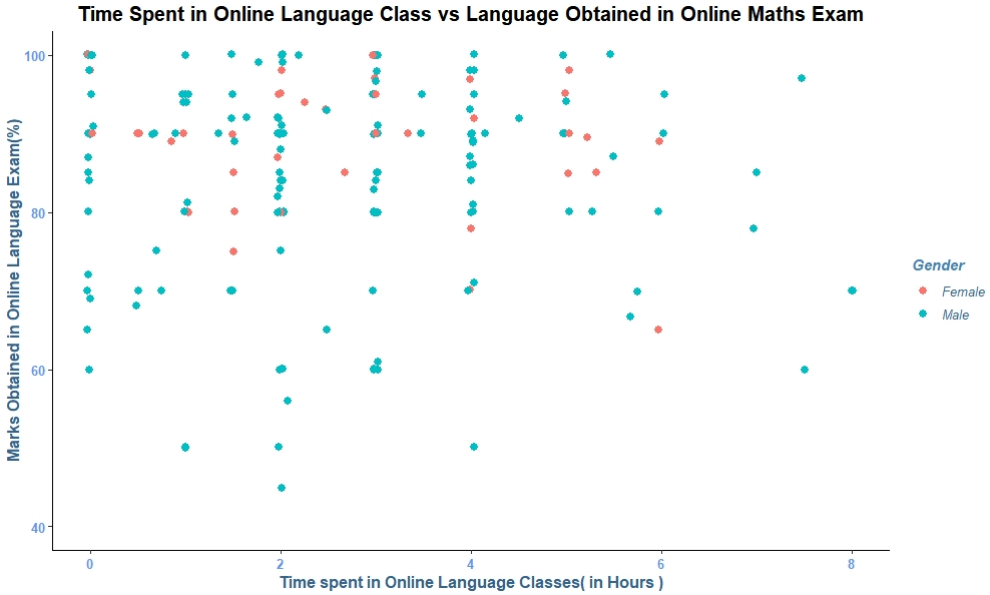
- For both Math and Language , higher time devotion has resulted in a somewhat higher marks average among the three groups .(this fact is also true for individual persons as marks obtained and time spent on that subject have a positive correlation which is evident from the scatter plots .)
- The time spent on Math for all three groups is considerably higher than that spent on Language . So this may be a factor by virtue of which average Math Scores are higher than the

corresponding Language Scores.

The following plots further help in understanding the situation :-



The above plot comprises all data points of the form (x,y),where 'x' represents the time spent on online language classes and 'y' represents the online language marks obtained for an individual .(The 1st year ,2nd year and class 12 students are represented by orange ,green and blue dots respectively).



The above plot comprises all data points of the form (x,y) where 'x' represents the time spent on online Math classes and 'y' represents the online math marks obtained for an individual .(The 1st

year , 2nd year and class 12 students are represented by orange, green and blue dots respectively)
.

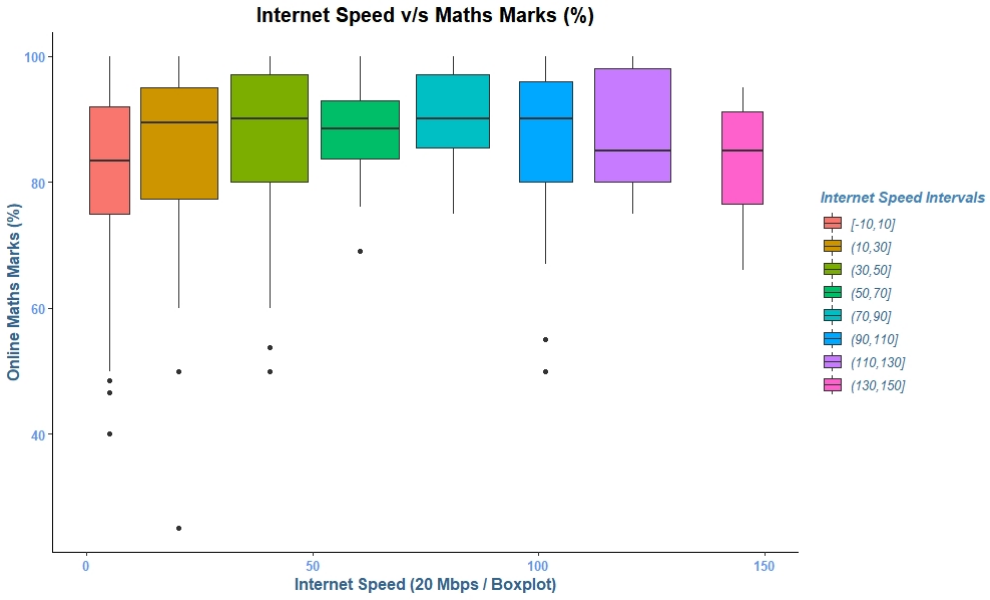
3.4.4 Average Marks obtained along Data Speed and Place of Residence :

Place of Stay	Average Internet Speed (Mb/s)	Online Mean Math Marks (%)	Offline Mean Math Marks (%)	Online Mean Language Marks (%)	Offline Mean Language Marks (%)	Marks Gain in Mean Maths Marks (%)	Marks Gain in Mean Language Marks (%)
Metro city	37.61062	82.75148	86.11484	74.24766	80.47852	-3.363359375	-6.23085937499999
Town	29.73722	81.69033	86.11889	74.59256	80.75000	-4.42855555555556	-6.15744444444445
Village	12.29750	72.00000	86.12500	62.50000	70.87500	-14.125	-8.375

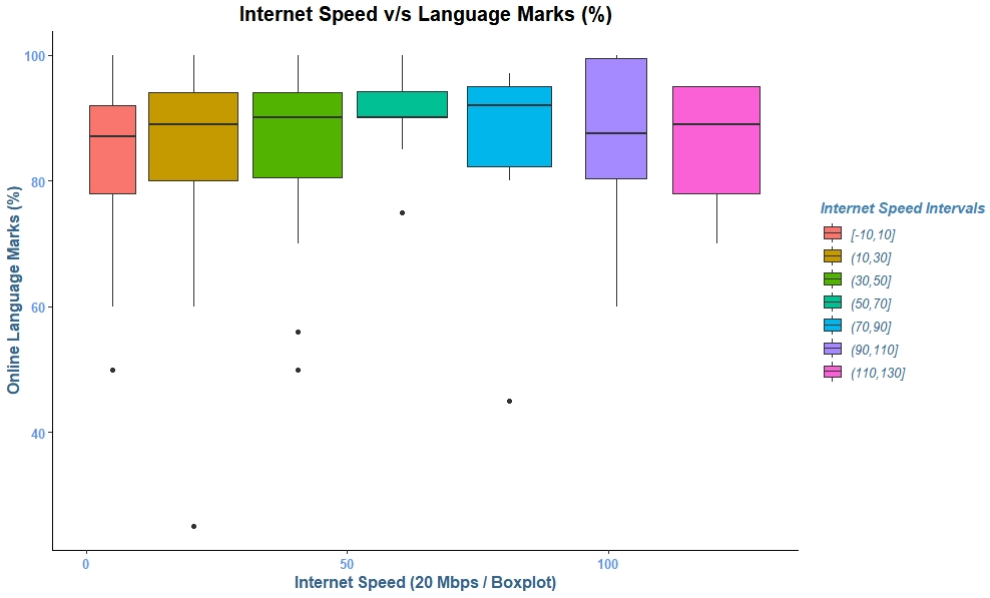
The above table gives us the Average Marks in either subject (Math and Language) for both Offline and Online modes and Average Net Speed for 3 categories of students (Belonging to Metro City ,Town or Village).
From the table , we can observe the following :-

- For Math,the average offline marks are close for all the 3 categories of students.But in the average online marks for students from the Villages are far less than that of their counterparts from Towns or Metro Cities. This may be attributed to the fact that Village Students are far less accustomed with the Online Setup as compared to their counterparts!
- The average marks in the online mode is less than the offline mode average for both subjects in all the categories .Thus we can infer that students of all the categories have performed poorly in the online mode than the offline one .
- The Average Net Speed decreases progressively from Metro City to Town to Village areas .Thus it is clear that the Metro City students get access to better quality of Online Classes than students belonging to either Towns or Villages
- Students belonging to regions with Higher Average Net Speed are tending to have Higher Average Online Marks.Thus the comparatively lower average score for students belonging to “Village” category may be attributed to the Lower Average Net Speed in the region .

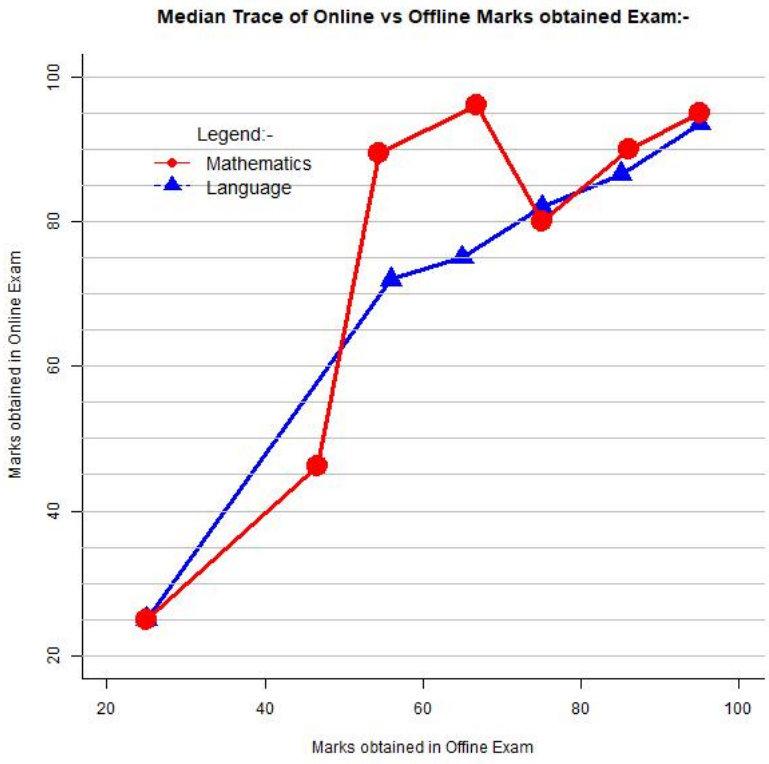
The following plots further help in understanding the situation :-



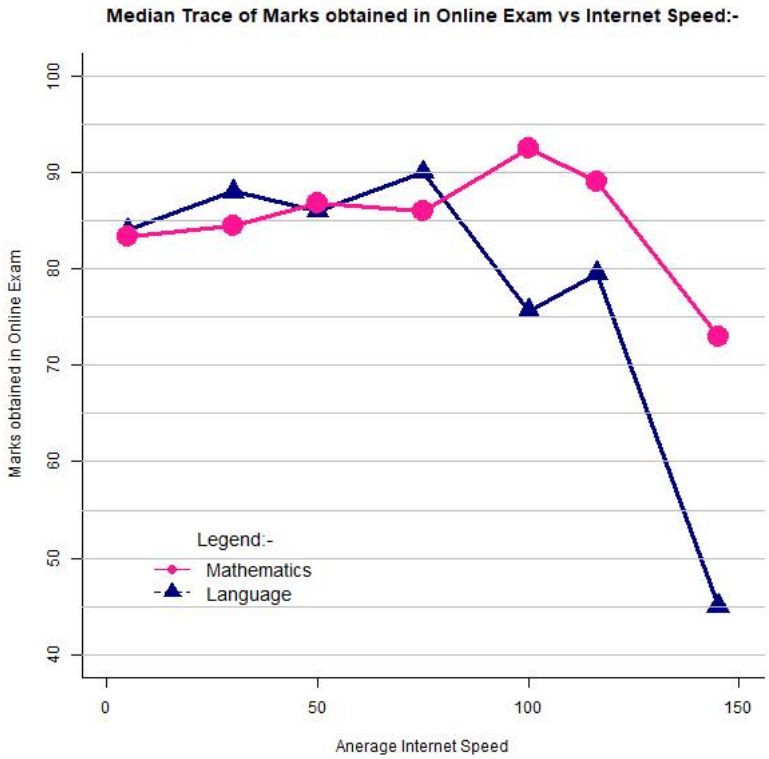
The above plot gives the distribution of Online Marks in Maths versus Internet Speed for an individual student.



The above plot gives the distribution of Online Marks in Language versus Internet Speed for an individual student. **From the following two box-plots we can see that there is a increase in the median marks of the individuals to some extent with the internet speed but then we see a little drop in the marks with further increase. So it gives us an idea that better internet speed helps in the performance of the students maybe for better and smoother experience in the classes and Tests**

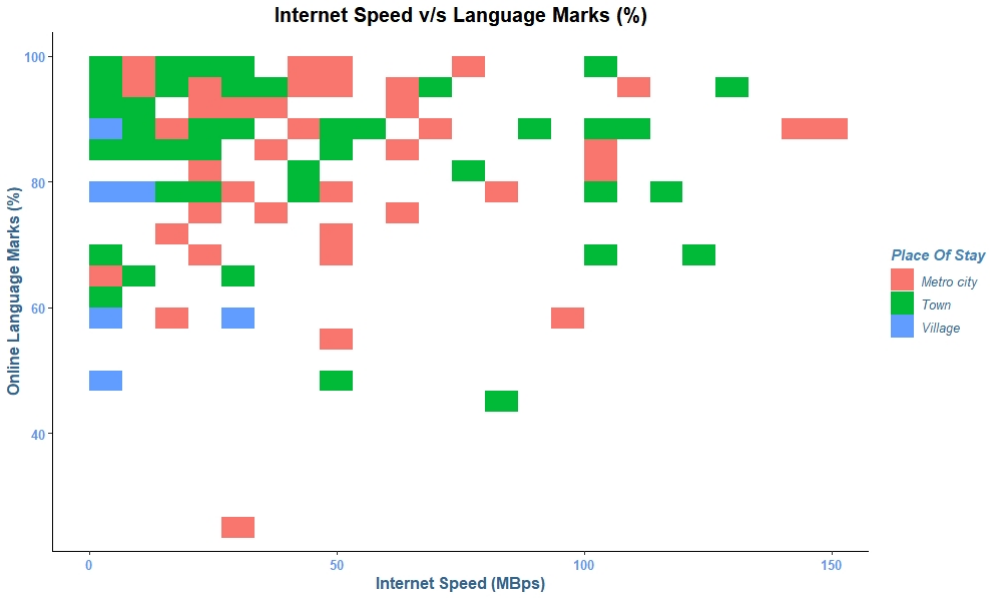


From the above median trace its clear that there is less parity between online and offline marks of maths than online and offline marks of Language.

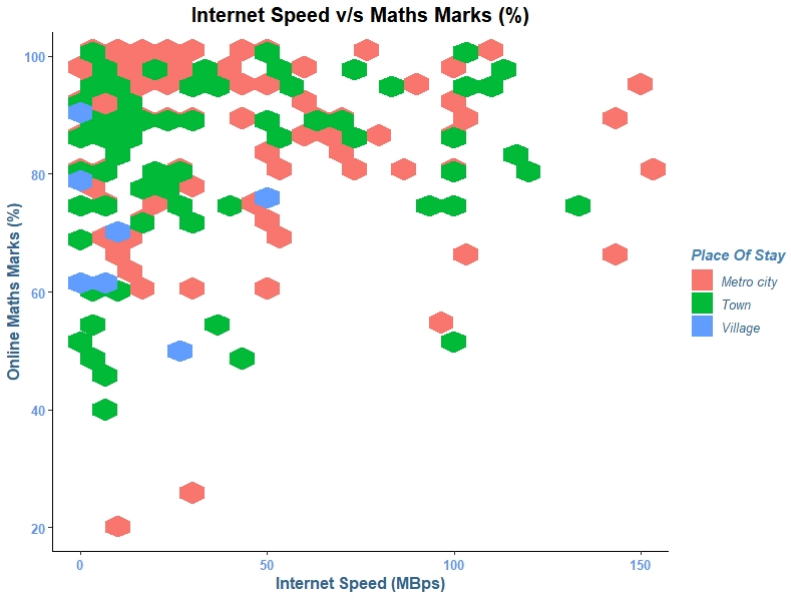


It seems the students with the fastest internet are the ones with the worst online exam marks.This might indicate the tendency of students to get distracted as they are unable to utilise their re-

sources fully. It also indicates that their Language marks gets far more affected than their Maths score.



The above plot gives the distribution of Online Marks in Language versus Average Internet Speed for an individual student.(The “Metro City”, “Town” and “Village” category students are respectively denoted by orange, green and blue colours.)



The above plot gives the distribution of Online Marks in Math versus Average Internet Speed for an individual student .(The “Metro City”, “Town” and “Village” category students are respectively denoted by pink,green and blue colours .)

From the above 2 scatter plots, it is quite clear that the “Village” category students have

the worst average internet speed which has in turn affected their online exam scores for either subject .

3.4.5 Marks obtained in online exams and devices used

Device Used	Average Internet Speed (Mb/s)	Online Mean Math Marks (%)	Offline Mean Math Marks (%)	Online Mean Language Marks (%)	Offline Mean Language Marks (%)	Marks Gain in Mean Maths Marks (%)	Marks Gain in Mean Language Marks (%)
Laptop/Desktop	36.91365	86.39662	88.31892	75.50500	81.23716	↓ -1.92229729729731	↓ -5.73216216216217
Mobile	17.75574	76.30064	85.08511	73.99277	79.38298	↓ -8.78446808510638	↓ -5.39021276595744
Mobile, Laptop/Desktop	34.01106	81.82753	85.31529	74.49412	80.27294	↓ -3.48776470588236	↓ -5.77882352941175
Mobile, Tablet	35.37500	62.12500	64.75000	41.25000	62.75000	↓ -2.625	↓ -21.5
Mobile, Tablet, Laptop/Desktop	50.22083	79.33333	85.25000	66.25000	79.83333	↓ -5.91666666666667	↓ -13.5833333333333
Tablet	26.27000	90.00000	99.00000	75.00000	86.00000	↓ -9	↓ -11
Tablet, Laptop/Desktop	120.46667	98.33333	98.33333	95.00000	91.66667	↓ 0	↑ 3.33333333333333

We try to understand if the nature of the device used influences the outcome of an online exam and if any such trend exists. As can be seen in the table above,we jot down the average marks scored by the users of that device.
A few trends can be found as follows:-

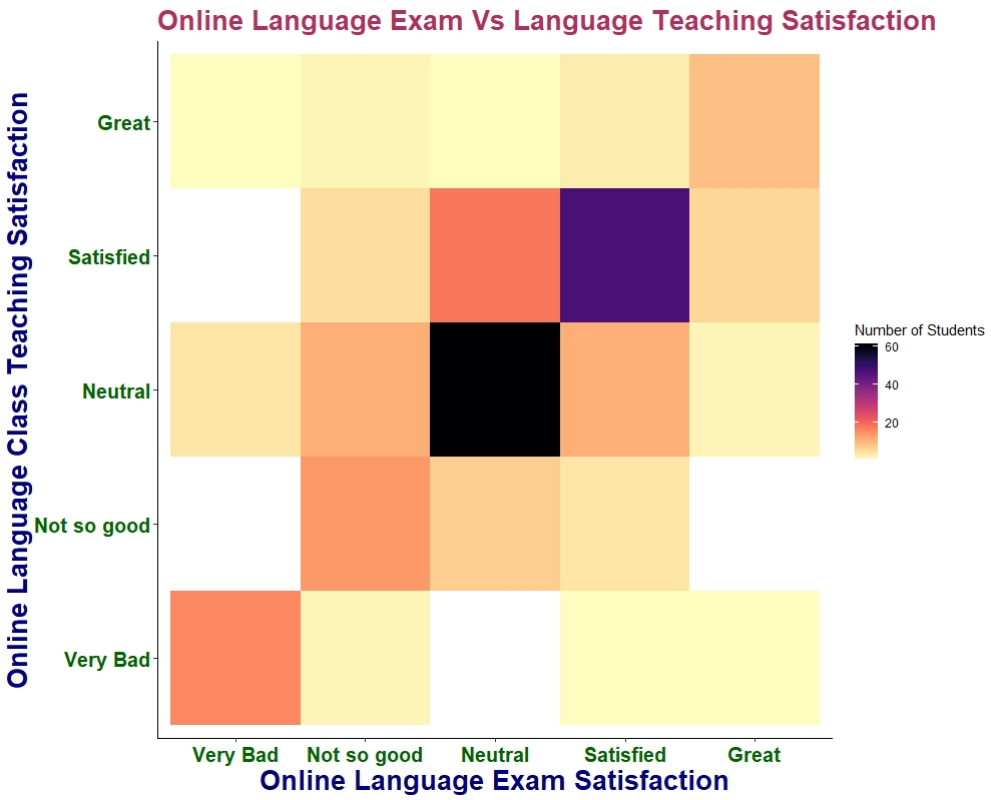
- There is a very serious fall in marks for those using mobile phones and tablets for both online and offline examinations pointing to a skewed grouping.
- There is a dip in online mathematics marks compared to offline for all device users with average marks of mobile-phone-only users suffering the most with a 8.7% decrease.
- A similar trend is seen with online language exam as well with the only exception of the tablet and PC group who were able to secure a 3.3% increase.

