DA332 : Data Visualization

Project Report

A Brief Analysis Of Enrollment Statistics

Prepared by
Laanith Chouhan
210150014

1. Introduction :

This report offers a detailed analysis of the variation of class wise and district wise enrolment over the years from 2012 to 2020. It shows how data visualization can be leveraged to find out meaningful insights from data.

2. Dataset :

The dataset for the project was obtained from the website of data.gov.in. The link for the dataset is given below :
Dataset Link

The whole dataset is regarded as a set of 8 CSV (Comma Separated Value) files each showing district-wise and class-wise enrolment of students from several regions.

3. Motivation :

Driven by a passion for data visualization and a keen interest in understanding educational dynamics, our project embarks on a journey to unravel the intricate patterns of student enrollment data.

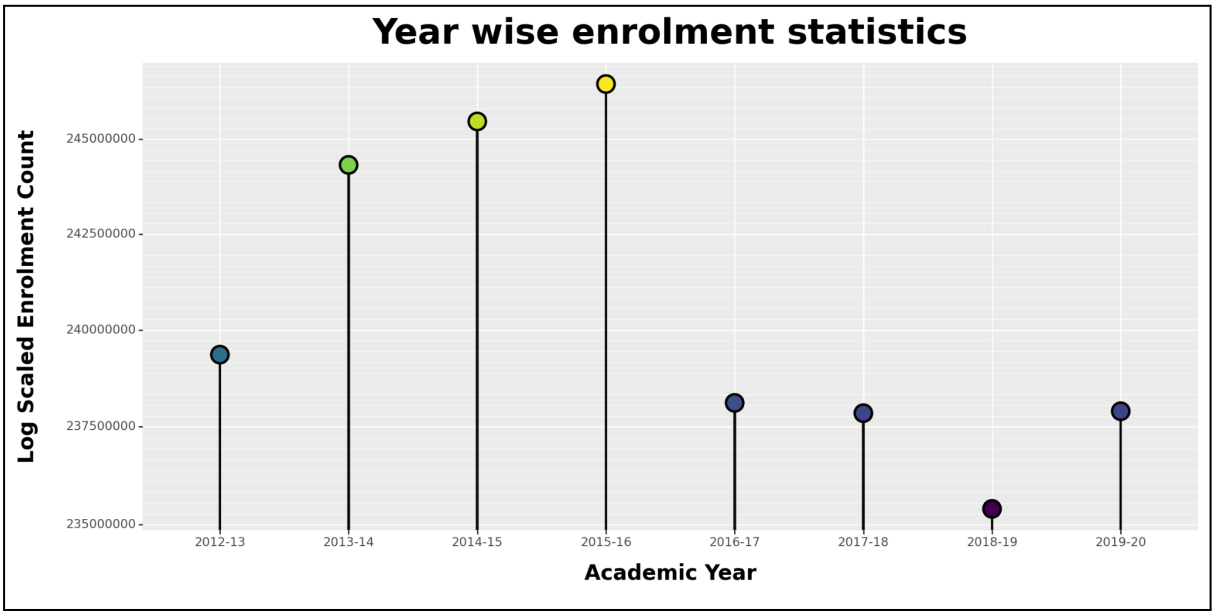
With a focus on class-wise and district-wise enrollment spanning nearly a decade, our endeavor is fueled by a deep-seated desire to transform raw data into actionable insights. Through meticulous analysis and innovative visualization techniques, we aspire to shed light on the underlying trends and disparities within the education landscape. Our commitment to leveraging data visualization as a tool for informed decision-making underscores our dedication to driving positive change and fostering equitable educational opportunities for all.

4. The Insights :

A. A primer : Year Wise Enrolment Statistics

For this purpose, a chart of categorical type is an apt one as data is categorical. So, a variant of bar-chart called the Lollipop chart which is custom made using Plotnine library, is leveraged to show the corresponding statistics.

