DMS692
Advanced Statistical Methods for Business Analytics

Project 2
Examining Changes in Educational Outcomes in Indian Schools (2017 & 2021)

Group No: 8

Team Members:

Bhavishya Gupta [220295]

Rishav Raj [220889] Rohan Nimesh [220907]

1. Introduction

Education is a critical determinant of a nation's development and prosperity. In India, with its vast and diverse population, monitoring educational outcomes is essential for formulating effective policies and interventions. Periodic assessment of academic performance helps identify trends, gaps, and areas of improvement in the education system.

This study aims to examine how academic performance in Indian schools has changed between 2017 and 2021, a period that includes the COVID-19 pandemic, which significantly disrupted educational activities worldwide. By analyzing performance metrics across various subjects and states, we seek to gain insights into the resilience and adaptability of the Indian education system during this challenging period.

2. Research Hypotheses

Based on our research objectives, we formulate the following hypotheses:

- **Primary Hypothesis:** There is a significant difference in academic performance metrics between 2017 and 2021 across Indian schools.
- **Secondary Hypothesis:** The trends in academic performance vary significantly across different states in India, suggesting regional disparities in educational outcomes.
- **Tertiary Hypothesis:** Performance metrics across different subjects are correlated and can be reduced to fewer underlying factors that represent broader academic competencies.

3. Data

3.1 Data source

The data used in this study come from a national educational assessment database, which includes performance metrics for students in class 8 in various states and districts in India for 2017 and 2021.

3.2 Dataset Overview

The dataset consists of 1450 observations and 14 variables, with no missing values. Each observation represents a district's average performance in a particular year.

3.3 Variables Description

- **Country**: All observations are from India (constant)
- **State**: The state or union territory within India
- **District**: The district within the state (732 unique values)
- **Year**: Calendar year of assessment (2017 or 2021)
- **Class**: All observations are for Class 8 students (constant)
- **Schools_Surveyed**: Number of schools included in the survey for that district
- **Students_Surveyed**: Number of students surveyed in that district
- **Performance Metrics**:
  - **AVG_L813**: Average score in Language (range: 13.82–76.36)
  - **AVG_M601**: Average score in Mathematics-I (range: 12.08–85.82)
  - **AVG_SCI703**: Average score in Science-I (range: 10.98–77.03)
  - **AVG_SST605**: Average score in Social Studies-I (range: 10.70–89.44)
  - **AVG_M801**: Average score in Mathematics-II (range: 6.25–69.93)
  - **AVG_SCI801**: Average score in Science-II (range: 8.75–81.11)
  - **AVG_SST704**: Average score in Social Studies-II (range: 3.24–82.77)

3.4 Data Cleaning and Preparation

The dataset was checked for missing values and none were found. A new variable `Year_Simple` was created to simplify the year format from "Calendar Year (Jan - Dec), 2017" to "2017" and "Calendar Year (Jan-Dec), 2021" to "2021" for easier analysis.

3.5 Descriptive Statistics

The performance metrics show various ranges and distributions:

Metric	Minimum	Mean	Median	Std Dev	Maximum
AVG_L813	13.82	53.75	54.01	8.15	76.36
AVG_M601	12.08	43.91	43.36	10.80	85.82
AVG_SCI703	10.98	40.25	39.19	7.91	77.03
AVG_SST605	10.70	43.69	41.96	11.47	89.44
AVG_M801	6.25	32.17	31.37	7.16	69.93
AVG_SCI801	8.75	47.96	47.38	9.51	81.11
AVG_SST704	3.24	40.89	40.92	12.55	82.77

Table 1: Descriptive Statistics for Performance Metrics

3.6 Data Visualizations

Several visualizations were created to better understand the data:

1. **Correlation Matrix**: Shows relationships between different performance metrics. Strong correlations (above 0.7) were observed between `AVG_L813`, `AVG_SCI703`, and `AVG_SCI801`.