Given the nature of the trends,there is a possibility that Students have performed poorly in the online mode of examination irrespective .However,the trends also suggest the presence of a lurking variable due to certain categories of students showing great variation compared to the general trend.

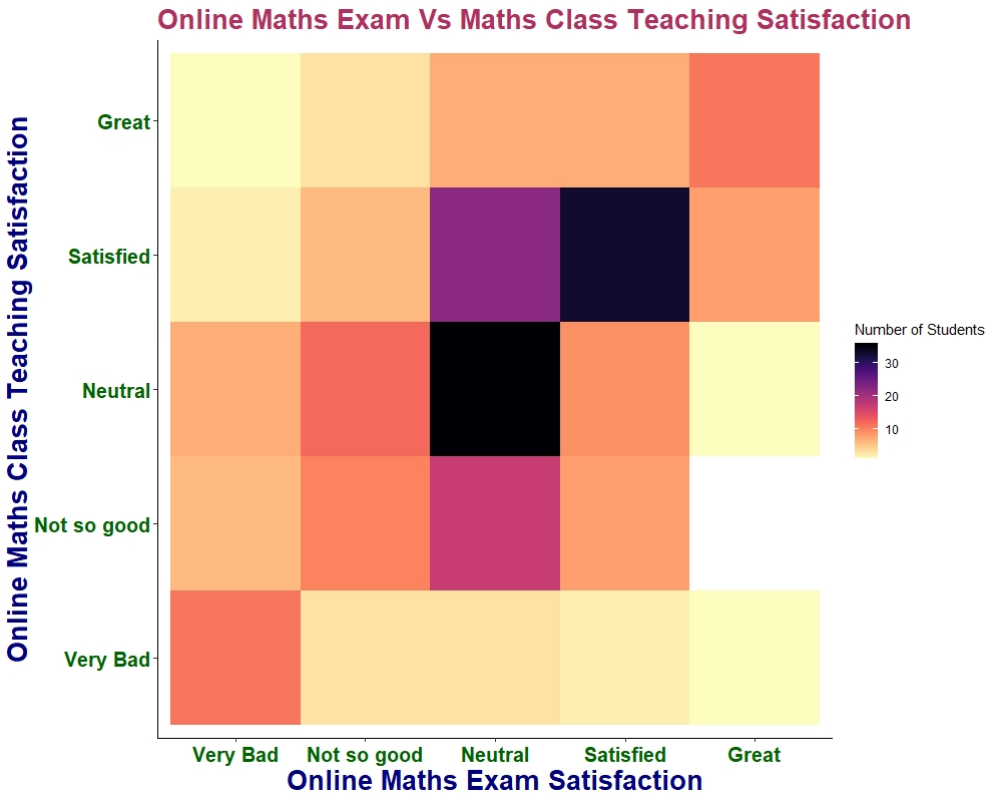
3.4.6 Likert Scale Analysis

A Likert scale is a psychometric scale commonly involved in research that employs questionnaires. It is the most widely used approach to scaling responses in survey research, such that the term is often used interchangeably with rating scale, although there are other types of rating scales.

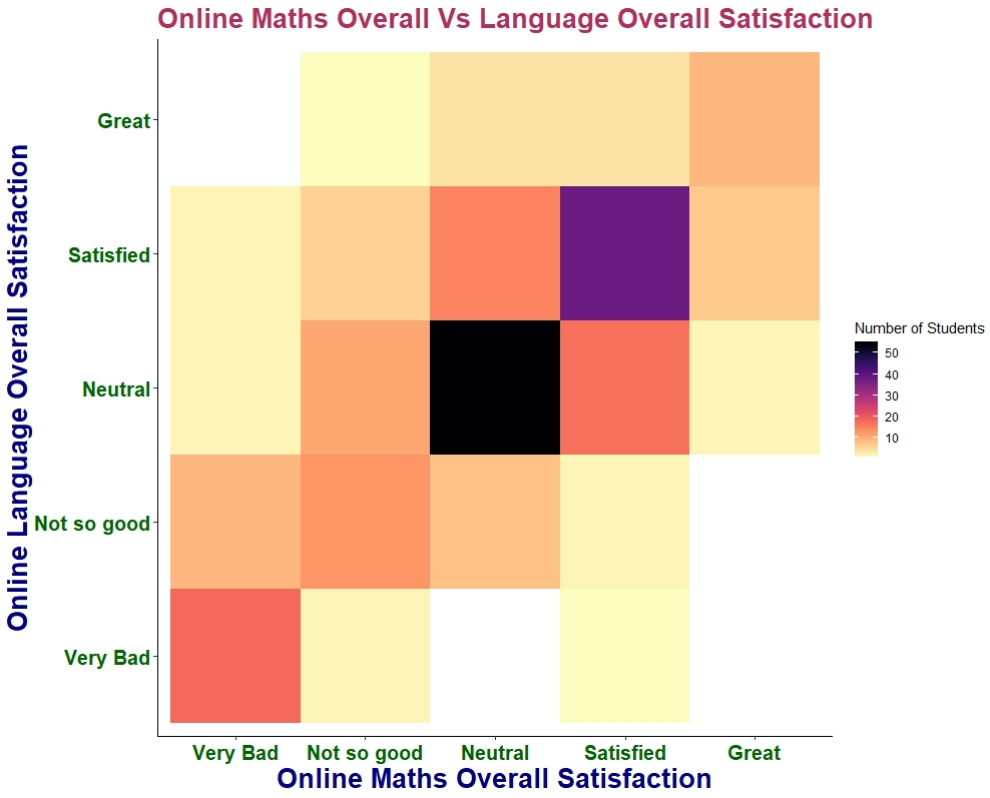
Satisfaction Level of Students for some aspects of the Online Setup:-



The above plot gives a distribution of the Satisfaction Levels of students with the Online Language Teaching (on vertical axis) and Exam conducted (on horizontal axis). Analysing the plot , we can see that most of the students have a Neutral satisfaction level for both Teaching and Exam and very few are critical for either of them .



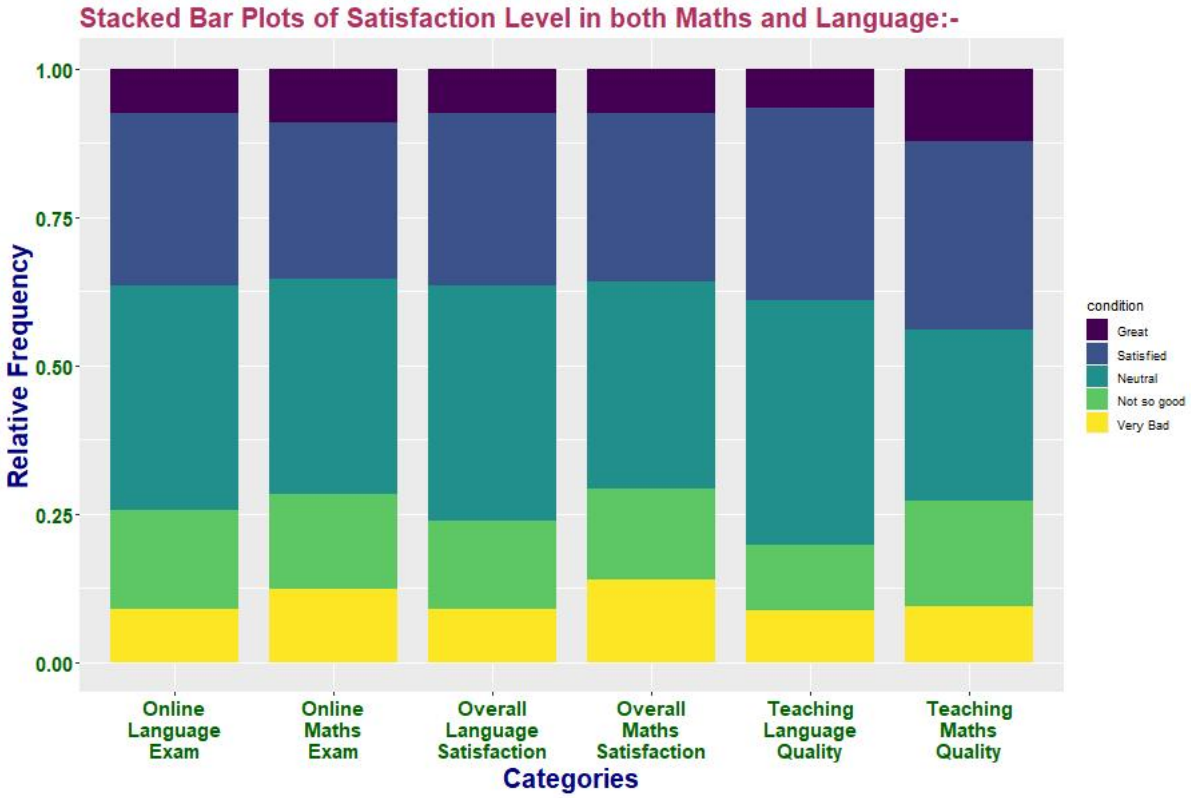
The above plot gives a distribution of the Satisfaction Levels of students with the Online Math Teaching (on vertical axis) and Exam conducted (on horizontal axis). Analysing the plot , we can again see that most of the students have a Neutral satisfaction level for both Teaching and Exam and very few are critical for either of them .



The above plot gives the distribution of students corresponding to their overall Satisfaction Levels(i.e. for both teaching and exams) for Both Online Language and Online Math classes . From the distribution it is quite clear that most of the students have Neutral Satisfaction Levels regarding either of the two subjects . The students who are critical of the online classes (as evident from their Satisfaction Levels) may be mostly the ones with lower scores in the exams which can either be attributed with the teaching quality or their unfamiliarity with the overall Online Setup !

Likert scales are categorical and ordinal variables which only give us a qualitative idea of the data. So first we have a look at some stacked bar plots of various satisfaction levels below.

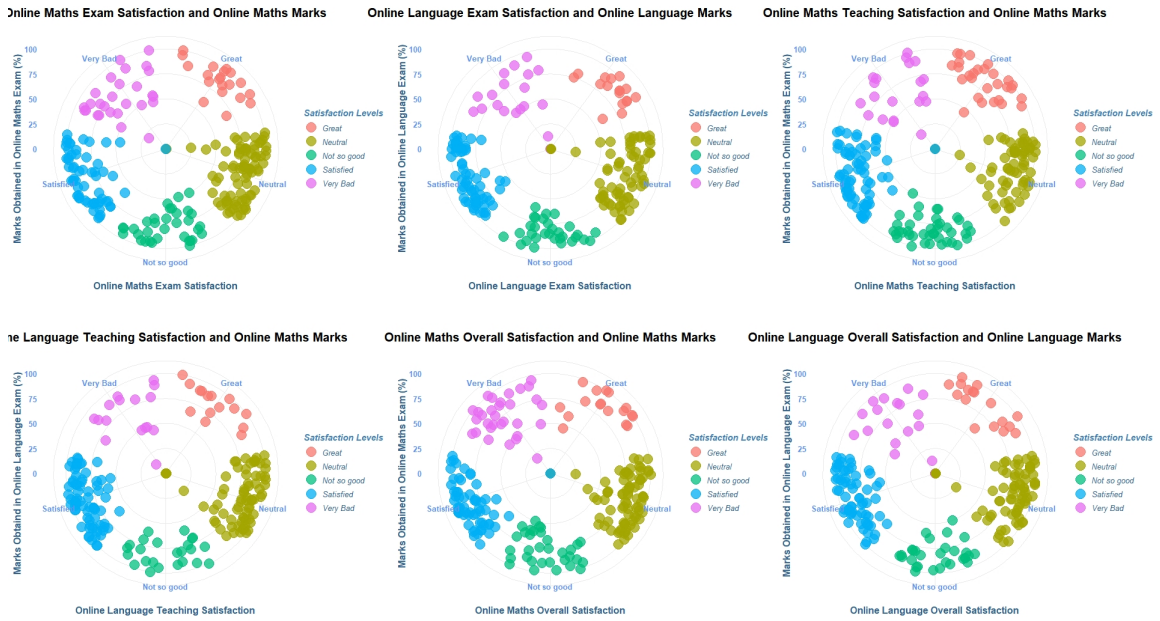
Distribution of Satisfaction levels in different likert scales:-



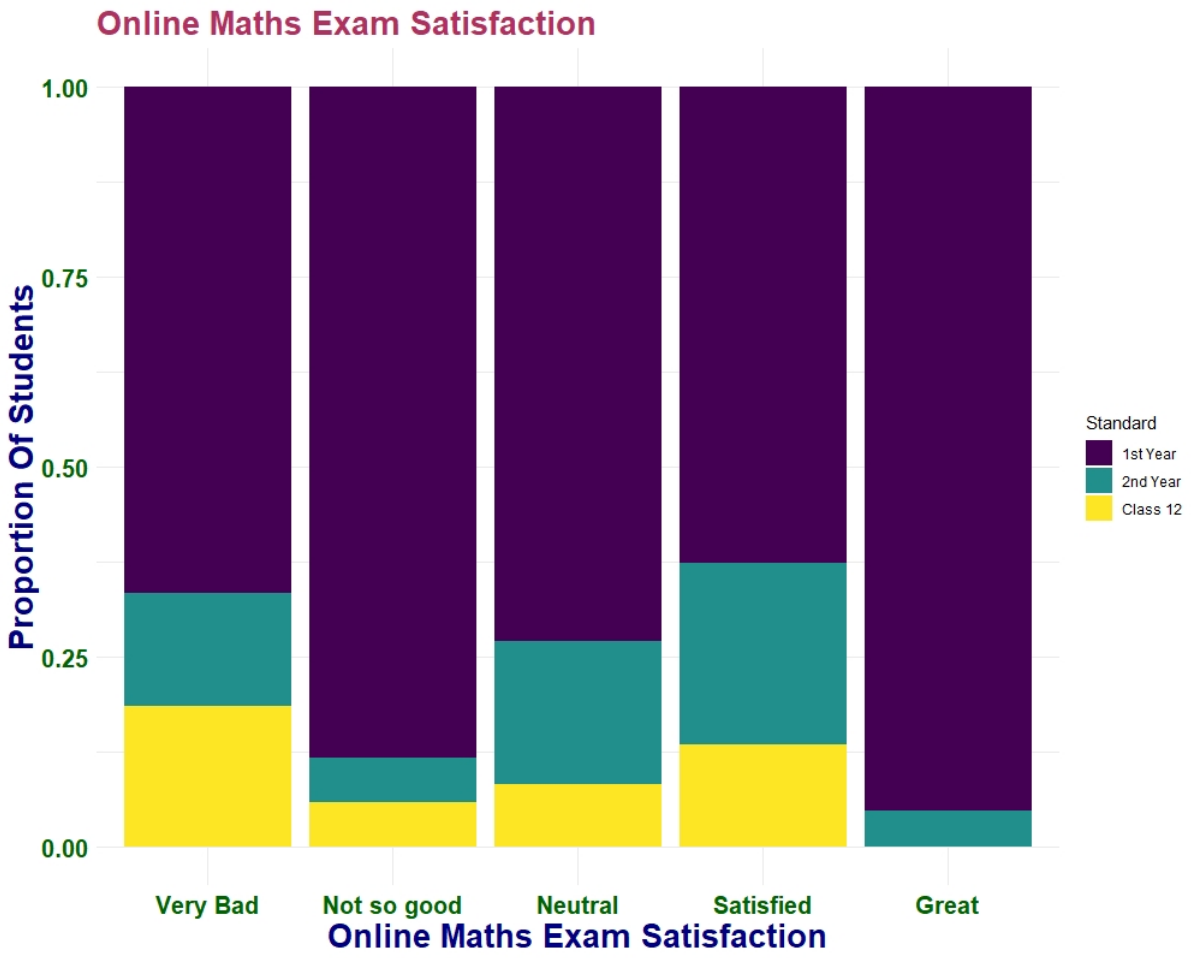
We can see that a little over 10% of the respondents are finding the online setup(online exam, teaching and overall) "very bad", the majority of them, nearly 30% each are in "Satisfied" and "Neutral" categories, which implies that as a whole students aren't having that much difficulty in the online setup, and again just like the "very bad" one we have 10% people in the "not so good" category and merely 5% who are actually finding it "Great".

Perception\Marks	M.Sat.Exam	M.Sat.Teach	M.Sat.Total	L.Sat.Exam	L.Sat.Teach	L.Sat.Total	Total
Great	21	29	18	18	16	18	120
Satisfied	59	71	62	66	76	68	402
Neutral	85	65	82	87	89	87	495
Not so good	34	41	33	34	25	32	199
Very Bad	27	20	31	21	20	21	140
Total	226	226	226	226	226	226	1356

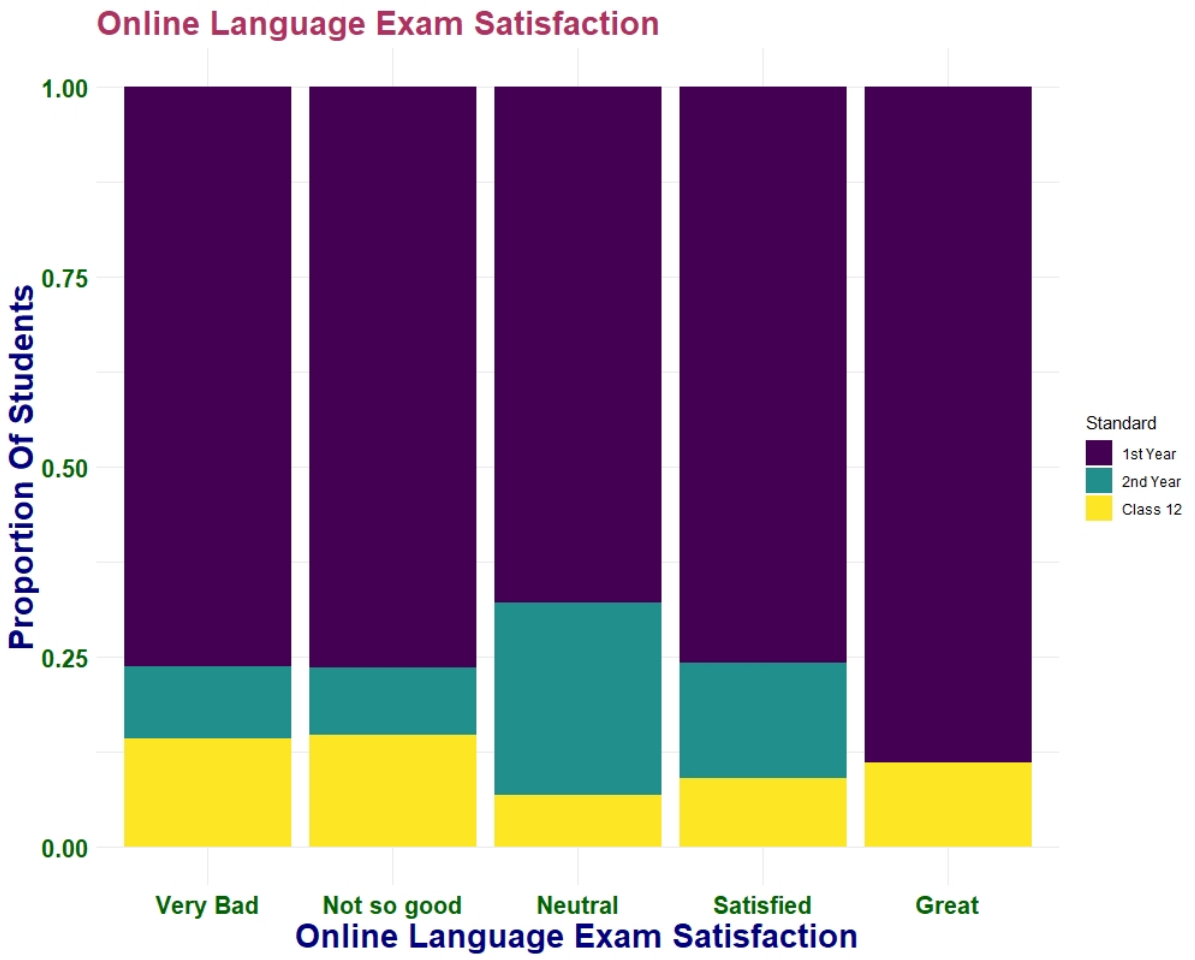
Relationship between marks and satisfaction levels:-