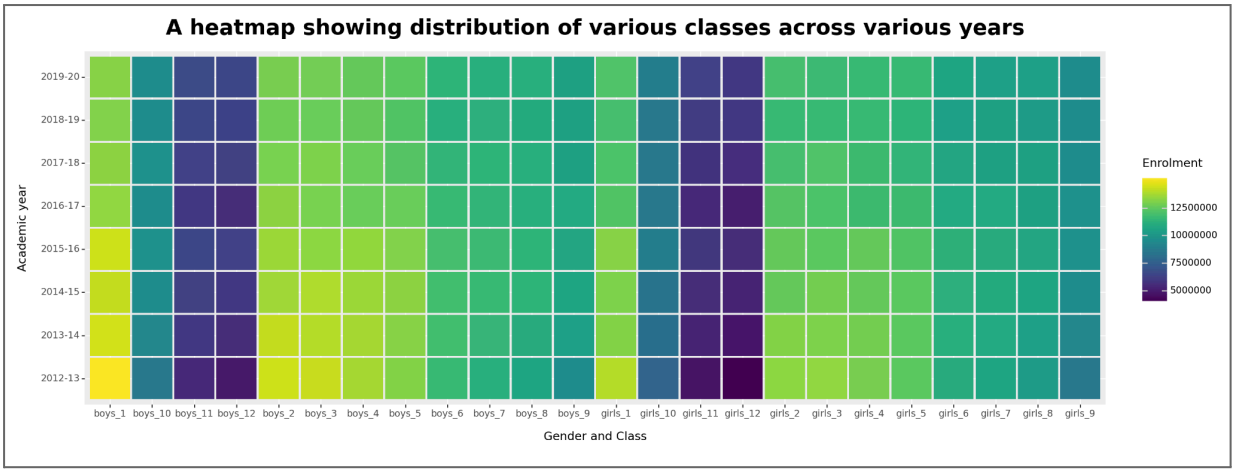
To enhance the differences between the data points, a Logarithmic scale was used.



B. A deeper inspection into the Distribution of Data

To have an insight into how the distribution of class-wise enrolment strength of students varied over the years, This is actually a plot showing distribution of two-categorical families, (academic year - although temporal, is regarded as categorical as only 8 such years were available in the dataset.).

A heatmap is the excellent solution for the above visualization problem. The plot that is obtained is shown in the next page :



Insights obtained :

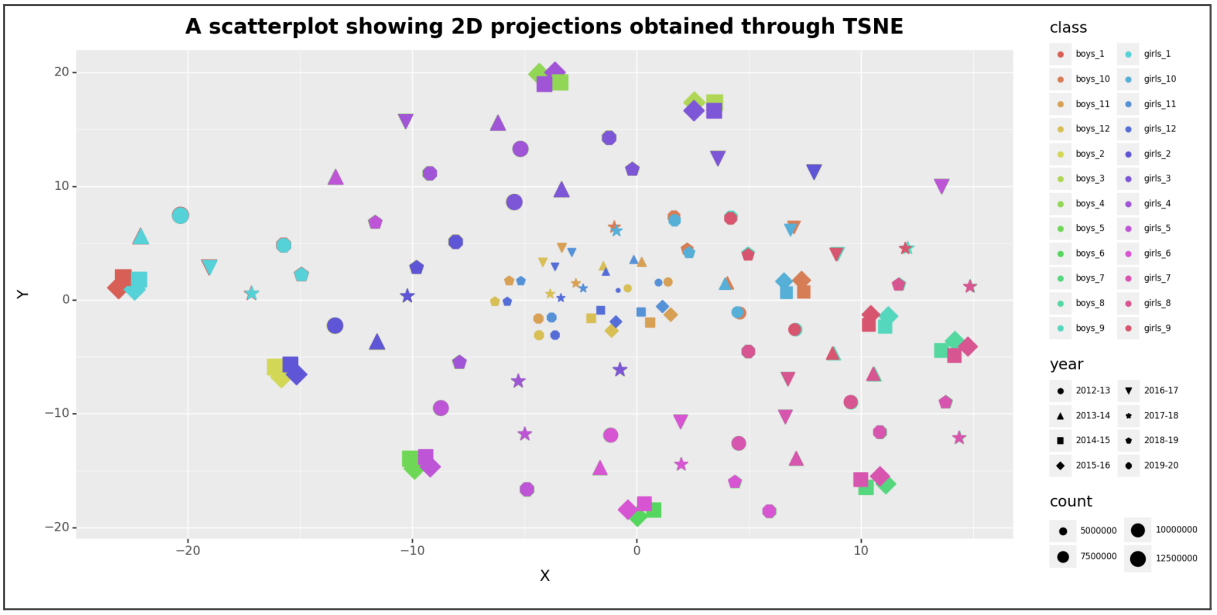
- The enrolment of boys and girls in class 12 is quite similar.
- If we compare a single class between boys and girls, we find their distributions over the years to be identical. This indeed states that the enrollment strength of boys and girls in all the classes is also the same.
- There is a higher rate of enrolment in primary classes (1 to 5) than all others.

