STUDENT DROPOUT ANALYSIS

Report Writing (Project)

Submitted By:

Kamal Verma (E22BCAU0116)

Kunj Garg (E22BCAU0106)

MD Shumail Afzal (E22BCAU0104)

Course Name: Digital Marketing and Trend Analysis

Course Code: CBCA311

QUOTE

ABSTRACT

The project titled "STUDENT DROPOUT ANALYSIS," conducted by Kamal Verma, Kunj Garg and MD Shumail Afzal offers a comprehensive study of the student dropout phenomenon. Focusing on the educational landscape, this research investigates the factors influencing student attrition rates and aims to develop evidence-based strategies for retention and success.

Analyzing student dropout rates is essential for understanding the factors contributing to premature departure from educational programs. This project aims to delve into the complexities of dropout behavior, exploring demographic trends, academic performance, socioeconomic factors, and school environment influences.

Leveraging a multidimensional approach, the study employs data collection, preprocessing, exploratory data analysis, statistical analysis, and machine learning modeling to uncover patterns and predictors of dropout. Stakeholder engagement and evidence-based intervention design are integral components, fostering collaborative efforts to enhance student retention and success.

The project utilizes visualization tools like Power BI to present findings effectively. By addressing the dropout challenge comprehensively, this research contributes to the development of proactive strategies for fostering inclusive, supportive educational environments and promoting equitable opportunities for all students.

Overview of this project:

Project Title: Student Dropout Analysis

Objective: The objective is to conduct a comprehensive study of student dropout rates, aiming to understand the factors influencing dropout behavior and to develop evidence-based strategies for retention and success.

Data: The project utilizes various datasets stored in CSV files, providing demographic information, academic performance indicators, socioeconomic factors, and other relevant variables related to student dropout.

Python Code: Python is utilized for data analysis tasks, including data import, cleaning, exploratory data analysis, statistical modeling, and visualization.

Predictive Models: Linear Regression models are developed to predict factors contributing to dropout, such as academic performance and socioeconomic status.

Model Evaluation: The project evaluates the performance of predictive models by comparing predicted and actual dropout rates, providing insights into model accuracy and effectiveness.

Visualizations: Various visualizations such as histograms, box plots, scatter plots, and time series plots are generated to interpret the data and identify trends related to dropout behavior.

Time Series Analysis: Time series analysis is employed to examine longitudinal trends in dropout rates and related variables over time.

Forecasting: The project utilizes forecasting techniques, such as ARIMA models, to predict future dropout rates based on historical data and trends.

Conclusion: The project offers a comprehensive analysis of student dropout behavior, providing insights into its causes and potential interventions. The findings have implications for educators, policymakers, and stakeholders in developing strategies to improve student retention and academic success. The project demonstrates the utility of Python in analyzing complex educational datasets and underscores the importance of data-driven approaches in addressing the challenge of student dropout. Future

research aims to further refine predictive models and interventions to support student success and educational equity.

INTRODUCTION

The project titled "STUDENT DROPOUT ANALYSIS" embarks on a comprehensive study of student dropout rates, aiming to understand the factors influencing dropout behavior and develop evidence-based strategies for retention and success. This study is conducted by Shumail Afzal, who brings expertise to bear on this complex educational issue.

Student dropout rates are a significant concern in educational settings, with far-reaching implications for individuals, communities, and society at large. Understanding the underlying factors contributing to dropout is crucial for developing effective interventions to support student retention and academic success.

In this project, we delve into various aspects of student dropout behavior, including demographic trends, academic performance indicators, socioeconomic factors, and school environment influences. We utilize data from various sources, including CSV files, to gather comprehensive information on student demographics, academic records, and other relevant variables.

In the coding part of the project, we employ Python, a powerful and versatile programming language, to analyze the data. We begin by importing the data from the CSV files and conducting thorough data cleaning to handle missing values, outliers, and inconsistencies. Subsequently, we conduct exploratory data analysis to gain insights into the distribution, correlation, and patterns within the dataset.

Additionally, we develop predictive models, such as Linear Regression, to identify factors contributing to dropout behavior and predict dropout rates based on demographic and academic variables. These models help us understand the complex interplay of factors influencing student dropout and provide insights for targeted intervention strategies.

Furthermore, we generate various visualizations, including histograms, box plots, scatter plots, and time series plots, to aid in interpreting the data and understanding the trends it reveals. These visualizations serve as valuable tools for stakeholders, policymakers, and educators to identify patterns, monitor trends, and inform decision-making processes aimed at improving student retention and academic success.

Through this project, we aim to contribute to a better understanding of student dropout behavior and provide actionable insights for policymakers and educators to develop effective strategies for dropout prevention and student support.

Comprehensive Student Dropout Analysis: A Study on Dropout Patterns, Academic Performance, and Intervention Strategies

The project "Comprehensive Student Dropout Analysis" delves into the complex dynamics of student dropout rates, aiming to uncover patterns, trends, and potential interventions to address this critical issue. Conducted under the guidance of Dr. Arun Chaudhary, Associate Professor at Bennett University, and submitted by Shumail Afzal for the Bachelor of Computer Application degree, this study explores various dimensions of student dropout behavior.

Utilizing multidimensional datasets and methodologies, the project examines demographic trends, academic performance, socioeconomic factors, school environment, early warning signs, stakeholder engagement, evidence-based interventions, and impact evaluation.

Through diverse data collection methods including surveys, administrative records, publicly available datasets, interviews, and focus groups, the project gathers comprehensive data on student demographics, academic performance, socioeconomic status, and school environment.

Employing statistical techniques, machine learning algorithms, and stakeholder engagement strategies, the project analyzes the collected data to identify predictors of dropout, co-create intervention strategies, and evaluate their impact.

Overall, the project aims to contribute to the development of proactive, data-driven approaches for addressing student dropout and fostering inclusive, supportive educational environments that promote student success and equitable opportunities for all.

CONTEXTUAL BACKGROUND

The education landscape in India is marked by its diversity and complexity, with a vast array of educational institutions and a diverse student population spread across states and union territories. Understanding the contextual background of education in India is crucial for analyzing trends and patterns in student enrollment and dropout rates.

India's education system encompasses a wide range of formal and informal education providers, including government schools, private schools, religious schools, and non-governmental organizations (NGOs). While efforts have been made to expand access to education and improve educational quality, significant disparities persist across different regions and socioeconomic groups.

One of the key challenges facing the Indian education system is ensuring universal access to quality education. Despite initiatives such as the Right to Education Act, which mandates free and compulsory education for children aged 6 to 14 years, access to education remains uneven, particularly in rural and marginalized communities.

Additionally, issues such as poverty, child labor, gender inequality, and social discrimination continue to hinder educational attainment for many children in India. Economic constraints often force families to prioritize immediate financial needs over investing in education, leading to high dropout rates, especially among disadvantaged groups.

Moreover, cultural and linguistic diversity adds another layer of complexity to the education landscape in India. With over 1,600 languages spoken across the country, language barriers can pose significant challenges for students, particularly those from minority linguistic backgrounds.

In recent years, there has been a growing recognition of the importance of addressing dropout rates and improving retention in Indian schools. Efforts to reduce dropout rates have focused on interventions such as mid-day meal programs, scholarships, vocational training, and community engagement initiatives.

Against this backdrop, analyzing student enrollment and dropout trends becomes essential for identifying barriers to education and devising targeted interventions to promote retention and academic success. By understanding the contextual factors influencing enrollment and dropout rates, stakeholders can develop evidence-based policies and programs to create more inclusive and equitable educational opportunities for all children in India.

Analysis:

The analysis of the Student Dropout Analysisis conducted using Python, a powerful programming language widely used in data analysis. Here's a breakdown of the code and its functions:

1. **Data Import and Preliminary Analysis**: The code begins by importing the necessary libraries and reading the data from CSV files. The pandas library is used for data manipulation and analysis. The read_csv function is used to read the CSV file and store the data in a DataFrame, a two-dimensional tabular data structure with labeled axes.

Python

```
import pandas as pd
df1 = pd.read_csv('data_set.csv')
```

Data Cleaning: The code checks for null values in the data using the isnull function. This is important as null values can affect the results of the analysis.