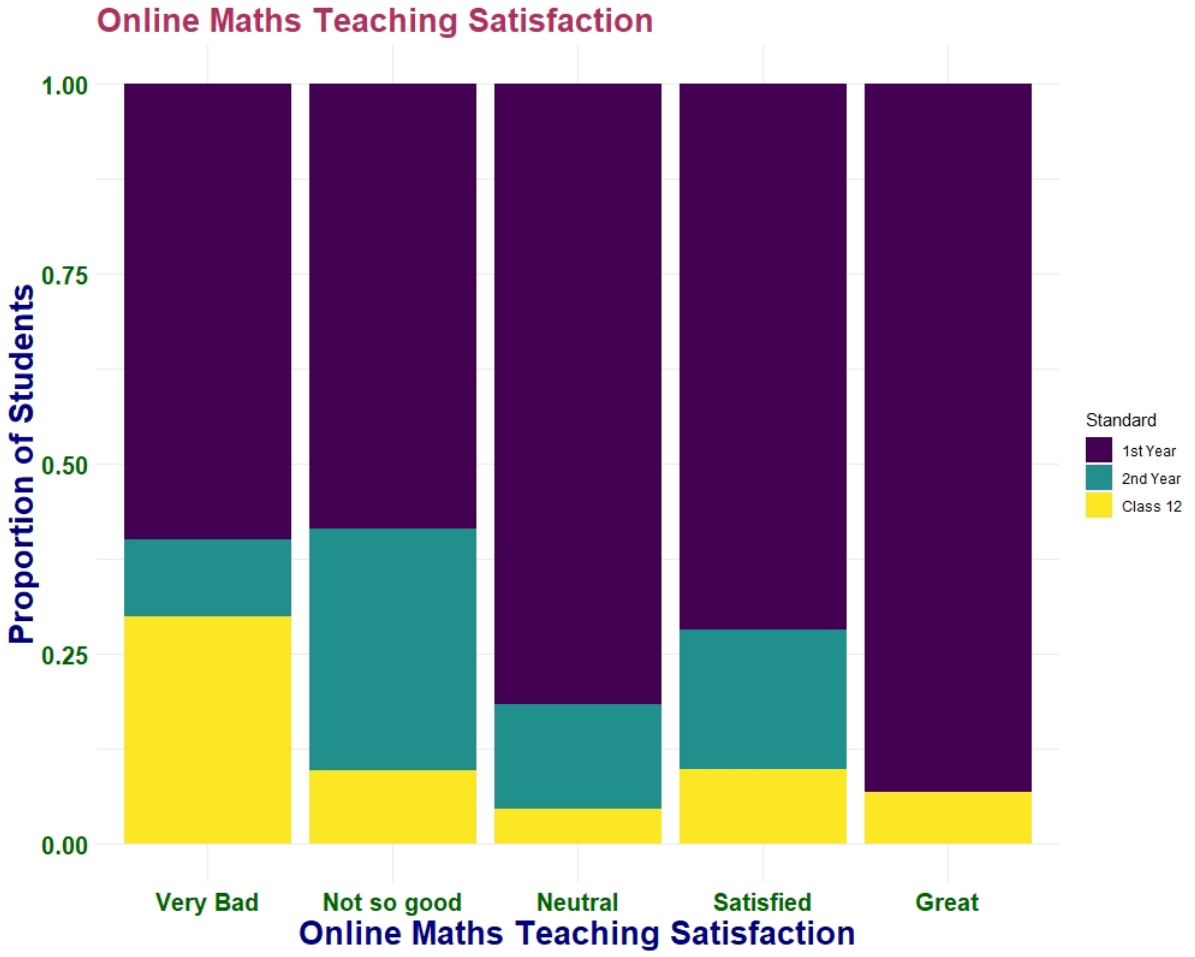
We can see that most of the students are either satisfied or have neutral feelings regarding the online class setup. Here the radial distance of a point from the centre indicates the marks scored by that particular student, One key observation is that even though the people scored really high(indicated by the points being near the periphery of the circle) they do no necessarily have "great" feelings towards those setup, many have chosen to even stay "neutral". In all of the plots the 3rd most chosen option is "Not so good" followed by a tie between "Great" and "Very Bad". The students who showed such feelings clearly aren't having a good time with the online setup, and their marks are more towards the mid-range or lower.



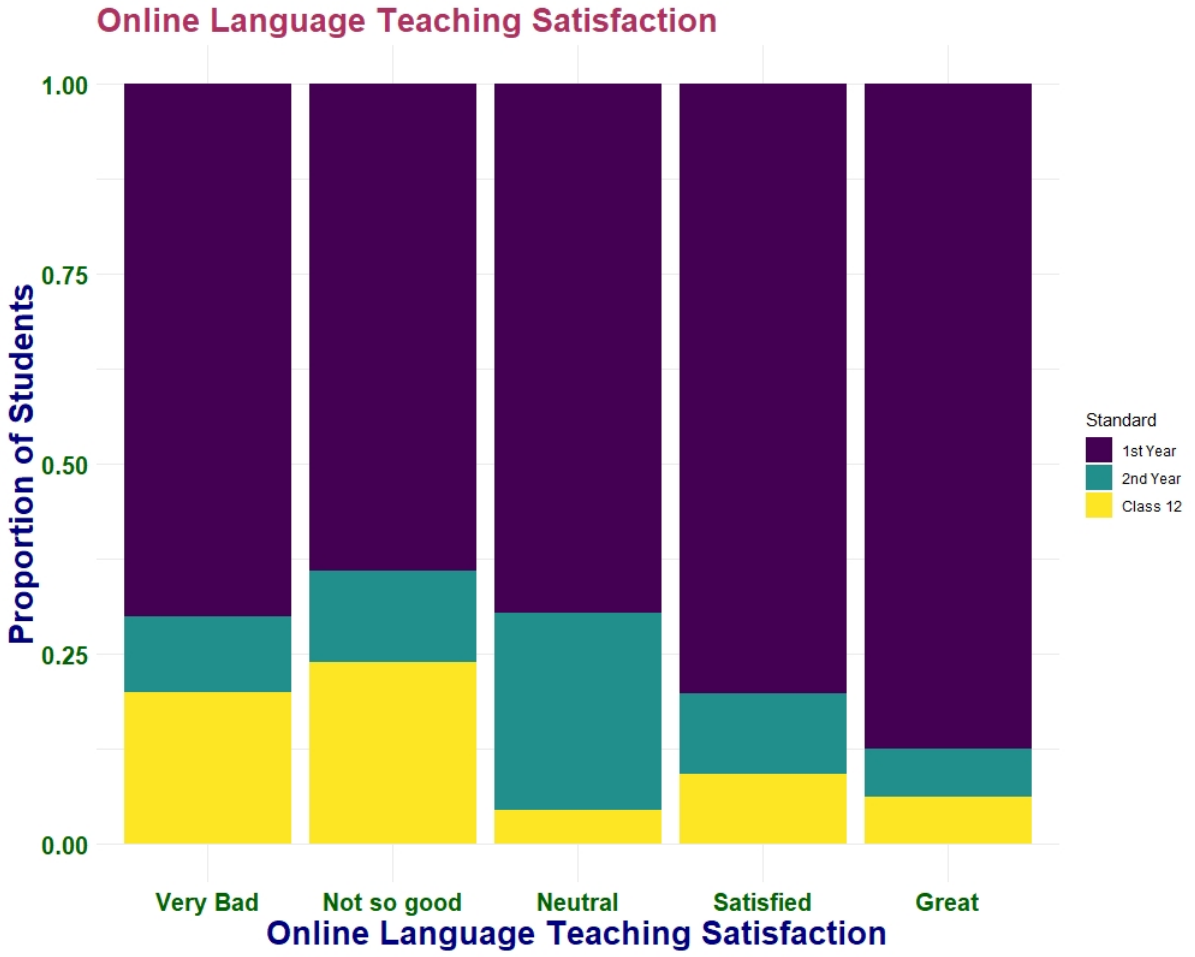
From the above Stacked Bar plot it seems students of class 12 dislike online classes the most and as these plots are not normalised it shows no. of 1st year students constitute major part of the our respondents.



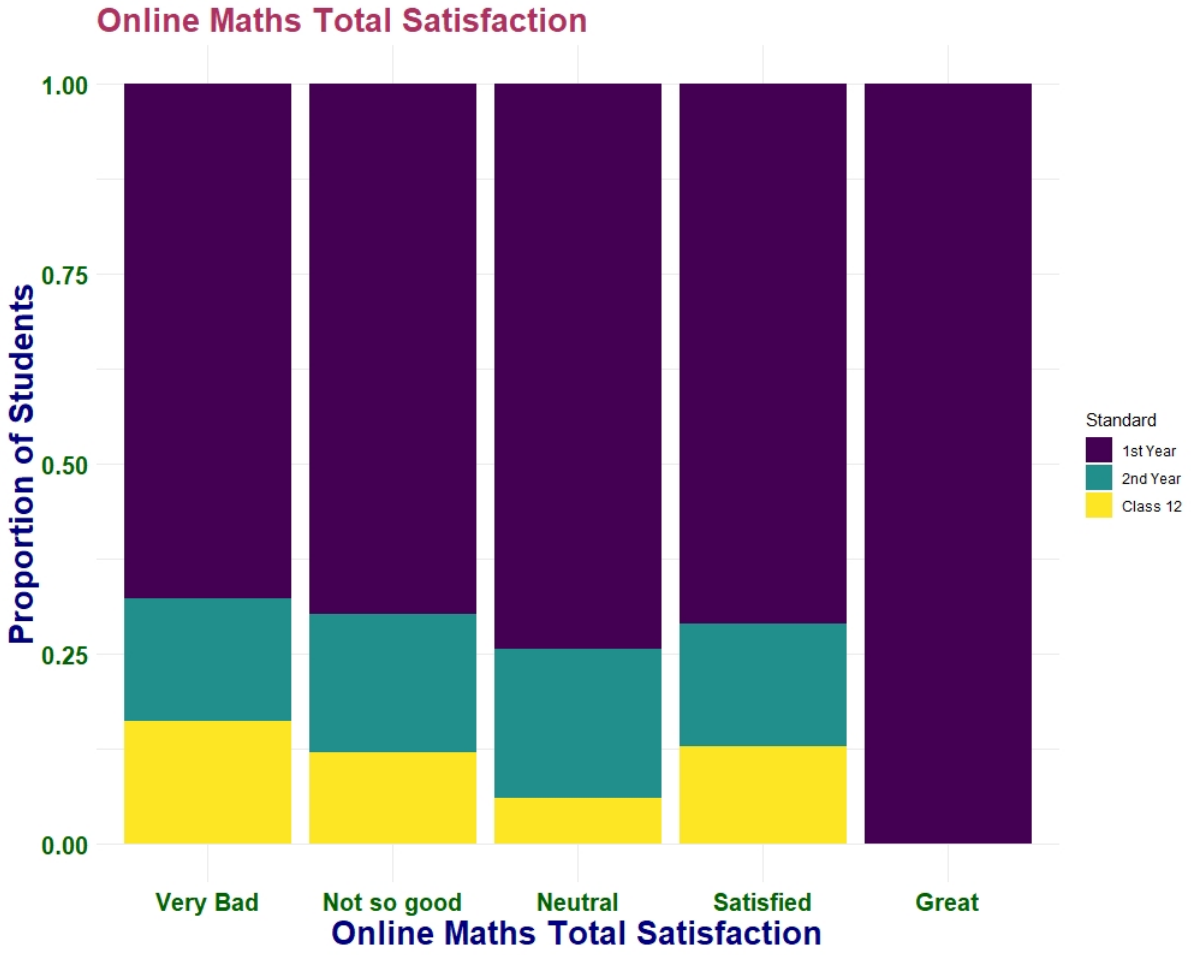
From the plot in the previous page, it is evident that since most of the students in our population were first year students, they have a substantial proportion for all categories. Most of the first year students seem to like the online classes in Language. The second year students are more or less satisfied by the online classes in Language, and it is noticeable that *none* of them have selected “Great". The class 12 students have an almost equal proportion for all the categories.



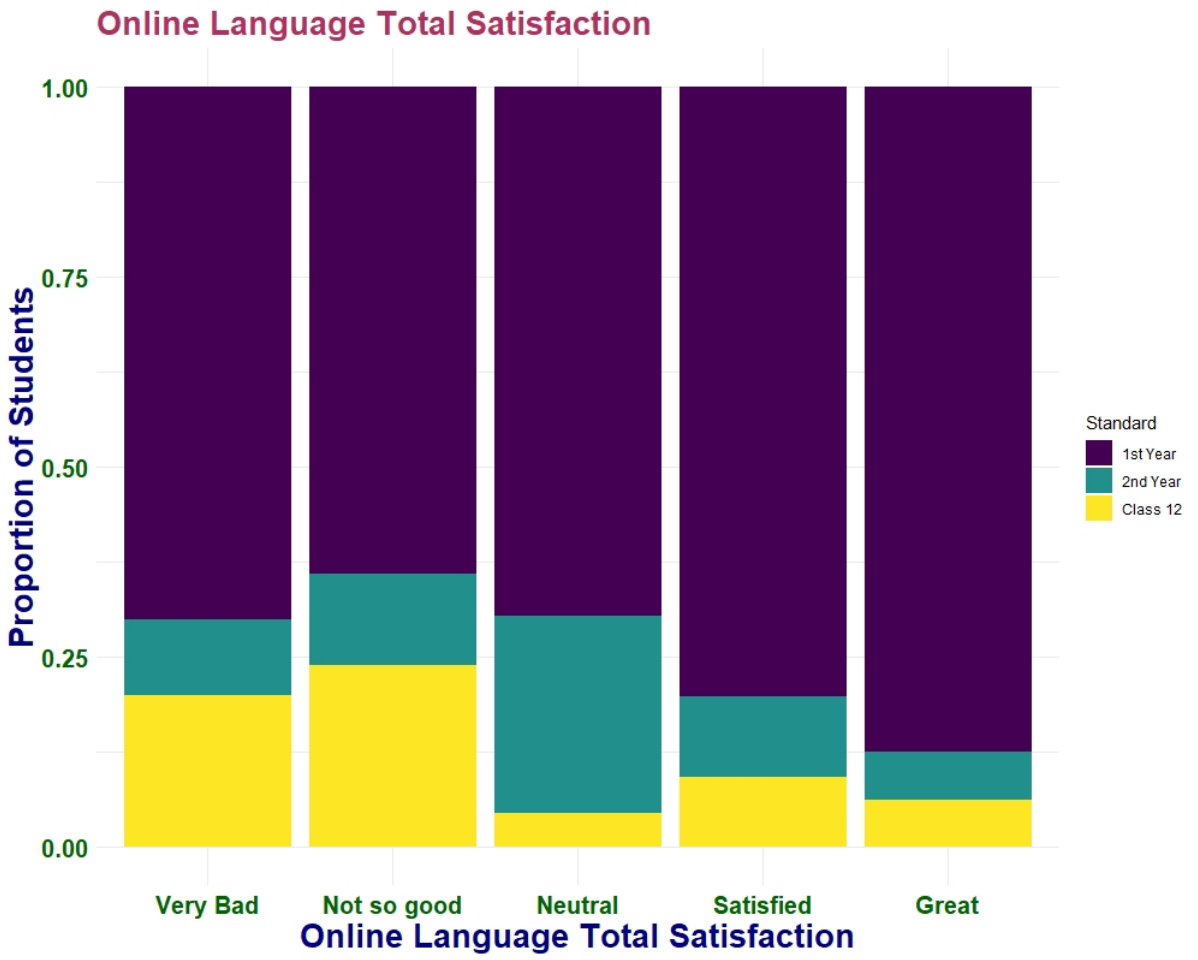
It is evident from the above plot, we conclude that most of the 1st year satisfied, and most of the class 12 students have answered “Very Bad" in the questionnaire.



Similarly, the above plot displays that most if the 1st year students are satisfied with the online classes. For the class 12 students, and 2nd year students, their opinions are mostly neutral and "Not so good" respectively.



Through these likert scales, we have been able to tap into the feelings of the people filling the form. The nature of the data entered by them reveal their thoughts while they answered the questionnaire. The greater number of satisfied and neutral reviews(intermediates) suggest that in-spite of their experience with offline classes, they have been able to adapt to the online setup (at least that's how they think it is). Even though the (uni-variate and bi-variate)graphs clearly suggests their difficulty in shifting to a medium due to the depreciation in the percentage of the average marks achieved, extreme views(great or very bad) are relatively low compared to intermediate views.



3.4.7 Conclusion:-

Hence , as evident from the previous discussions , we can make further plots of sets of seemingly related variables and study their relationship by analysing the plots . But these inferences seem a bit incomplete as they lack a strong Mathematical Framework . Thus , we need to use Statistical tools like Regression and Correlation Ideas to get a further rigorous insight into our Data !

3.5 Bivariate 2

The students in this group were Vaibhav Sherkar, Arghya Sarkar, Soumava Mondal, Sanskar Lalwani, Arunav Bhowmick, Aditya Narayan Sharma, Pranav Nair.

The group worked on bivariate analysis.

3.5.1 Bivariate analysis

Bivariate analysis is the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out whether there exists a relationship and if it exists, the strength of this relationship between two variables viz. x and y. The variables can be quantitative or categorical in nature. The bivariate analysis dealing with both the variables as quantitative has been covered in detail in the next section of regression. In this section, we will mostly explore the relationship between categorical and quantitative variables. The main categorical variables which we will be analysing are :

- Gender
- Place of residence
- Devices used for the purpose of online classes
- Satisfaction level

The main quantitative variables which we will be analysing are :

- Average internet speed
- Time spent in online classes
- Percentage marks in offline mode
- Percentage marks in online mode

The plots which we have used for analysing the data are :

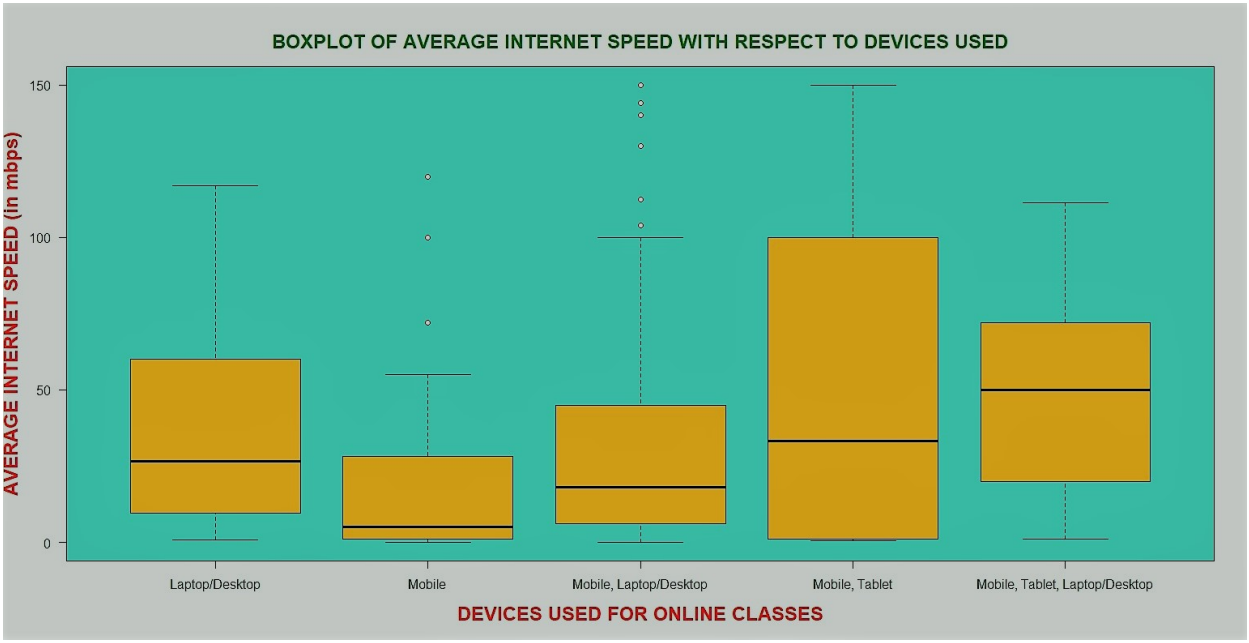
- Side by side boxplots
- Labelled Scatter Plots

3.5.2 Different types of plot used

- **Side by side boxplot :** Also known as parallel or comparative boxplot, this is a statistical tool used to compare the levels of a categorical variable with the help of a quantitative variable.
- **Labelled Scatter Plots :** Scatterplots display the relationship between two numerical variables. We have further labelled them to show their variation along with a categorical variable.

3.5.3 Average internet speed vs device used

We have made comparative vertical boxplots of average internet speed against the devices used for online classes. The speed was measured in mbps. The students entered the devices that they use for online classes. Their entries could have been a combination of multiple devices as well.



