Viz.It Report

A Comprehensive Report analysing the Student Academic Dataset.

“Data is a precious thing and will last longer than the systems themselves.

If you think about it carefully; what we remember, what we think, what we see, is all Data”

01

Introduction

A basic understanding of the given dataset.

02

Visualizations

Analysing different factors affecting the output using Visualizations.

03

Inference

Understanding visualisations and drawing inferences, Final Report.

04

End Report

Credits and Report close

01. Introduction

About Viz.It

A flagship project by **Consulting & Analytics Club of IIT Guwahati (India)** to understand the relation between the grades of a first year student at IIT Guwahati with their previous background and their activities at the campus.

Student Academic Dataset.

About the dataset:

- This dataset consists of data of almost 400 students at IITG
- Data is collected without compromising any piece of information which could be helpful in better understanding the factors affecting student grades
- This data was collected via door-to-door survey collection in all hostels by representatives of C&A club evenly
- An online survey was avoided to have no fake information to be input by the students

The anonymity and authenticity of each input was taken care of

More about the Dataset

This dataset consists of various collections of data such as:

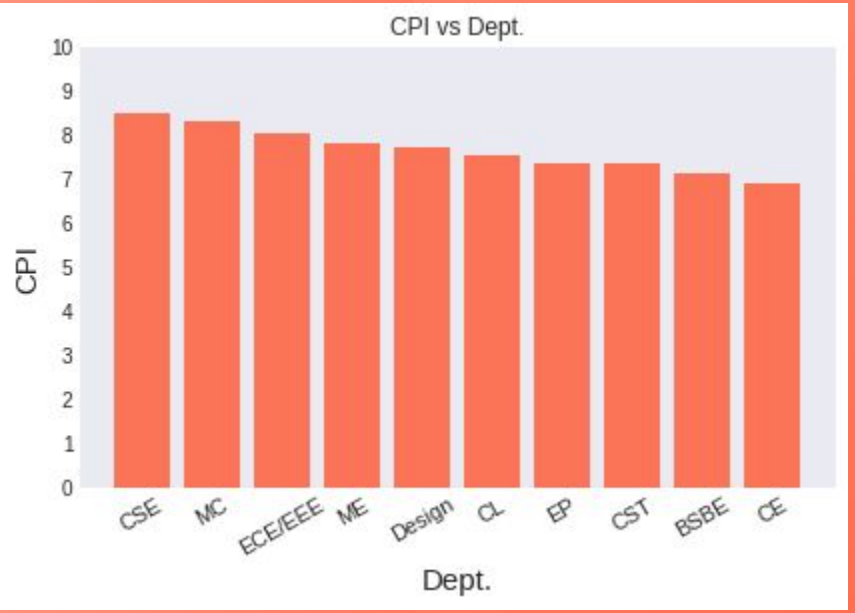
- The attendance percentage category of the student
- The relationship status of the student
- Average study time of students at campus
- How's the relations of student with his/her Teaching Assistant, etc.

Primary aim of this project:

This project is to understand the relation between the grades of a first year student at IIT Guwahati with their previous background and their activities at the campus.
Hence this project is aimed to help the next batch of incoming freshers to better prioritize their activities at campus and analyse trends based on inputs too.

02. Visualisations

Dept. vs CPI



This bar graph shows the arithmetic average of CPI corresponding to each Dept.

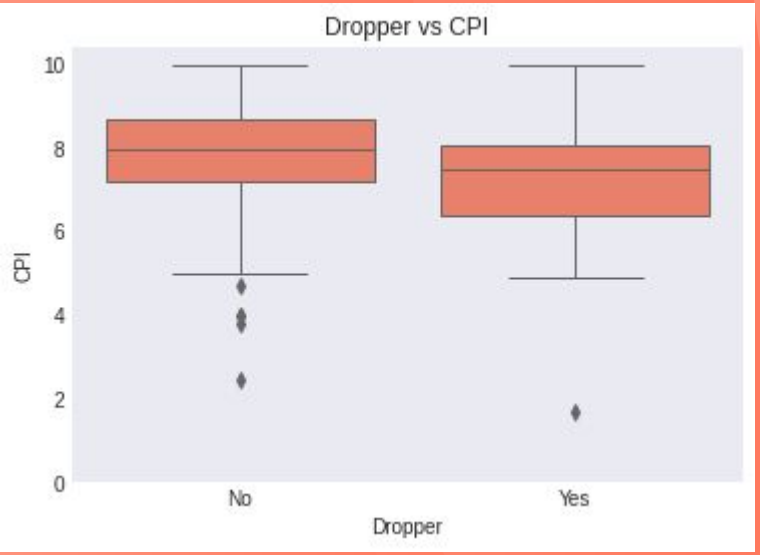
From the graph, we see that the average CPI of CSE Dept. stands the highest with a value of 8.5, while the average value of Civil Dept. is around 6.91.

The spread in CPI values is almost the same in every Dept.

Below given are the maximum values of CPI with respect to each Dept.