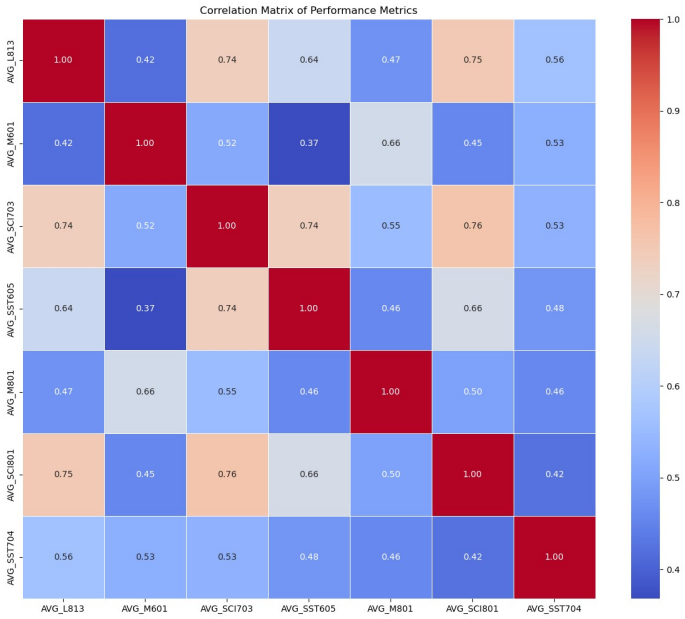
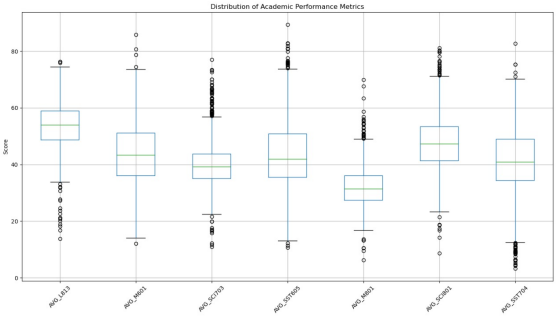
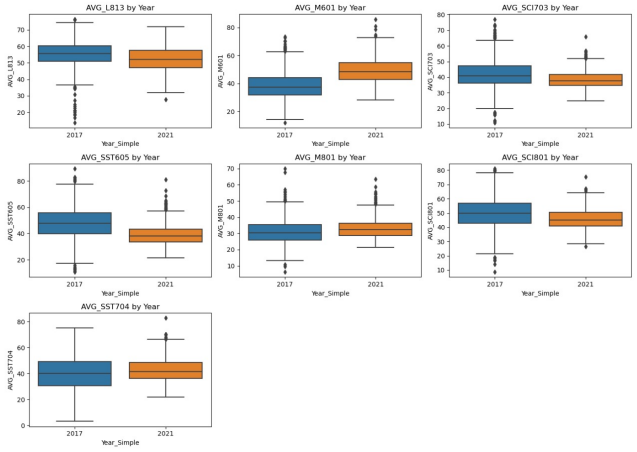


Figure 1: Correlation Matrix of Performance Metrics

2. **Boxplots**: Displays the distribution of each performance metric, revealing varying ranges and potential outliers.



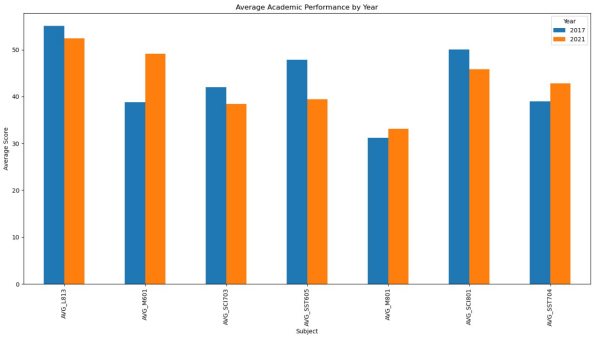
(a) Distribution of Academic Performance Metrics