The mean average internet speed corresponding to each device is :-

- Laptop/Desktop - 36.914
- Mobile – 17.76
- Mobile, Laptop/Desktop – 34.01
- Mobile, Tablet – 52.962
- Mobile, Tablet, Laptop/Desktop – 47.23

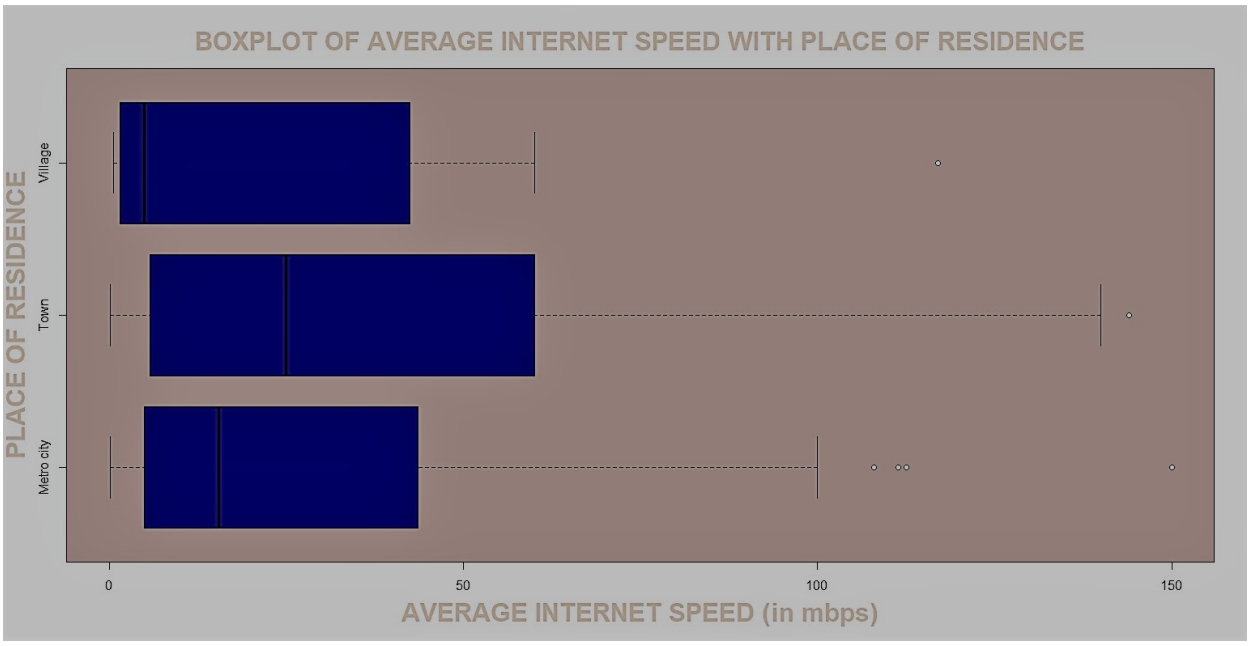
As we can observe from the boxplots, mobile phone users in general have very slow internet speed whereas tablet and laptop users have a considerably higher internet speed. Students who use a tablet and a mobile phone seem to have the highest internet speed but this can be due to the fact that we received less entries corresponding to this combination.

Note : There were some students who failed to enter their internet speed and these datapoints were left blank. Consequently, we have ignored these. We have also left out the entries corresponding to “Tablet” , “Tablet, Laptop/Desktop” as these had very low frequency and an appro-

priate boxplot could not be made.

3.5.4 Place of residence and average internet speed

We have made comparative horizontal boxplots of average internet speed against the place of residence of the student.



The summary for the respective places of residences is as follows :-

- Village :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.60	1.50	5.00	30.09	42.50	117.00

- Town :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.10	5.73	25.00	37.25	60.00	144.00

- Metro City :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.18	5.00	15.00	29.95	42.00	150.00

As we can see, the mean internet speed is almost same for students belonging to village and metro city and it is slightly less than those for students belonging to towns.

However, there is a considerable difference when comparing the median and quartiles of village and town/metros. This is consistent with the fact that urban areas will have better connectivity that rural areas.

Note : As before, we have ignored the entries where connection speed was left blank.

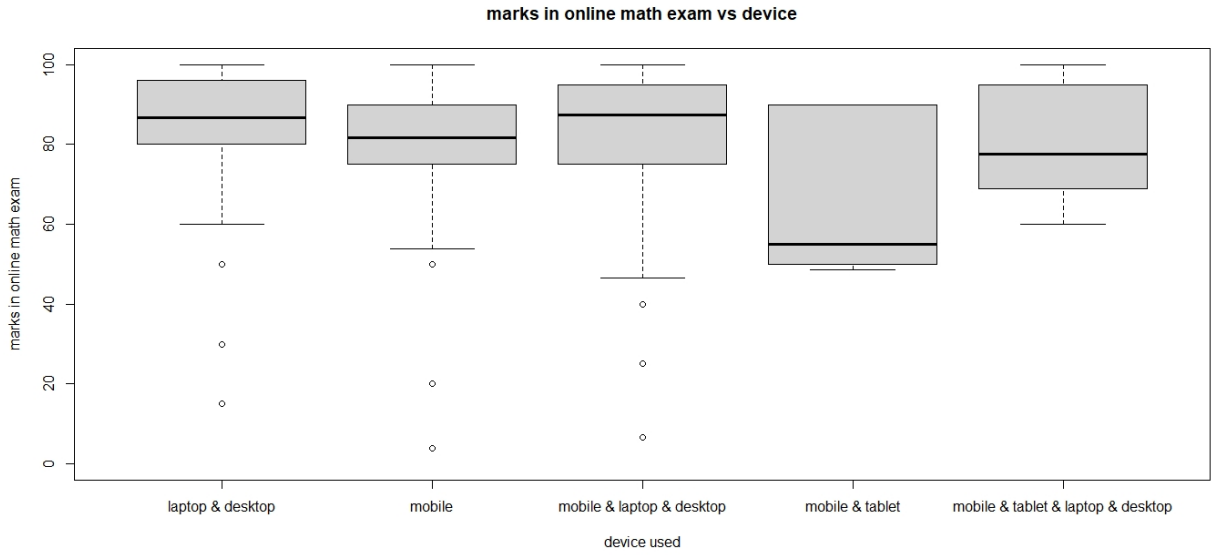
3.5.5 Online exam marks with devices used

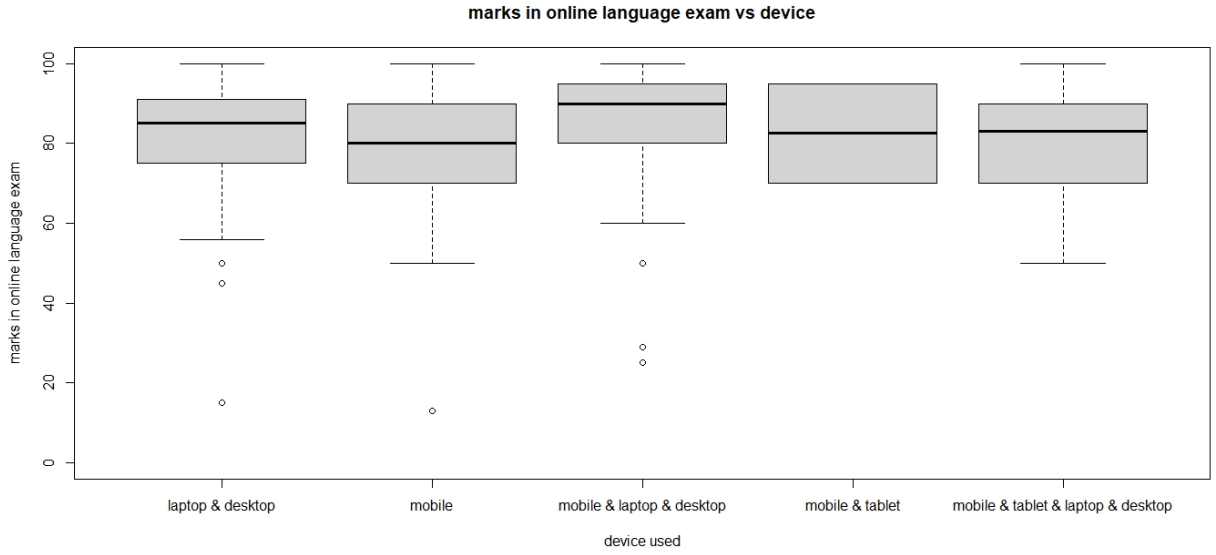
The data we collected from the survey was analysed and the boxplot of device name vs. online examination marks was drawn. The devices were :-

- Laptop/Desktop
- Mobile
- Mobile, Desktop/Laptop
- Mobile, Tablet
- Mobile, Laptop/Desktop & Tablet
- Tablet
- Tablet, Desktop/Laptop

However, the last two categories had very few entries to draw a meaningful boxplot and hence we ignored them. Also, some students had entered their marks as zero since they were yet to give an exam and thus, we have ignored these as well.

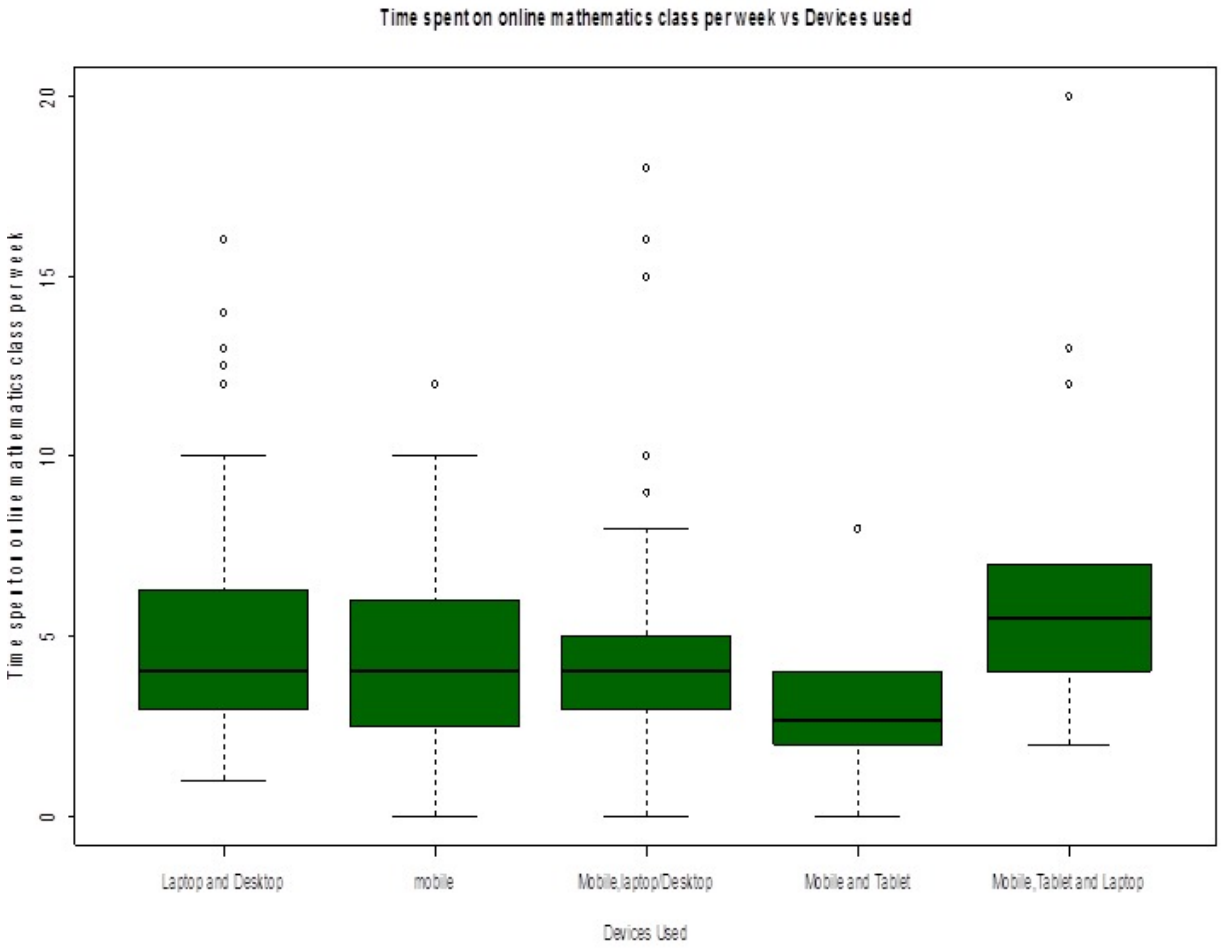
3.5.6 Analysis of boxplots





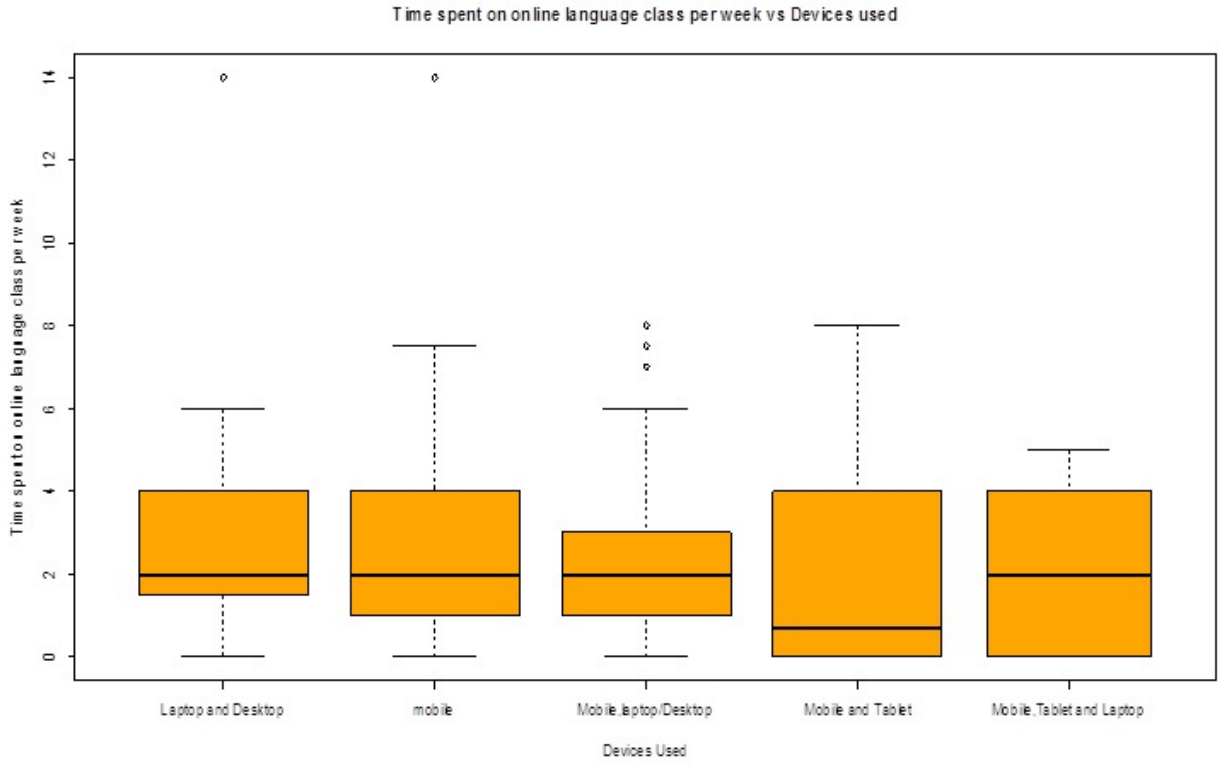
From the above diagram it appears that no matter what device a student is using his/her marks on the online exam roughly lies in 80%-95%. It seems from the diagrams that marks of online exams (both in mathematics and languages) are not dependent on their device's role.

3.5.7 Time spent on online mathematics classes per week with devices used



Mobile and Tablet users usually spend less time on online classes as compared to other categories. Median time spent for users of all three devices is greater as compared to others. This may be due to availability of more options.

3.5.8 Time spent on online languages classes per week with devices used

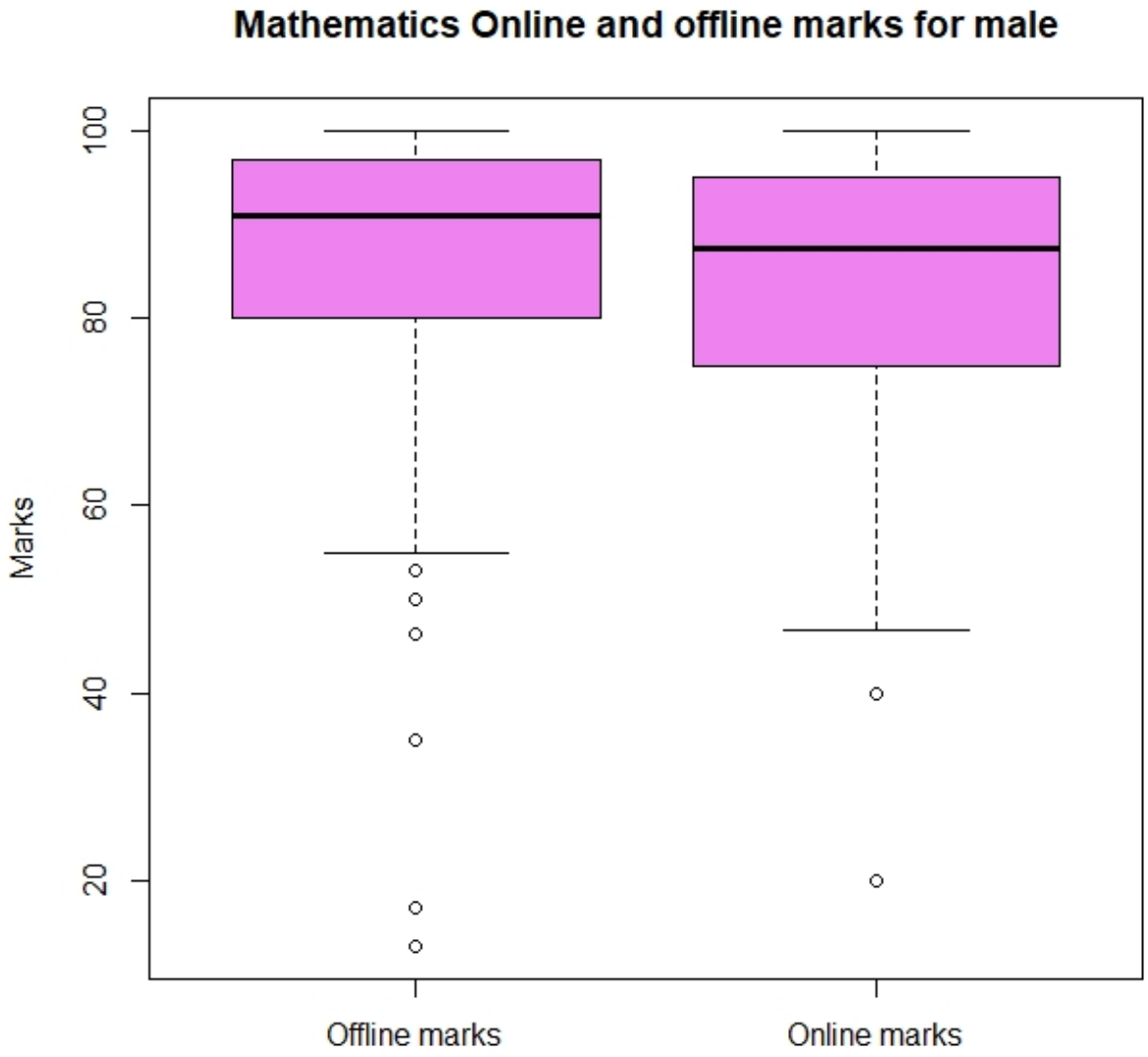


As we can see from the graph, median values are almost equal except for Mobile and Tablet group, which have a lesser value which was also the case for mathematics classes.

3.5.9 Gender wise distribution of online and offline marks

Here, we have split the data according to gender of student. We have considered one gender and one subject at a time and compared the distribution of marks in online and offline exam.
Some students entered the marks of only one subject (as they have not opted for the other subject) or provided only offline marks (as they did not appear for online exam). In such cases, the marks entered were 0.
We have discarded all data points with 0 marks and then considered the data for making box plots.