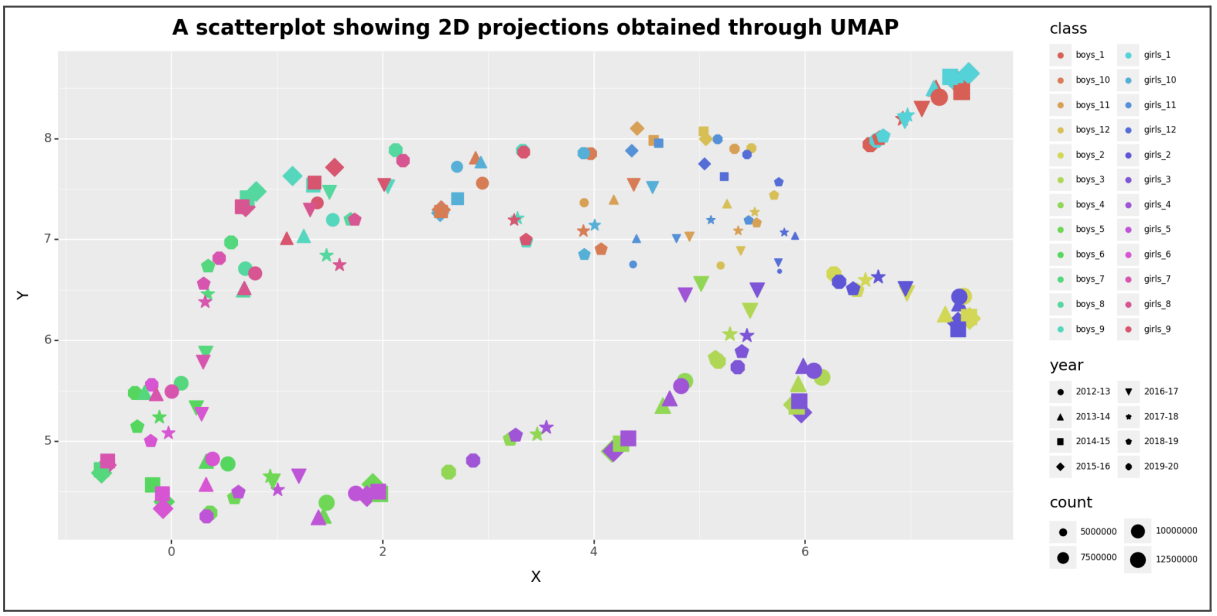
C. A more intriguing insight :

After visualizing the heatmap, I wanted to further look into the distribution of classes over all years over all the districts in India, but the problem is that it is not possible to have a look at them altogether as there were 700 districts roughly in India, which leads to a huge dimensional vector.

In the pursuit of an effective method for visualizing high-dimensional data, it has been established that t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are the foremost algorithms. These techniques excel in projecting complex, high-dimensional data onto lower-dimensional spaces, facilitating the visualization of embeddings and clusters within the dataset.