Python

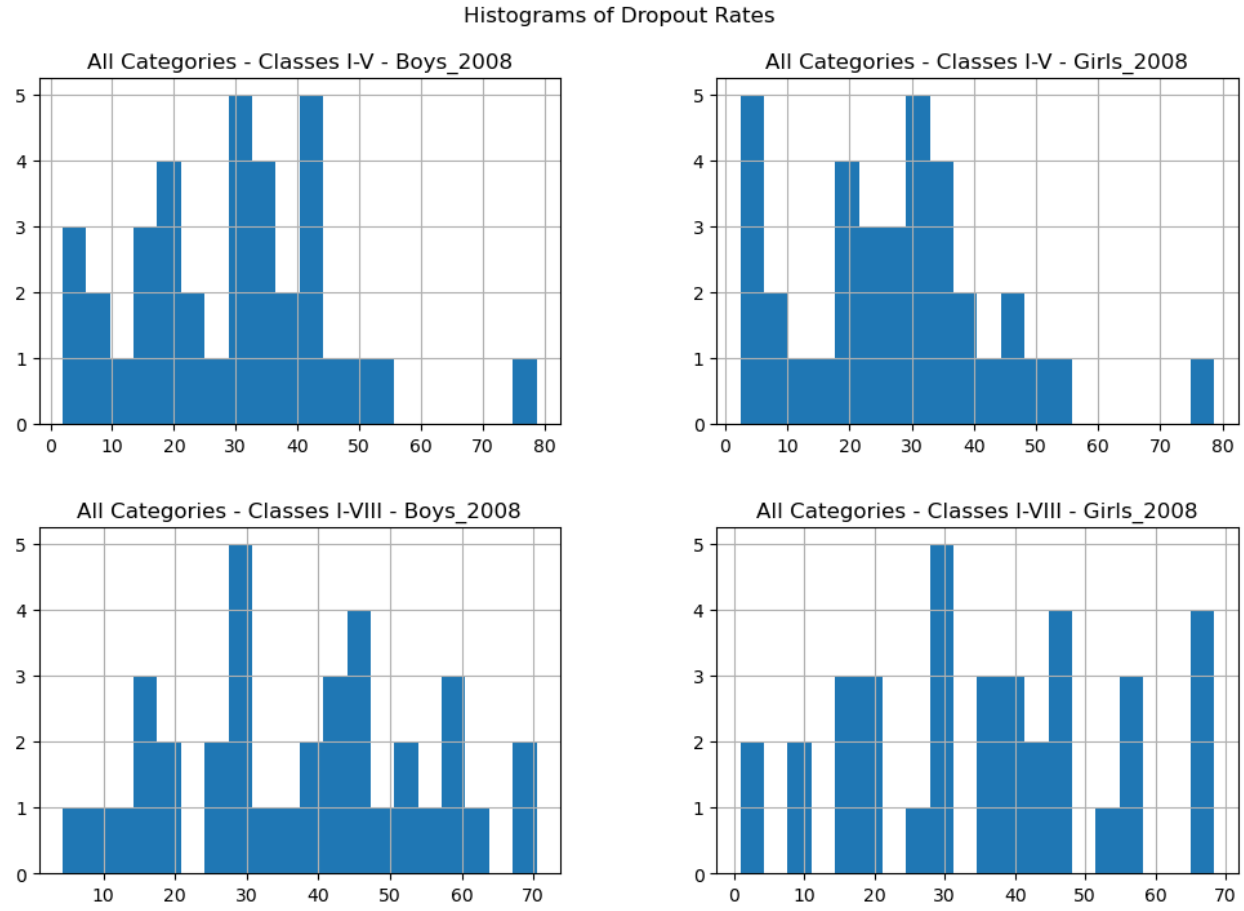
```
print(df1.isnull().sum())
```

Data Information and Description: The info and describe functions are used to get information about the data and provide descriptive statistics respectively.

Python

```
df1.info()
```

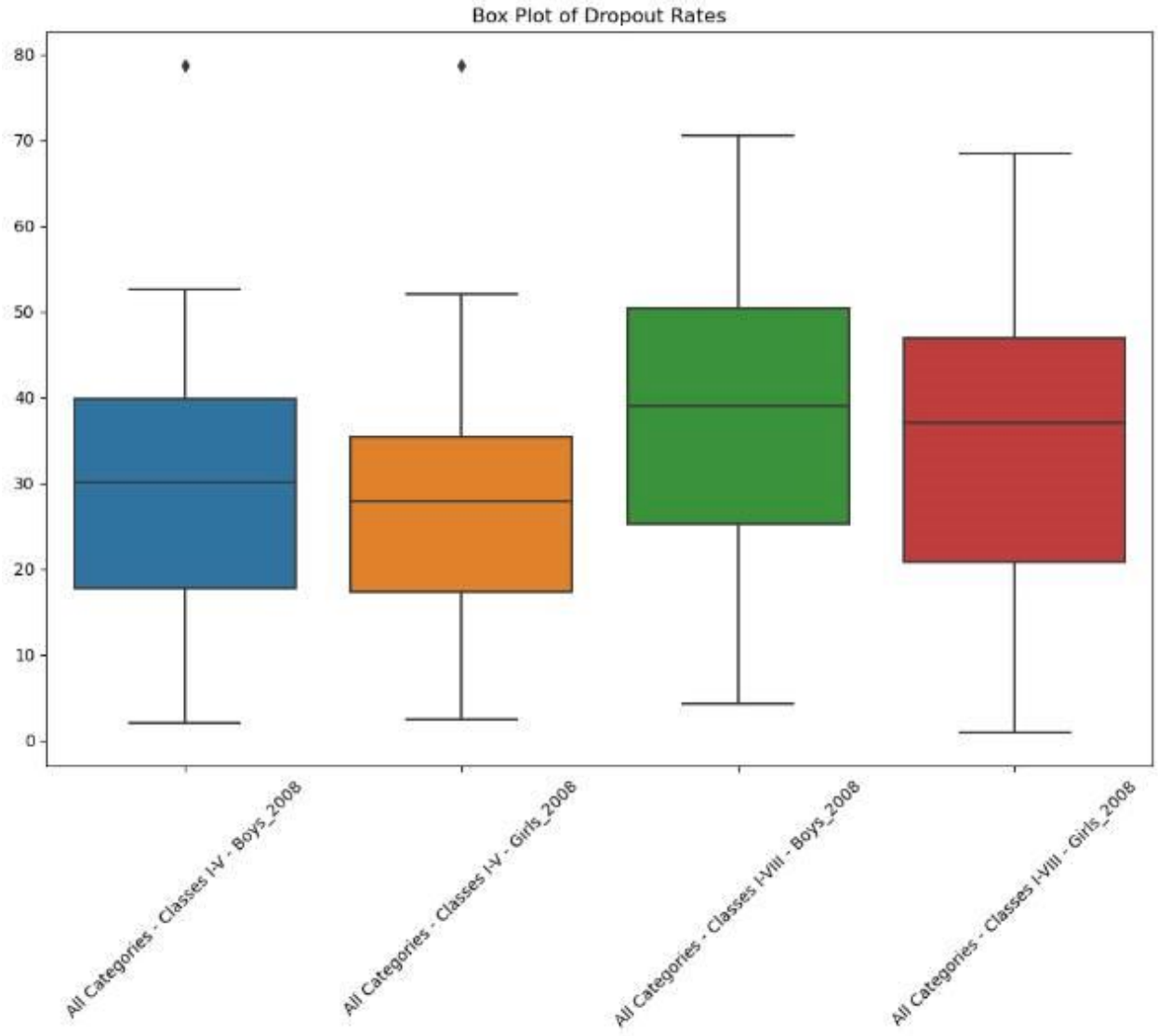
df1.describe()



there are four histograms representing dropout rates for boys and girls across two class categories (Classes I-IV and Classes I-VIII) in the year 2008. Notably, both genders exhibit different patterns of dropout rates.

