Name: Shreyas Kasture **PRN:** 22070521032

Department: CSE **Division:** A

Symbiosis Institute Of Technology, Nagpur

॥वसुधैव कुटुम्बकम्॥

Data Science

CA-1

Computer Science and Engineering Batch 2022-26

Semester-VII **Course Code : 0705210707**

Vidya Insights

Transforming Educational Data into Actionable Insights

Dataset: https://ndap.niti.gov.in/dataset/7057

Github Repository: https://github.com/Shreyas521032/National-Achievement-Survey-learning-outcomes

Live Deployed Project: https://shreyas-national-achievement-survey.streamlit.app/

About the Dataset:

Overview 📈

The **National Achievement Survey (NAS)** is a large-scale assessment conducted across India to evaluate the learning outcomes of students in various classes. This dataset provides a detailed analysis of student performance in key subjects like **Language, Mathematics, Science, and Social Science** for class 8. The NAS survey assesses competencies rather than rote memorization, following a competency-based framework aligned with the National Education Policy (NEP).

Data is collected through standardized tests administered to a representative sample of students. The survey measures students' understanding, application of concepts, and critical thinking skills. The dataset includes performance trends, subject-wise analysis, and comparisons with state and national averages, which can help identify learning gaps. The insights from this data are valuable for designing targeted policies, improving teacher training programs, and addressing learning deficiencies effectively.

Dataset Profile 📊

- **Data Published By:** Ministry of Education
- **Sector:** Education and Training
- **Dataset Hosted By:** National Data and Analytics Platform (NDAP)
- **Geographical Coverage:** State, District
- **Time Granularity:** Yearly
- **Frequency:** Yearly
- **Year Range:** CY 2017 - CY 2021
- **Date Updated:** July 09, 2025

Data Structure and Indicators 📊

The dataset comprises **1405 rows**, with the following structure:

Key Dimensions

- **Location:**
    - Country (1 unique value)

    - State (36 unique values)
    - District (737 unique values)
- **Time:**
    - Year (2 unique values)
- **Other Dimensions:**
    - Class (1 unique value: Class 8)

Key Indicators

The primary indicators in this dataset are:

- **Number of schools surveyed:** The count of schools that participated in the survey within a district.
    - **Min:** 8
    - **Max:** 1380
    - **Mean:** 223.14
- **Number of students surveyed:** The total number of students who were assessed in a district.
    - **Min:** 96
    - **Max:** 36224
    - **Mean:** 5188.01
- **Average performance of students (%):** The dataset contains dozens of indicators measuring the average performance (in percentage) on specific learning outcomes across different subjects. These are calculated as a weighted average based on the total population from the census. Performance varies significantly across different competencies and districts.

Detailed Learning Outcomes 🎯

This section provides a breakdown of each learning outcome assessed in the NAS survey.

Language (L)

- **L813:** Read textual/non-textual materials with comprehension and identify the details, characters, main idea, and sequence of ideas and events.
    - **Significance:** Essential for developing strong reading comprehension and analytical skills.

Mathematics (M)

- **M601:** Solves problems involving large numbers by applying appropriate operations.
- **M606:** Solves problems on daily life situations involving addition and subtraction of fractions/decimals.
- **M620:** Finds out the perimeter and area of rectangular objects in the surroundings.
- **M621:** Arranges given/collected information in the form of a table, pictograph, and bar graph and interprets them.
- **M702:** Interprets the division and multiplication of fractions.
- **M705:** Solves problems related to daily life situations involving rational numbers.

- **M706:** Uses exponential form of numbers to simplify problems involving multiplication and division of large numbers.
- **M707:** Adds/subtracts algebraic expressions.
- **M710:** Solves problems related to the conversion of percentage to fraction and decimal and vice versa.
- **M717:** Finds out the approximate area of closed shapes by using a unit square grid/graph sheet.
- **M719:** Finds various representative values for simple data from daily life contexts like mean, median, and mode.
- **M721:** Interprets data using a bar graph.
- **M801:** Generalizes properties of addition, subtraction, multiplication, and division of rational numbers through patterns.
- **M802:** Finds rational numbers between two given rational numbers.
- **M803:** Proves divisibility rules of 2, 3, 4, 5, 6, 9, and 11.
- **M804:** Finds squares, cubes, square roots, and cube roots of numbers using different methods.
- **M808:** Uses various algebraic identities in solving problems of daily life.
- **M812:** Verifies properties of a parallelogram and establishes the relationship between them through reasoning.
- **M818:** Finds the surface area and volume of cuboidal and cylindrical objects.
- **M819:** Draws and interprets bar charts and pie charts.

Science (SCI)

- **SCI703:** Classifies materials and organisms based on properties/characteristics.
- **SCI704:** Conducts simple investigations to seek answers to queries.
- **SCI705:** Relates processes and phenomena with causes.
- **SCI708:** Measures and calculates temperature, pulse rate, speed of moving objects, etc.
- **SCI710:** Plots and interprets graphs.
- **SCI711:** Constructs models using materials from surroundings and explains their working.
- **SCI801:** Differentiates materials, organisms, and processes.
- **SCI804:** Relates processes and phenomena with causes.
- **SCI805:** Explains processes and phenomena.
- **SCI807:** Measures angles of incidence and reflection, etc.
- **SCI811:** Applies learning of scientific concepts in day-to-day life.
- **SCI813:** Makes efforts to protect the environment.