CSE 10.0	MC 10.0	ECE/EEE 10.0	ME 10.0	Des. 9.1
CL 9.6	EP 9.1	CST 9.49	BSBE 9.28	CE 9.23

Dropper vs CPI

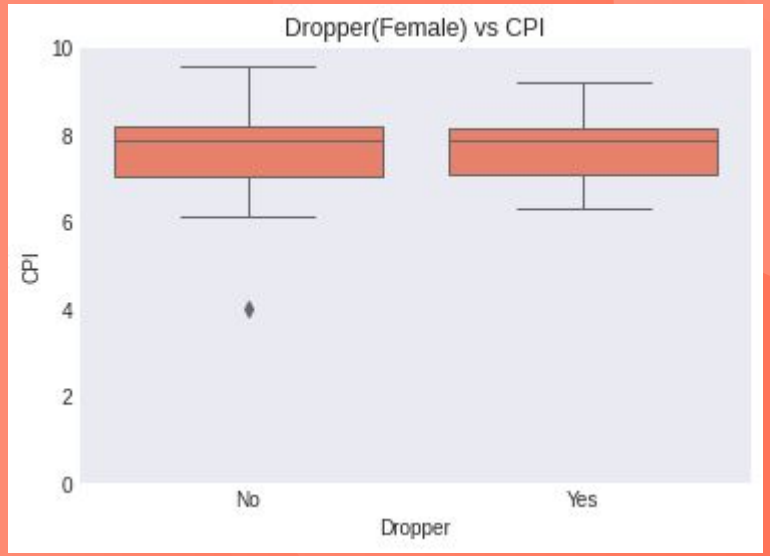


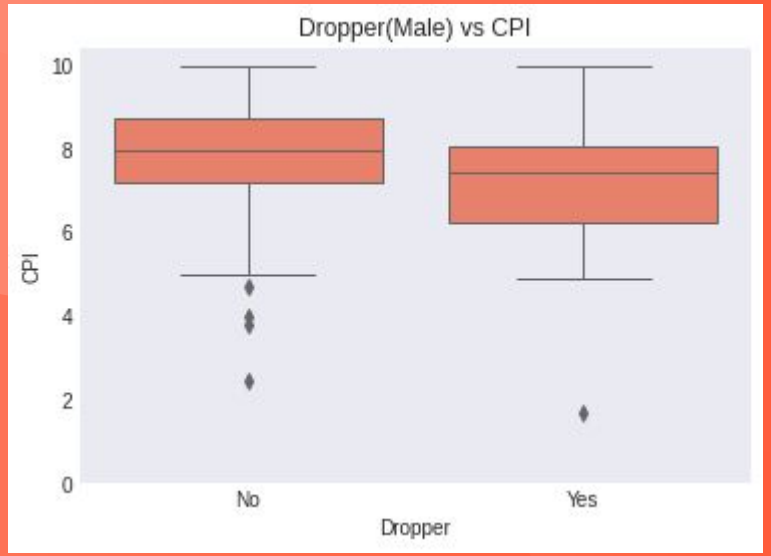
Yes	10.0
No	10.0

- The spread in values of CPI in both ‘Yes’ and ‘No’ is almost same, while it’s seen that a person who is not a dropper tends to have a better CPI.
- From the plot, it is seen that more than **Two-third** among non-droppers score more than **Half** of students who are droppers.
- Plotting similar graphs for Male and Female, the graph looks similar wrt Male, while the Dropper effect is negligible on Females
- On the bottom left, is the maximum values of CPI in each category (i.e. ‘Yes’ and ‘No’)

*Droppers are those students who took an additional year in JEE preparation before getting into the college.

Dropper vs CPI (cont'd)





For Female Students

For Male Students

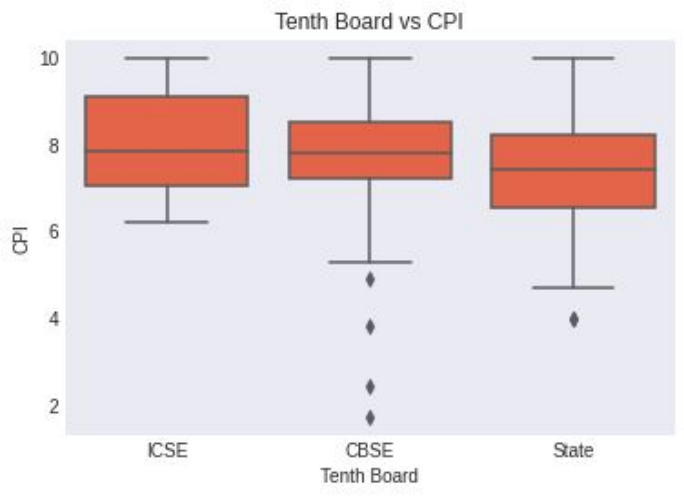
Surprisingly, being a dropper, for female students has almost no effect to a mild negative effect as compared to total population plot(in previous slide). The reason for this is not being reflected in the previous slide and the strong similarity between being a dropper(male) and being a dropper in general can be attributed to the high Male population at the college.

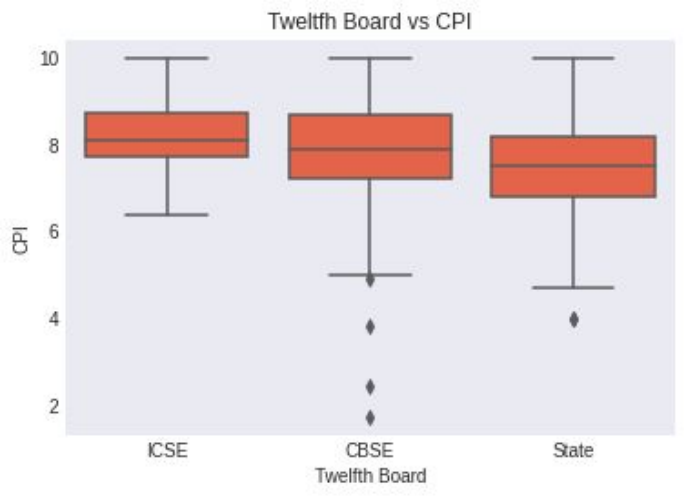
Parent's occupation vs CPI

Father's occupation	CPI (Avg)	CPI (Max)
Govt	7.58	10.0
Technical	7.68	10.0
Business	7.80	9.93
Management & Finance	8.07	9.93
Medical	8.25	9.93
Else	7.40	10.0

- Student's CPI is **almost equal** with respect to different **mother's occupations**.
- It's interesting how Father's occupation has an effect on student's CPI.
- Almost **75%** of students whose fathers are from medical background score more than **50%** of students whose fathers work as Govt. employees.