So, an effort to project these onto 2-dimensional space and then to use a scatter plot to visualize clusters is made . Let us have a look at the results :





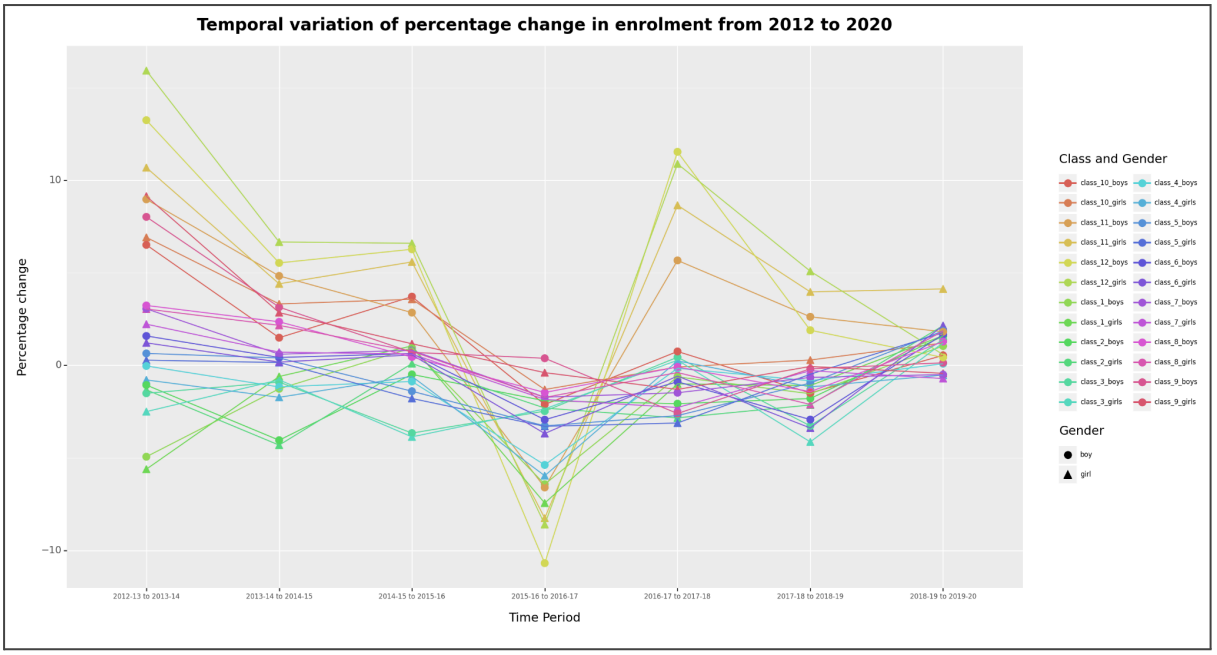
Insights obtained :

- Although it does not appear in the images here, the points with similar shape, same size and same class are scattered very close to each other or are overlapping, as shown in the plots.
- This actually indicates that not only the sums are similar (as shown in the heatmap) but also the vectors are very similar to each other indeed. Which proves that the distribution of the same class in the same year across the districts does not vary much among the genders.

D. An Anomaly :

Now try to have a look at the below image and analyze the plot.

The plot shows temporal variation of percentage change in net-enrolment in India, separated by classes and genders , plotted using simple geom_line in plotnine, together with grouping.



We can find an interesting anomaly in the plot, there is a drop in the enrollment rates of several classes, namely class-12 boys , class-12 girls and class-11 girls, along with some other classes.

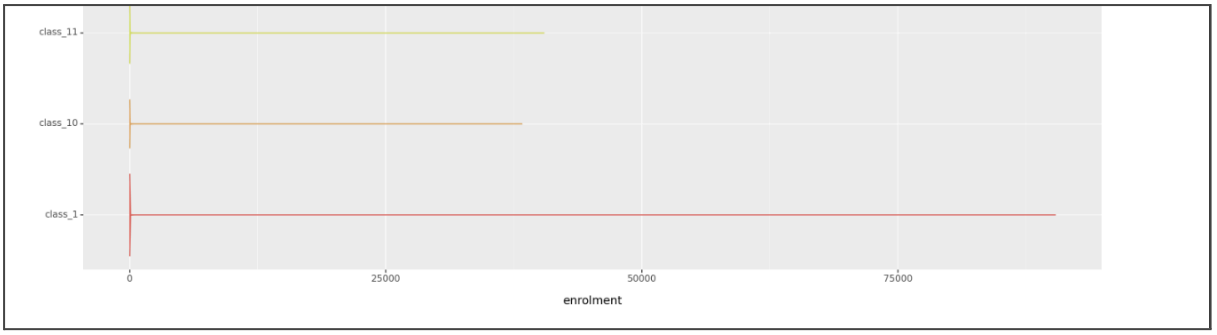
I tried to reason why there had been such a sudden anomaly in the trend, but could not come up with a satisfactory cause. But looking further into it had made me find a fact that "The average annual dropout rate was the highest for secondary schools in India was about 17.06 percent in 2016, which was the highest in India."