Social Science (SST)

- **SST605:** Identifies latitudes, longitudes, poles, equator, tropics, and neighboring countries on a globe and map.
- **SST610:** Locates important historical sites and places on an outline map of India.

- **SST625:** Describes the functioning of rural and urban local government bodies in sectors like health and education.
- **SST703:** Explains preventive actions to be undertaken in the event of disasters.
- **SST704:** Describes the formation of landforms due to various factors.
- **SST722:** Explains the significance of equality in a democracy.
- **SST726:** Describes the process of election to the legislative assembly.
- **SST731:** Explains the functioning of media with appropriate examples from newspapers.
- **SST733:** Differentiates between different kinds of markets.
- **SST734:** Traces how goods travel through various marketplaces.
- **SST802:** Describes major crops, types of farming, and agricultural practices in their own area/state.
- **SST805:** Locates the distribution of important minerals (e.g., coal, mineral oil) on the world map.
- **SST807:** Justifies the judicious use of natural resources.
- **SST809:** Draws interrelationships between types of farming and development in different regions of the world.
- **SST810:** Distinguishes the modern period from the medieval and ancient periods through the use of sources.
- **SST815:** Explains the origin, nature, and spread of the revolt of 1857 and the lessons learned from it.
- **SST816:** Analyses the decline of pre-existing industries and the development of new ones during the colonial period.
- **SST818:** Analyses issues related to caste, women, widow remarriage, child marriage, and social reforms.
- **SST823:** Applies the knowledge of Fundamental Rights to find out about their violation, protection, and promotion.
- **SST827:** Describes the process of making a law (e.g., Domestic Violence Act, RTI Act, RTE Act).
- **SST831:** Identifies the role of Government in providing public facilities (water, sanitation, road, electricity).
- **SST833:** Draws a bar diagram to show the population of different countries/India/states.

Project Documentation:

1. Introduction

This project analyzes educational performance data across different districts and states in India. The dataset is based on Annual Status of Education Report (ASER) - like surveys, focusing on student learning outcomes in various subjects. This documentation outlines the process from data loading and cleaning to exploratory data analysis, with a specific focus on data from the Calendar Year 2021.

2. Data Loading

This section details how the raw data was loaded into a pandas DataFrame. The data was read from the CSV file named /content/DATASETDATASCIENCE.csv using the pd.read_csv function. The initial shape of the DataFrame and the first few rows were displayed to confirm successful loading and to get a preliminary look at the data structure.

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

```
# Replace with your actual file name
file_name = '/content/DATASETDATASCIENCE.csv'

try:
    df = pd.read_csv(file_name)
    print(f"Successfully loaded '{file_name}'. It has {df.shape[0]} rows.")
except FileNotFoundError:
    print(f"ERROR: The file '{file_name}' was not found. Please upload it and
    check the name.")

# See the first 5 rows
df.head()
```

```
Successfully loaded '/content/DATASETDATASCIENCE.csv'. It has 1405 rows.
```

	Country	State	District	Year	Class	Number Of Schools Surveyed (UOM:Number), Scaling Factor:1	Number Of Students Surveyed (UOM:Number), Scaling Factor:1	Average Performance Of Students In L813 Learning Outcome (UOM:% (Percentage)), Scaling Factor:1	Average Performance Of Students In M601 Learning Outcome (UOM:% (Percentage)), Scaling Factor:1	Average Performance Of Students In M606 Learning Outcome (UOM:% (Percentage)), Scaling Factor:1	...	Average Performance Of Students In Sst807 Learning Outcome (UOM:% (Percentage)), Scaling Factor:1	Average Performance Of Students In Sst809 Learning Outcome (UOM:% (Percentage)), Scaling Factor:1	Average Performance Of Students In Sst810 Learning Outcome (UOM:% (Percentage)), Scaling Factor:1	Average Performance Of Students In Sst815 Learning Outcome (UOM:% (Percentage)), Scaling Factor:1
0	India	Andaman and Nicobar Islands	Nicobars	Calendar Year (Jan - Dec), 2021	8	80.0	1412.0	37.58	39.72	33.55	...	15.46	24.81	10.79	16.42
1	India	Andaman and Nicobar Islands	North and Middle Andaman	Calendar Year (Jan - Dec), 2021	8	132.0	2428.0	50.02	47.57	43.83	...	33.16	36.51	23.13	33.29
2	India	Andaman and Nicobar Islands	South Andamans	Calendar Year (Jan - Dec), 2021	8	188.0	4588.0	57.06	53.56	45.73	...	35.39	36.60	17.93	29.83
3	India	Andhra Pradesh	Ananthapuramu	Calendar Year (Jan - Dec), 2021	8	264.0	7000.0	47.09	45.60	44.65	...	34.64	29.59	24.51	26.15
4	India	Andhra Pradesh	East Godavari	Calendar Year (Jan - Dec), 2021	8	216.0	5736.0	48.83	43.50	47.11	...	39.76	32.18	21.34	25.52

5 rows × 62 columns

3. Data Cleaning and Preparation

This section describes the crucial steps taken to prepare the data for analysis:

Cleaning Column Names: A custom function clean_col_names was used to simplify the long and complex column names by removing parenthetical information and replacing spaces with underscores. This makes the columns easier to reference in code.