Boards vs CPI





Avg CPI	ICSE	CBSE	State
Tenth	8.03	7.77	7.36
Twelfth	8.15	7.81	7.49

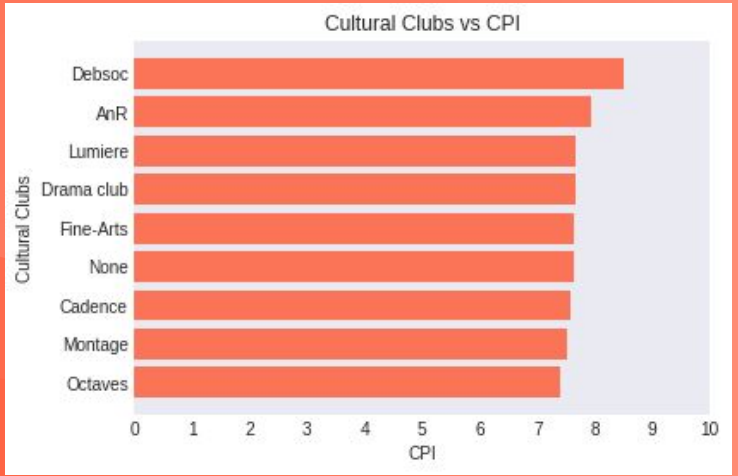
Both of the above plots follow a similar trend, where students of **ICSE** have a **higher chance of scoring high**, then CBSE and then State. These are just trends, moreover the max value in each category is **10.0**

Clubs vs CPI

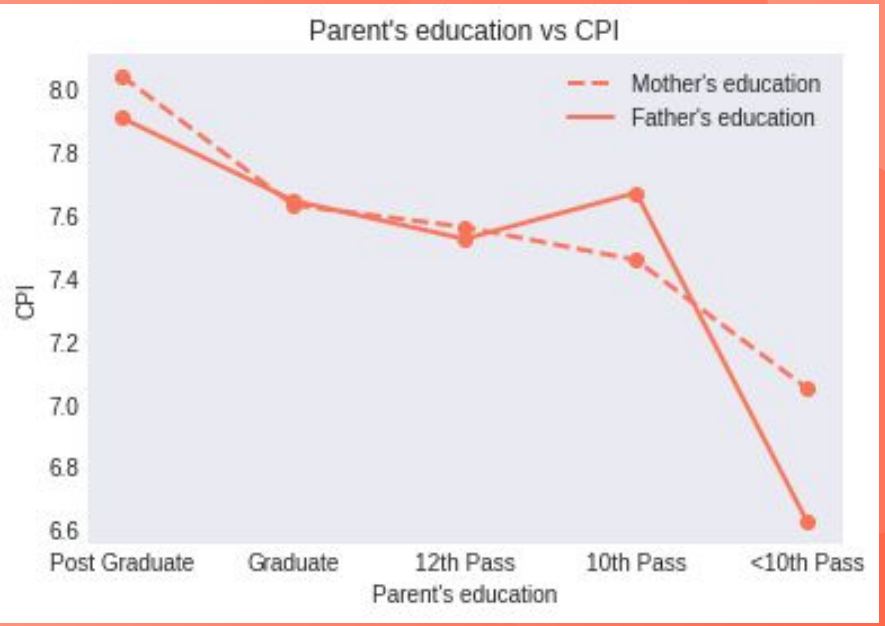
These plots debunk the myth that taking part in a club affects students' grades, as the average CPI of students belonging to many clubs is more than the average CPI of students who are not a part of any club.

Moreover, the highest CPI among students in almost every club is well over **9.0**.

*CPI of students belonging in multiple clubs are taken in every club that the student belongs to

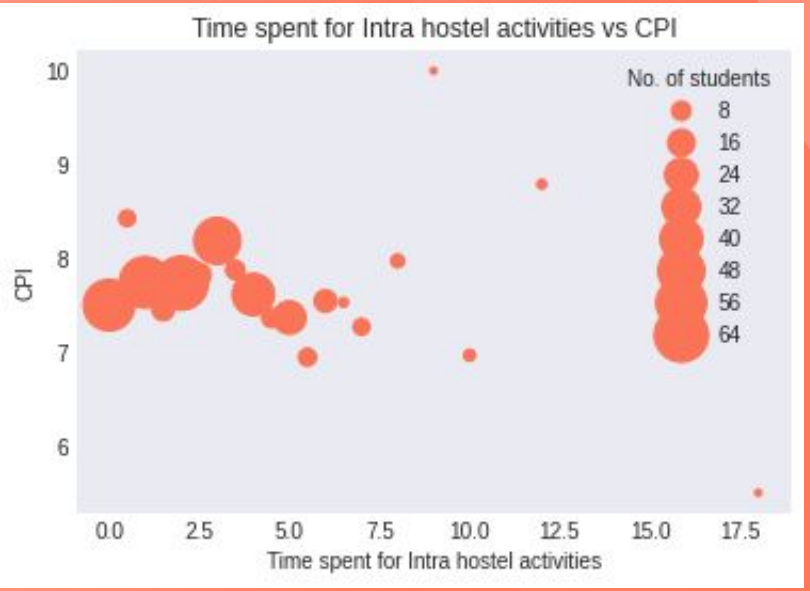


Parent’s education vs CPI (Student)



- Points on the plot correspond to average values in each category.
- Graph shows that, if parents are **more educated**, students are likely to have a **higher CPI**.
- Of course, this is just a trend and not to be mistaken with a general statement.

