Data Science Toolbox: Python Programming

PROJECT REPORT

(Project Semester January-April 2025)

(District-Level Performance Grading Index (PGI) 2021-22: An Exploratory Data Analysis)

Submitted by

Medha Jha

Registration No: 12305679

Section: K23SM

Course Code: INT375

Under the Guidance of

Dr. Mrinalini Rana Mam

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

<u>CERTIFICATE</u>

This is to certify that Medha Jha bearing Registration no. 12305679 has completed INT375 project titled, **“**Health Infrastructure During COVID 19**”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor:

Designation of the Supervisor:

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 12-04-25

DECLARATION

I, Medha Jha, student of B-Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-25 Signature:

Registration No. 12305679

Name of the student: Medha Jha

1. Introduction

1.1. Problem Statement

To what extent does the District-Level Performance Grading Index (PGI) data for 2021-22 reveal disparities in educational performance across different districts in India, and what targeted interventions can be designed to promote equitable access to quality education for all students, regardless of their geographic location?

1.2. Description

- **Genesis and Evolution of PGI:** The Performance Grading Index (PGI) was introduced by the Department of School Education and Literacy (DoSEL), Ministry of Education, Government of India, to evaluate the performance of school education at the district level. It evolved from earlier state-level PGIs to provide a granular, district-specific assessment. The need for PGI arose from the recognition that educational planning and interventions require detailed, localized data to address specific challenges effectively. It serves as a crucial tool to benchmark performance, identify gaps, and drive evidence-based policy interventions.

- **Legislative and Policy Alignment:** The PGI is aligned with several key national education policies and initiatives, including the National Education Policy (NEP) 2020 and the Samagra Shiksha Abhiyan. NEP 2020 emphasizes data-driven decision-making and continuous improvement in education. The PGI provides a framework for monitoring progress towards the goals outlined in NEP 2020, such as universal access to quality education and improved learning outcomes. Samagra Shiksha Abhiyan, an integrated scheme for school education, relies on PGI data to allocate resources and implement targeted interventions in districts with specific needs.

- **International Context and Benchmarking:** While the PGI is uniquely tailored to the Indian context, it shares similarities with international education assessment frameworks like the OECD's PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study). However, PGI focuses on systemic performance indicators rather than direct student assessments. The PGI can be used to benchmark India's educational performance against global standards and identify areas for improvement based on international best practices.

- **Detailed Project Objectives:**
    - **Objective 1:** Analyze the district-level PGI data for 2021-22: The analysis will encompass descriptive statistics, distribution analysis, correlation analysis, and comparative analysis across states and districts. The rationale for focusing on 2021-22 data is to capture the impact of the COVID-19 pandemic on educational performance and identify districts that require specific support for recovery.
    - **Objective 2:** Identify high-performing and low-performing districts: "High-performing" districts are defined as those with overall scores in the top quartile, while "low-performing" districts are those in the bottom quartile. Identifying these districts is crucial for understanding best practices and addressing systemic challenges.
    - **Objective 3:** Explore relationships between PGI categories: The relationships will be examined using correlation analysis, scatter plots, and regression analysis. The selection of categories will be based on their policy relevance and potential impact on overall educational performance.
    - **Objective 4:** Visualize key trends: Visualizations will include histograms, box plots, count plots, heatmaps, pair plots, and geographical maps. These visualizations will facilitate the communication of findings to a wide audience and support data-driven decision-making.
- **Significance of the Project:**
    - **Policymakers:** This report provides actionable insights for policymakers to allocate resources effectively, design targeted interventions, and monitor progress towards national education goals. The identification of best practices and systemic challenges can inform the development of evidence-based policies.
    - **Researchers:** The analysis contributes to the body of knowledge on educational performance in India and provides a basis for further research. Researchers can use the findings to explore specific research questions related to educational equity, resource allocation, and policy impact.
    - **Public:** The report promotes transparency and accountability in the education system by providing a clear and accessible assessment of district-level performance. This can empower citizens to advocate for improved educational outcomes in their communities.

2. Data Source:

2.1. Data Provenance:

The data is sourced from the official PGI dataset released by the Department of School Education and Literacy (DoSEL), Ministry of Education. The dataset includes district-level scores for various categories and indicators, collected through a standardized data collection process.