Extracting Year: The year was extracted from the 'Year' column, which contained descriptive text, and converted into an integer format for easier filtering and analysis.

Creating Summary Scores: Average performance scores were calculated for key subjects (Math, Science, and Social Studies - SST) by averaging the scores across the various learning outcome columns related to each subject.

Calculating Overall Performance: An 'Overall_Performance' score was computed for each district by taking the average of the Math, Science, and SST summary scores.

Filtering for 2021 Data: The analysis was focused specifically on the data from the year 2021 by creating a new DataFrame df_2021, which is a subset of the cleaned data.

```python
# --- 2.1: Clean Column Names ---
# This function makes column names shorter and easier to use in code
def clean_col_names(df):
    cols = df.columns
    new_cols = []
    for col in cols:
        new_col = col.split('(')[0].strip().replace(' ', '_')
        new_cols.append(new_col)
    df.columns = new_cols
    return df

df = clean_col_names(df)
df['Year'] = df['Year'].astype(str).str.extract(r'(\d{4})').astype(int)

# --- 2.2: Create Summary Scores for Subjects ---
# We calculate an average score for Math, Science, and Social Studies (SST)
math_cols = [col for col in df.columns if '_In_M' in col]
science_cols = [col for col in df.columns if '_In_Sci' in col]
sst_cols = [col for col in df.columns if '_In_Sst' in col]

df['Math_Performance'] = df[math_cols].mean(axis=1)
df['Science_Performance'] = df[science_cols].mean(axis=1)
df['SST_Performance'] = df[sst_cols].mean(axis=1)
df['Overall_Performance'] = df[['Math_Performance', 'Science_Performance', 'SST_Performance']].mean(axis=1)

# --- 2.3: Focus on the most recent year's data (2021) ---
df_2021 = df[df['Year'] == 2021].copy()

print("Data is now clean and summary scores have been created.")
# Display the new summary columns
df_2021[['State', 'District', 'Overall_Performance', 'Math_Performance', 'Science_Performance']].head()
```

```
Data is now clean and summary scores have been created.
```

	State	District	Overall_Performance	Math_Performance	Science_Performance
0	Andaman and Nicobar Islands	Nicobars	28.996854	31.4395	26.298333
1	Andaman and Nicobar Islands	North and Middle Andaman	36.163354	33.3290	39.883333
2	Andaman and Nicobar Islands	South Andamans	37.321576	34.3445	41.162500
3	Andhra Pradesh	Ananthapuramu	33.874722	35.7450	33.019167
4	Andhra Pradesh	East Godavari	34.892015	35.8615	34.795000

4. Exploratory Data Analysis (EDA)

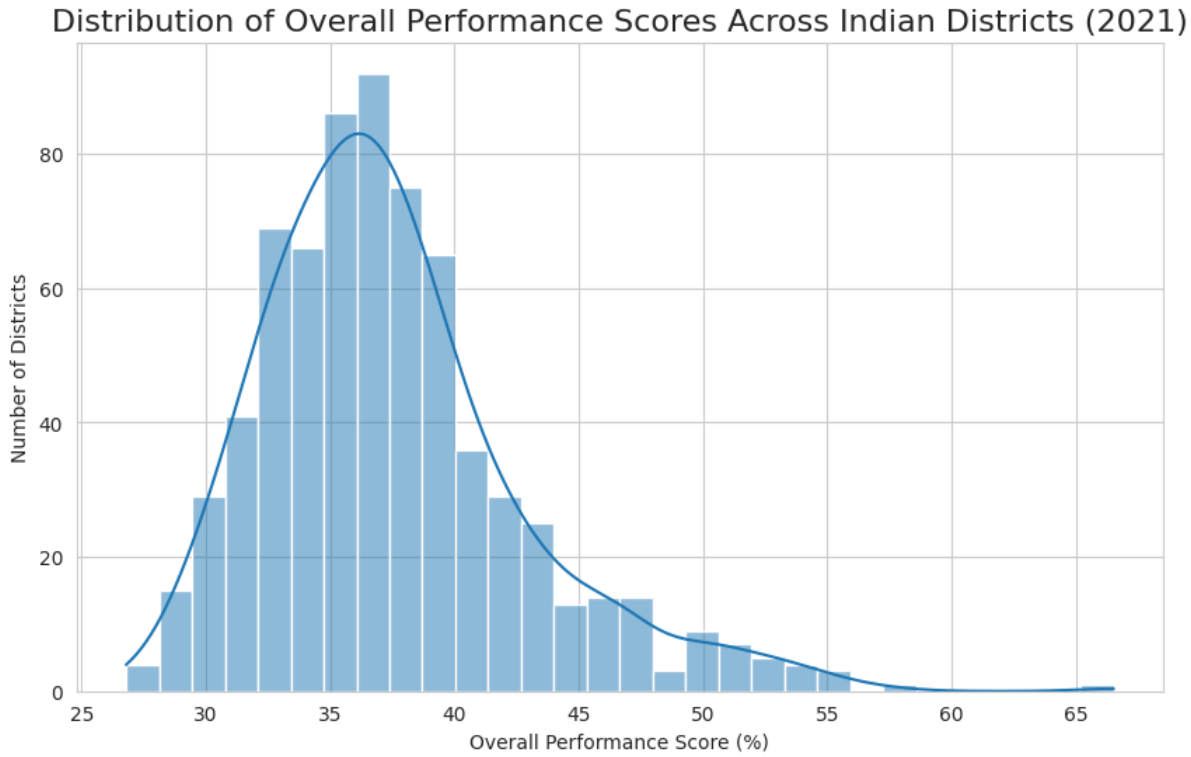
This section presents the initial exploration and visualization of the prepared data:

National Summary Statistics: Descriptive statistics (.describe()) were calculated and displayed for the overall and subject-specific performance scores across all districts in 2021, providing a national overview of performance metrics (mean, standard deviation, min, max, quartiles).

```
--- National Performance Summary (2021) ---
       Overall_Performance  Math_Performance  Science_Performance  \
count           706.000000        706.000000           706.000000
mean             37.425093         35.926949            38.303162
std               5.318112          6.422964             5.090959
min              26.786535         24.211500            26.298333
25%              33.848948         31.366250            34.759375
50%              36.597359         35.014750            37.603333
75%              39.739771         38.692250            41.190625
max              66.518803         72.113000            60.727500

       SST_Performance
count       706.000000
mean         38.045167
std           5.040998
min          27.198182
25%          34.632614
50%          37.393409
75%          40.313523
max          66.715909
```

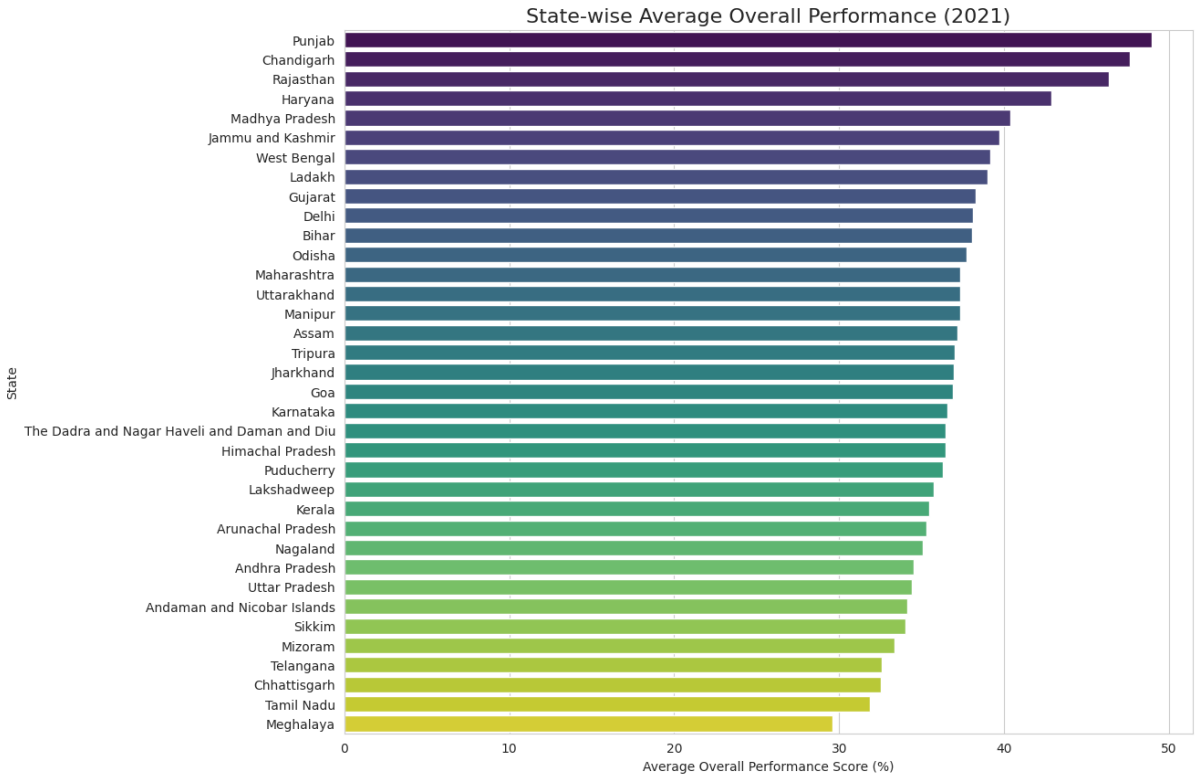
Distribution of Overall Performance: A histogram with a Kernel Density Estimate (KDE) plot was generated to visualize the distribution of 'Overall_Performance' scores across districts, showing the frequency of different performance levels.