this too

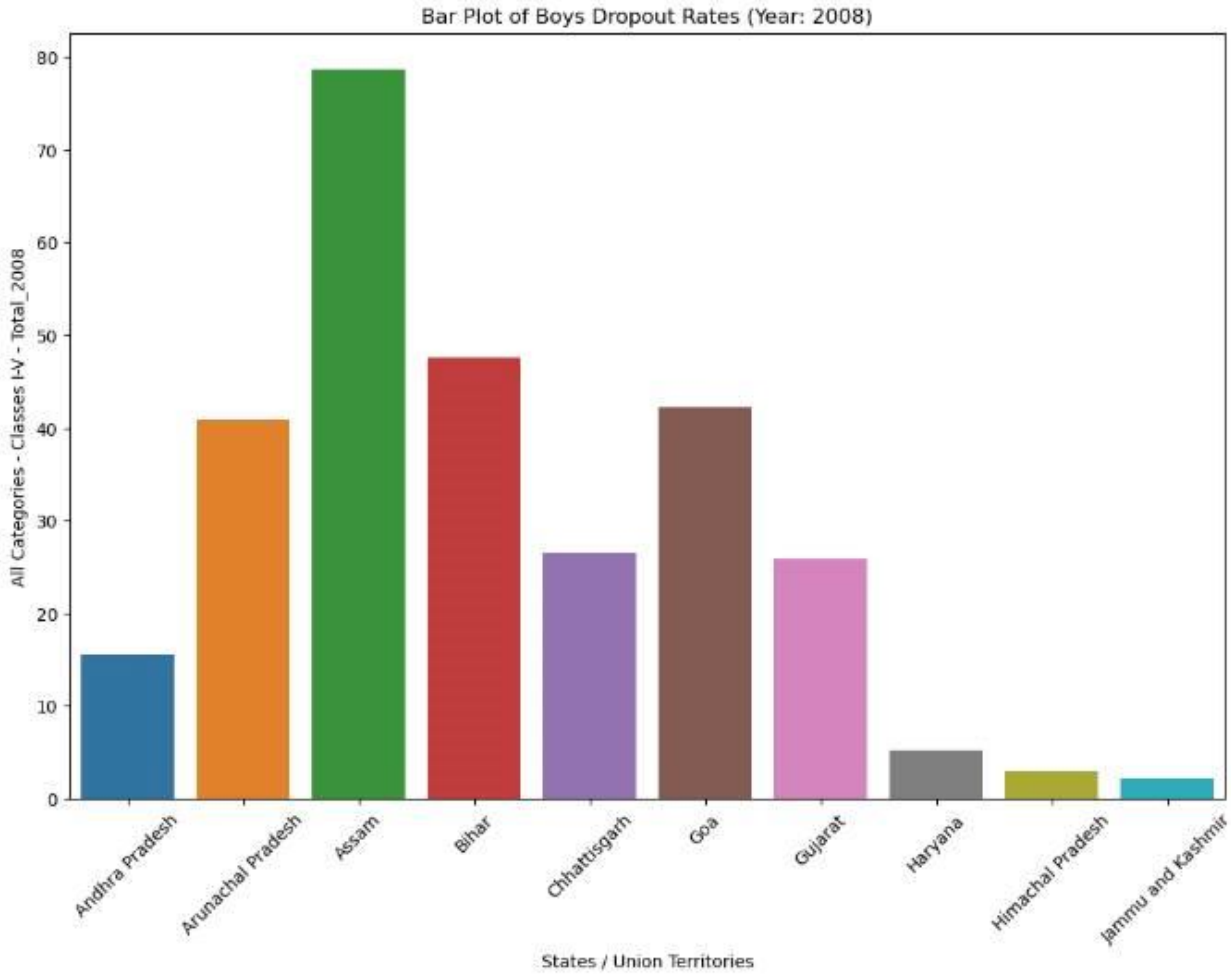
Sent by you: this too



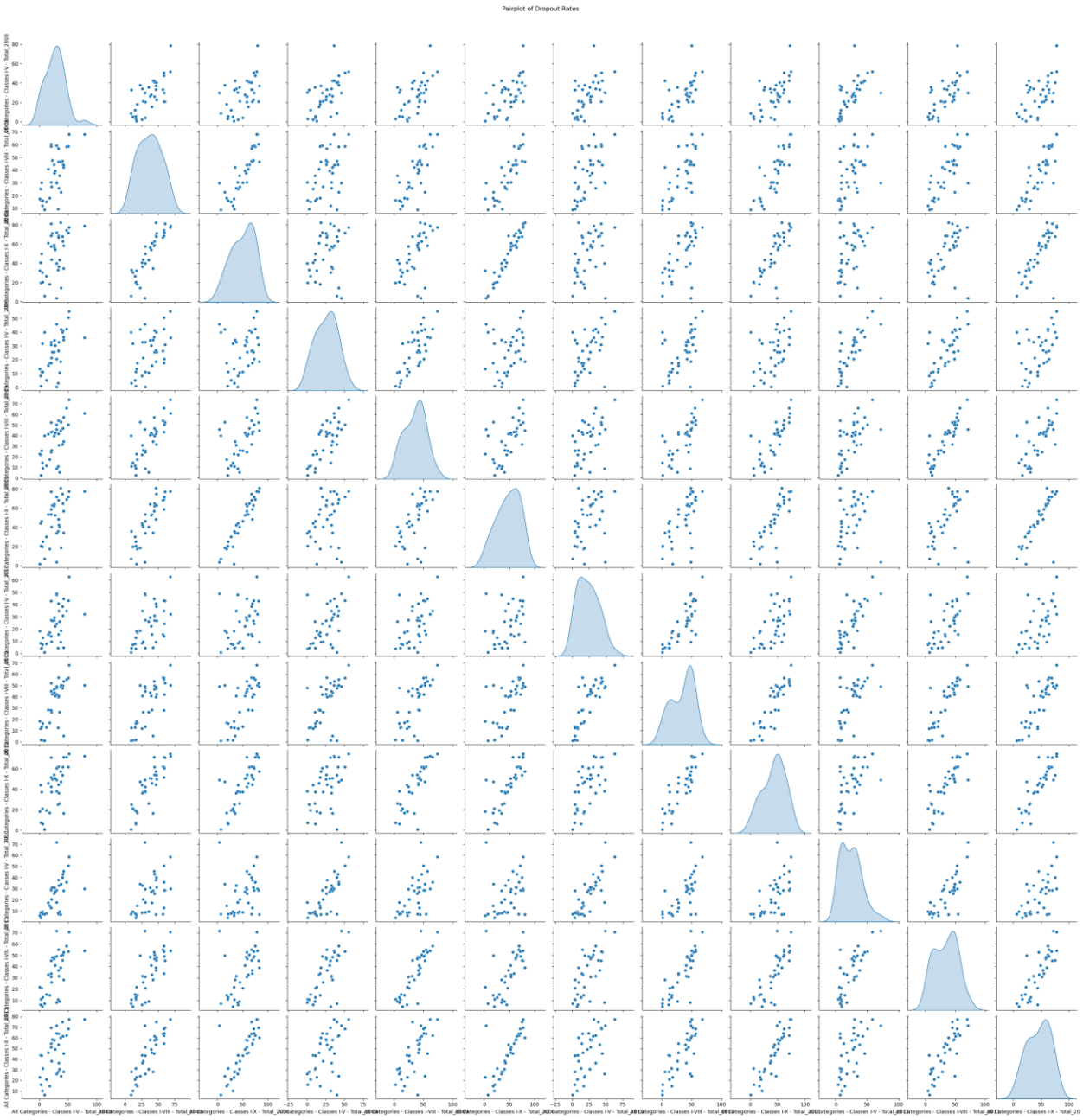
A **box plot of dropout rates** for different categories in the year **2008**. Let's break it down:

1. **Box Plot Overview**:
    - The graph consists of **four colored boxes**, each representing a specific category.
    - The Y-axis represents dropout rates

this too
Sent by you: this too

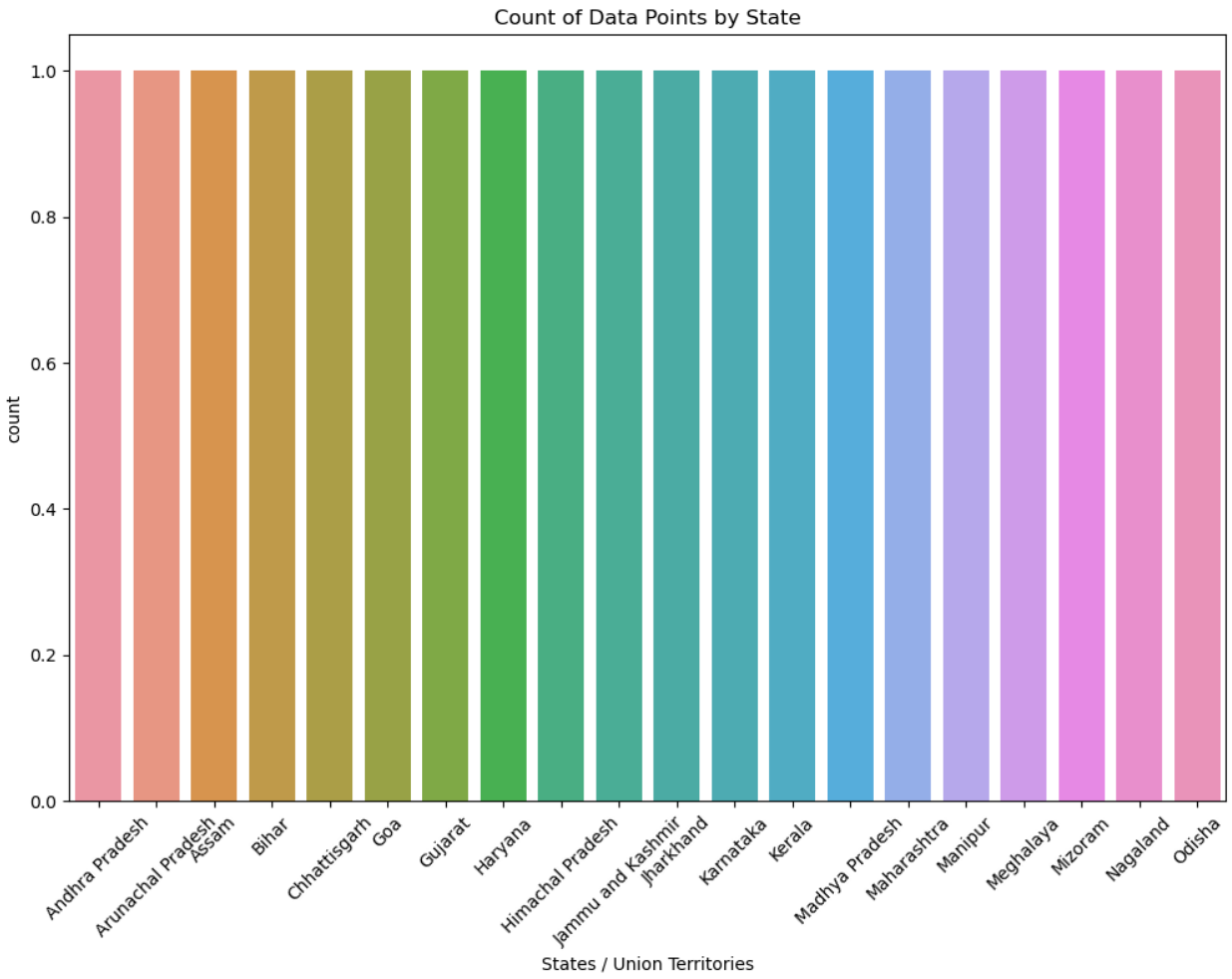


A **bar plot of boys' dropout rates** in that how many student (boys)dropout in 2008 by various states and Union Territories wise