3.5.10 Online and offline marks in mathematics for male students



Summary for Offline Marks

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Dev.
13.00	80.00	91.00	87.05	97.00	100.00	14.34898

Summary for Online Marks

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Dev.
20.00	75.00	87.5	83.04	95.00	100.00	14.83306

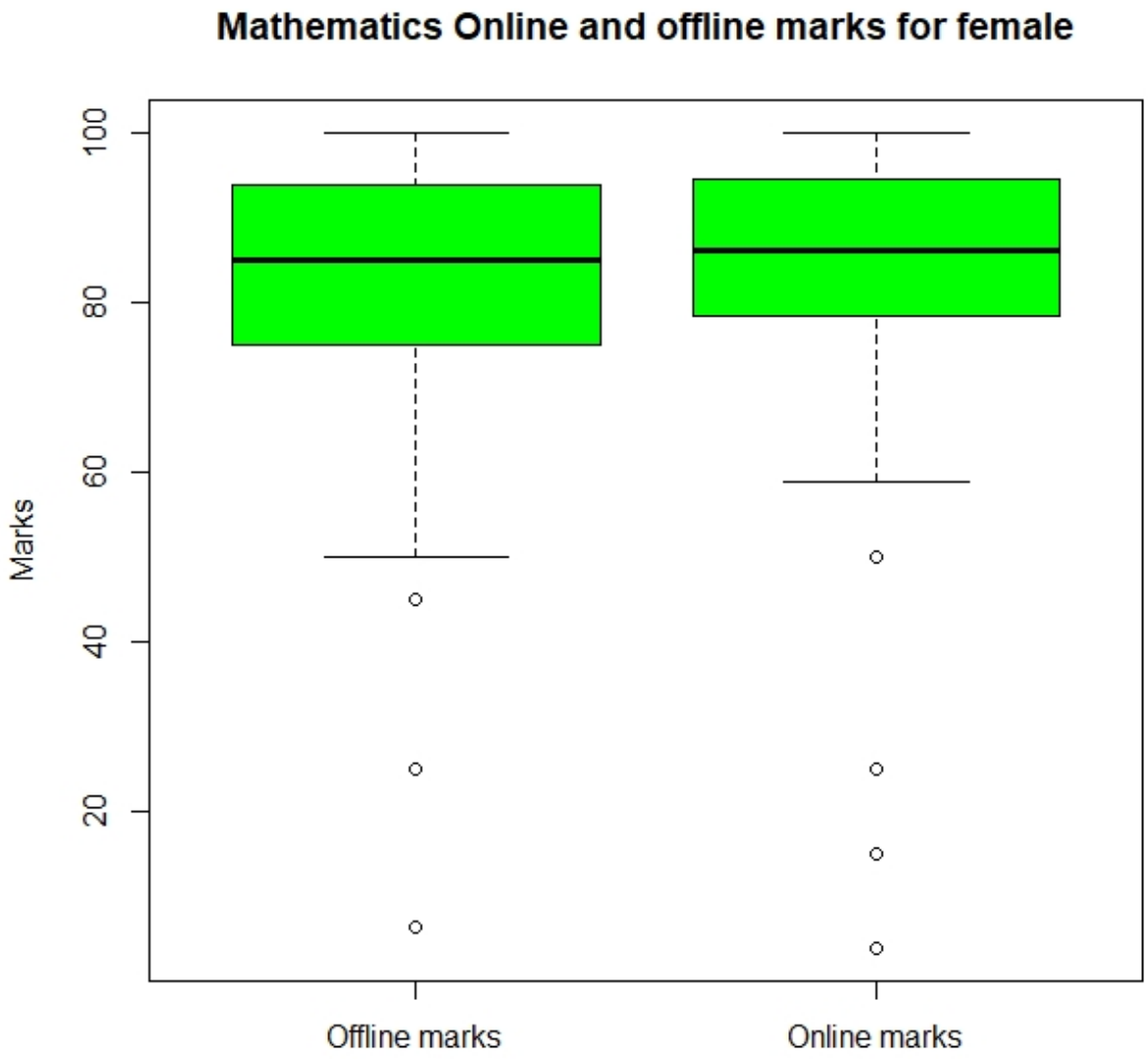
It is evident from values of standard deviation and increase in value IQR that variability of online marks is more as compared to offline marks.

In offline mode, marks are more concentrated above 80.

As we can see, the values of Q_1, Q_3, median and mode have decreased, when the mode of examination is changed from offline to online.

It seems that offline exams in Mathematics are more beneficial for males.

3.5.11 Online and offline marks in mathematics for female students



Summary of offline marks

Min	1st Quartiles	Median	Mean	3rd Quartiles	Max	Std. Deviation
6.50	75.00	85.00	80.46	94.00	100.00	19.45035

Summary of Marks in offline exams

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Dev.
4.00	78.50	86.17	80.68	94.50	100.00	20.57504

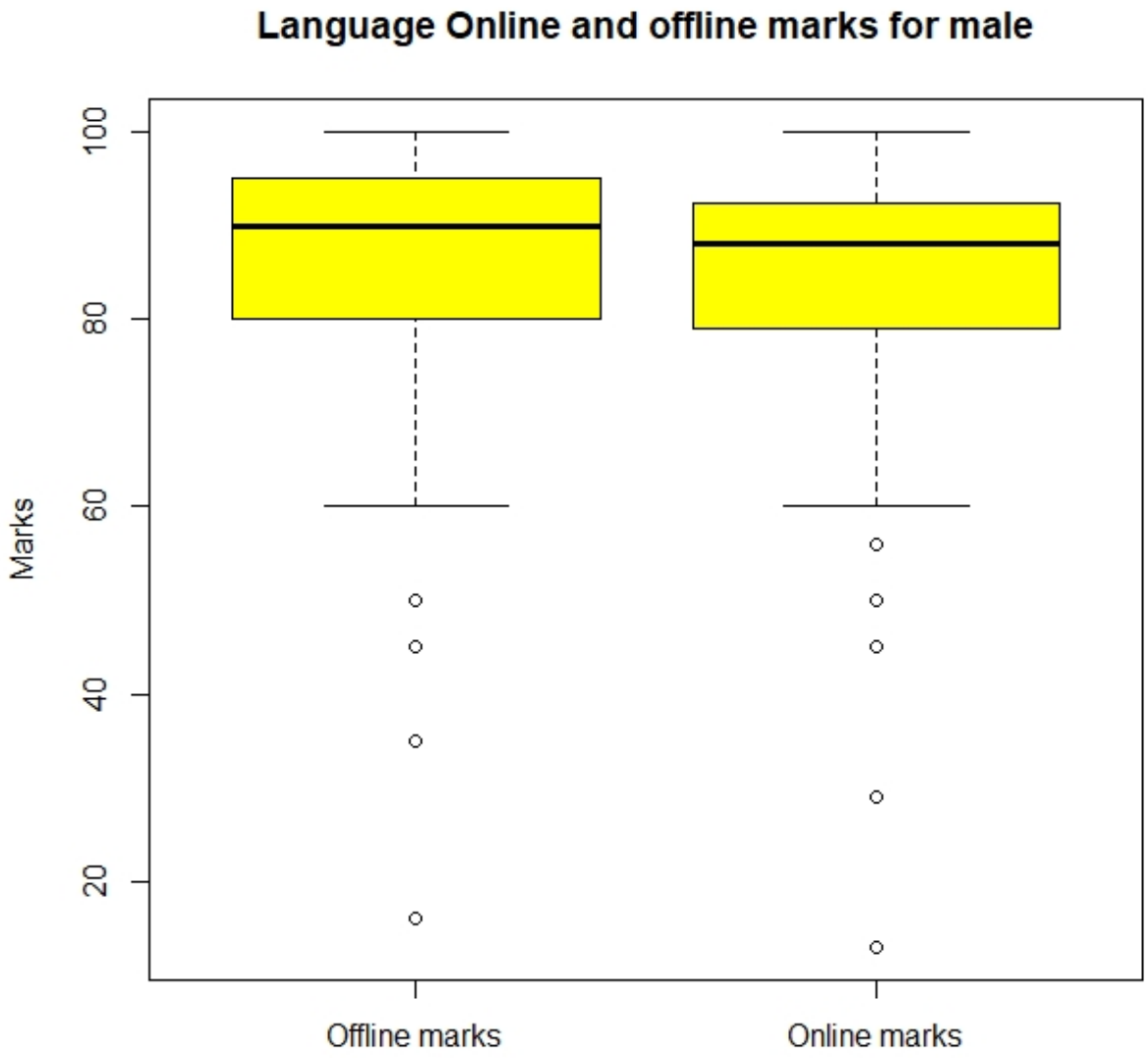
Variability on online marks is greater than that on offline marks.

In online mode, marks are more concentrated above 80.

Values of Q_1, Q_3, median and mean had increased slightly as the mode of the examination is changed from offline to online.

It seems that online exams are more beneficial for females than offline exams.

3.5.12 Online and offline marks in languages for male students



Summary of offline Marks

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Dev.
16.00	80.00	90.00	85.44	95.00	100.00	12.76549

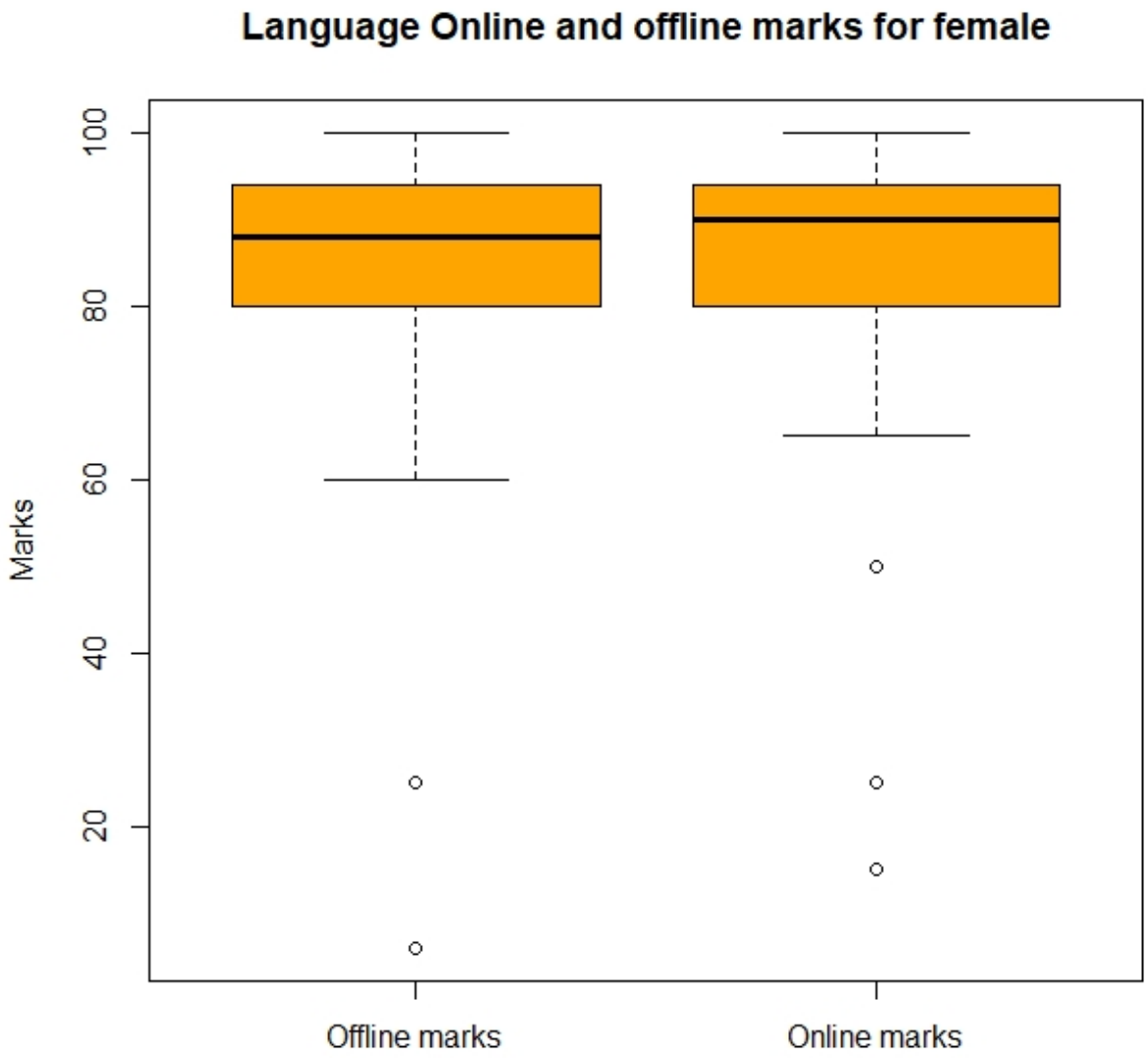
Summary of online marks

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Dev.
13.00	79.50	88.00	83.05	92.25	100.00	14.73801

The value of Q_1, Q_3 median and mode have decreased when the mode of the examination is changed from offline to online.

IQR for marks in online exam in language is lesser than IQR for marks in offline exams in language.

3.5.13 Online and offline marks in languages for female students



Summary of offline Marks

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Dev.
6.00	80.00	88.00	84.21	94.00	100.00	16.20772

Summary of online marks

Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Dev.
15.00	80.00	90.00	84.56	94.00	100.00	16.46006

The values of Q_1 and Q_3 are the same in both the cases. The mean is slightly more when the mode of the exam is online.

Median marks in language in offline exam are lesser than median marks in language in online exam in case of females.

We see from approximately same value of IQR and standard deviation that Variability in marks in both modes is same.

3.5.14 Bivariate Frequency Tables

Devices	Overall Satisfaction Level (Maths)					Total
	Very Bad	Not So Good	Neutral	Satisfied	Great	
Mobile	7	4	15	18	3	47
Laptop/Desktop	11	10	24	23	6	74
Tablet	0	0	1	0	0	1
Mobile, Laptop/Desktop	8	17	35	18	7	85
Mobile, Tablet	2	1	0	1	0	4
Tablet, Laptop/Desktop	0	0	1	1	1	3
Mobile, Tablet, Laptop/Desktop	3	1	6	1	1	12
Total	31	33	82	62	18	226

Devices	Overall Satisfaction Level (Language)					Total
	Very Bad	Not So Good	Neutral	Satisfied	Great	
Mobile	7	5	20	13	2	47
Laptop/Desktop	5	12	25	25	7	74
Tablet	0	1	0	0	0	1
Mobile, Laptop/Desktop	4	12	34	28	7	85
Mobile, Tablet	2	0	1	1	0	4
Tablet, Laptop/Desktop	0	0	1	1	1	3
Mobile, Tablet, Laptop/Desktop	3	2	6	0	1	12
Total	21	32	87	68	18	226

Conditional distribution of overall satisfaction level (Maths) for different devices

Devices	Overall Satisfaction Level (Maths)					Total
	Very Bad	Not So Good	Neutral	Satisfied	Great	
Mobile	14.89	8.51	31.91	38.29	6.38	100
Laptop/Desktop	14.86	13.51	32.43	31.08	8.10	100
Tablet	0	0	100	0	0	100
Mobile, Laptop/Desktop	9.41	20	41.17	21.17	8.23	100
Mobile, Tablet	50	25	0	25	0	100
Tablet, Laptop/Desktop	0	0	33.33	33.33	33.33	100
Mobile, Tablet, Laptop/Desktop	25	8.33	50	8.33	8.33	100

Conditional distribution of overall satisfaction level (Language) for different devices

Devices	Overall Satisfaction Level (Language)					Total
	Very Bad	Not So Good	Neutral	Satisfied	Great	
Mobile	14.89	10.63	42.55	27.65	4.25	100
Laptop/Desktop	6.75	16.21	33.78	33.78	9.45	100
Tablet	0	100	0	0	0	100
Mobile, Laptop/Desktop	4.70	14.11	40	32.94	8.23	100
Mobile, Tablet	50	0	25	25	0	100
Tablet, Laptop/Desktop	0	0	33.33	33.33	33.33	100
Mobile, Tablet, Laptop/Desktop	25	16.67	50	0	8.33	100

Devices	Online Exam Satisfaction Level (Maths)					Total
	Very Bad	Not So Good	Neutral	Satisfied	Great	
Mobile	9	6	14	12	6	47
Laptop/Desktop	4	14	27	22	7	74
Tablet	0	1	0	0	0	1
Mobile, Laptop/Desktop	8	10	39	23	5	85
Mobile, Tablet	2	0	1	1	0	4
Tablet, Laptop/Desktop	0	1	1	0	1	3
Mobile, Tablet, Laptop/Desktop	4	2	3	1	2	12
Total	27	34	85	59	21	226

Devices	Online Exam Satisfaction Level (Language)					Total
	Very Bad	Not So Good	Neutral	Satisfied	Great	
Mobile	8	8	15	13	3	47
Laptop/Desktop	5	8	32	22	7	74
Tablet	0	1	0	0	0	1
Mobile, Laptop/Desktop	4	16	31	27	7	85
Mobile, Tablet	1	1	1	1	0	4
Tablet, Laptop/Desktop	0	0	2	0	1	3
Mobile, Tablet, Laptop/Desktop	3	0	6	3	0	12
Total	21	34	87	66	18	226

Devices	Online Teaching Quality Satisfaction Level (Maths)					Total
	Very Bad	Not So Good	Neutral	Satisfied	Great	
Mobile	5	7	14	18	3	47
Laptop/Desktop	5	16	17	26	10	74
Tablet	0	0	1	0	0	1
Mobile, Laptop/Desktop	6	14	29	22	14	85
Mobile, Tablet	2	1	0	1	0	4
Tablet, Laptop/Desktop	0	0	1	1	1	3
Mobile, Tablet, Laptop/Desktop	2	3	3	3	1	12
Total	20	41	65	71	29	226

Devices	Online Teaching Quality Satisfaction Level (Language)					Total
	Very Bad	Not So Good	Neutral	Satisfied	Great	
Mobile	8	3	17	18	1	47
Laptop/Desktop	4	7	29	29	5	74
Tablet	0	1	0	0	0	1
Mobile, Laptop/Desktop	3	11	38	25	8	85
Mobile, Tablet	2	0	1	1	0	4
Tablet, Laptop/Desktop	0	0	0	2	1	3
Mobile, Tablet, Laptop/Desktop	3	3	4	1	1	12
Total	20	25	89	76	16	226

3.5.15 Online and offline marks with overall satisfaction level

As we are studying the relationship between two quantitative variables, we have introduced the measure of correlation r, to explore the relationship between the variables.

Here,

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}.$$

where s_x and s_y are the standard deviations of x and y respectively, and are given by,

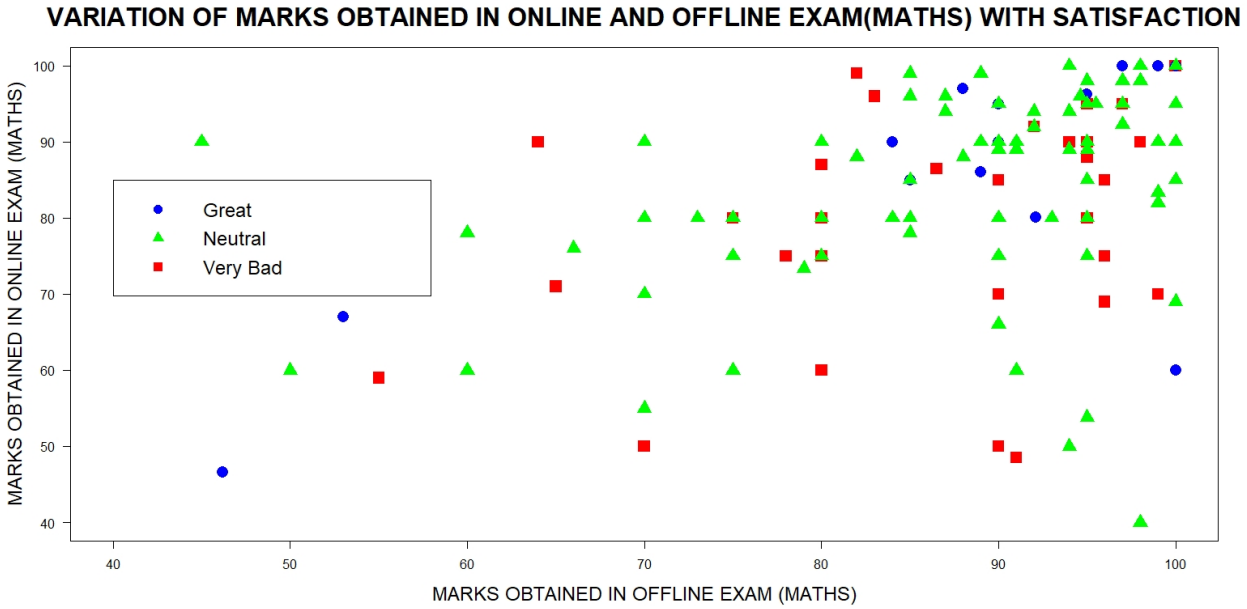
$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

and

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}.$$

Higher the value of $|r|$, higher is the correlation between the two variables.

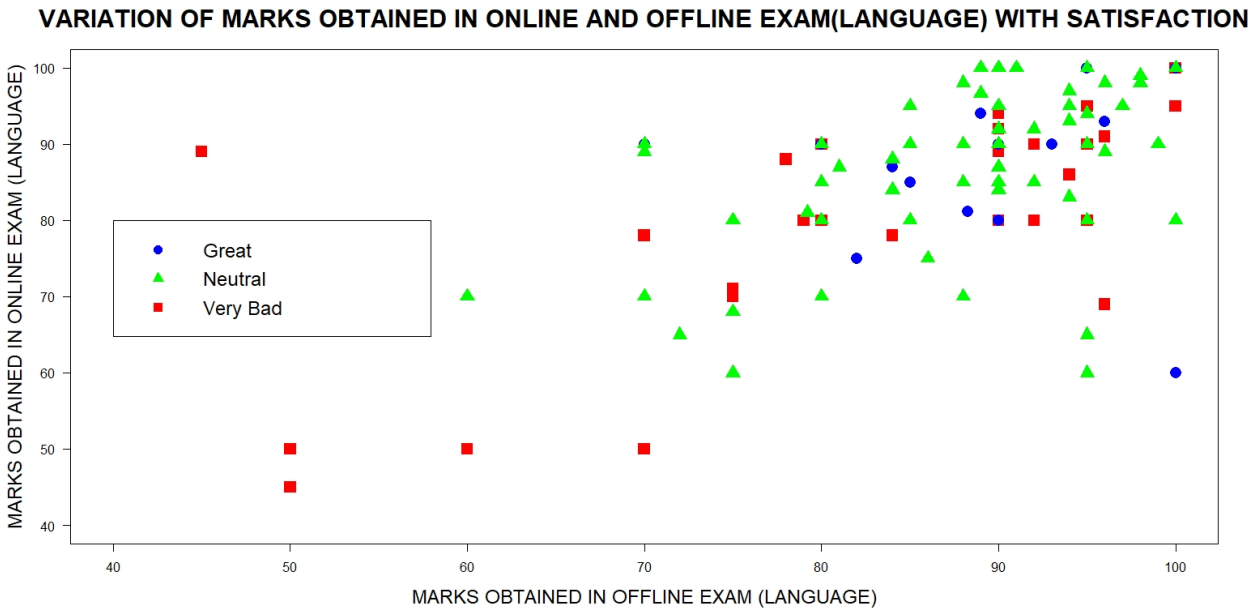
We have plotted the marks obtained in online examination along the y-axis and marks obtained in offline examination along the x-axis. We have considered the variation of these variables with respect to the overall satisfaction level of the student in online classes. We have considered only three levels, the extremes and the central one. We have done this for both Math and language.