Time spent on intra hostel activities vs CPI



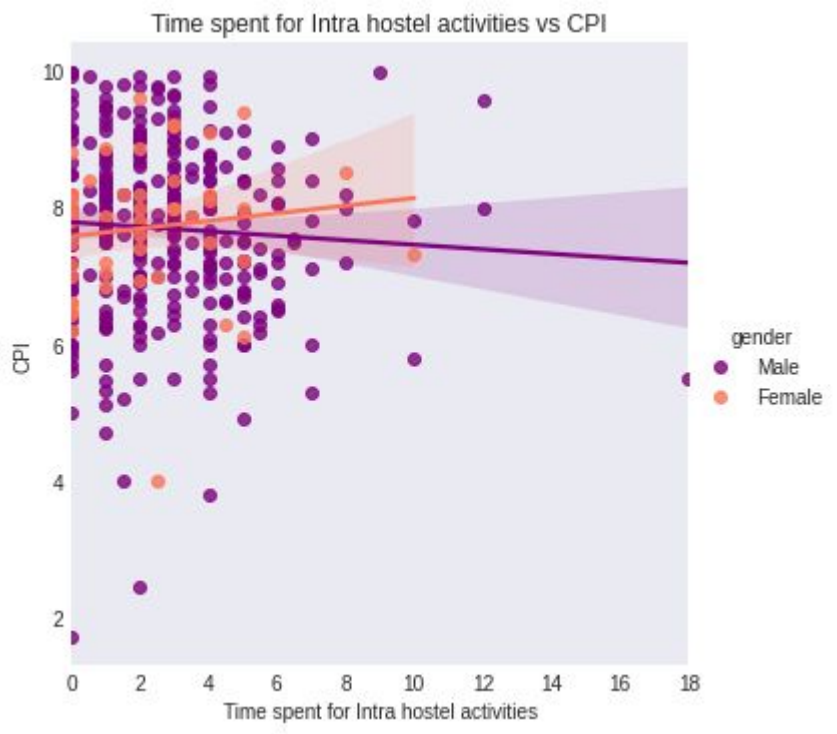
H_0 = There is no statistically significant relationship between time spent on intra hostel activities and CPI.
H_1 = There is a statistically significant relationship between time spent on intra hostel activities and CPI.

Significance Level(alpha) = 0.05
P-Value = 0.43
P-Value > alpha = 0.05
Hence, we <u>can not</u> reject H_0

R squared = 0.001 (Not significant, i.e. almost no correlation)

Though this visualization isn't significant enough, we get an idea on how CPI might change wrt time spent on intra hostel activities.
We understand that CPI doesn't really get affected with time spent on intra hostel activities contrary to what we've been hearing.

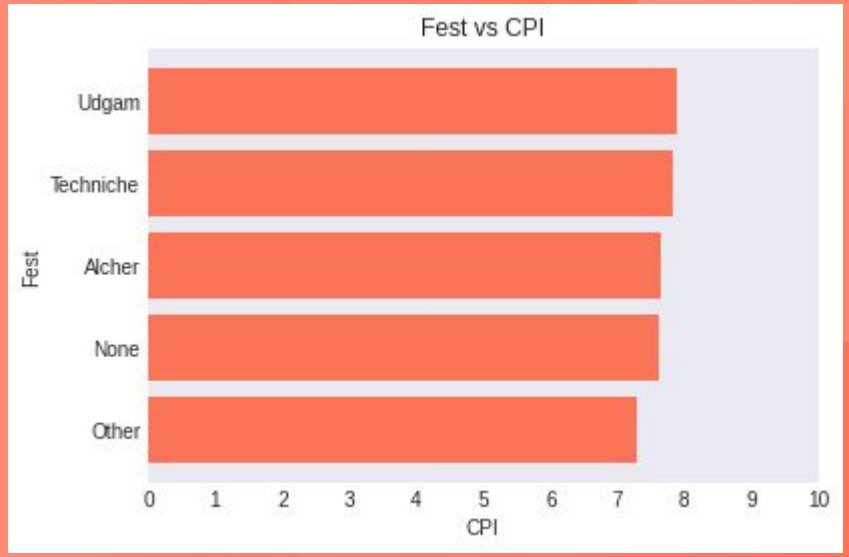
Time spent on intra hostel activities vs CPI (cont'd)



It is seen that, Males who spend more time on intra hostel activities tend to have a lower grade while Females who spend more time on intra hostel activities tend to have a higher grade(though a negligible effect).
The Scattered plot speaks more than the lines in the plot, it is seen that the data is almost evenly scattered, making almost no effect.

*Again, this visualisation doesn't help us in predictions (as statistical values like P-Value and R squared etc. suggest). This is just for understanding the trends this year.

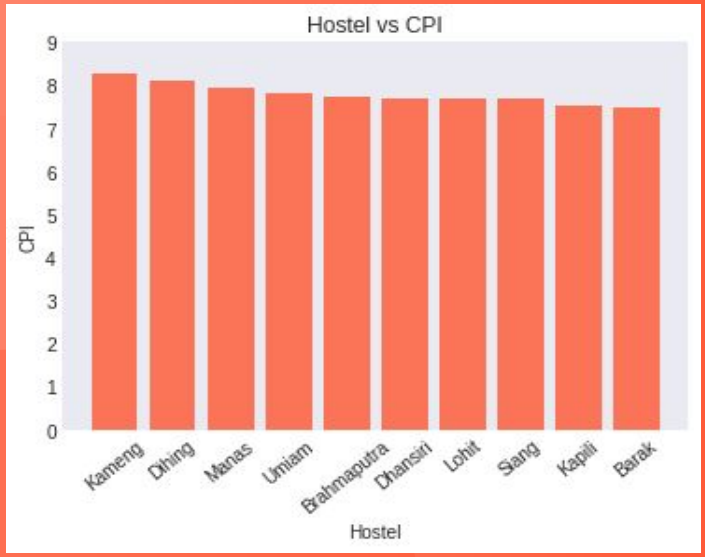
❖ Fest vs CPI



This debunks another myth which says, people working in Fests end up having a lesser CPI.