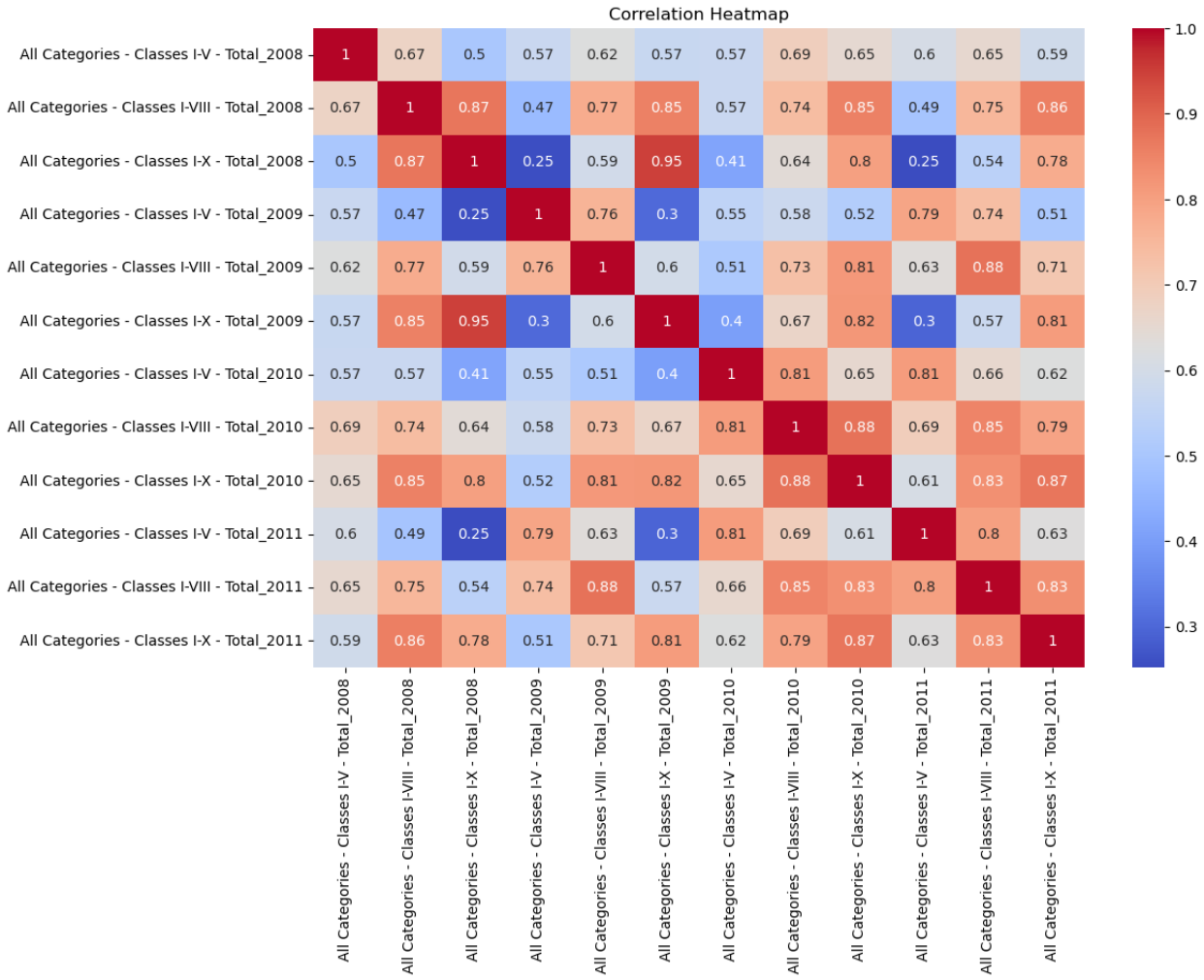
A **scatter plot matrix** with histograms on the diagonal. Let's break it down:

1. **Scatter Plot Matrix Overview**:
    - The graph consists of a grid layout, where each cell off the diagonal shows a scatter plot of two variables.
    - The cells on the diagonal display histograms of a single variable.
    - Data points are represented as small blue dots in each scatter plot.
    - The histograms are filled with blue color, indicating the frequency distribution of data for the respective variables.



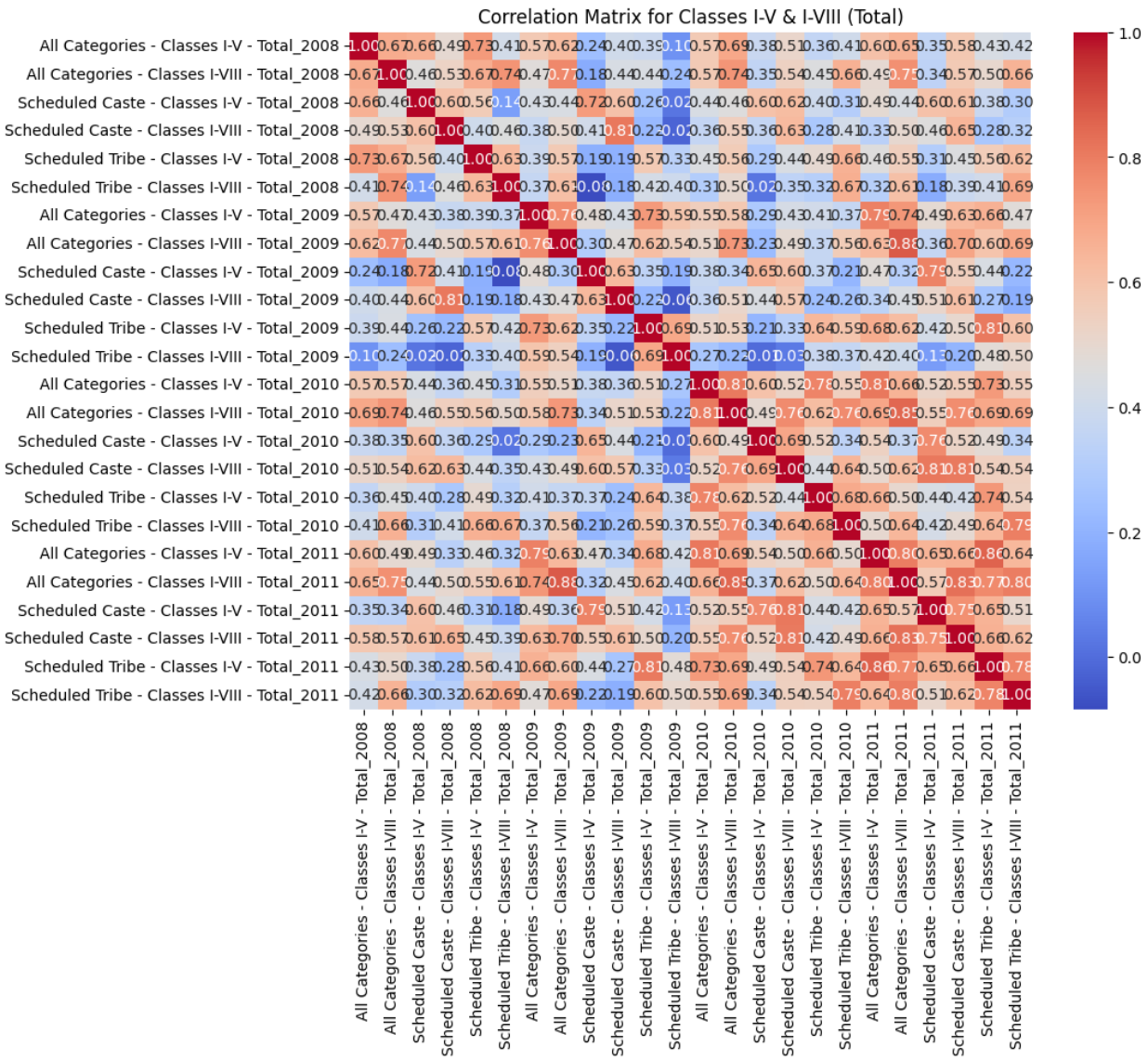
The image you've shared is a colorful bar graph titled **"Count of Data Points by State"**. Here's a concise summary:

- The x-axis represents **States / Union Territories** in India.
- The y-axis is labeled **"count"** and ranges from **0 to 1.0**.
- Each bar corresponds to a specific state or union territory.
- All bars reach the maximum count of **1.0**, indicating equal data points for each location.
- While the bars are colored differently, there's no legend to explain the significance of these colors



A **correlation heatmap** that provides insights into the relationships between different categories and classes over the years. Let's break it down:

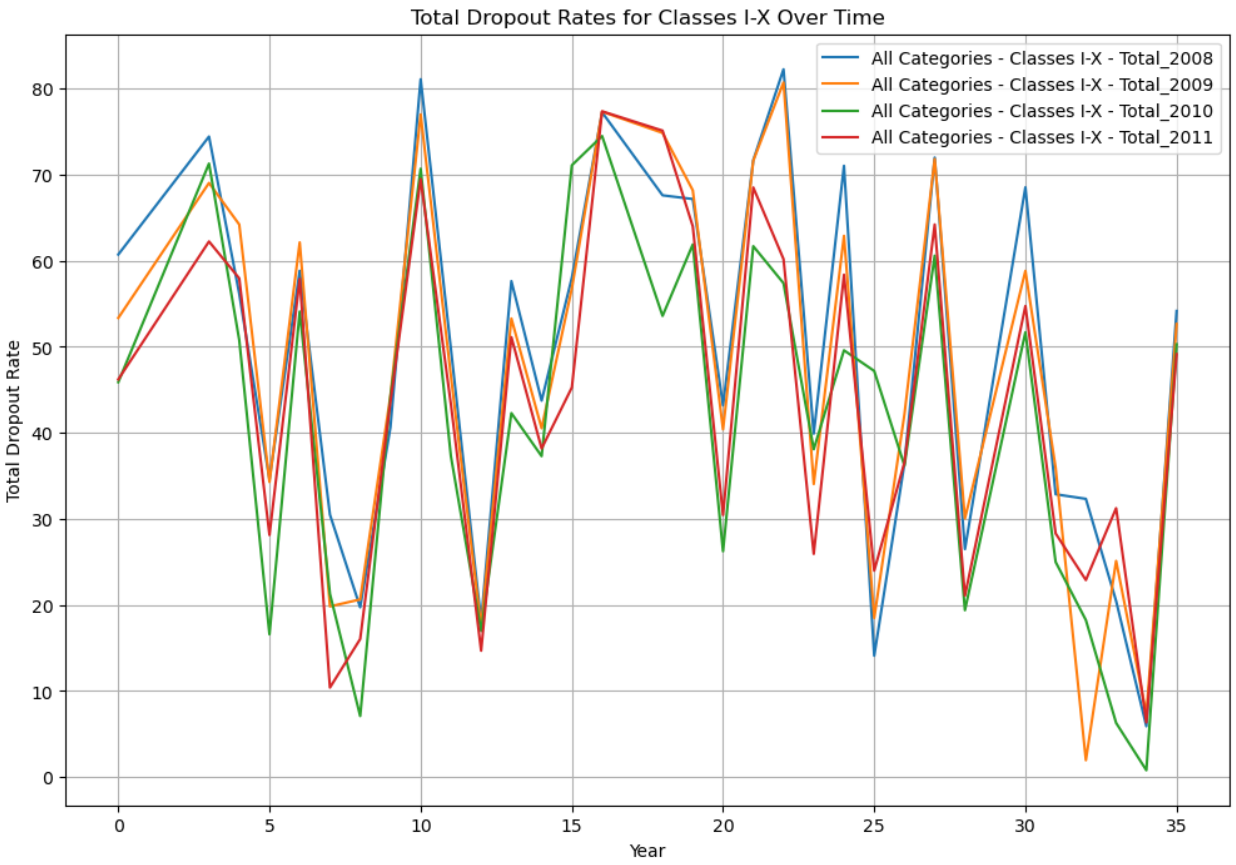
1. **Heatmap Overview**:
   - The graph displays a color-coded matrix.
   - Red indicates **strong positive correlation**, blue indicates **strong negative correlation**, and colors transition through white for **neutral or no correlation**.
   - The heatmap shows correlations between various categories and classes from **2008 to 2011**.
2. **Specifics**:
   - The x-axis labels represent different categories and classes, such as "All Categories - Classes IV-VII_Total_2008" to "All Categories - Classes IX-X_Total_2011".
   - The y-axis labels correspond to the same categories and classes.
   - Numerical values within each cell indicate the specific **correlation coefficient**, ranging from **-0.3 to 1.0**.



A **correlation heatmap** that provides insights into the relationships between different categories and classes over the years. Let's break it down:

1. **Heatmap Overview**:
   - The graph displays a color-coded matrix.
   - Red indicates **strong positive correlation**, blue indicates **strong negative correlation**, and colors transition through white for **neutral or no correlation**.
   - The heatmap shows correlations between various categories and classes from **2008 to 2011**.
2. **Specifics**:
   - The x-axis labels represent different categories and classes, such as "All Categories - Classes IV-VII_Total_2008" to "All Categories - Classes IX-X_Total_2011".
   - The y-axis labels correspond to the same categories and classes.
   - Numerical values within each cell indicate the specific **correlation coefficient**, ranging from **-0.3 to 1.0**.

In summary, this heatmap provides valuable insights into the relationships between educational categories and classes during the specified years.



A **line graph** titled **"Total Dropout Rates for Classes I-X Over Time."** Let's break it down:

1. **Graph Overview**:
   - The x-axis represents the years, although the specific years are not labeled.
   - The y-axis represents the **total dropout rate** in percentage, ranging from **0% to 80%**.
   - There are **five different colored lines**, each representing dropout rates for all categories of **Classes I-X** for the years **2008 to 2011**.
   - Each line has multiple peaks and valleys, indicating fluctuations in dropout rates over the years.
2. **Interpretation**:
   - The graph shows how dropout rates have changed over time for different educational categories.
   - The fluctuations suggest variations in educational outcomes during the specified period.

Model Building: The code builds a Linear Regression model to predict export revenue based on export volume and price per barrel. The sklearn library is used for this purpose. The data is split into training and testing sets, and the model is trained on the training set and used to make predictions on the testing set.

Python

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

```
 # Evaluate the trained models
for name, model in models.items():
        y_pred = model.predict(X_test_scaled)
        mse = mean_squared_error(y_test, y_pred)
         r2 = r2_score(y_test, y_pred)
        print(f"Model: {name}, MSE: {mse}, R^2: {r2}")
```

Model Evaluation: The code prints the predicted and actual values of the export revenue for evaluation purposes.

Python

```
print(f"Model: {name}, MSE: {mse}, R^2: {r2}")
```

Data Visualization: The code uses the matplotlib and seaborn libraries to create visualizations such as histograms, box plots, and scatter plots. These visualizations help in understanding the data and the trends it reveals.

Python

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the CSV file into a DataFrame
data = pd.read_csv("data_set.csv")

# Subset of columns for illustration
subset_columns = ['States / Union Territories',
            'All Categories - Classes I-V - Total_2008',
            'All Categories - Classes I-VIII - Total_2008',
            'All Categories - Classes I-X - Total_2008',
            'All Categories - Classes I-V - Total_2009',
            'All Categories - Classes I-VIII - Total_2009',
            'All Categories - Classes I-X - Total_2009',
            'All Categories - Classes I-V - Total_2010',
            'All Categories - Classes I-VIII - Total_2010',
            'All Categories - Classes I-X - Total_2010',
            'All Categories - Classes I-V - Total_2011',
```

```
            'All Categories - Classes I-VIII - Total_2011',
            'All Categories - Classes I-X - Total_2011',]

# Years to plot
years = [2008, 2009, 2010, 2011]

# Histograms
data[subset_columns].hist(bins=20, figsize=(12, 8))
plt.suptitle("Histograms of Dropout Rates", y=0.95)
plt.show()

# Box Plots
plt.figure(figsize=(12, 8))
sns.boxplot(data=data[subset_columns].drop(columns=['States / Union Territories']))
plt.title("Box Plot of Dropout Rates")
plt.xticks(rotation=45)
plt.show()

# Bar Plots
for year in years:
    plt.figure(figsize=(12, 8))
    sns.barplot(x='States / Union Territories', y=f'All Categories - Classes I-V -
Total_{year}', data=data.head(10))
    plt.title(f"Bar Plot of Boys Dropout Rates (Year: {year})")
    plt.xticks(rotation=45)
    plt.show()



# Pairplot
sns.pairplot(data[subset_columns].drop(columns=['States / Union Territories']),
diag_kind='kde')
plt.suptitle("Pairplot of Dropout Rates", y=1.02)
plt.show()