State-level Performance Ranking: The average 'Overall_Performance' was calculated for each state, and the states were ranked based on these averages.

```
--- State Rankings by Average Overall Performance (2021) ---
State
Punjab                                          48.984702
Chandigarh                                      47.653646
Rajasthan                                       46.337079
Haryana                                         42.867666
Madhya Pradesh                                  40.397288
Jammu and Kashmir                               39.727462
West Bengal                                     39.146810
Ladakh                                          39.001154
Gujarat                                         38.258212
Delhi                                           38.106105
Bihar                                           38.052472
Odisha                                          37.711416
Maharashtra                                     37.347620
Uttarakhand                                     37.326392
Manipur                                         37.325056
Assam                                           37.165769
Tripura                                         37.003203
Jharkhand                                       36.928697
Goa                                             36.906987
Karnataka                                       36.596513
The Dadra and Nagar Haveli and Daman and Diu    36.451687
Himachal Pradesh                                36.443608
Puducherry                                      36.289088
Lakshadweep                                     35.712894
Kerala                                          35.450517
Arunachal Pradesh                               35.304222
Nagaland                                        35.085922
Andhra Pradesh                                  34.529514
Uttar Pradesh                                   34.438673
Andaman and Nicobar Islands                     34.160594
Sikkim                                          34.001261
Mizoram                                         33.365081
Telangana                                       32.574445
Chhattisgarh                                    32.533783
Tamil Nadu                                      31.881073
Meghalaya                                       29.583090
Name: Overall_Performance, dtype: float64
```

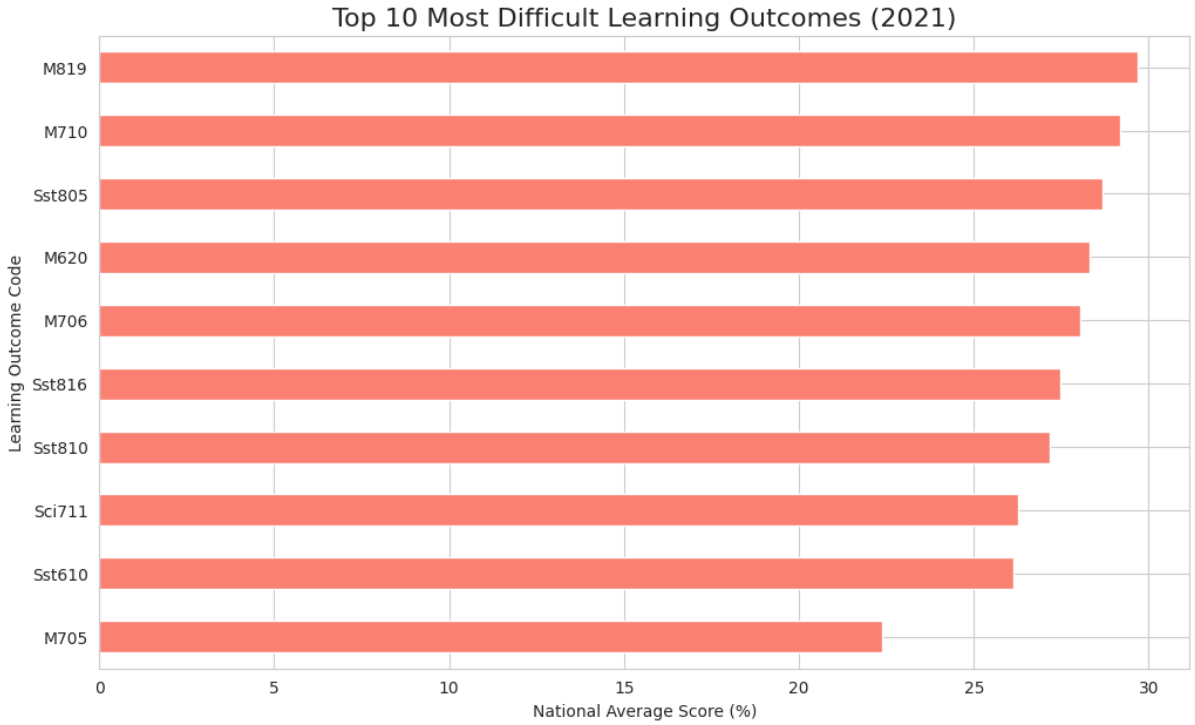
Visualization of State Rankings: A horizontal bar chart was created to visually compare the average overall performance across different states, highlighting the states with the highest and lowest average scores.



Identifying Most Difficult Topics: The 10 learning outcomes with the lowest national average scores were identified as the most difficult topics and presented in a table.

```
--- Top 10 Most Difficult Topics Nationally (Lowest Average Score) ---
M705      22.385850
Sst610    26.116657
Sci711    26.259915
Sst810    27.169533
Sst816    27.474398
M706      28.044292
M620      28.323265
Sst805    28.661197
M710      29.167252
M819      29.681558
dtype: float64
```