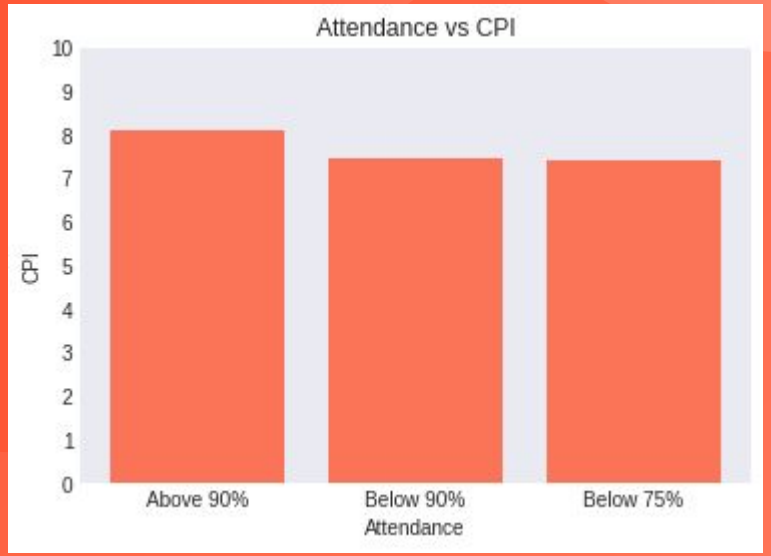
*CPI of students belonging in multiple Fests are taken in every Fest that the student belongs to

❖ Hostel vs CPI



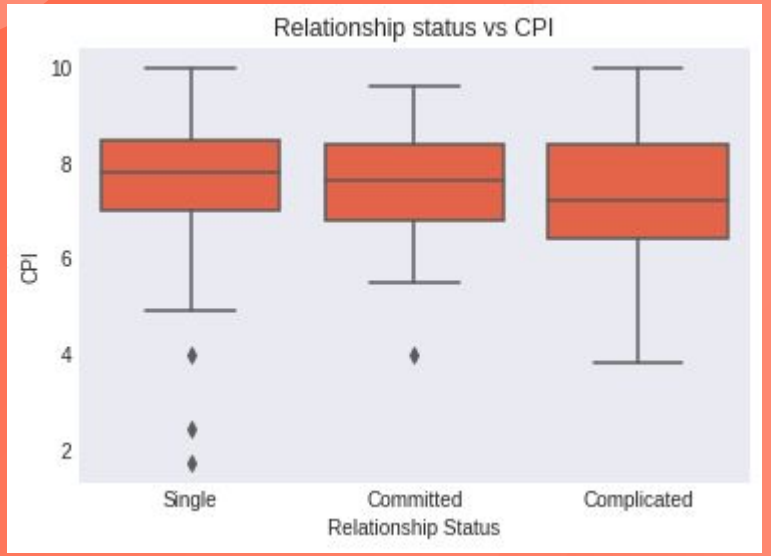
The above plot shows that students of Kameng hostel have the best average in CPI among all Hostels.
While, the maximum values of CPI of each hostel are close to each other.

Attendance vs CPI



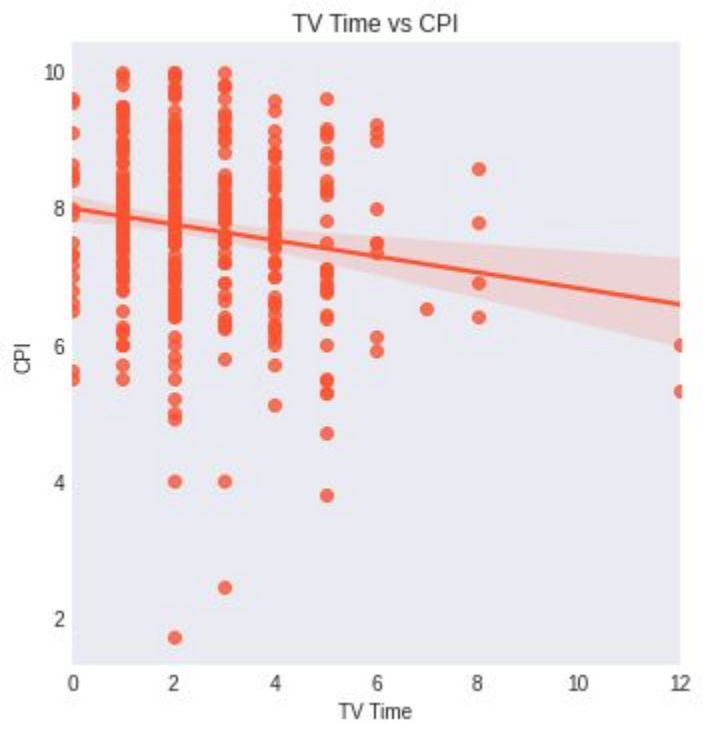
Students who maintain attendance greater than 90% tend to have a better grade than others by a significant amount.

Relationship Status vs CPI



Students whose relationship status is ‘single’, tend to have better grades (though the tendency is minute as there’s no significant difference).

TV Time vs CPI



H_0 = There is no statistically significant relationship between TV Time and CPI.
H_1 = There is a statistically significant relationship between TV Time and CPI.

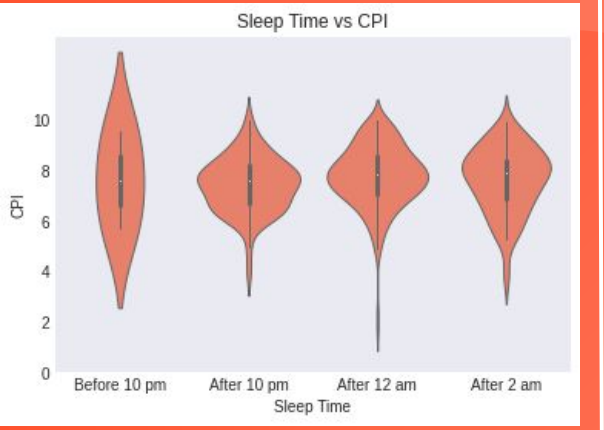
Significance Level(alpha) = 0.05
Observed P-Value = 0.001
Observed P-Value < 0.05

We reject the Null Hypothesis in the favor of Alternative Hypothesis

RMSE = 1.22
This shows that the values are widely scattered.
Still, there is a relation between TV Time and CPI

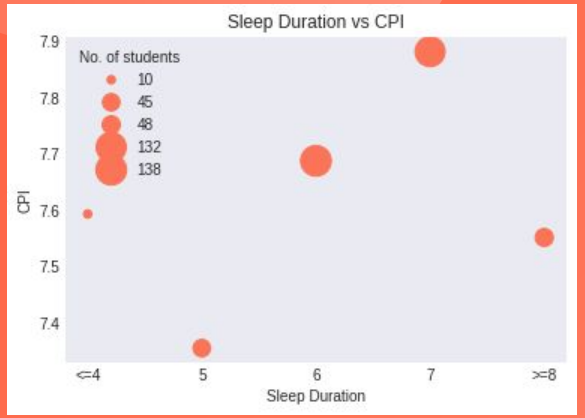
* TV Time = The average time spent in gaming/TV series/movies in rooms daily(in hours)

Sleep Time vs CPI



The values are almost similar with a negligible increase as we move right.
Hence, a violin plot helps us to better understand the data distribution.