The value of $|r|$ for the three levels of satisfaction is :-

- Great – 0.9185
- Neutral – 0.5148
- Very Bad – 0.6457

It is evident from high value of $|r|$ that online and offline marks are highly correlated in the case of students who are highly satisfied (which have given 'Great' level of satisfaction).



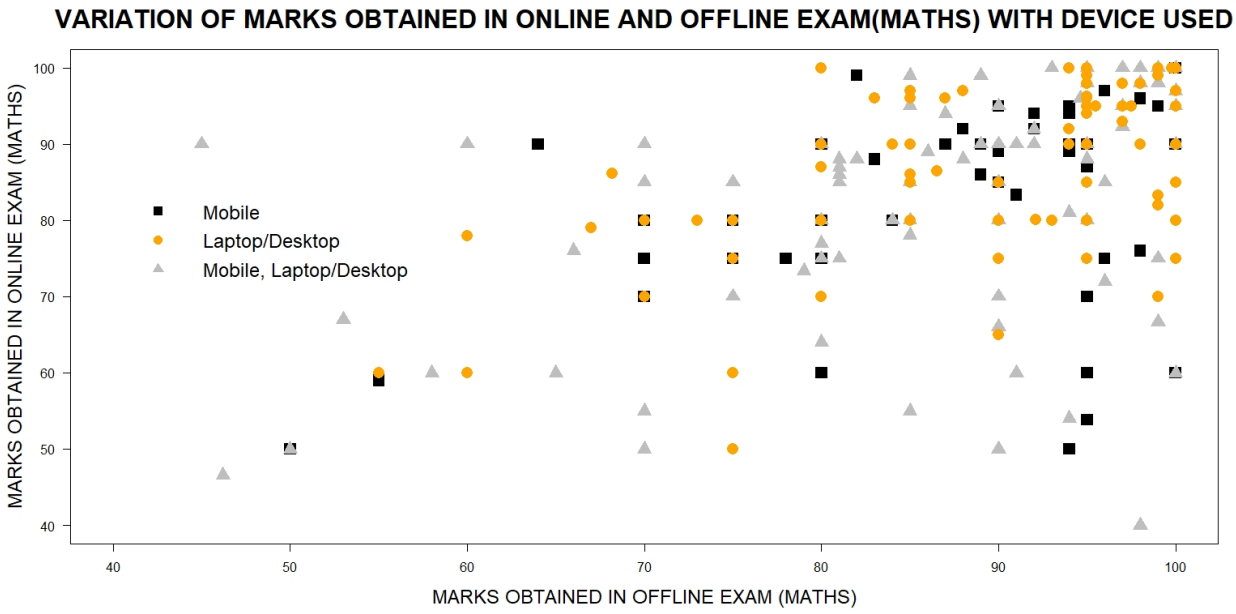
The value of $|r|$ for the three levels of satisfaction is :-

- Great – 0.4918
- Neutral – 0.5830
- Very Bad – 0.7130

On the other hand, online and offline marks in language have a higher correlation for the students who were displeased (which have given 'Very Bad' level of satisfaction) with online classes.

3.5.16 Online and offline marks with devices

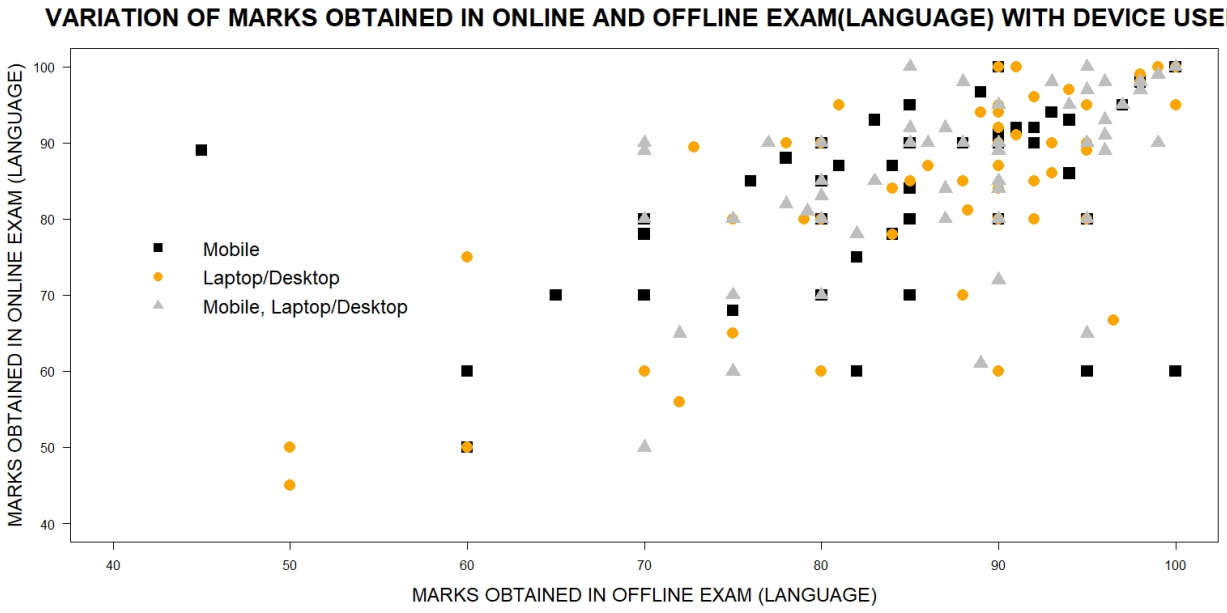
As before, we have plotted the marks obtained in online examination along the y-axis and marks obtained in offline examination along the x-axis. We have considered the variation of these variables with respect to the device used for attending online classes. We only consider three only combinations, Mobile; Laptop/Desktop; Mobile and Laptop/Desktop since these comprise over 90% of the total datapoints.



We have made the plots for language and device. The value of $|r|$ for the three types of combination of devices is :-

- Mobile – 0.6303
- Laptop/Desktop – 0.6529
- Mobile, Laptop/Desktop – 0.5983

Here, we can see from the values of $|r|$, all the three categories have more or less the same correlation.



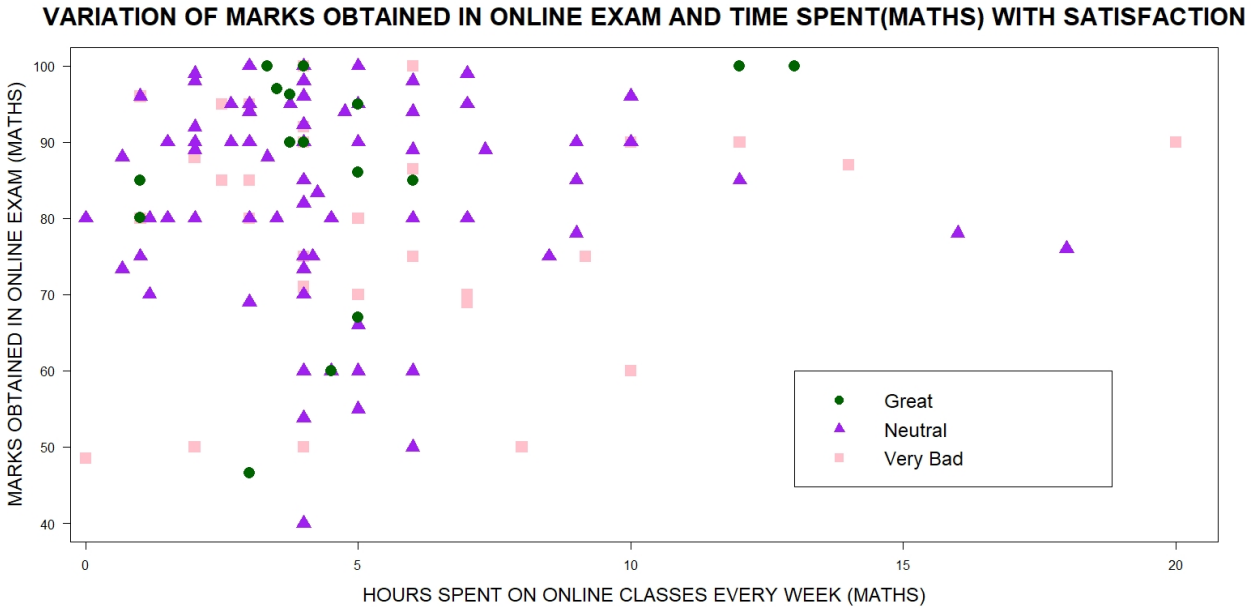
The value of $|r|$ for the three types of combination of devices is :-

- Mobile – 0.6059
- Laptop/Desktop – 0.5648
- Mobile, Laptop/Desktop – 0.7311

For language, we can see that online and offline marks have a higher correlation for Mobile, Laptop/Desktop case.

3.5.17 Online marks and time spent with overall satisfaction level

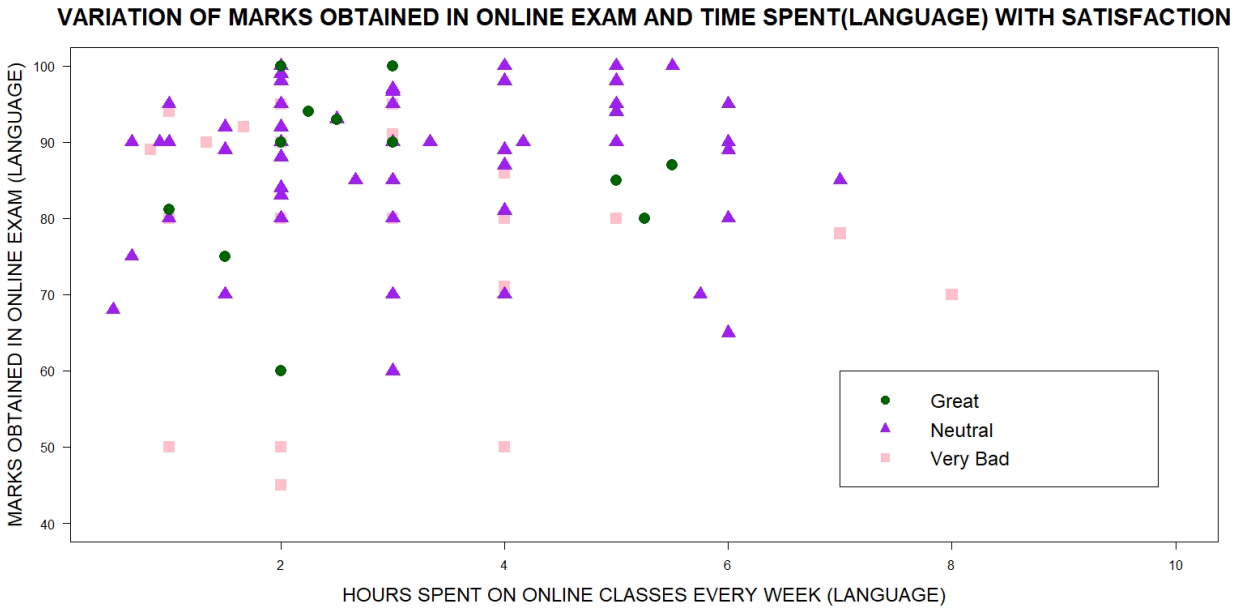
We have plotted the marks obtained in online examination along the y-axis and time spent (in hours) in attending online classes every week along the x-axis. We have considered the variation of these variables with respect to the overall satisfaction level of the student in online classes. We have considered only three levels, the extremes and the central one. We have done this for both Math and language.



The value of $|r|$ for the three levels of satisfaction is :-

- Great – 0.3004
- Neutral – 0.0374
- Very Bad – 0.1568

Very low values of $|r|$ in all 3 cases signify that that time spent on mathematics and marks obtained in mathematics are not correlated irrespective of student's satisfaction.



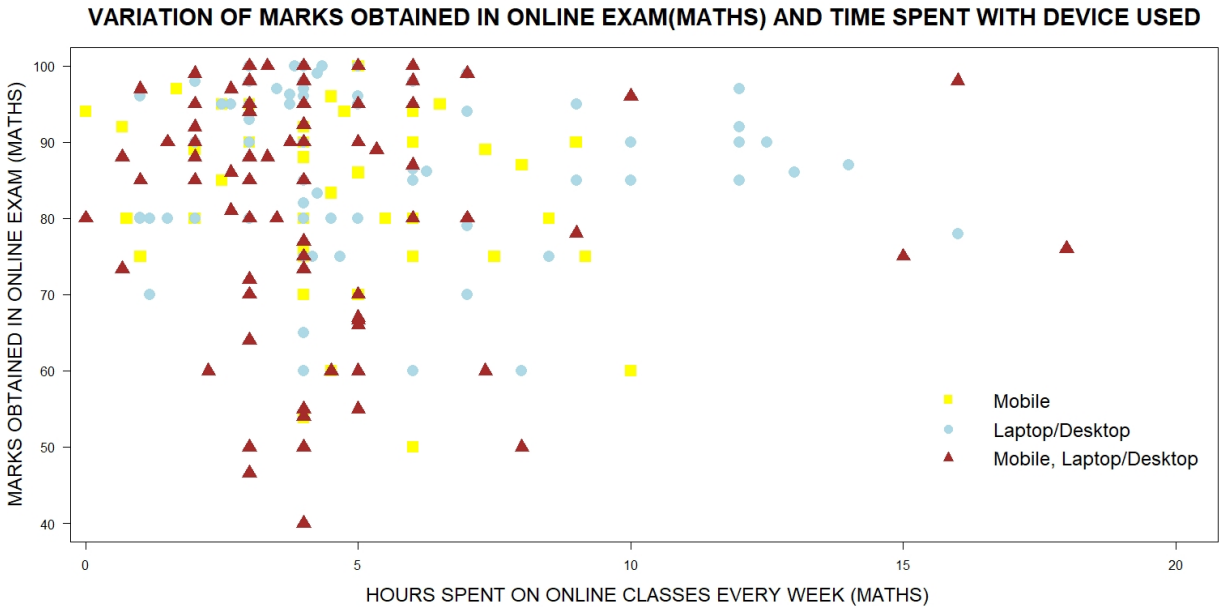
The value of $|r|$ for the three levels of satisfaction is :-

- Great – 0.1609
- Neutral – 0.3536
- Very Bad – 0.3453

From low values of $|r|$, we can say that hours spent in language are not much correlated with marks obtained in language. Students which score high marks had responded as 'Great' or 'Neutral' level of satisfaction.

3.5.18 Online marks and time spent with device used

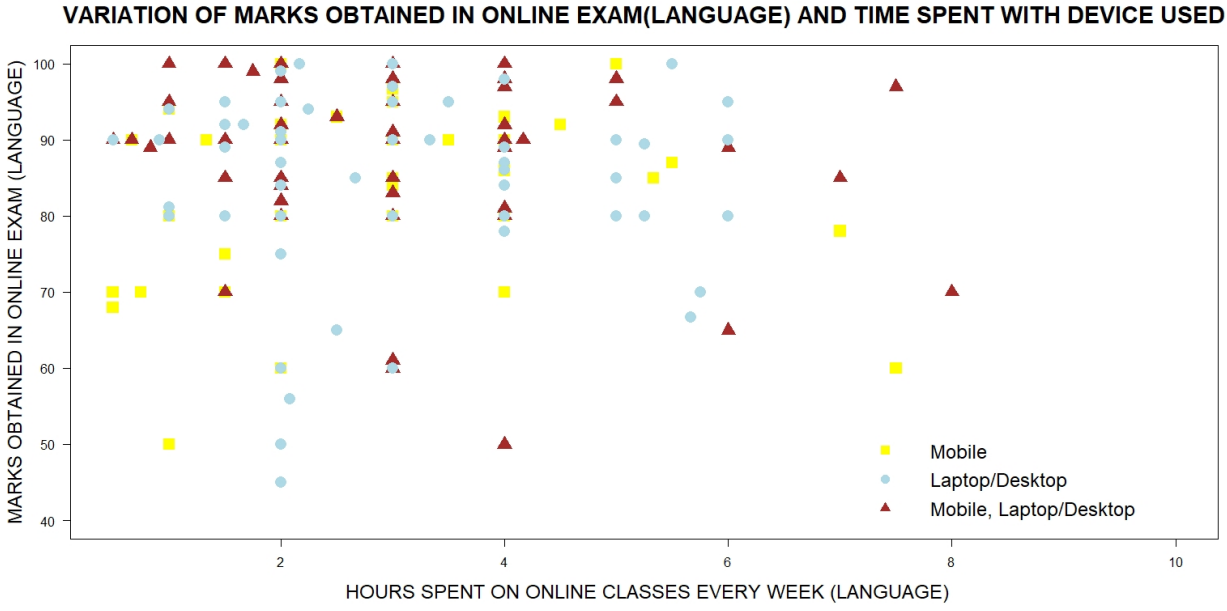
As before, we have plotted the marks obtained in online examination along the y-axis and time spent (in hours) in attending online classes every week along the x-axis. We have considered the variation of these variables with respect to the device used for attending online classes. We only consider three only combinations, Mobile; Laptop/Desktop; Mobile and Laptop/Desktop since these comprise over 90% of the total datapoints. We have made the plots for language and device.



The value of $|r|$ for the three types of combination of devices is :-

- Mobile – 0.0141
- Laptop/Desktop – 0.0244
- Mobile, Laptop/Desktop – 0.0842

It is evident from very low values of $|r|$ that Time spent on mathematics and marks obtained in mathematics are not correlated irrespective of device used.



The value of $|r|$ for the three types of combination of devices is :-

- Mobile – 0.0244
- Laptop/Desktop – 0.2998
- Mobile, Laptop/Desktop – 0.2687

It is evident from very low values of $|r|$ that Time spent on language and marks obtained in language are not correlated irrespective of device used.

3.6 Regression

The members in this group were Shreeja Bhakat, Anushka De, Semanti Dutta, Samahriti Mukherjee, Prisha Reddy and Aytijhya Saha.

This group worked on **Regression**. In this section, we will analyse the following continuous variables :

- Hours devoted in Online Mathematics Classes
- Hours devoted in Online Language Classes
- Marks in Mathematics in Offline mode of Examination
- Marks in Language in Offline mode of Examination
- Marks in Mathematics in Online mode of Examination
- Marks in Language in Online mode of Examination
- Internet Speed

A regression line summarizes the relationship between two quantitative variables, i.e., it shows how a response variable y changes as an explanatory variable x changes.

A regression line is also used to **predict** the value of y for a given value of x. After calculating the direction and strength of linear relationship between two quantitative variables in the previous sections, we now move to **fitting a line** to the data.

Here we use the method of least squares for determining the regression line. The **least-squares**

regression line of y on x minimizes the sum of squares of errors or residuals, i.e, the line makes the sum of squares of the vertical distances of the data points from the line as small as possible.

Residuals are defined by :

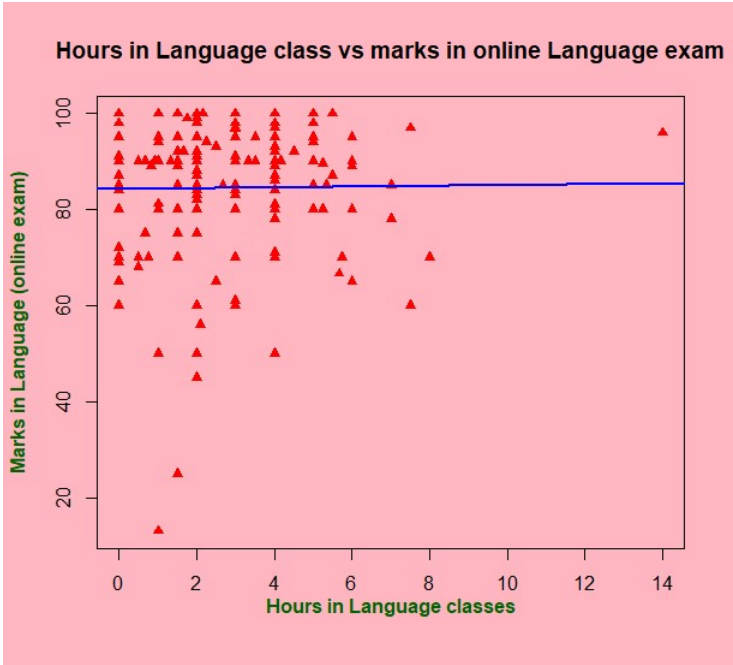
$$e_i = (\text{observed y value})_i - (\text{predicted y value})_i$$

The absolute value of e_i is given by the length of the segment for each data point.