- This reason actually explains why there was such an enrollment rate in that year. But the reasons for the sudden increase in the dropout rate are still unknown.

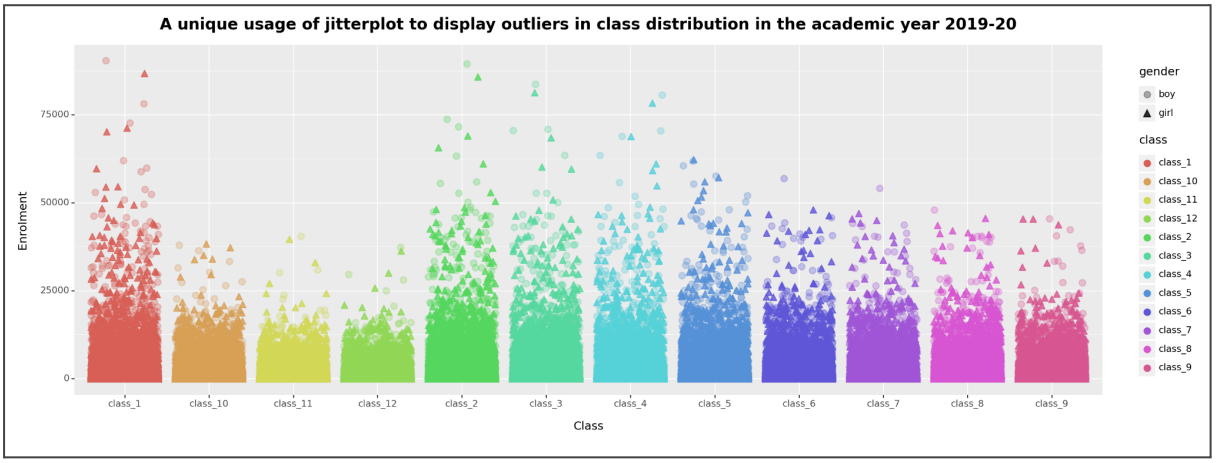
E. A look into the Outliers :

Analyzing the distribution of data among districts presents challenges, particularly due to **zero class enrollment** in many districts and hence this was reflected a lot in the data. Box plots and Violin plots appeared to be heavily biased towards these, where it is just a T shaped distribution.

. Traditional visualization methods such as box plots and violin plots are biased towards these districts, resulting in skewed distributions. To address this, a jitter plot is employed for visualization. An example :



To address this, a jitter plot is employed for visualization.