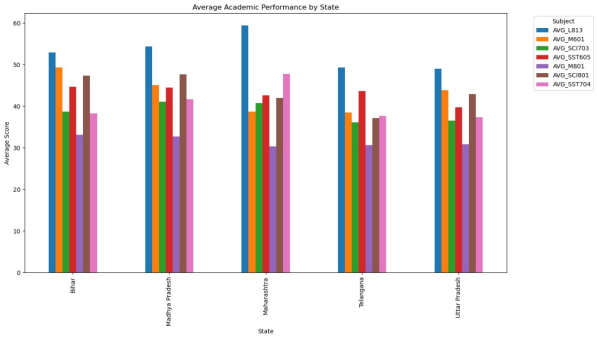
(b) Yearly Trends in Subject Performance Metrics: 2017 vs 2021

Figure 2: Comparison of Performance Metrics by year and overall distribution

3. **Bar Charts**: Comparison of average performance by year and by state, highlighting temporal and regional differences.



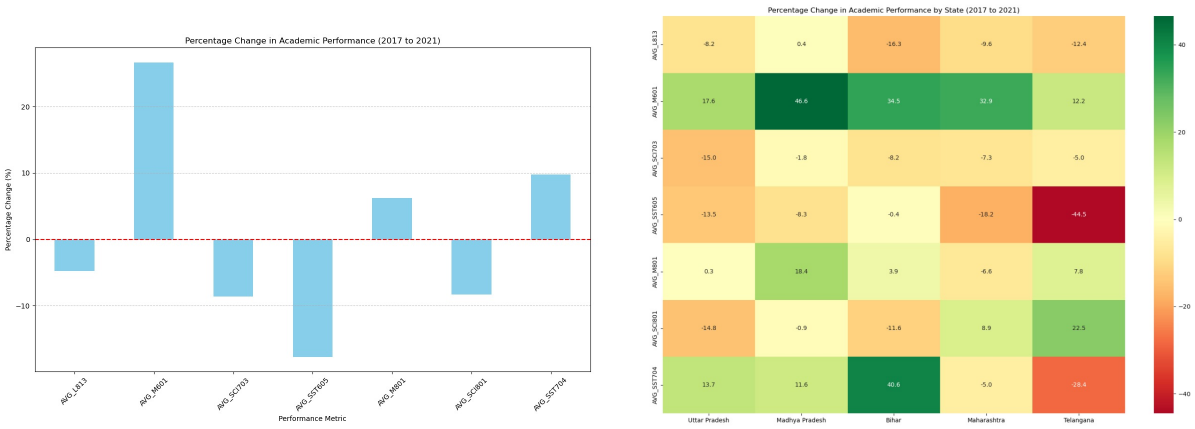
(a) Average Performance by Year



(b) Average Performance by State

Figure 3: Average Performance Change Across Time and Region

4. **Percentage Changes Plots:** Comparison of percentage change in academic performance by year and state, highlighting temporal and region differences.



(a) Percentage Change by Year (2017→2021) (b) Percentage Change by State (2017→2021)

Figure 4: Percentage Changes in Academic Performance Metrics : Time vs. State

Key Observations:
Figure 4.a: Percentage Change by Year (2017→2021)

- Three subjects showed improvement: AVG_M601 (+26.67%), AVG_M801 (+6.19%), AVG_SST704 (+9.73%).
- Four subjects showed decline: AVG_L813 (–4.76%), AVG_SCI703 (–8.63%), AVG_SST605 (–17.73%), AVG_SCI801 (–8.37%).