3.6.1 Hours Devoted in Language Subject and Marks in Language

Let us consider marks in Language to be the response variable (say y) and hours devoted in language to be the explanatory variable (say x).
The value of $|r|$ is 0.0001429.
So 0.01429% values of y are explained by the values of x, so the least square linear regression line is not at all a good fit to the data.
The equation of least square linear regression line of y on x is:

$$Y = 0.08339811x + 84.20971709$$

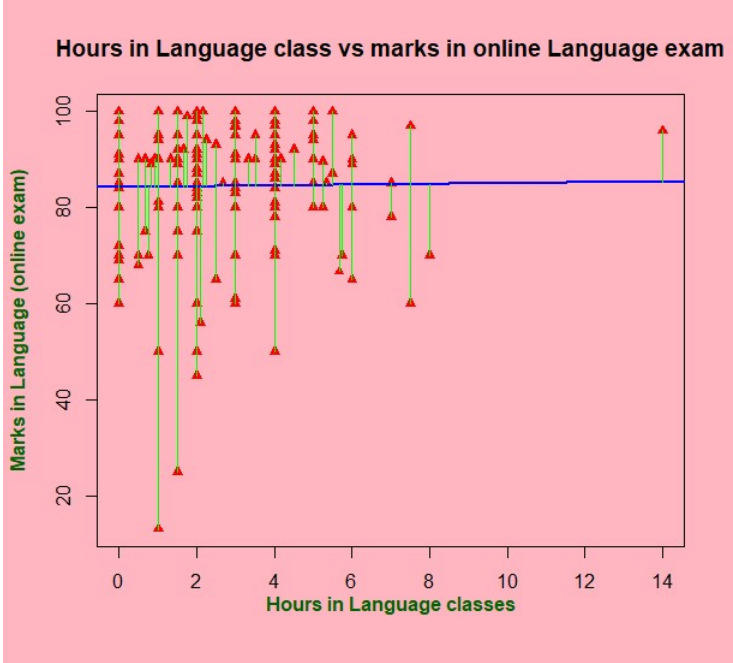


After seeing the scatter plot of hours devoted in language and marks in online language exam, we can say that the points are randomly scattered, and there is no specific pattern, and after $x = 8$, there is only one point which is leverage point. Having no specific pattern, it seems

that the least square linear regression line would not be a good model for the data.

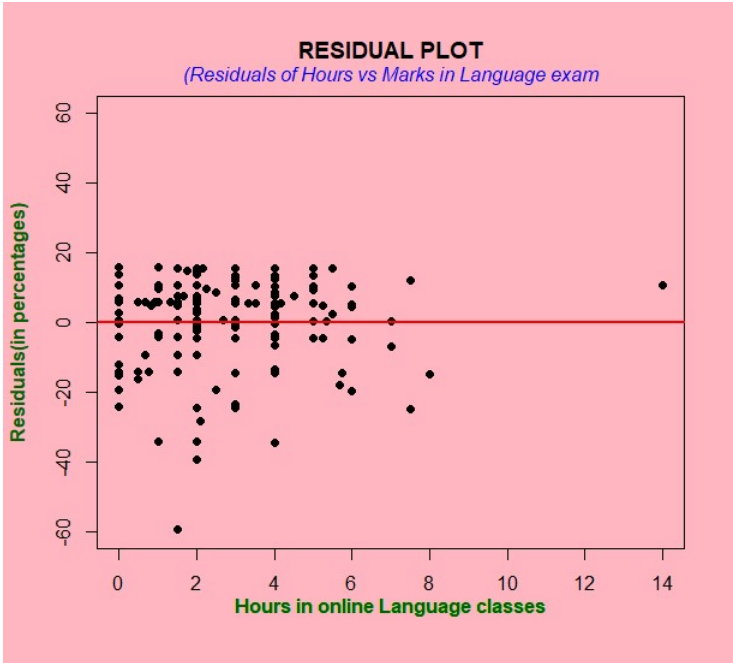
Deviation of points from the regression line

The deviations of the data points from the regression line can be better visualized in the following plot :



Residual Plot

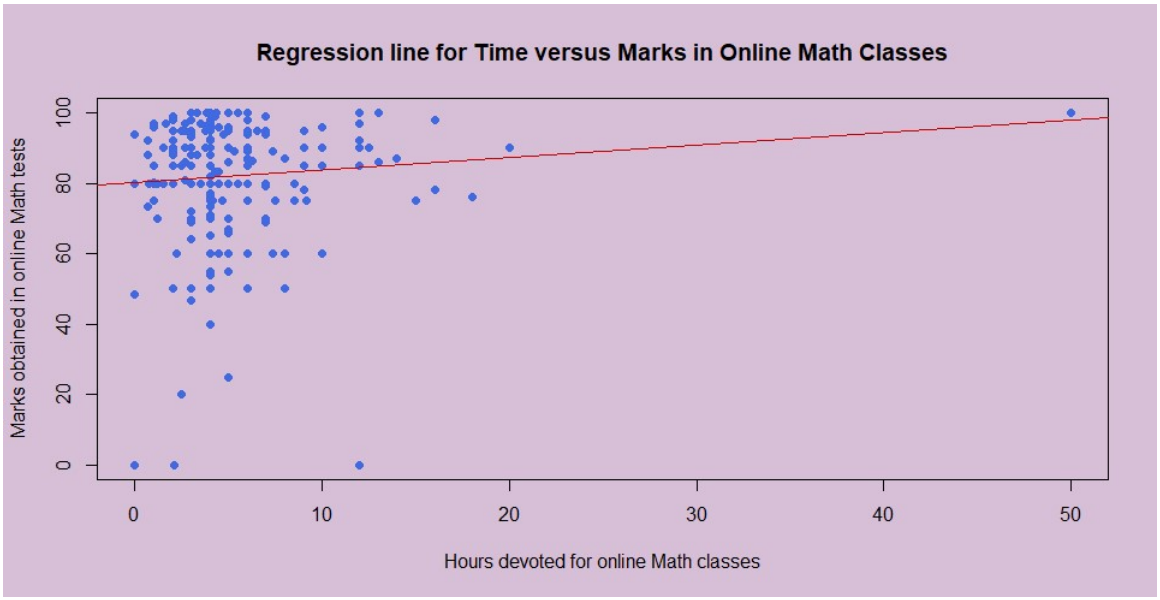
Now we plot the residuals against the explanatory variable, i.e., hours devoted in Language. We can see that after $x = 4$, the variance decreases, so errors are not homoscedastic. So as x increases prediction of y becomes less accurate.



3.6.2 Time Devoted for Online Math Classes and Marks Obtained in Online Math Tests

From the below graph, we must note that most of the data points are clustered at the top-left side of the plot, and thus, we can say that most students have spent not more than 10 hours on online math classes per week. Thus, the data is not uniformly distributed. Also, we observe that data points, with the same x value (time devoted for online math classes) vary largely on their y values(marks obtained in the online math tests). Due to this very observation, we can predict that the regression coefficient r for the given data would be small, since there is no proper linear relation between the two variables.

The least squares linear regression is as shown below :



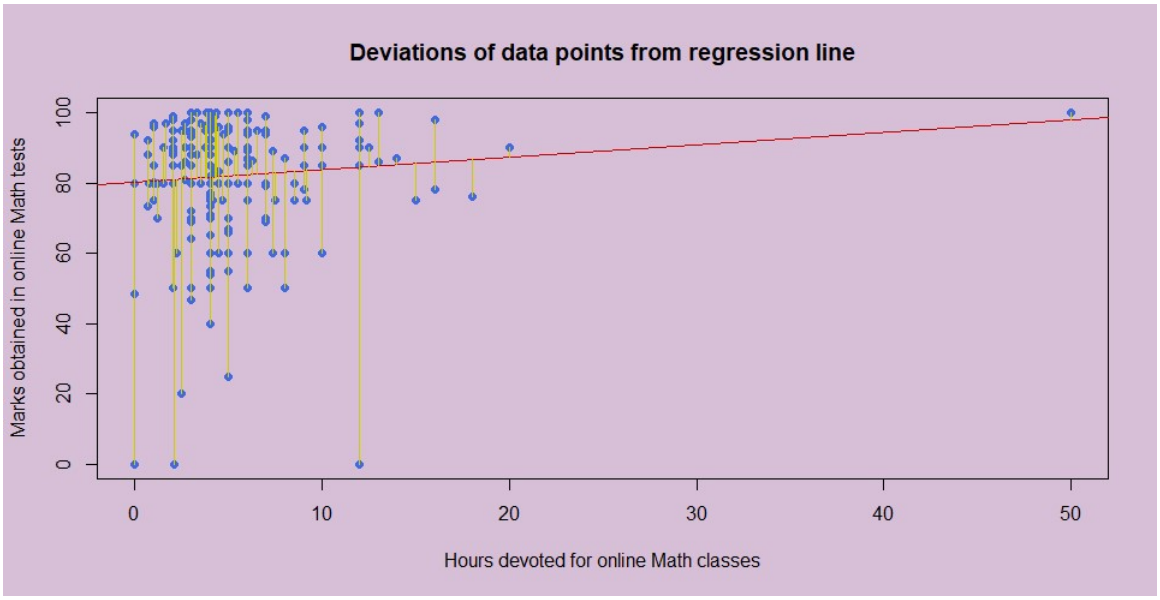
The value of r^2 for the above plot is 0.007513.
And the equation of the above line is given by:

$$y = 0.3582x + 80.1509$$

And therefore, this graph is not helpful in predicting the marks based on the time spent for online math classes.
The above point can be easily explained by the fact that only 0.7513% of the variation of y is explained, and the r^2 value should nearly be 0.5 for the graph to be useful in predicting marks based on time spent for online classes.
This low r^2 value is due to the fact that range of marks obtained by students who devoted the same amount of time for those classes varies from 20 to 100, or even from 0 to 90+ scores in some

cases. These wide variations result in the low r^2 value.

We can visualise this better in the following graph:



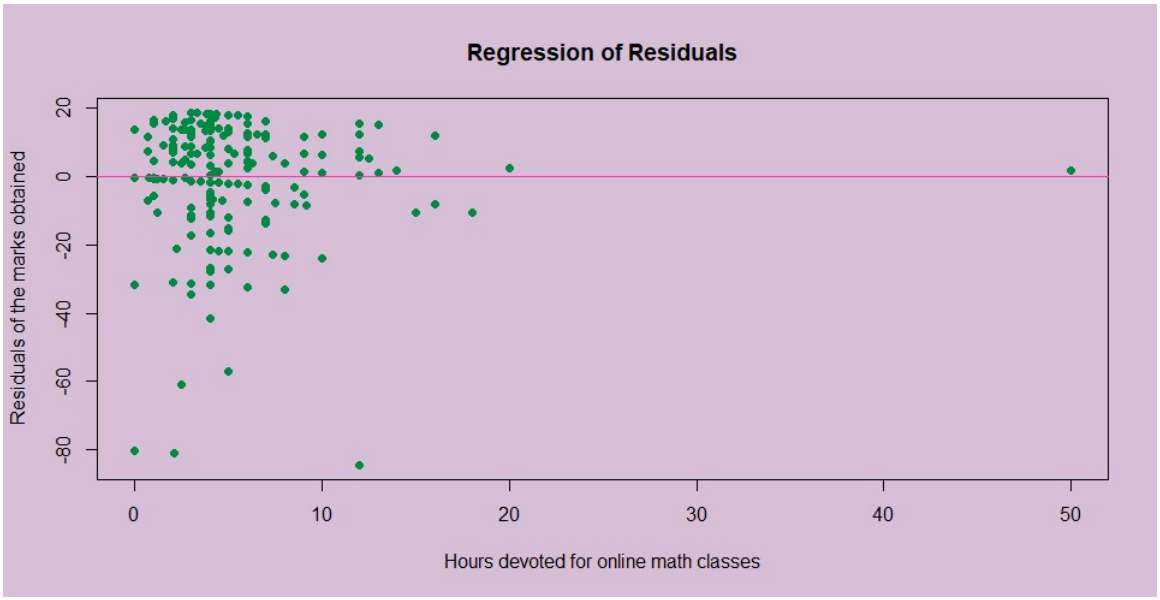
Note that in the above plot, the line represents the least squares linear regression line. And thus, the y-value gives the actual marks of a particular student, while the y-value, for the same x, on the line, gives the predicted marks based on the data.

In the above graph, we observe that most line are moderately or extremely long. This, as mentioned earlier, is due to the fact that most observations belong to same class interval (i.e., lie within the range 0 to 10 hours) but largely vary in the marks obtained.

Due to the lengthy segments, which result in the low r^2 value, we can be assured that this plot cannot be useful to predict the online marks based on the time devoted for online classes, since the predicted values do not give good proper approximations of the actual marks obtained.

Residual Plots

The residual plot for the above data is plotted by the (x_i, e_i) points, and is as shown below:



Notice that the regression line of the residual plot is $y = 0$, as it should be, in general. Also notice that the line clearly passes through at least 2 of the data points.

The points might seem to be very close to the regression line, but if observed properly, the residuals of the marks obtained vary largely and thus the graph had to be compressed for complete representation.

Here, we note that the number of points above the regression line might be slightly more than the number of points below it, but the distance of the points below the line seem to be much greater than the points above it.

3.6.3 The Marks obtained in Language in Online versus Offline Examination

Assuming that an offline examination correctly predicts the merit of the student, we want to see if the marks obtained in online examination correlates with his/her caliber. Thus we have 'Offline marks obtained in language' as our explanatory variable and 'Online marks obtained in language' as our response variable.

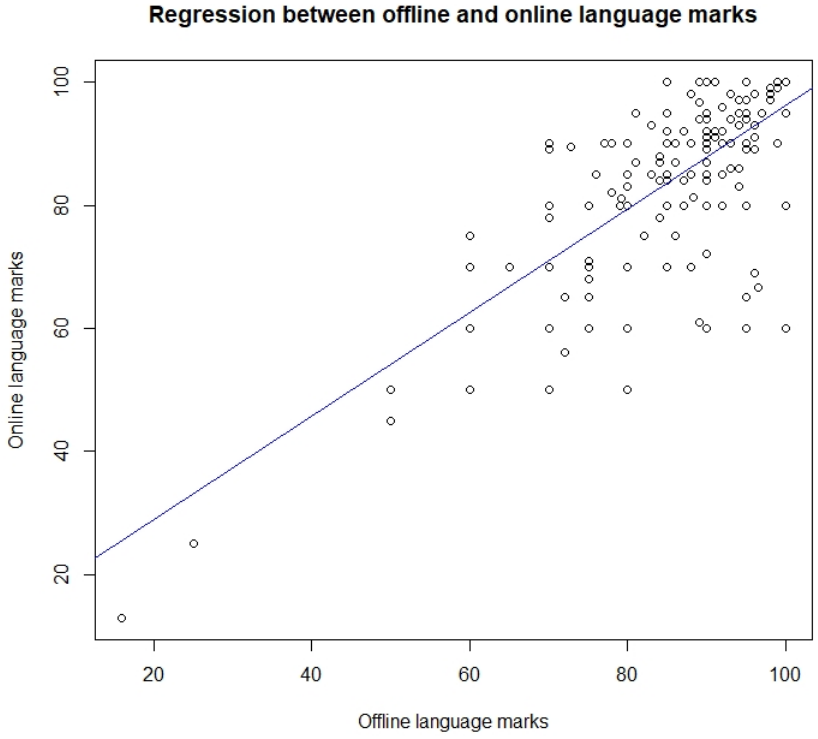
Hypothesis: There is a linear relationship between the online marks and offline marks obtained. Let variable y_i denote Online marks obtained in language by i^{th} student and x_i denote Offline marks obtained in language by i^{th} student.

Note: Our X (or Y) variable obtain zero as a value because there are students who did not have an offline (or online) examination when the survey was taken. These data points do not provide adequate information, and thus, we have eliminated all the data points with zero value in either x or y variable. So, we finally work upon 197 data points.

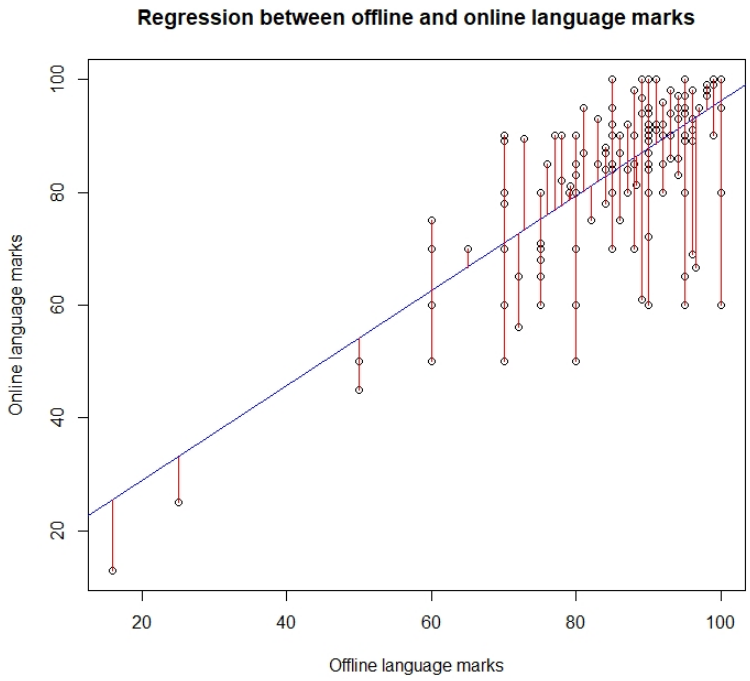
The equation of the least squared linear regression line :

$$y = 9.0265 + 0.8093x$$

The squared correlation coefficient, i.e, $r^2 = 0.4312$. We have a moderately high r^2 which indicates that more than 43% of the variability in y can be explained by x. Since r^2 is moderately high, we can safely claim that our linear regression model gives us a good fitted line. Thus our initial hypothesis is true.



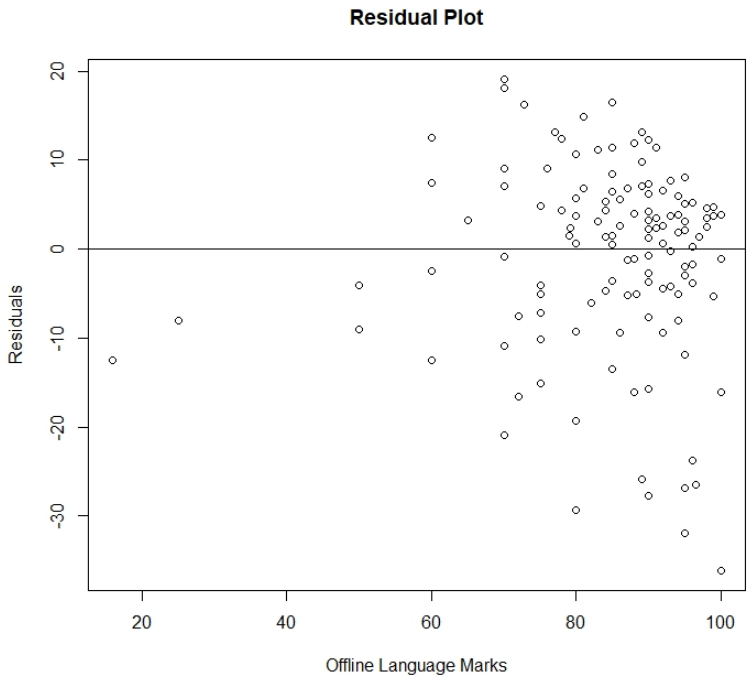
To verify further, let us look at the deviations of the predicted values of y from the observed values of y.



We see that magnitude of most of the deviations are small. Also, the magnitude of negative deviations are comparatively greater than the magnitude of the positive deviations.

Residual plots

Now let us look at the residual plot :



From the above plot, we see that the residuals are randomly scattered. This proves that our

linear regression line is indeed a good fit. There are almost equal number of data points above and below the zero-line.
Few things that we observe from the above graphs are

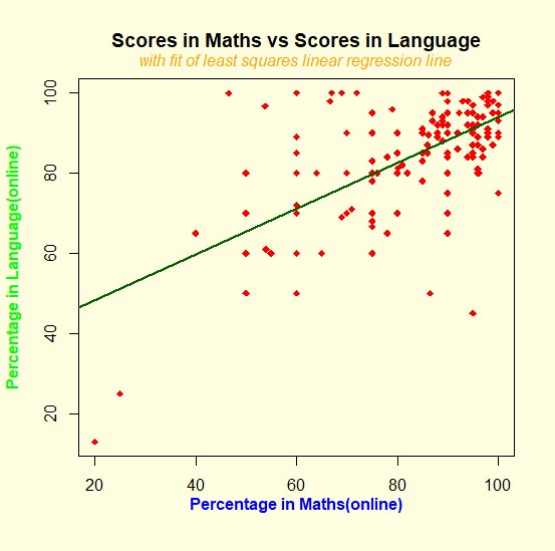
- Most of the data points are clustered around the region [60,100] in x-axis and [40,100] in y-axis.
- Considering our regression line is a good fit, the y-axis intercept is 9.02 which is positive. Thus initially, a student gets more marks in an online examination compared to the offline examination. But the slope is 0.8, thus the marks obtained in online examination gradually decreases with respect to the marks obtained in offline examination.
- The deviations in the residual plot shows us that the magnitude of negative errors are comparatively greater than the magnitude of the positive errors. But majority of the residuals range from [-10,10].
- Overall, the marks awarded to the students in an online examination does not differ largely from its offline counterpart.