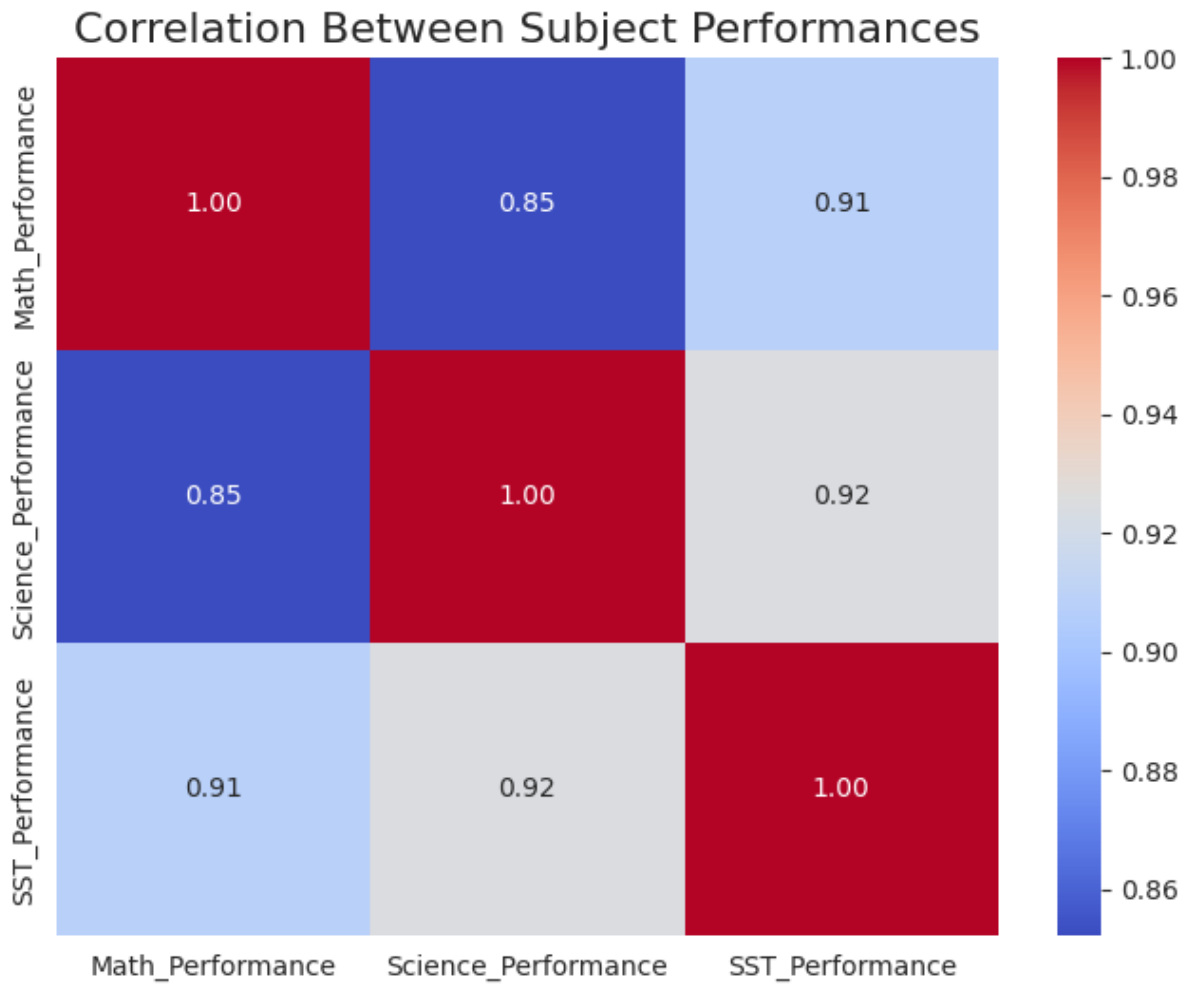
Visualization of Most Difficult Topics: A horizontal bar chart was plotted to visually represent the national average scores for the top 10 most difficult learning outcomes, making it easy to compare their relative difficulty.



Correlation Between Subject Performances: The correlation matrix between 'Math_Performance', 'Science_Performance', and 'SST_Performance' was calculated to understand the relationships between student performance in these subjects.

```
--- Correlation Matrix Between Subjects ---
                   Math_Performance  Science_Performance  SST_Performance
Math_Performance           1.000000             0.852117         0.907612
Science_Performance        0.852117             1.000000         0.924590
SST_Performance            0.907612             0.924590         1.000000
```

Visualization of Correlation Matrix: A heatmap was generated to visually represent the correlation matrix, showing the strength and direction of the relationships between subject scores.



5. Key Findings and Insights

Based on the exploratory data analysis, several key insights were gained:

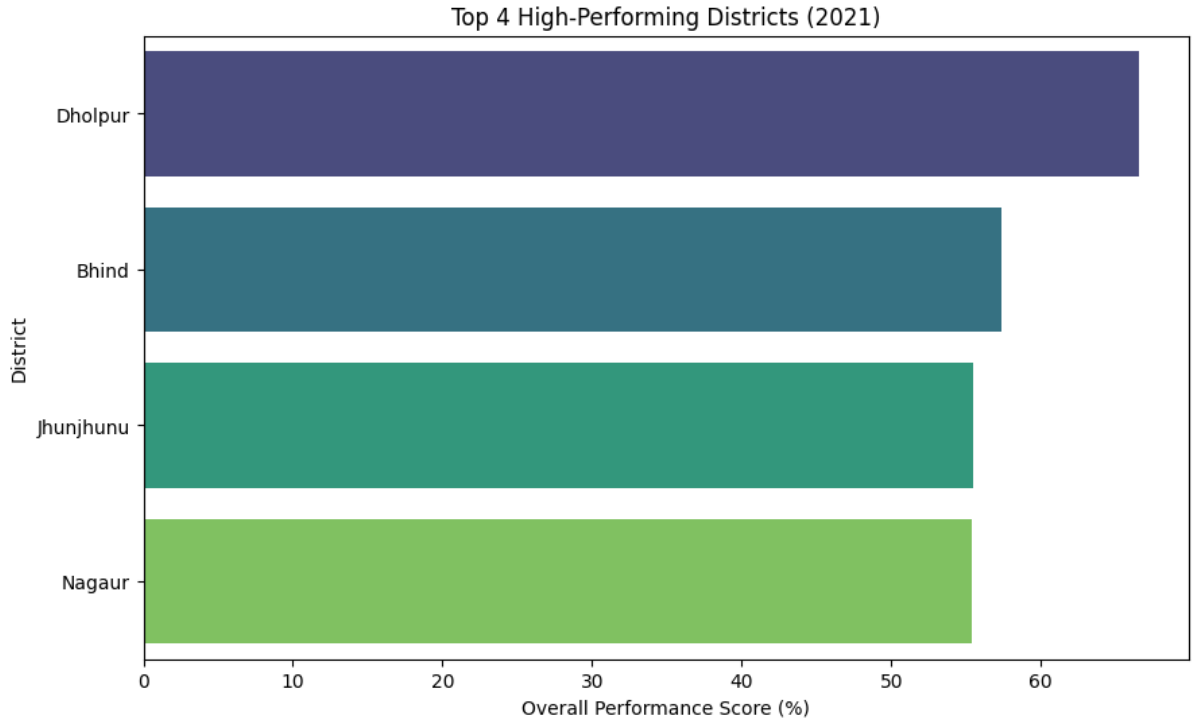
The national average overall performance and the variability across districts were observed from the summary statistics.

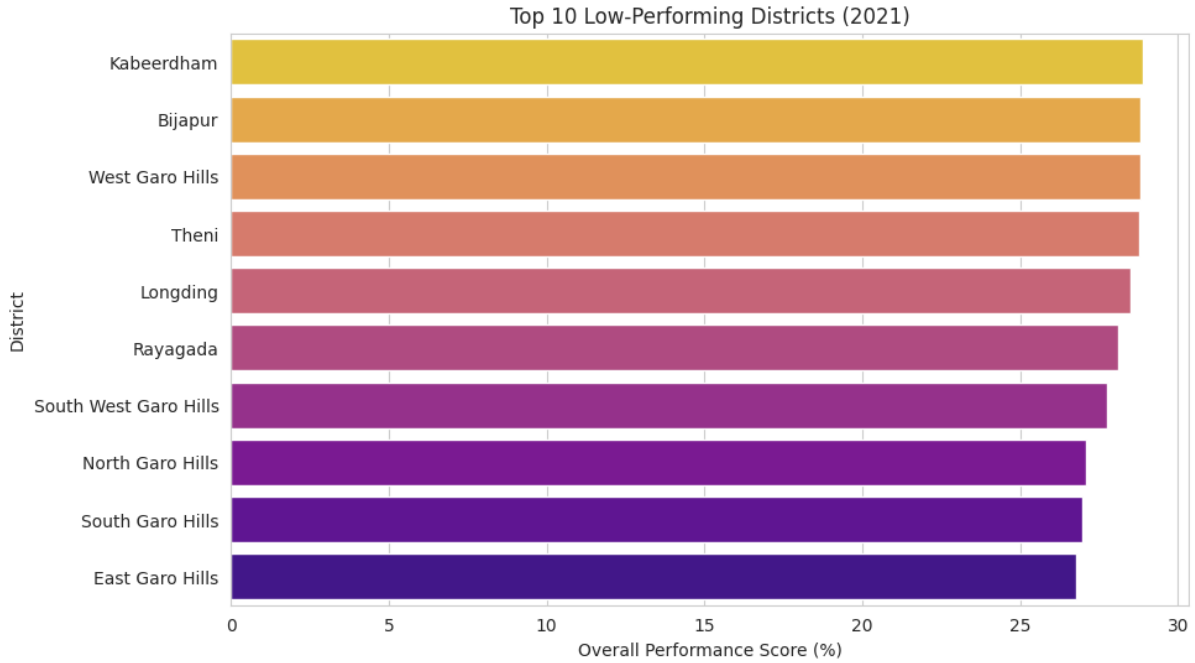
Significant differences in average performance exist across states, with some states consistently performing higher (e.g., Punjab) and others lower (e.g., Meghalaya).

Specific learning outcomes were identified as particularly challenging for students nationally.