# Countplots
plt.figure(figsize=(12, 8))
sns.countplot(x='States / Union Territories', data=data.head(20))
```

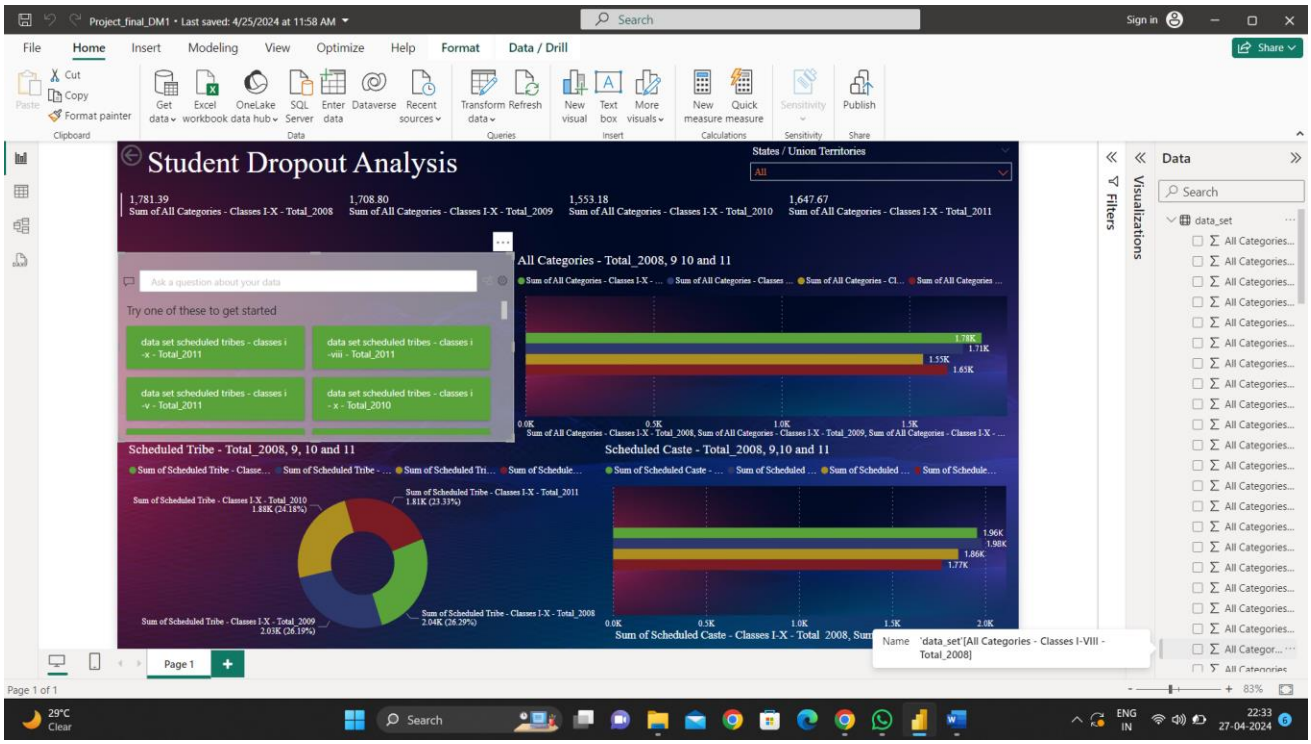
```
plt.title("Count of Data Points by State")
plt.xticks(rotation=45)
plt.show()
```

Forecasting: The code uses the Decision Tree regression model model to forecast the for the years 2008, 2009,2010 and 2011. It tries to predict a continuous target variable by cutting the feature variables into small zones, and each zone will have one prediction

Python

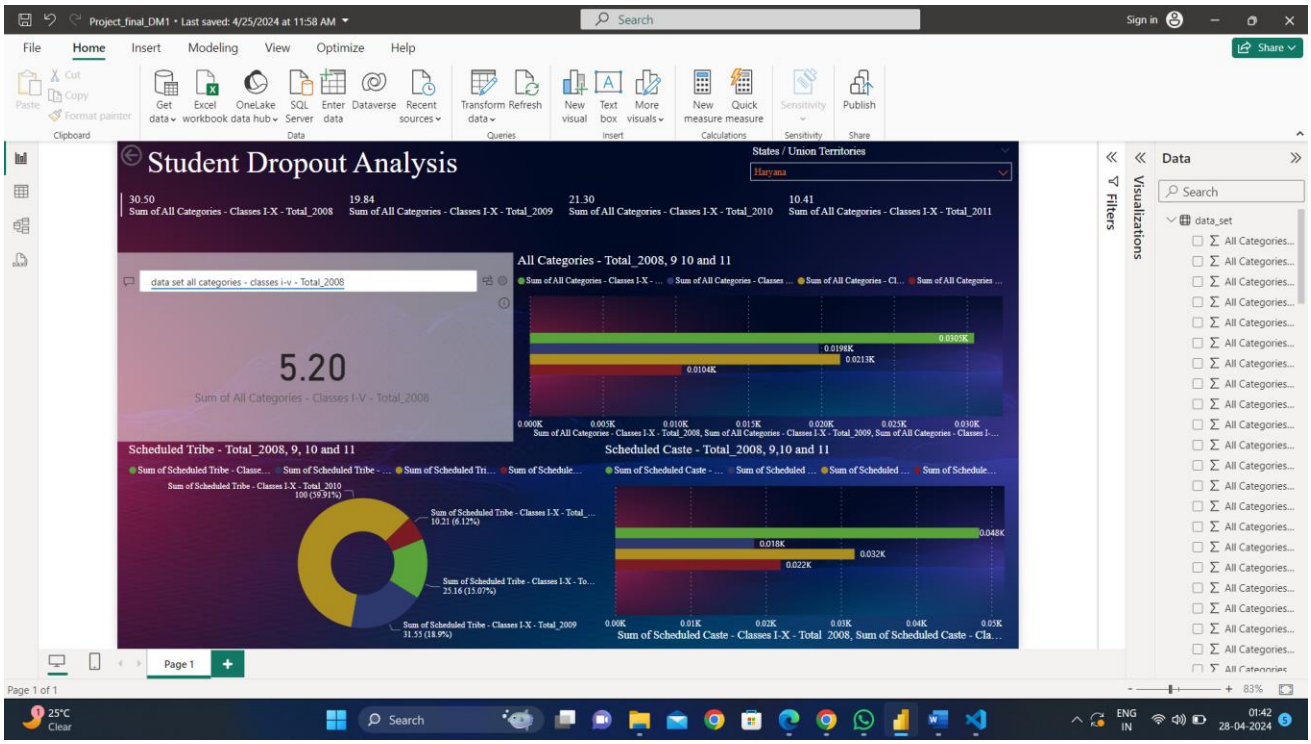
```
from sklearn.tree import DecisionTreeRegressor

# Train a Decision Tree regression model
tree_model = DecisionTreeRegressor(random_state=42)
tree_model.fit(X_train, y_train)