Figure 4.b: Percentage Change by State (2017→2021)

Table 2: State-wise Percentage Change in Academic Performance (2017–2021)

Metric	UP	MP	Bihar	Maharashtra	Telangana
AVG_L813	-8.16%	+0.42%	-16.26%	-9.56%	-12.39%
AVG_M601	+17.57%	+46.55%	+34.45%	+32.94%	+12.17%
AVG_SCI703	-15.05%	-1.78%	-8.24%	-7.27%	-5.01%
AVG_SST605	-13.49%	-8.28%	-0.44%	-18.20%	-44.54%
AVG_M801	+0.27%	+18.41%	+3.94%	-6.62%	+7.84%
AVG_SCI801	-14.77%	-0.89%	-11.61%	+8.86%	+22.49%
AVG_SST704	+13.70%	+11.60%	+40.62%	-4.99%	-28.37%

- Madhya Pradesh exhibited the most positive changes across several metrics, especially in mathematics (AVG_M601: +46.55%).
- Telangana displayed mixed results, with strong improvements in some metrics (e.g., AVG_SCI801: +22.49%) but severe declines in others (e.g., AVG_SST605: –44.54%).

4. Methodology

To address our research hypotheses, we employed several multivariate statistical methods: Principal Component Analysis (PCA), Factor Analysis, Multivariate Analysis of Variance (MANOVA), and Analysis of Variance (ANOVA). Prior to applying these methods, we conducted assumption tests to ensure their validity which will be the next point in the report.

4.1 Principal Component Analysis (PCA)

PCA was used to reduce the dimensionality of the seven academic performance metrics and uncover underlying patterns.

The PCA model is expressed as:

$$X = WP + \varepsilon$$

Where:

- X: Matrix of standardized academic performance metrics
- W: Matrix of loadings (eigenvectors)
- P: Matrix of principal components
- ε: Residual error matrix

4.2 Factor Analysis

Factor Analysis was applied to model latent constructs underlying correlations among performance metrics. The model is:

$$X_i = \lambda_{i1}F_1 + \lambda_{i2}F_2 + \ldots + \lambda_{im}F_m + \varepsilon_i$$

Where:

- X_i: The i-th observed variable
- λ_{ij}: Loading of X_i on the j-th factor
- F_j: The j-th common factor
- ε_i: The unique factor (residual error) for X_i

4.3 Multivariate Analysis of Variance (MANOVA)

MANOVA was used to examine whether the mean vectors of the performance metrics differed significantly across:

- Year (2017 vs. 2021)
- State

- Interaction between Year and State

The MANOVA model is defined as:

$$Y = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Where:

- Y: Matrix of dependent variables (academic metrics)
- μ: Overall mean vector
- α_i: Effect of the i-th level of factor A (Year)
- β_j: Effect of the j-th level of factor B (State)
- $(\alpha\beta)_{ij}$: Interaction effect between Year and State
- ε_{ijk}: Residual error term

4.4 Analysis of Variance (ANOVA)

To investigate subject-wise year effects, we conducted one-way ANOVAs for each academic metric:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Where:

- Y_{ij}: Observation j in year group i
- μ: Overall mean
- α_i: Fixed effect of year
- ε_{ij}: Residual error

4.5 Assumptions Testing Summary

Method/Assumption	Test Used	Status	Evidence
MANOVA			
Multivariate Normality	Shapiro-Wilk	✓*	Mixed results, large sample size (N=1450) provides robustness
Homogeneity of Covariance	Box-M Test	✓*	M = 201.4, p <0.001
Independence	Study design	✓	District-level measures are independent
No Multicollinearity	Correlation matrix	✓	Present but below 0.9 threshold
ANOVA			
Normality	Shapiro-Wilk	✓*	Some non-normality in 2017 data, large sample size provides robustness
Homogeneity of Variance	Levene's test	✓*	Mixed results across metrics
Independence	Study design	✓	Random sampling at district level
Absence of Outliers	Boxplots	✓	Limited influence of outliers