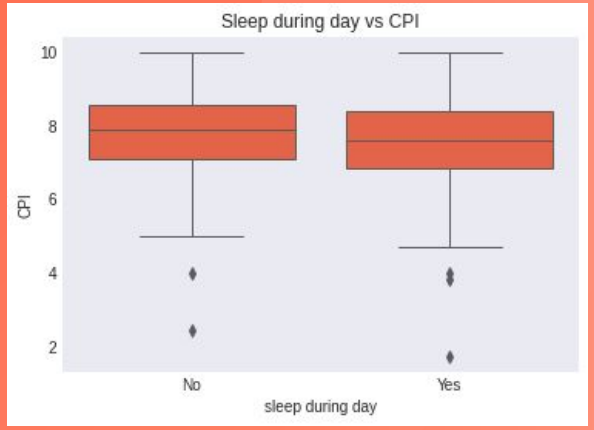
Sleep Duration vs CPI



Interestingly, the above plot shows us a positive correlation as we move right. Though this correlation is statistically insignificant, it gives us a basic idea on how the plot changes wrt sleep duration.

*bubbles correspond to average CPI values.

Sleep during Day vs CPI



The above plot shows that, people who don't sleep during day tend to score better grades.

Avg CPI for 'Yes' = 7.53
Avg CPI for 'No' = 7.81

CPI vs

- **TA Relation : How's the relations of student with his/her Teaching Assistant (1- bad, 5-good)**
- **Lab: Labs taken seriously (1-bad, 5-good)**
- **Tut : Tutorials taken seriously (1-bad, 5-good)**
- **Quiz : Quizzes taken seriously (1-bad, 5-good)**
- **Midsem : Midsem taken seriously (1-bad, 5-good)**
- **Endsem : Endsem taken seriously (1-bad, 5-good)**
- **Branch Change : How serious is the student to change their stream? (1-bad, 5-good)**
- **Attention in class : Attention in classes (1-bad, 5-good)**

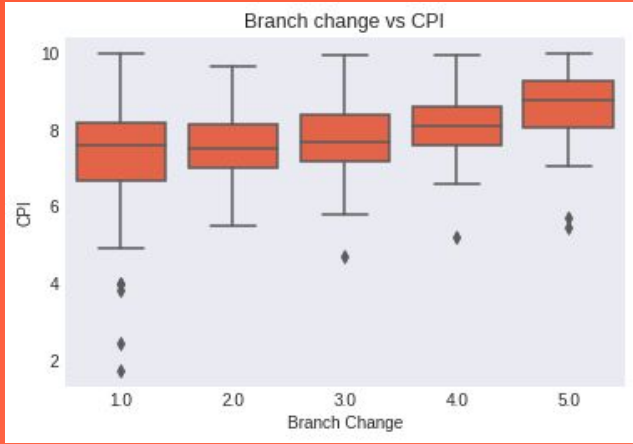
**correlation doesn't necessarily imply causation, we need to have more data and deeper insights to confirm it as a causation.

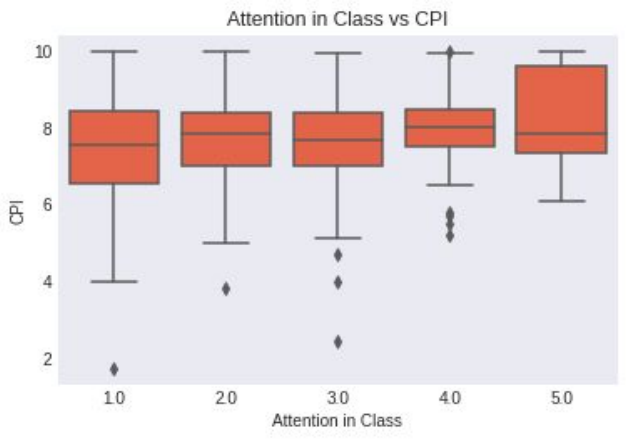
All these come under ordinal variables since we can never know if the difference between 5(good) and 4 is same as difference between 4 and 3.
For instance, let's assume that the difference between any two consecutive numbers is equal on these scales. Then;
There would be these moderately correlated values with CPI:

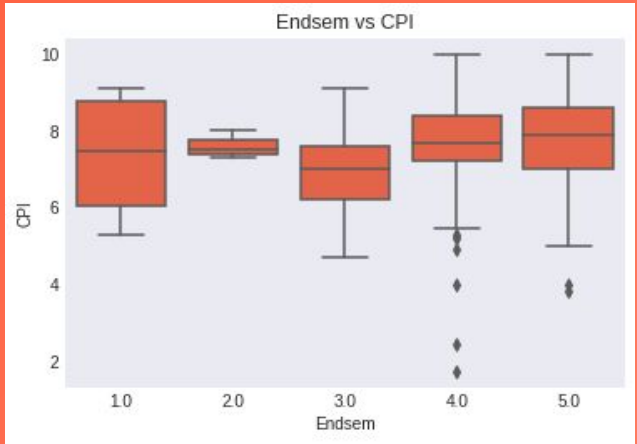
Factors	Correlation const	P-Value
Branch change	0.285149	2.07 * 10^-8
Midsem	0.248030	1.23 * 10^-6
Quiz	0.216345	2.50 * 10^-5
Endsem	0.182151	0.0004
Tut	0.168593	0.001
Lab	0.142047	0.006
Attention in Class	0.133687	0.009