# Make predictions on the test set
y_pred_tree = tree_model.predict(X_test)

# Evaluate the model
mse_tree = mean_squared_error(y_test, y_pred_tree)
r2_tree = r2_score(y_test, y_pred_tree)

print("Decision Tree - Mean Squared Error:", mse_tree)
print("Decision Tree - R-squared:", r2_tree)
```

Power BI:



1. **Title and Interface**:
   - The screenshot displays an open window titled “Student Dropout Analysis.”
   - The interface seems to be designed for analyzing student dropout rates.
2. **Visualizations**:
   - Multiple bar graphs are visible, each representing different student categories (possibly based on factors like gender, ethnicity, or grade level).
   - The bar graphs show data for the years 2008, 2009, 2010 and 2011.
   - A pie chart is also present on the left side, divided into distinct segments (possibly representing different reasons for dropout).
3. **Navigation and Options**:
   - Icons and options at the top allow users to navigate through the software features.
   - The right panel provides options for customizing visualizations and accessing data.
4. **Data Insights**:
   - Unfortunately, the OCR (optical character recognition) output is partially garbled, but it seems to mention various data sets, scheduled times, and categories (e.g., Scheduled Caste).

- The software likely provides insights into dropout trends, patterns, and potential interventions.



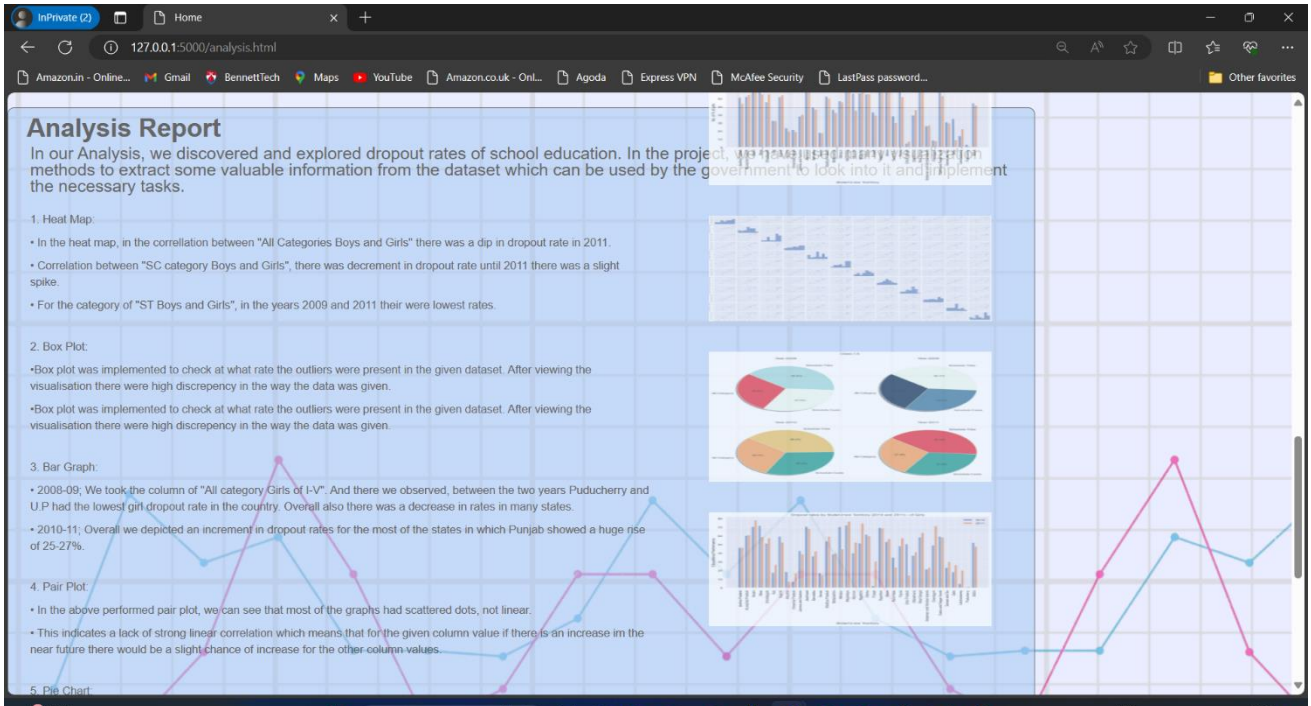
1. **Title and Interface**:
   - The screenshot displays an open window titled “Student Dropout Analysis.”
   - The interface seems to be designed for analyzing student dropout rates of Haryana State.
2. **Visualizations**:
   - Multiple bar graphs are visible, each representing different student categories (possibly based on factors like gender, ethnicity, or grade level).
   - The bar graphs show data for the years 2008, 2009, 2010 and 2011.
   - A pie chart is also present on the left side, divided into distinct segments (possibly representing different reasons for dropout).
   - In this image shows the visualization of Haryana.
3. **Navigation and Options**:
   - Icons and options at the top allow users to navigate through the software features.
   - The right panel provides options for customizing visualizations and accessing data.

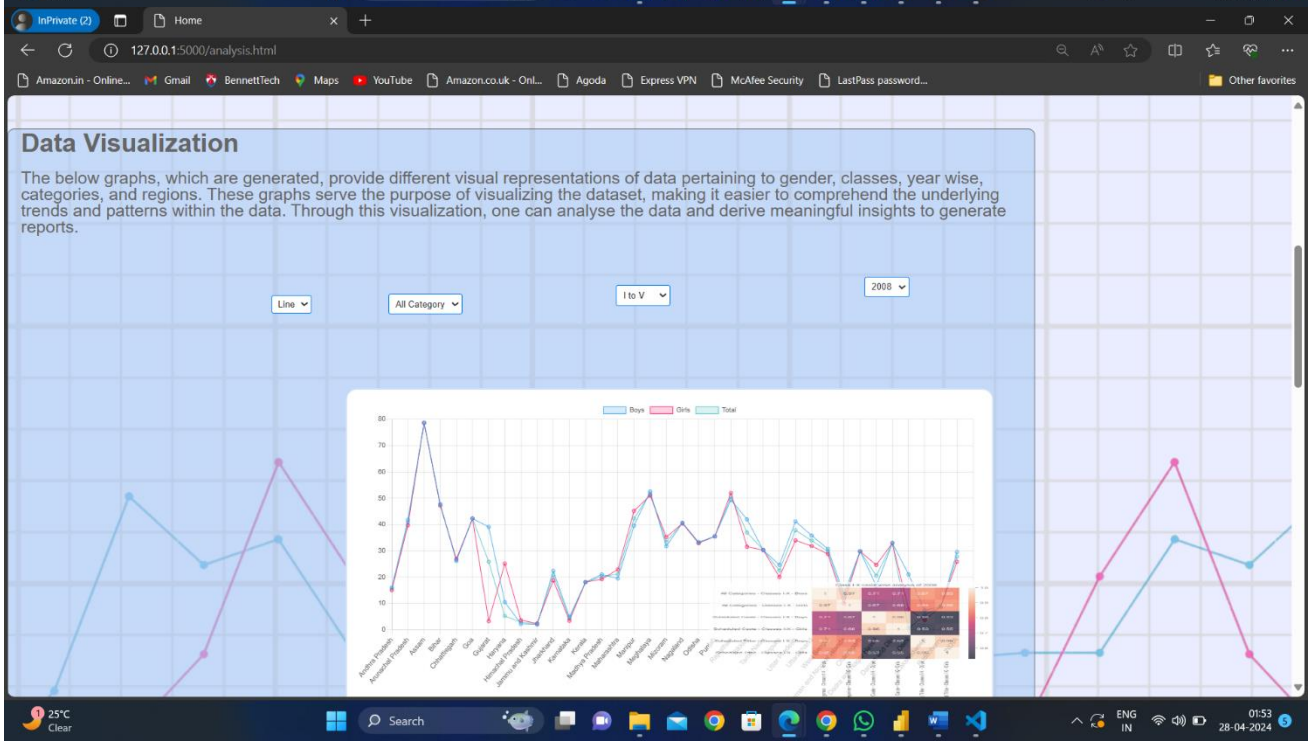
- The Top right corner provides slicer options for select the States and Union Territories for visual analysis
-

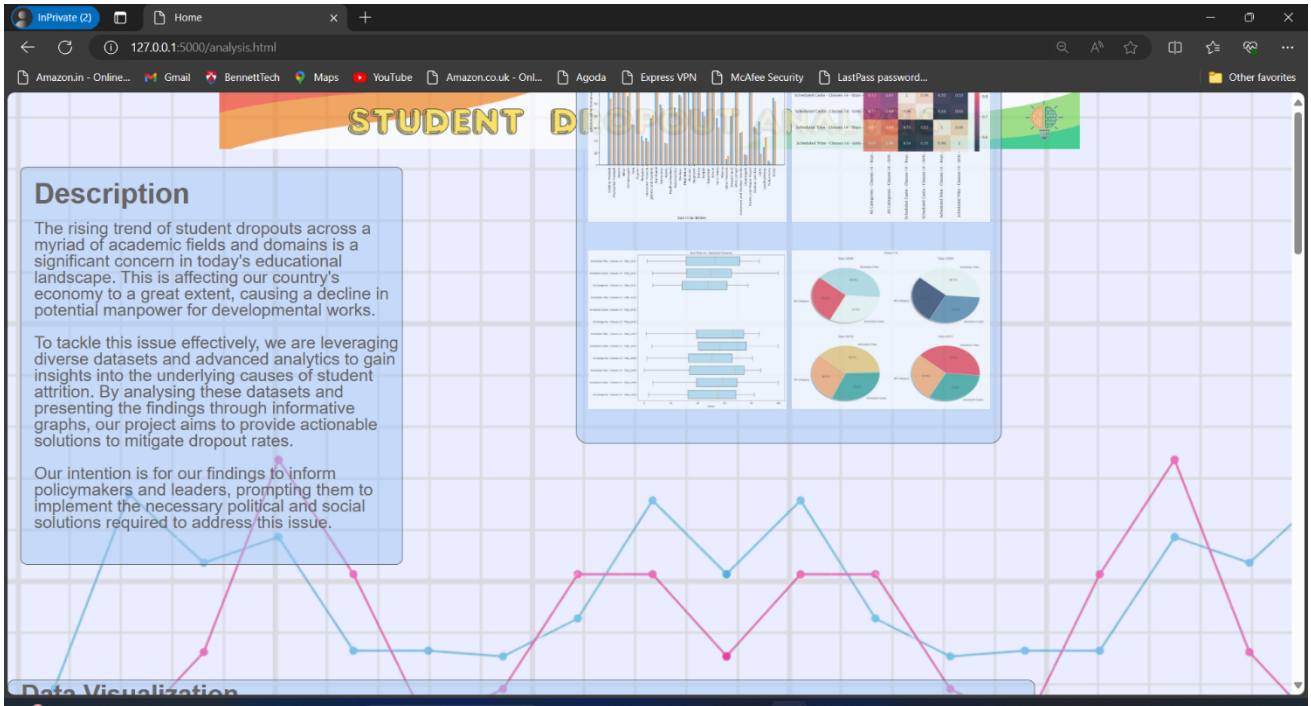
4. **Data Insights**:
    - Unfortunately, the OCR (optical character recognition) output is partially garbled, but it seems to mention various data sets, scheduled times, and categories (e.g., Scheduled Caste).
    - The software likely provides insights into dropout trends, patterns, and potential interventions.

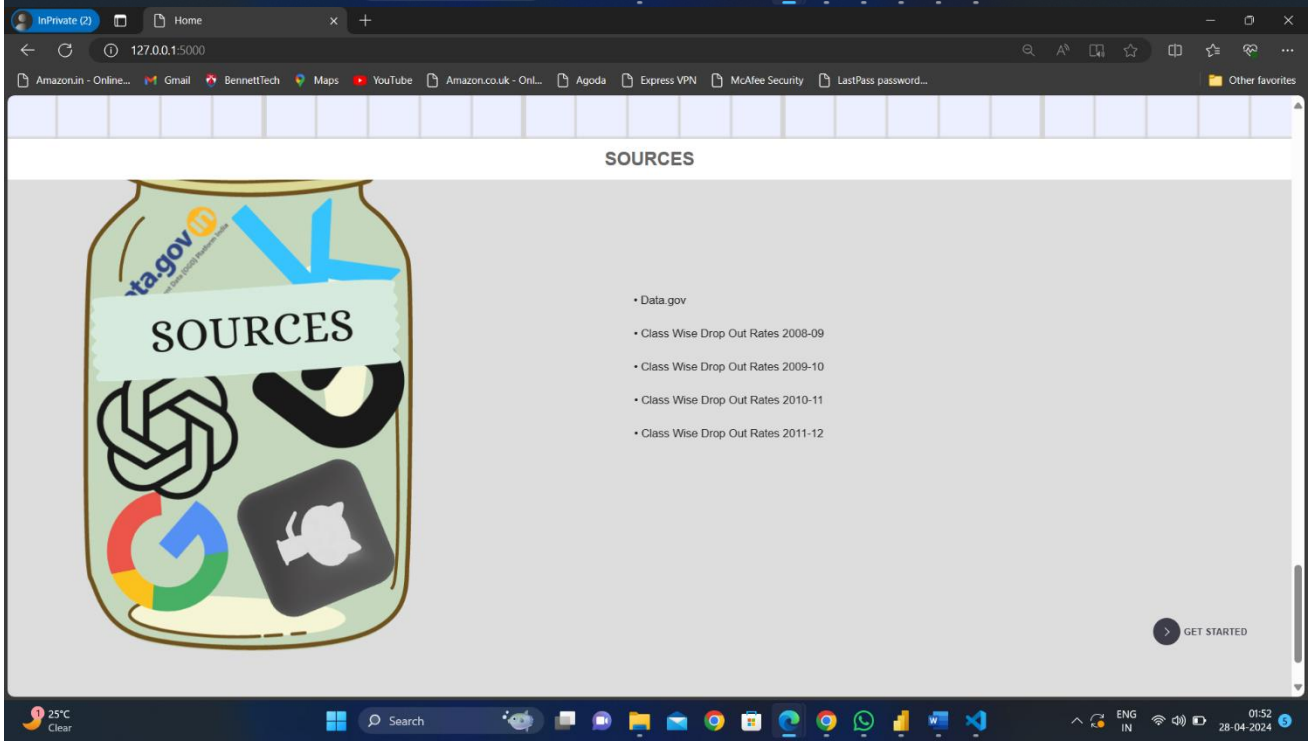
Web Page:

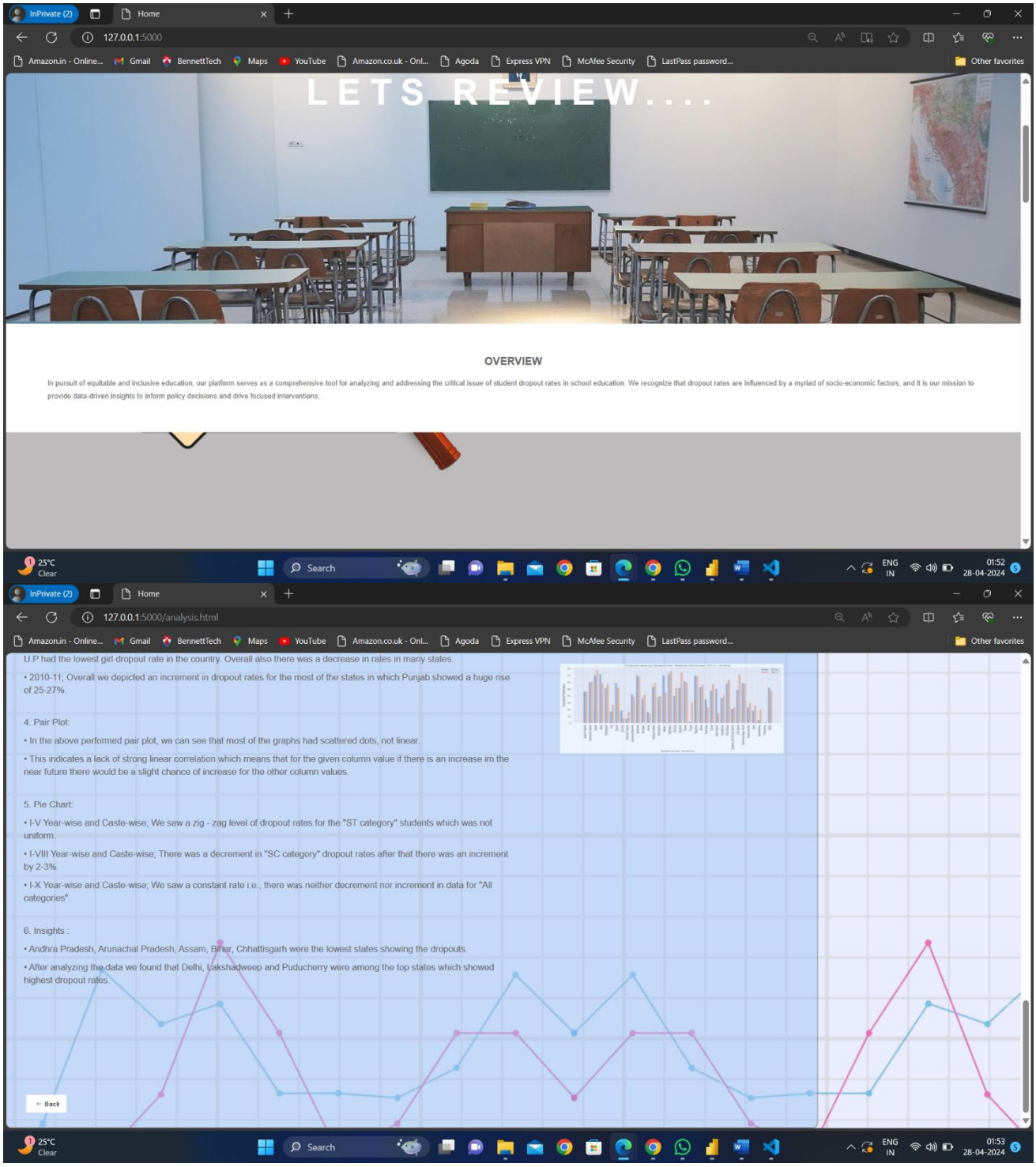
1.


InPrivate (2)
Home
127.0.0.1:5000/analysis.html
Amazon.in - Online...
Gmail
BennettTech
Maps
YouTube
Amazon.co.uk - Onl...
Agoda
Express VPN
McAfee Security
LastPass password...
Other favorites
Analysis Report
In our Analysis, we discovered and explored dropout rates of school education. In the project, we used various methods to extract some valuable information from the dataset which can be used by the government to look into it and implement the necessary tasks.
1. Heat Map:
• In the heat map, in the correllation between "All Categories Boys and Girls" there was a dip in dropout rate in 2011.
• Correlation between "SC category Boys and Girls", there was decrement in dropout rate until 2011 there was a slight spike.
• For the category of "ST Boys and Girls", in the years 2009 and 2011 their were lowest rates.
2. Box Plot:
•Box plot was implemented to check at what rate the outliers were present in the given dataset. After viewing the visualisation there were high discrepency in the way the data was given.
•Box plot was implemented to check at what rate the outliers were present in the given dataset. After viewing the visualisation there were high discrepency in the way the data was given.
3. Bar Graph:
• 2008-09; We took the column of "All category Girls of I-V". And there we observed, between the two years Puducherry and U.P had the lowest girl dropout rate in the country. Overall also there was a decrease in rates in many states.
• 2010-11; Overall we depicted an increment in dropout rates for the most of the states in which Punjab showed a huge rise of 25-27%.
4. Pair Plot:
• In the above performed pair plot, we can see that most of the graphs had scattered dots, not linear.
• This indicates a lack of strong linear correlation which means that for the given column value if there is an increase im the near future there would be a slight chance of increase for the other column values.
5. Pie Chart:
25°C
Clear
Search
ENG
IN
01:53
28-04-2024


InPrivate (2)
Home
127.0.0.1:5000/analysis.html
Amazon.in - Online...
Gmail
BennettTech
Maps
YouTube
Amazon.co.uk - Onl...
Agoda
Express VPN
McAfee Security
LastPass password...
Other favorites
Data Visualization