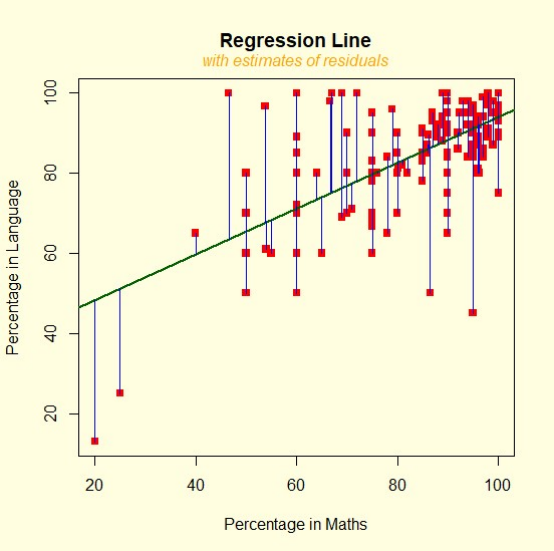
3.6.4 Marks in Language and marks in Mathematics (Online Examination)

Let us consider marks in Math to be the explanatory variable (say x) and marks in Language to be the response variable (say y).
The value of r^2 in the above plot is 0.389. Here 38.9% of the variability in percentage marks in Language is explained by the linear regression line of y on x. The least squares linear regression line gives a moderate fit to the data.
The equation of the least squares linear regression line is:

$$y = 0.5715516x + 36.9029431$$

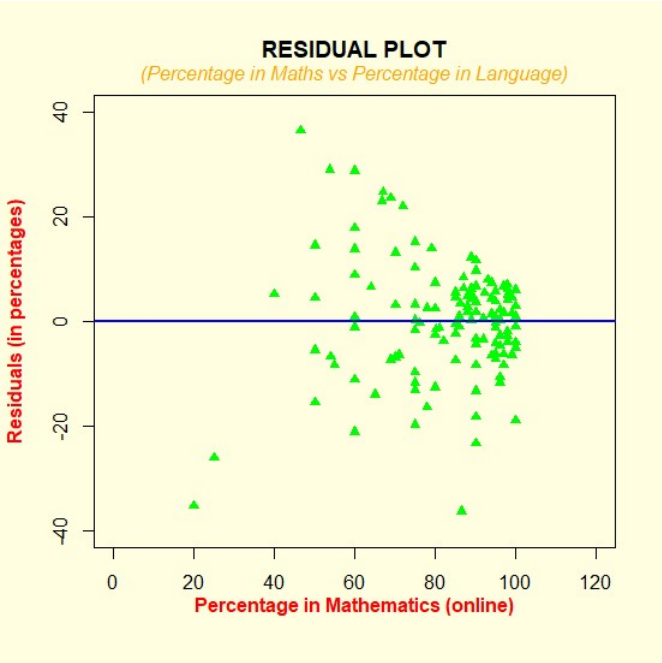
with $a = 36.9029431$ as the y-intercept and $b = 0.5715516$ as the regression coefficient of y on x.





Residual Plot

Now we plot the residuals against the explanatory variable ,i.e., marks in Mathematics.



Comments: From the residual plot it can be concluded that value (absolute value) of residuals increases in the range 40 %-80 % . However in the range 80%-100% ,the residuals (absolute value) seem to be concentrated around 0-20%. On an average the (absolute) value of residuals decreases with increase in marks in Mathematics.

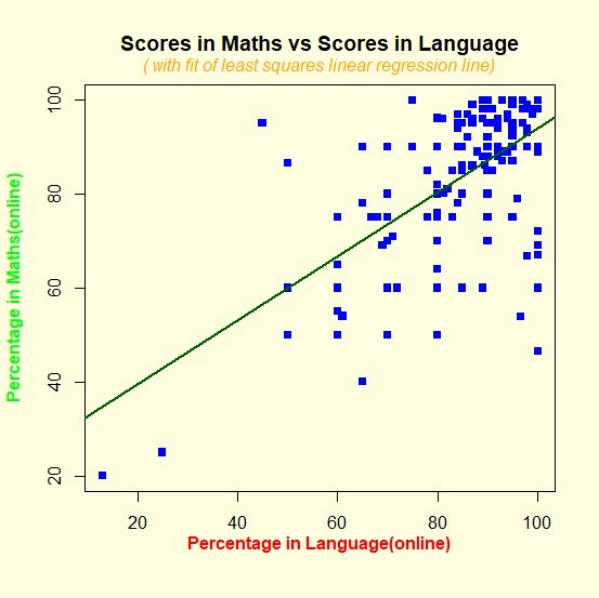
Thus the linear regression line gives a more accurate estimate of marks in Language when the marks in Mathematics are in the range (80%-100%).

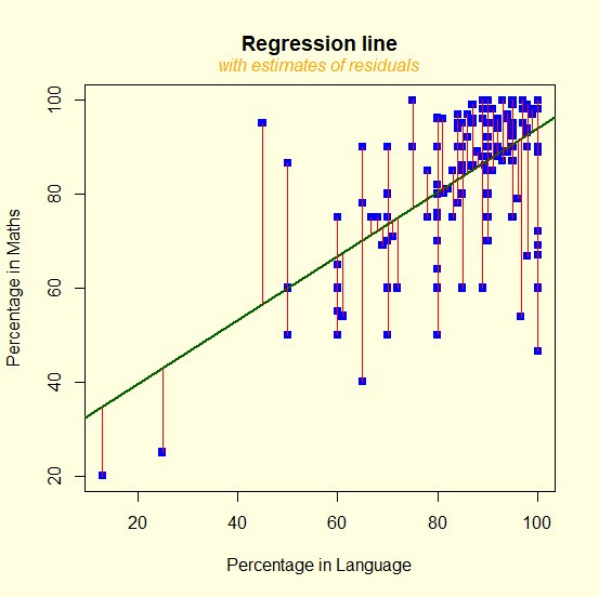
Now let us consider marks in Language to be the explanatory variable and marks in Math to be the response variable. Let x_i denote the marks in Language and y_i denote the marks in Mathematics for i^{th} student .

The value of r^2 is 0.389. Here 38.9% of the variability in percentage marks in mathematics is explained by the least squares linear regression linear on y on x. The value of r^2 being same in both the cases because of the symmetric nature of the correlation coefficient r.
The least squares linear regression line gives a moderate fit to the data.
The equation of the least squares regression line is:

$$x = 25.8603271 + 0.6806884y$$

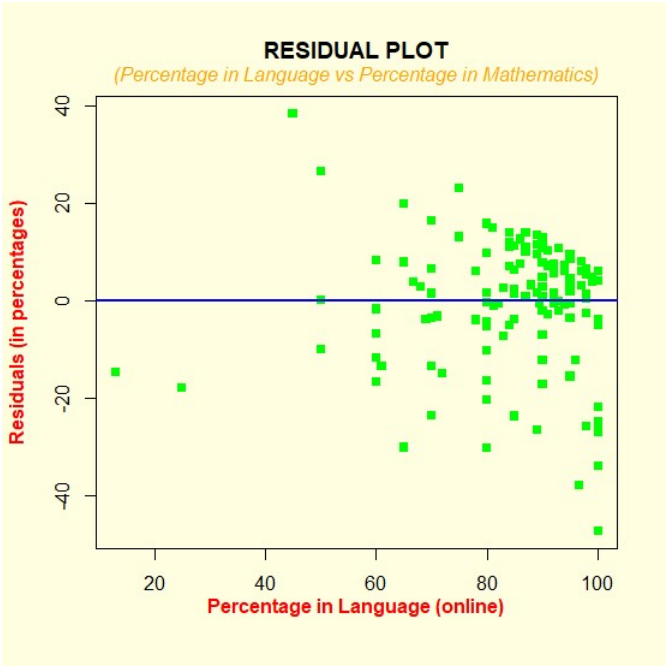
with $a = 25.8603271$ as the y-intercept and $b = 0.6806884$ as the regression coefficient of y on x.





Residual Plots

Now we plot the residuals against the explanatory variable ,i.e., marks in Language.



Here the value of residuals (absolute value) is quite a lot scattered. In the range 60%-100% (of scores in Language) ,the are many data points showing negative residual .The maximum value of residuals (absolute value) are be-tween 40-45%.

3.6.5 Marks obtained in Mathematics in Online versus Offline Examination

We take the 'Offline marks obtained in Mathematics' as our explanatory variable and 'Online marks obtained in Mathematics' as our response variable, as performance of a student in an online examination can be predicted from his/her performance in the offline examination.

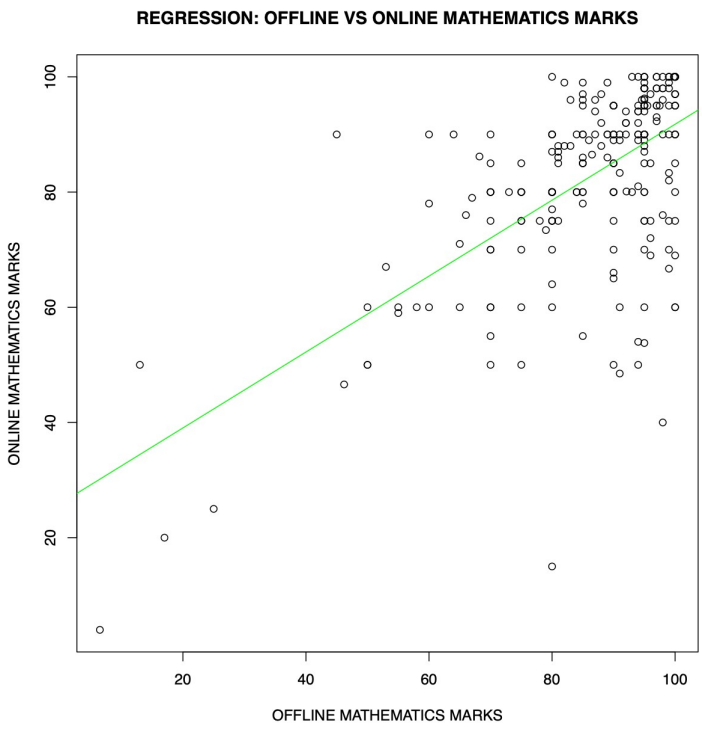
Hypothesis: There is a linear relationship between the online marks and offline marks obtained.

Note: When the survey was conducted, there where some students who had not attended an online or offline examination. Hence we got few 0 values in either of the two variables. These data points did not provide us any information in the study and thus, we delete these points.

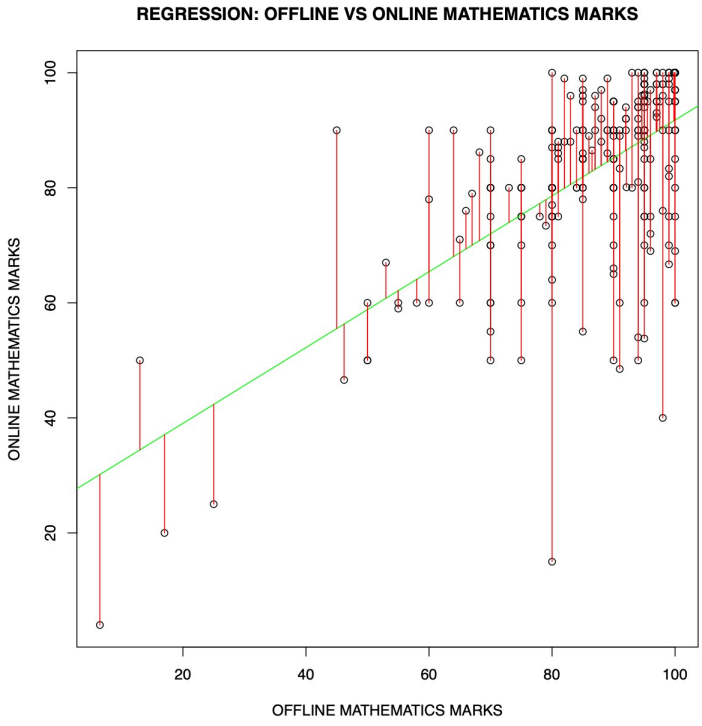
The equation of the least squared linear regression line :

$$y = 0.65915x + 25.86783$$

The squared correlation coefficient, i.e, $r^2 = 0.375$ We have a low r^2 value of 0.375 which indicates that 37.5% of the variability in y can be explained by x. Since the value of r^2 is low, we can say that our linear regression line is not a very good fitted line.



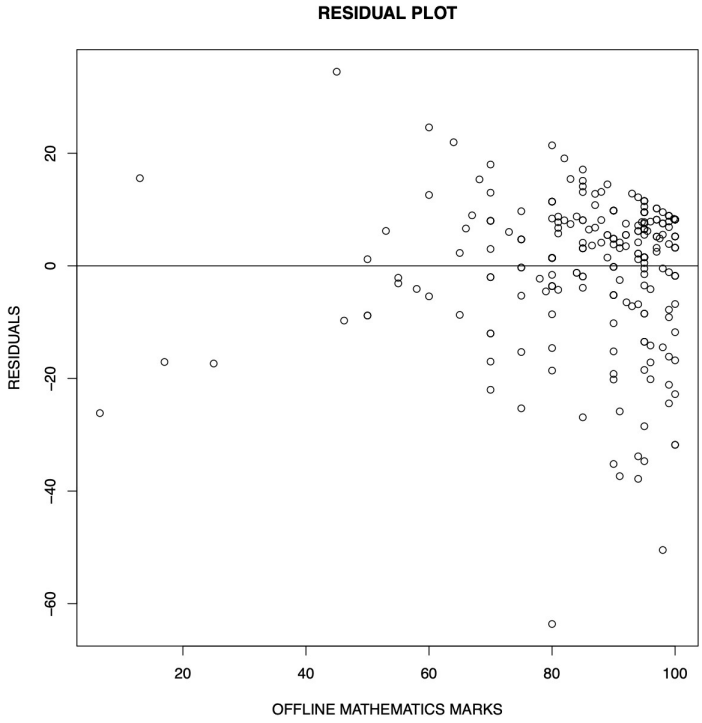
To verify further, let us look at the deviations of the predicted values of y from the observed values of y.



We see that magnitude of some of the deviations are quite large. For example: the data point (80,15) has a very large negative deviation.

Residual Plots

Now let us look at the residual plot :



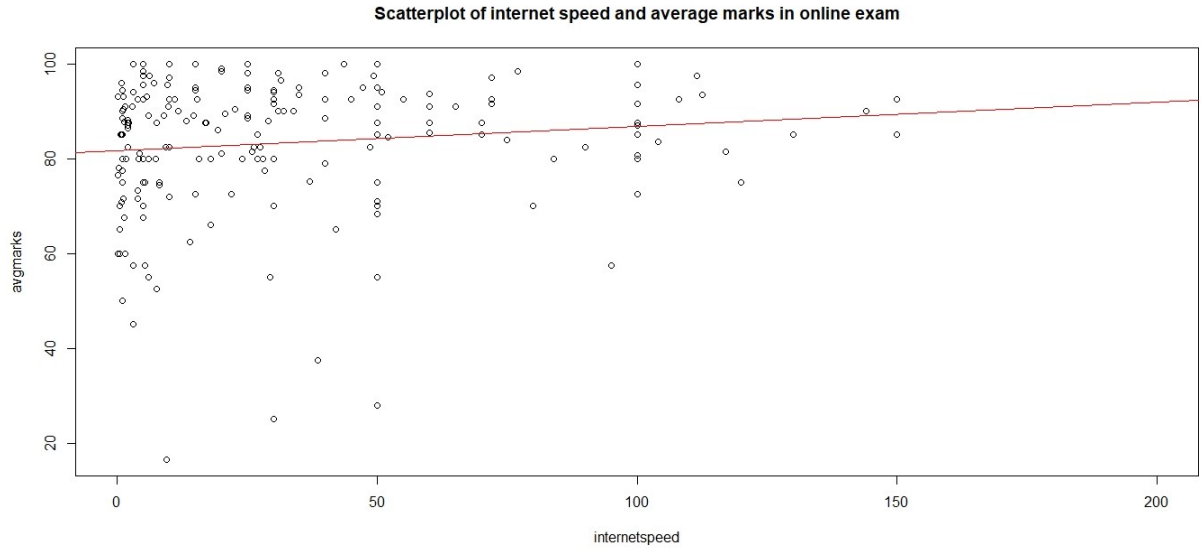
From the above plot, we see that the residuals are more or less randomly scattered. We see that there are relatively more points above the zero line as compared to below the zero line. We also notice that there are more data points in the region where the offline Mathematics marks is between 60 and 100.

3.6.6 Average marks in online exam & internet speed

Let us consider average marks in online exam to be the response variable and internet speed(in Mbps unit) to be the explanatory variable.

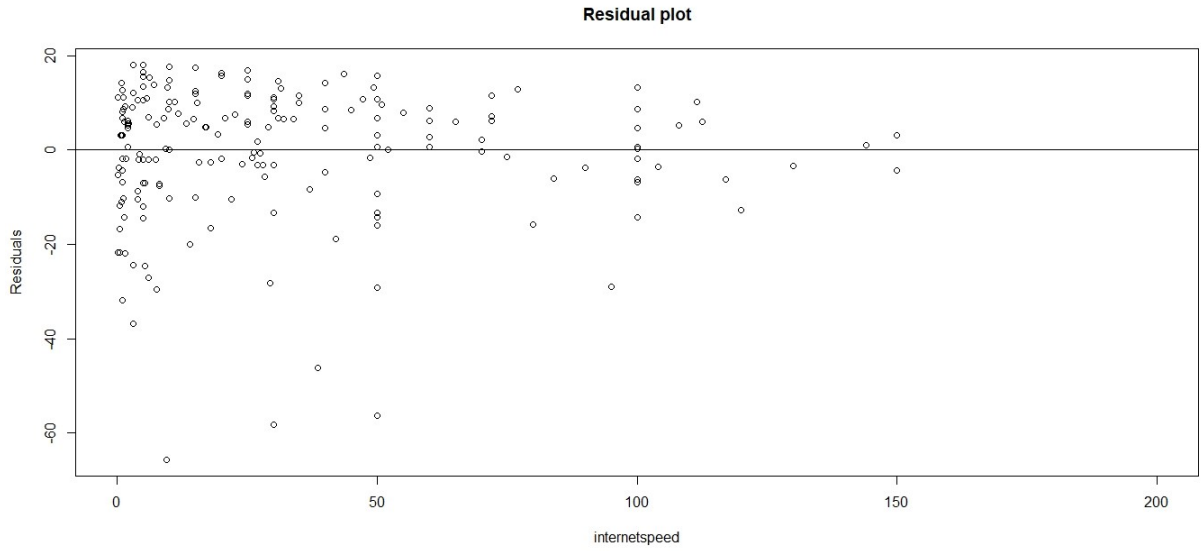
The value of r^2 is 0.01603. So 0.01603% values of y are explained by the values of x, so the least square linear regression line is not a good fit to the data. The equation of least square linear regression line of y on x is:

$$y = 0.05054x + 81.75431$$



Residual Plot

Now we plot the residuals against the explanatory variable, i.e., internet speed. We can see that there are many data points with large negative residuals. Also, variation of residuals decreases as X increases, so errors are not homoscedastic.



4 Conclusion

"One worthwhile task carried to a successful conclusion, is better than 50 half-finished tasks."

- B. C. Forbes

From the exploratory data analysis that was carried out, we reached several conclusions, the details whereof have been elaborated already. Moreover, explicitly classifying the types of conclusions that were reached is a rather Herculean task. So, we refrain from listing the already stated conclusions here once again. However, we raise a pertinent question at this point.

Can offline classes be permanently replaced?

Some may argue that pedagogy is far more fervent when physical. That is not a pedantic perspective though, but it still is inconclusive if online classes are in the interests of both the teachers and the students.

On the other hand, whether offline classes can be replaced entirely is a question that remains yet to be answered. From the data collected, it seems likely that most people would argue against it. However, several instances of online classes are already in practice, such as distance learning programmes provided by various overseas universities. To reach an accurate conclusion that resonates with the masses, a survey of a much larger magnitude involving a much wider demographic is needed. It might be challenging, but it remains entirely within the realms of possibility. As is evident, even elementary statistics supplies more than enough input for us to make a highly accurate educated guess. It is only a matter of time before the same statistics leads us to the answers we so desperately seek.