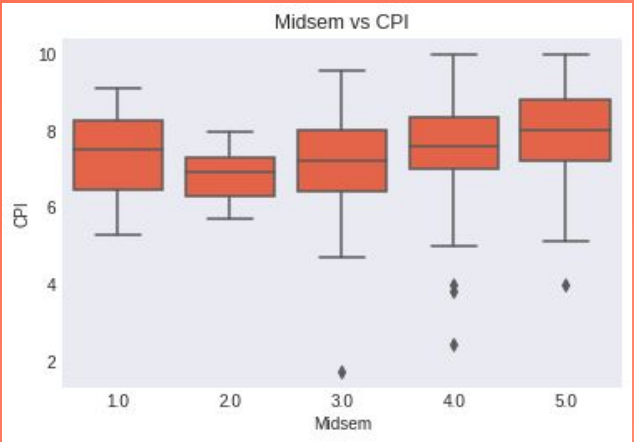
A positive correlation is observed

Null hypothesis can be rejected in all of the above cases and we can see a mild correlation between these variables and CPI.
Now, what if the difference between any two consecutive numbers on these scales is not equal?
Then, the statistical data we calculated would go wrong but the trends would remain same. So, we get an idea about the factors affecting CPI the most.

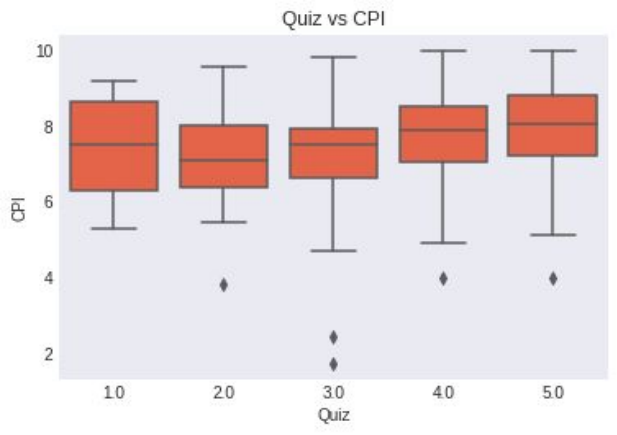


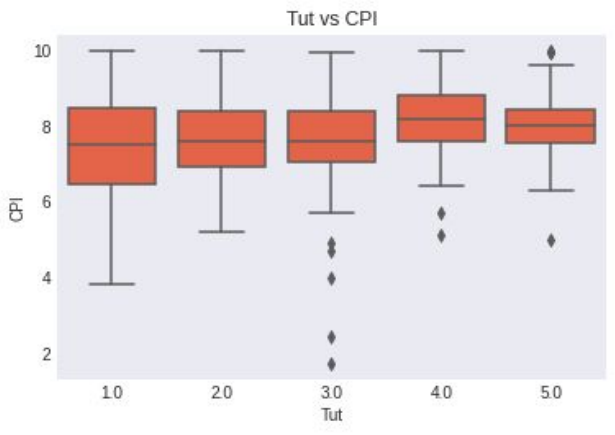


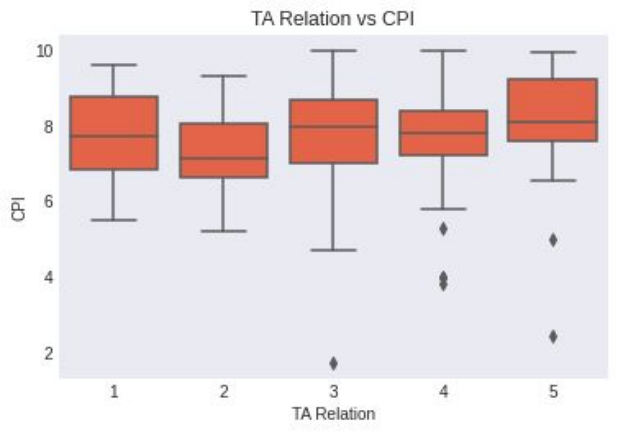


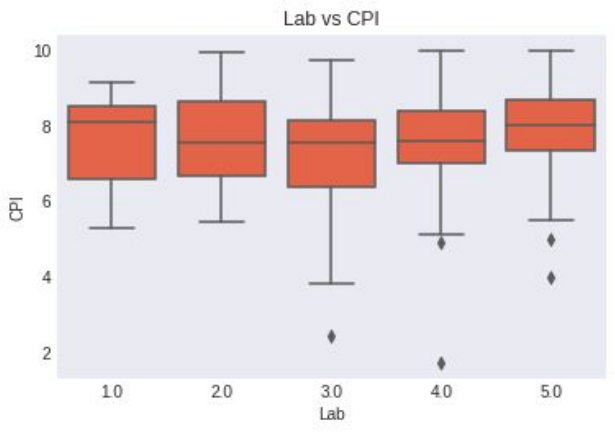


This is more or less what we expected, higher values as we go to the right.









Almost similar trends with a little variations.