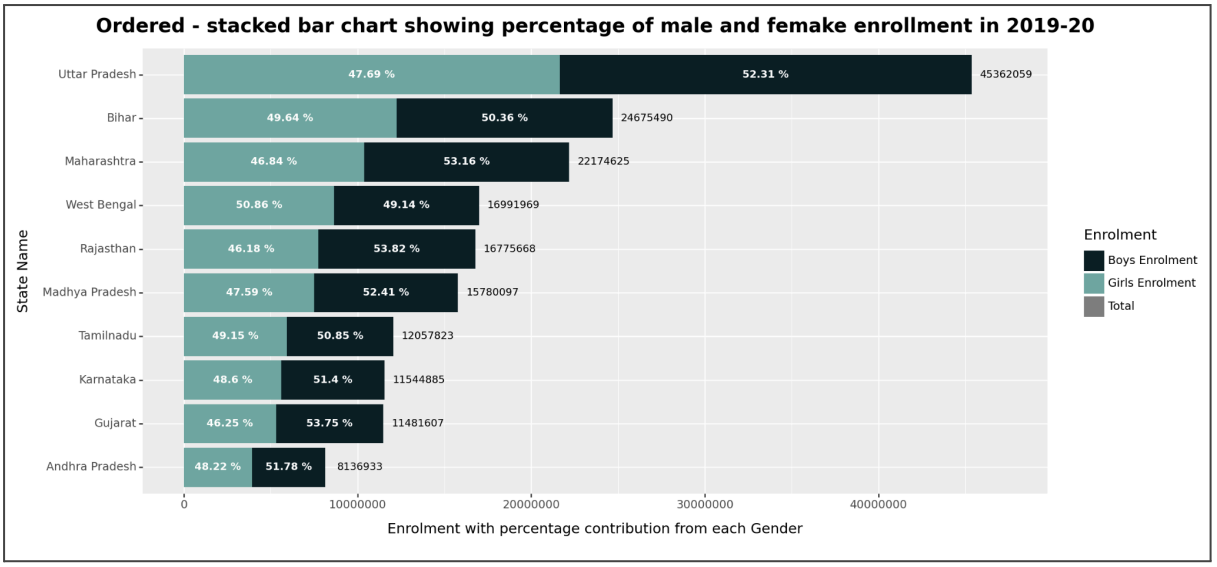
For the ease of noticing, a separate shape and alpha (opacity) are assigned to points belonging to boy and girl categories.

Insights obtained :

- The maximum enrolment of boys and girls in classes 11th and 12th does not cross the mark of about 37,500. The enrollment actually decreases as students study higher and higher.

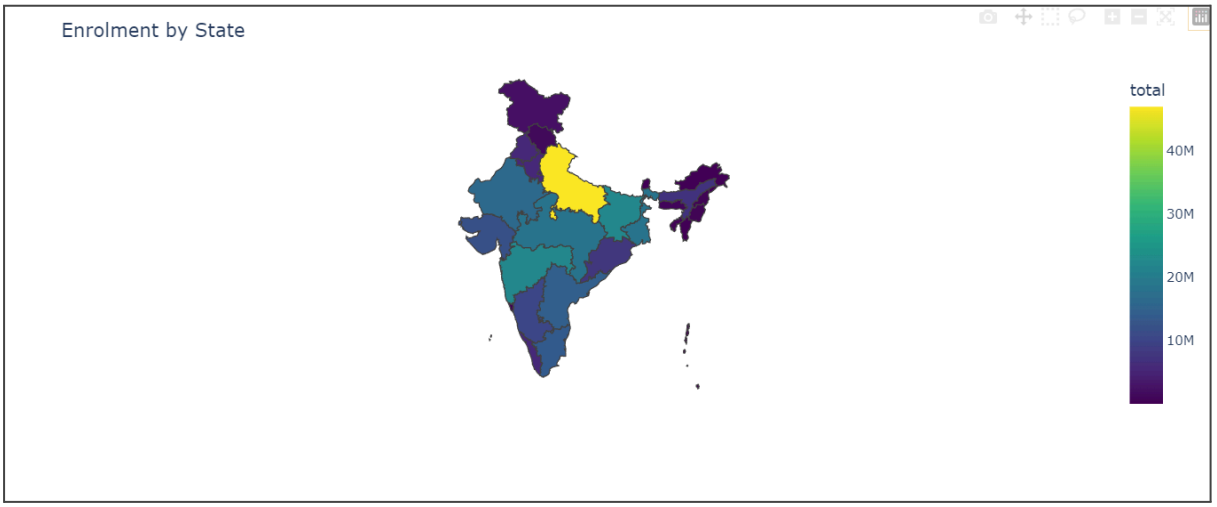
F. Percentage contribution of Genders :

Visualization of the percentage of boys and girls enrolled in the top 10 states in the country is achieved through an ordered bar chart.