The below graphs, which are generated, provide different visual representations of data pertaining to gender, classes, year wise, categories, and regions. These graphs serve the purpose of visualizing the dataset, making it easier to comprehend the underlying trends and patterns within the data. Through this visualization, one can analyse the data and derive meaningful insights to generate reports.
Line
All Category
I to V
2008
Boys
Girls
Total
25°C
Clear
Search
ENG
IN
01:53
28-04-2024


InPrivate (2)
Home
127.0.0.1:5000/analysis.html
Amazon.in - Online...
Gmail
BennettTech
Maps
YouTube
Amazon.co.uk - Onl...
Agoda
Express VPN
McAfee Security
LastPass password...
Other favorites
STUDENT
Description
The rising trend of student dropouts across a myriad of academic fields and domains is a significant concern in today's educational landscape. This is affecting our country's economy to a great extent, causing a decline in potential manpower for developmental works.
To tackle this issue effectively, we are leveraging diverse datasets and advanced analytics to gain insights into the underlying causes of student attrition. By analysing these datasets and presenting the findings through informative graphs, our project aims to provide actionable solutions to mitigate dropout rates.
Our intention is for our findings to inform policymakers and leaders, prompting them to implement the necessary political and social solutions required to address this issue.
Data Visualization
25°C
Clear
Search
ENG
IN
01:53
28-04-2024


InPrivate (2)
Home
127.0.0.1:5000
Amazon.in - Online...
Gmail
BennettTech
Maps
YouTube
Amazon.co.uk - Onl...
Agoda
Express VPN
McAfee Security
LastPass password...
Other favorites
SOURCES
SOURCES
• Data.gov
• Class Wise Drop Out Rates 2008-09
• Class Wise Drop Out Rates 2009-10
• Class Wise Drop Out Rates 2010-11
• Class Wise Drop Out Rates 2011-12
GET STARTED
25°C
Clear
Search
ENG
IN
01:52
28-04-2024



1. **Webpage Purpose**:
    - The webpage serves as an interactive platform where stakeholders can explore and understand the data visually.
    - It provides a user-friendly interface for accessing key insights and trends.
2. **Graphical Representations**:

    - Our webpage features various **graphs and charts** that summarize complex data.
    - These visualizations include **bar graphs**, **line charts**, and **pie charts**.
    - Each graph is carefully designed to convey specific information related to student dropout rates.
3. **Interactivity**:
    - Users can **interact** with the graphs by hovering over data points or clicking on specific elements.
    - For instance, clicking on a bar in a bar graph might reveal detailed information about dropout rates for a particular category or year.
4. **Text Explanations**:
    - Alongside the visualizations, we provide **text explanations**.
    - These explanations clarify the context, highlight significant findings, and guide users through the data.
    - Users can gain deeper insights by reading these textual descriptions.
5. **Navigation and Backtracking**:
    - The webpage includes navigation options, such as a **“Back” button**, allowing users to explore different sections.
    - Users can backtrack to previous views or explore related data points easily.

Conclusion:

In conclusion, our analysis of student enrollment and dropout trends in Indian states and union territories from 2008 to 2011 provides valuable insights into the dynamics of the education system in India. Through our examination of data encompassing various categories, including enrollment by gender, class, and social category, we have gained a deeper understanding of the challenges and opportunities within the Indian education landscape.

Our findings highlight the complexity and diversity of the education sector in India, characterized by disparities in access, quality, and outcomes across different regions and demographic groups. While significant progress has been made in expanding access to education, particularly through initiatives such as the Right to Education Act, persistent challenges such as poverty, gender

inequality, and social discrimination continue to hinder educational attainment for many children.

The data also underscore the importance of addressing dropout rates as a critical issue facing the Indian education system. High dropout rates not only impede individual academic and socio-economic progress but also pose broader challenges to national development and social cohesion. Efforts to reduce dropout rates must be multifaceted, encompassing interventions at the policy, institutional, and community levels to address the root causes of dropout and promote retention.

Moving forward, our analysis serves as a foundation for evidence-based policymaking and programmatic interventions aimed at improving educational outcomes and promoting inclusive and equitable access to education for all children in India. By harnessing the insights gleaned from our study, policymakers, educators, and stakeholders can collaborate to design targeted strategies that address the unique needs and challenges of diverse student populations and foster a more inclusive and vibrant education system in India.