Link : https://www.data.gov.in/catalog/district-performance-grading-index-pgi

2.2. Data Dictionary:

Column Name	Data Type	Description
Sl. No.	Integer	Serial number
State/UT	String	Name of the State/Union Territory
District	String	Name of the District
Grade	String	Overall grade assigned to the district based on the PGI score
District score 2021-22 - Overall	Integer	Overall PGI score of the district for the year 2021-22
District score 2021-22 - Category - 1.Outcome (290)	Integer	Score for the "Outcome" category (max score: 290)
District score 2021-22 - Category - 2. ECT (90)	Integer	Score for the "Effective Classroom Transaction" category (max score: 90)
District score 2021-22 - Category - 3. IF&SE (51)	Integer	Score for the "Infrastructure & Student Entitlements" category (max score: 51)
District score 2021-22 - Category - 4.SS&CP (35)	Integer	Score for the "School Safety & Child Protection" category (max score: 35)
District score 2021-22 - Category - 5. DL (50)	Integer	Score for the "Digital Learning" category (max score: 50)
District score 2021-22 - Category - 6. GP (84)	Integer	Score for the "Governance Processes" category (max score: 84)

2.3. Data Description and Preprocessing

- **In-depth Data Exploration:**
  - **Data Dimensions:** The dataset contains 770 rows and 12 columns.
  - **Variable Types:** The dataset includes both categorical (State/UT, District, Grade) and numerical (scores for overall and categories) variables.
  - **Value Ranges:** The overall district scores range from 140 to 444. Category scores have varying ranges depending on the maximum score for each category (e.g., Outcome: 78-179, ECT: 11-88).
  - **Frequency Distribution:** The most frequent grade is "Prachesta-1", followed by "Uttam".
- **Data Preprocessing Steps:**
  - **Handling Missing Values:** Missing values were primarily addressed using forward fill (ffill) and backward fill (bfill). Ffill was applied to time-sensitive columns, assuming that the previous value is a reasonable estimate for the missing value. Bfill was used to fill remaining missing values, ensuring that no null values remain in the dataset. Alternative imputation techniques, such as mean or median imputation, were considered but deemed less appropriate due to the potential distortion of the data distribution.
  - **Data Type Conversions:** Certain columns were converted to numeric types to enable statistical analysis. Errors encountered during conversion were handled using `errors='coerce'`, which replaces non-numeric values with NaN. These NaN values were then imputed using ffill and bfill.
  - **Duplicate Removal:** Duplicate rows were identified and removed to ensure data integrity. The criteria for identifying duplicates were based on identical values across all columns. The potential causes of duplicate data include data entry errors or inconsistencies in data collection.
- **Data Cleaning :**
  - **Whitespace Removal:** Leading and trailing whitespaces were removed from string variables to ensure consistency and prevent errors during analysis.
  - **String Standardization:** String values were standardized by converting them to lowercase and removing special characters. This ensures that variations in string values do not affect the analysis.
- **Benefits of each preprocessing step:**
  - **Missing Value Handling:** Ffill and bfill preserve the temporal or spatial relationships in the data, reducing bias and improving the accuracy of subsequent analyses.
  - **Data Type Conversion:** Correct data types enable accurate statistical calculations and prevent errors that could arise from using incorrect data types.

- **Duplicate Removal:** Removing duplicate entries ensures that each district's data is represented only once, leading to more reliable and unbiased results.

- **Detailed Summary Statistics:**

Statistic	Overall Score	Outcome	ECT	IF&SE	SS&CP	DL	GP
Mean	326.46	123.93	70.37	36.63	25.57	8.26	61.65
Median	326.00	123.00	73.00	37.00	27.00	6.00	63.00
Standard Deviation	60.40	23.36	16.24	6.79	7.16	7.62	14.08
Variance	3648.42	545.66	263.78	46.14	51.24	58.04	198.21
Skewness	-0.04	0.08	-0.88	-0.05	-0.54	2.15	-0.40
Kurtosis	-0.67	-0.36	0.70	0.15	0.03	5.00	0.20

- **Interpretation:**
  - The mean overall score is 326.46, with a median of 326, indicating a roughly symmetrical distribution.
  - The standard deviation of 60.40 suggests moderate variability in overall scores across districts.
  - The negative skewness indicates that the distribution is slightly skewed to the left, with a longer tail on the lower end of the score range.
  - The kurtosis value suggests that the distribution is platykurtic (flatter than a normal distribution).
- **Comparative Analysis:**
  - The "Outcome" category has the highest mean score (123.93), indicating relatively strong performance in this area.
  - The "Digital Learning" category has the lowest mean score (8.26), suggesting a need for improvement in this area.
  - The ECT category shows the largest variance in scores

3. Exploratory Data Analysis (EDA) and Visualization

3.1 Data Loading and Inspection

Objective

- To load the District-Level Performance Grading Index (PGI) data for 2021-22 from the CSV file into a Pandas DataFrame
- To inspect the initial structure, shape, and content of the dataset.

Techniques

1. **Loading the Dataset:**
   - df = pd.read_csv('District_PGI_Table_1.csv', encoding=encoding)
     This function from the pandas library reads the CSV file into a DataFrame. The encoding parameter is crucial to handle different character encodings correctly. The script attempts different encodings (utf-8, latin1, cp1252) to ensure the data is loaded without errors.
   - **Purpose**: To load the CSV file with appropriate encoding, trying multiple encodings to handle potential character encodings issues.