G. A Chloropleth Map :

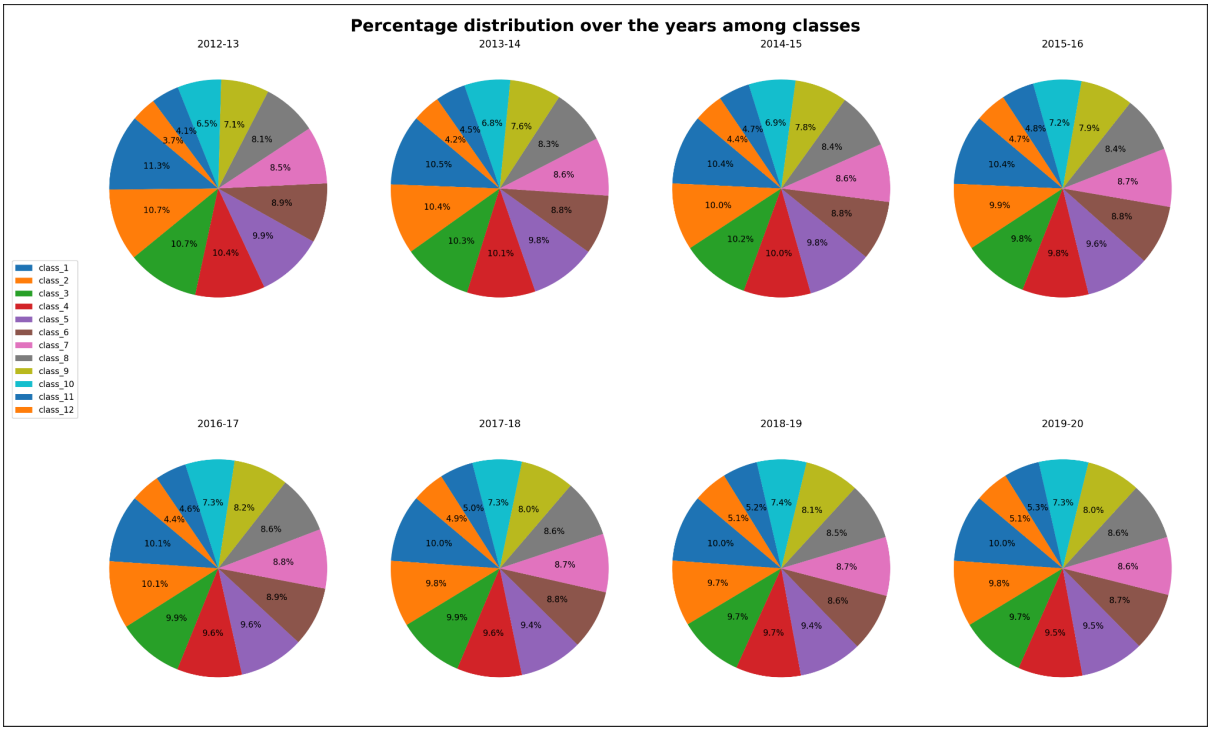
State-wise enrollment statistics are visualized using a Choropleth map created with the Plotly library. The interactive map highlights variations in enrollment rates across different states in India.



As we can see, UP is the highest populated state, as well as the state with the largest rate of enrollment. Bihar stands next to it. The north eastern states have a very low rate of enrollment (in thousands !!!) whereas the other regions of India have minimum enrolment rate in the range of a million at least.

H. Distribution through Pie Charts:

Pie charts are utilized to visualize class-wise distribution percentages over the years. Insights reveal that the distribution of percentages remains relatively constant over time.



Insights obtained :

- The distribution of percentages actually remained constant over the years, with a very little difference between each year.

Conclusion :

Data visualization is a great tool for analyzing, interpreting, and gaining insights and reasoning from data. Libraries like Plotnine, Matplotlib, Seaborn, Plotly in Python, and D3.js in JavaScript actually give us a lot of power to manipulate and visualize data as much as we like. Through the use of customizable plots, charts, and interactive graphics, these libraries enable us to convey complex information effectively to diverse audiences, fostering better understanding and informed decision-making.

The code for the Project is found in : https://github.com/Laanith/DataVisualization