✓= Met, ✓* = Partially met but analysis robust to violations

Method/Assumption	Test Used	Status	Evidence
PCA			
Linearity	Correlation matrix	✓	Strong correlations between variables (r >0.7)
Sampling Adequacy	KMO test	✓	KMO = 0.862 (excellent)
Suitability	Bartlett's test	✓	χ^2 = 6276.29, p <0.001
Variance Explained	PC analysis	✓	First PC: 62.31%, First 3 PCs: 83.77%
Factor Analysis			
Sample Size	N >150	✓	N = 1450, far exceeding requirements
Correlation Structure	Correlation matrix	✓	Values range 0.37 - 0.76, ideal for the factor analysis
Factorability	KMO & Bartlett's	✓	KMO = 0.862, Bartlett's: p <0.001
Variable Distribution	Visual inspection	✓*	Some normality violations exist but not critical

✓= Met, ✓* = Partially met but analysis robust to violations

5. Results

5.1 Principal Component Analysis Results

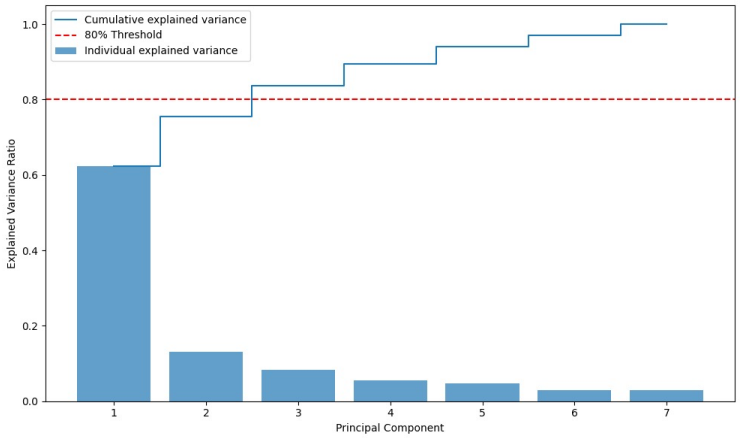


Figure 5: Explained Variance By Principal Components

The PCA revealed that the first principal component explains 62.31% of the total variance in the performance metrics. The first two components together explain 75.42% of the variance, and the first three components account for 83.77%.

Metric	PC1	PC2	PC3
AVG_L813	0.844	-0.272	-0.104
AVG_M601	0.696	0.594	0.099
AVG_SCI703	0.889	-0.197	0.080
AVG_SST605	0.798	-0.329	0.006
AVG_M801	0.731	0.454	0.311
AVG_SCI801	0.840	-0.290	0.208
AVG_SST704	0.710	0.233	-0.646

Table 3: PCA Component Loadings

The first principal component captures overall academic proficiency, with strong positive loadings across all subjects.The second principal component distinguishes mathematics from language and science, highlighting a math versus non-math dimension.The third principal component contrasts social studies with advanced math and science, reflecting a trade-off between these domains.

5.2 Factor Analysis Results

The Bartlett test of sphericity yielded a chi-square value of 6276.29 ($p < 0.001$), indicating sufficient correlation among variables for factor analysis. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.862, which is considered excellent.

Using the Kaiser criterion (eigenvalues >1), one factor was retained, explaining approximately 62% of the total variance.

Table 4: Factor Loadings (1-Factor Model)

Metric	**Factor 1**
AVG_L813	-0.822
AVG_M601	-0.625
AVG_SCI703	-0.890
AVG_SST605	-0.759
AVG_M801	-0.666
AVG_SCI801	-0.817
AVG_SST704	-0.643

The negative signs are due to the rotation method and do not affect interpretation. The single factor reflects general academic performance across all subjects.