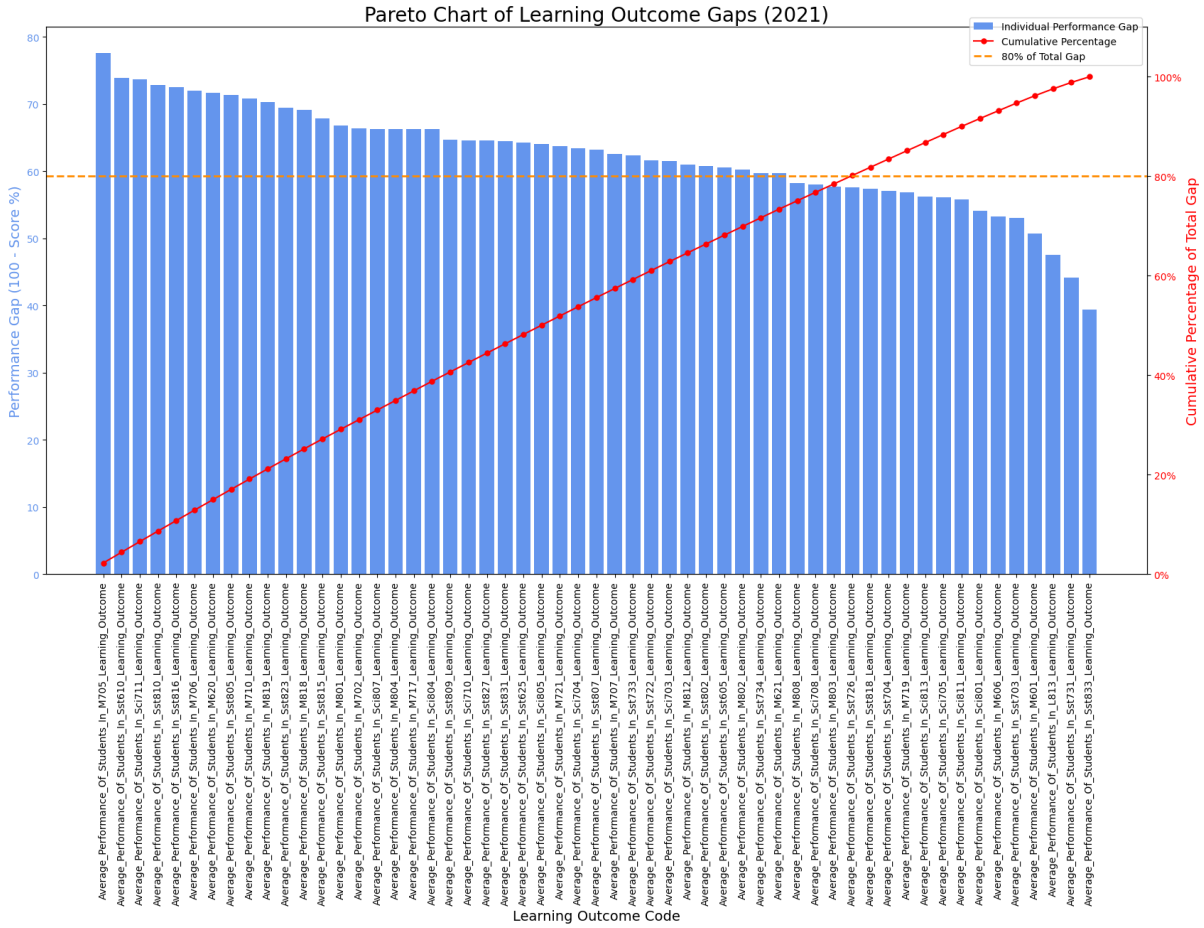
There are strong positive correlations between performance in Math, Science, and SST, suggesting that districts where students perform well in one subject tend to perform well in others.

The analysis allowed for the identification of "bright spot" districts (high-performing outliers) and "hotspot" districts (low-performing outliers) that warrant further investigation.





A Pareto chart was used to visualize the cumulative performance gap across the most difficult learning outcomes, highlighting the few outcomes contributing most to the overall performance deficit.



Problem Statement:

In India's vast and diverse educational landscape, ensuring uniform learning outcomes across regions remains a major challenge. The National Achievement Survey (NAS) highlights significant disparities in student performance across districts and states, with certain subjects and learning outcomes consistently underperforming despite national educational reforms. However, the current analysis often stops at descriptive statistics, lacking predictive foresight. There is a critical need for an intelligent, data-driven solution that can not only identify low-performing regions but also **predict future student performance** based on factors like the number of schools, students surveyed, past performance trends, and subject-specific competencies. By leveraging machine learning techniques on the NAS dataset, we can develop predictive models that forecast educational outcomes, uncover key drivers of performance gaps, and inform targeted interventions, ultimately helping policymakers, educators, and stakeholders design effective, evidence-based strategies to improve learning outcomes and bridge regional disparities in education.

Conclusion:

The Vidya Insights project effectively transforms complex educational data from the National Achievement Survey (NAS) into meaningful insights, focusing on student learning outcomes across India. Through comprehensive data cleaning, analysis, and visualization, the study reveals significant disparities in district and state-level performance, highlights critical learning gaps, especially in subjects like Math and Science and uncovers key learning outcomes that pose challenges nationwide. The strong correlations between subject performances suggest systemic factors influencing student achievement. These findings can guide targeted educational interventions, policy reforms, and resource allocation to improve learning outcomes, making this project a vital tool for data-driven educational planning and reform.