03. Drawing Inferences

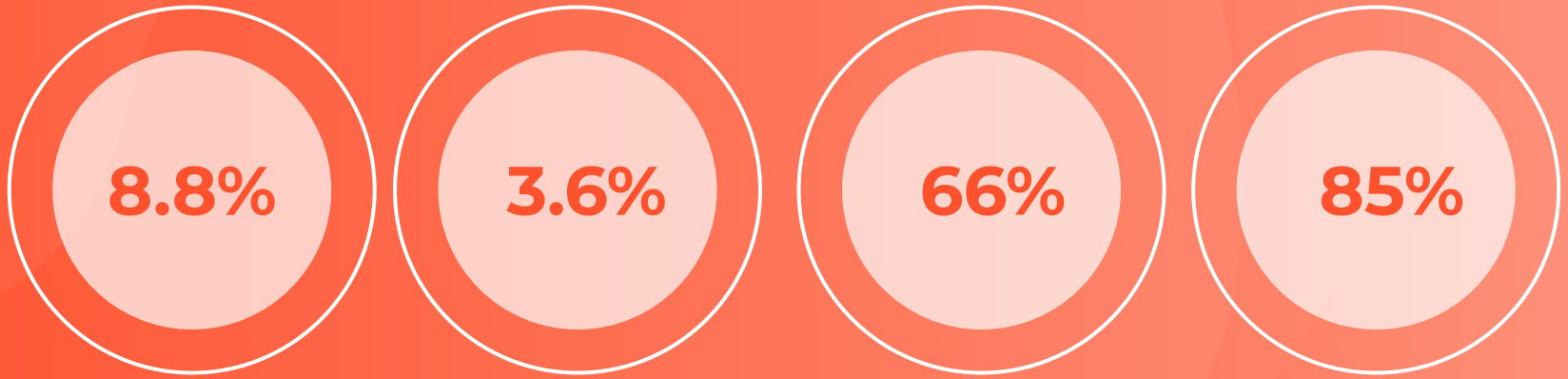
Before moving ahead.

In the visualisations involving past data or data which you didn't get to choose(dropper, parents occupation etc.), the maximum values of CPI of respective categories are put up so as to remind you the fact that trends are not always true for everyone.

It is to show that, no matter what trends suggest, you've every chance to defy them and score high

Students who doesn't sleep in the day, at an average, scores 3.6% more than a student who sleeps in the day

An average student, who has a branch change interest of 5 scores more than almost 85% of students who do not have a branch change interest of 5



Students who has attendance over 90%, at an average scores 8.8% more than who has attendance less than 90%

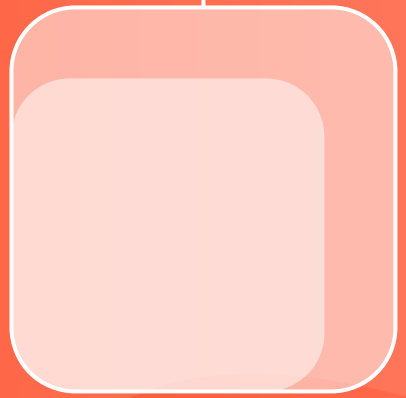
An average student, who has 'midsem' Seriousness of 5 scores more than almost 66% of students who do not have 'Midsem' seriousness of 5. (surprisingly, 'endsem' seriousness doesn't give us an inference like this.)

Trends for things you don’t get to choose while in campus

16%

A student, whose parent is a postgraduate scores 16%(absolute value) more than a sudent whose parent is <10th pass

66%



More than 66% of non-droppers score more than half of the droppers

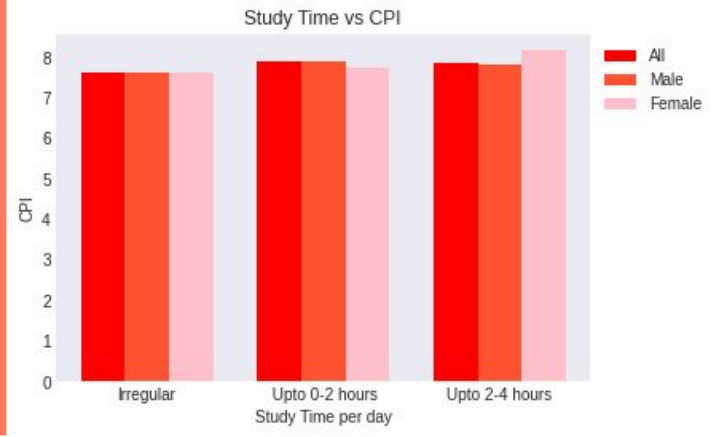
75%

More than 75% of students whose fathers are medical professionals score more than half of the students whose fathers are govt. employees

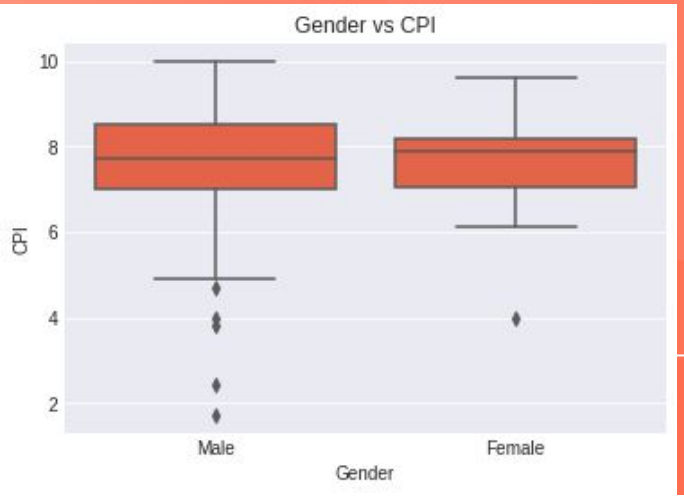
Other Important Inferences were more or less drawn near the visualisations itself.

Now, let us have a brief look on why few data visualisations were not included earlier

- Few factors like gender, study time, etc. have almost no effect on CPI.
- Also, Time spent outside, family size etc. gave an evenly scattered visualisation, which implies CPI might not be directly dependent on Family size, Time Spent Outside etc.



Over here, males and females have an opposite effect, making the total result not significant.



Almost similar plots. No significant information.

04. End Report

It’s such a great experience to work with a real dataset. It was really interesting as few visualisations were giving a significant inference and few weren’t. It opened up doors to experiment with the dataset, analyse a real world data to find interesting insights, learn more about data analytics and visualisations. I’d like to thank C&A - IITG for this opportunity.

Thank You!

Jaswanth Gudiseva
g.jaswanth@iitg.ac.in
linkedIn