5.3 MANOVA Results

Effect of Year on Academic Performance

Table 5: MANOVA Tests For Year

Intercept	Value	Num DF	Den DF	F Value	p-value
Wilks' lambda	0.0374	7.000	1442.000	5303.0749	$< .0001$
Pillai's trace	0.9626	7.000	1442.000	5303.0749	$< .0001$
Hotelling–Lawley trace	25.7431	7.000	1442.000	5303.0749	$< .0001$
Roy's greatest root	25.7431	7.000	1442.000	5303.0749	$< .0001$
Year_Simple	**Value**	**Num DF**	**Den DF**	**F Value**	**p-value**
Wilks' lambda	0.3548	7.000	1442.000	374.5329	$< .0001$
Pillai's trace	0.6452	7.000	1442.000	374.5329	$< .0001$
Hotelling–Lawley trace	1.8181	7.000	1442.000	374.5329	$< .0001$
Roy's greatest root	1.8181	7.000	1442.000	374.5329	$< .0001$

Interpretation:

- As p-values are less than 0.001 across all MANOVA tests, indicating that the year has a statistically significant impact on the combined performance metrics.
- This indicates that there is a significant difference in academic performance between 2017 and 2021

Effect of State on Academic Performance

Table 6: MANOVA Tests for State (Top 5 States Shown)

Intercept	Value	Num DF	Den DF	F Value	p-value
Wilks' lambda	0.0699	7.000	455.000	865.0918	$< .0001$
Pillai's trace	0.9301	7.000	455.000	865.0918	$< .0001$
Hotelling–Lawley trace	13.3091	7.000	455.000	865.0918	$< .0001$
Roy's greatest root	13.3091	7.000	455.000	865.0918	$< .0001$
State	Value	Num DF	Den DF	F Value	p-value
Wilks' lambda	0.2703	28.000	1641.948	25.6457	$< .0001$
Pillai's trace	0.9699	28.000	1832.000	20.9429	$< .0001$
Hotelling–Lawley trace	1.8845	28.000	1127.787	30.5423	$< .0001$
Roy's greatest root	1.3844	7.000	458.000	90.5774	$< .0001$

Interpretation:

- As p-values are less than 0.001 across all MANOVA tests, indicating that the state (top-5 states) has a statistically significant impact on the combined performance metrics.
- This suggests that academic performance varies significantly across different states.

Interaction Effect of Year and State on Academic Performance

Table 7: MANOVA Tests for Year × State Interaction

Intercept	Value	Num DF	Den DF	F Value	p-value
Wilks' lambda	0.1127	7.000	450.000	506.2854	$< .0001$
Pillai's trace	0.8873	7.000	450.000	506.2854	$< .0001$
Hotelling–Lawley trace	7.8756	7.000	450.000	506.2854	$< .0001$
Roy's greatest root	7.8756	7.000	450.000	506.2854	$< .0001$
Year_Simple:State	Value	Num DF	Den DF	F Value	p-value
Wilks' lambda	0.2036	28.000	1623.9203	32.1791	$< .0001$
Pillai's trace	1.0866	28.000	1812.000	24.1351	$< .0001$
Hotelling–Lawley trace	2.6442	28.000	1115.2875	42.3837	$< .0001$
Roy's greatest root	2.1592	7.000	453.000	139.7337	$< .0001$

Interpretation:

- As p-values are less than 0.001 across all MANOVA tests, indicating that the interaction between state and year is also significant.
- This indicates that the change in academic performance from 2017 to 2021 varied across states.

5.4 ANOVA Results for Individual Performance Metrics

ANOVA - Effect of Year on AVG_L813:

Source	Sum of Squares	df	F	PR(¿F)
Year_Simple	2484.490750	1.0	38.369148	< 0.001
Residual	93761.337536	1448.0	NaN	NaN

Table 8: ANOVA Results for AVG_L813

ANOVA - Effect of Year on AVG_M601:

Source	Sum of Squares	df	F	PR(¿F)
Year_Simple	38800.867268	1.0	431.584552	< 0.001
Residual	130179.950941	1448.0	NaN	NaN

Table 9: ANOVA Results for AVG_M601

ANOVA - Effect of Year on AVG_SCI703:

Source	Sum of Squares	df	F	PR(¿F)
Year_Simple	4767.466244	1.0	80.356247	< 0.001
Residual	85908.580330	1448.0	NaN	NaN

Table 10: ANOVA Results for AVG_SCI703

ANOVA - Effect of Year on AVG_SST605:

Source	Sum of Squares	df	F	PR(¿F)
Year_Simple	26131.466012	1.0	230.004877	< 0.001
Residual	164511.132531	1448.0	NaN	NaN

Table 11: ANOVA Results for AVG_SST605

ANOVA - Effect of Year on AVG_M801:

Source	Sum of Squares	df	F	PR(¿F)
Year_Simple	1351.463155	1.0	26.829567	< 0.001
Residual	72938.883028	1448.0	NaN	NaN

Table 12: ANOVA Results for AVG_M801

ANOVA - Effect of Year on AVG_SCI801:

Source	Sum of Squares	df	F	PR(¿F)
Year_Simple	6348.484768	1.0	73.675329	< 0.001
Residual	124771.833483	1448.0	NaN	NaN

Table 13: ANOVA Results for AVG_SCI801

ANOVA - Effect of Year on AVG_SST704:

Source	Sum of Squares	df	F	PR(¿F)
Year_Simple	5228.327975	1.0	33.922424	< 0.001
Residual	223174.466301	1448.0	NaN	NaN

Table 14: ANOVA Results for AVG_SST704

Interpretation:

- All p-values are less than 0.001, indicating that the year has a statistically significant impact on each of the performance metrics (AVG_L813, AVG_M601, AVG_SCI703, AVG_SST605, AVG_M801, AVG_SCI801, and AVG_SST704).
- This suggests that the differences in performance across the years are unlikely to be due to chance and are meaningful in terms of the factors associated with each metric.

6. Discussion

Interpretation of Findings

- **Overall Trends:** The data confirms our primary hypothesis — academic performance significantly changed between 2017 and 2021. The direction of change varied by subject.
- **Subject-Specific Patterns:** Mathematics-I showed the strongest improvement (+26.67%), suggesting effective interventions. Social Studies-I declined the most (-17.73%), which requires further investigation.
- **Regional Disparities:** MANOVA and ANOVA results support secondary hypothesis, showing significant performance differences by state and interactions over time.
- **Latent Factors:** PCA and factor analysis support the tertiary hypothesis - a dominant factor explains most variance in scores, representing general academic ability.
- **COVID-19 Impact:** Changes likely reflect disruptions caused by the pandemic. Math may have benefited from self-learning, while science and social studies declined due to practical/lab components.

Implications for Education Policy

- **Support Declining Subjects:** Interventions needed in Language, Science-I, Science-II and Social Studies-I.
- **Learn from Math Success:** Extend best practices of math education to the other subjects.
- **Reduce Regional Inequities:** Offer targeted resources to underperforming states.
- **Build Resilience:** Future policies should prepare for disruptions with flexible delivery for all subjects.

Limitations

- Contextual factors like policy shifts or socio economics were not included.
- The sample, though large, may not represent all Indian districts equally.

Future Research Directions

- Longitudinal studies can track ongoing post-pandemic trends.
- Mixed-method studies can explain the qualitative drivers behind shifts.
- Policy evaluations can link reforms to performance outcomes.
- Integrating socioeconomic data can clarify the underlying inequalities.

Conclusion

This study demonstrates significant changes in academic performance between 2017 and 2021 among Indian Class 8 students. The subject-specific and regional trends provide critical insights for educational policy and suggest targeted strategies to strengthen academic outcomes in the future.

References

- GitHub Repository: Code used for generating results and visualizations.
- Dataset Source: NDAP – National Data & Analytics Platform (NITI Aayog).
- Johnson, R. A., & Wichern, D. W. (2018). *Applied Multivariate Statistical Analysis* (7th ed.). Pearson. ISBN: 978-0134995397.