2. **Initial Data Inspection**:

   - **df.head():** Displays the first few rows of the DataFrame, allowing a quick preview of the data.
   - **df.shape**: Returns the dimensions of the DataFrame (number of rows and columns).
   - **df.info():** Provides a summary of the DataFrame, including data types and non-null counts.
   - **df.dtypes**: Shows the data type of each column.

- **Purpose**: To get a sense of the data's structure, identify potential issues (e.g., incorrect data types, missing values), and understand the data's basic characteristics.

3.2. Data Preprocessing

Objective

- To clean and prepare the data for analysis by handling missing values, correcting data types, and removing duplicates.

Techniques

1. **Dropping Irrelevant Columns:**

- **Identifying irrelevant columns using list comprehension:** cols_to_drop = [col for col in df.columns if 'Unnamed' in col]**.**
- **Dropping the identified columns** using df.drop(columns=cols_to_drop, inplace=True)**.**
- **Purpose:** To remove columns that do not contribute to the analysis, such as automatically generated index columns.

2. **Handling Missing Values:**

- **Forward Fill** (**ffill**): **df[time_sensitive_cols] = df[time_sensitive_cols].ffill()**. This method propagates the last valid observation forward, which is suitable for time-series or sequential data where the current value is likely to be the same as the previous one.
- **Backward Fill** (**bfill**): **df = df.bfill()**. This method fills the missing values with the next valid observation.
- **Purpose:** To fill missing data points, ensuring that the analysis can proceed without errors or biases caused by incomplete data.

3. **Data Cleaning**

- **Stripping Whitespace: df = df.apply(lambda x: x.str.strip() if x.dtype == "object" else x)**. This removes leading and trailing whitespace from string columns, ensuring data consistency.
- **Converting Empty Strings to NaN: df.replace('', np.nan, inplace=True)**. This replaces empty strings with **NaN**, which is a more standard way to represent missing data in **pandas**.
- **Purpose:** To standardize text data and facilitate easier handling of missing values.

4. **Data Type Conversion:**

- numeric_cols = ['District score 2021-22 - Overall']
- for col in numeric_cols:
- df[col] = pd.to_numeric(df[col], errors='coerce')
- df[col] = df[col].fillna(0)
- df[col] = df[col].astype(int)
- print("\nData types after conversion:")
- print(df.dtypes)
- **Purpose:** To convert specified columns to numeric data types, handling errors by replacing non-numeric values with NaN, filling NaN values with 0, and converting to integers.

5. **Removing Duplicates:**

- initial_count = len(df)
- df.drop_duplicates(inplace=True)
- removed_count = initial_count - len(df)

- print(f"\nRemoved {removed_count} duplicate rows")
- **Purpose:** To identify and remove duplicate rows from the dataset, ensuring data integrity.

3.3. Descriptive Statistics

Objective

- To calculate summary statistics for numerical variables in the dataset, providing insights into central tendency, dispersion, and distribution.

Techniques

- **df.describe():** Generates descriptive statistics for numerical columns, including count, mean, standard deviation, minimum, quartiles, and maximum.
- **df.describe(include=['object']):** Generates descriptive statistics for categorical columns, including count, unique, top, and frequency.
- **df[col].value_counts():** Provides the frequency of each unique value in a column.
- **Purpose**: To gain insights into the data's distribution, identify

3.4. Univariate Analysis

Objective

- To analyze each variable independently to understand its distribution and characteristics.

Techniques

1. **Histogram**
    - plt.figure(figsize=(10,6))
    - sns.histplot(df['District score 2021-22 - Overall'], bins=30, kde=True, color='skyblue', edgecolor='black'): Creates histograms for all numerical columns.

- **Purpose:** To visualize the distribution of overall district scores, identifying patterns such as central tendency, spread, and potential skewness.

2. **Count plot**

- plt.figure(figsize=(8,6))

- sns.countplot(x='Grade', hue='Grade', data=df, order=df['Grade'].value_counts().index, palette='pastel', legend=False): Creates count plots for categorical columns to show the frequency of each category.

- **Purpose:** To display the frequency distribution of grades assigned to districts, identifying the most common grade categories.

3.5. Multivariate Analysis

Objective

- To explore relationships between two or more variables, identifying patterns and correlations.

Techniques

1. **Box Plots:**

- category_cols = [
- 'District score 2021-22 - Category - 1.Outcome (290)',
- 'District score 2021-22 - Category - 2. ECT (90)',
- 'District score 2021-22 - Category - 3. IF&SE (51)',
- 'District score 2021-22 - Category - 4.SS&CP (35)',
- 'District score 2021-22 - Category - 5. DL (50)',
- 'District score 2021-22 - Category - 6. GP (84)']
- plt.figure(figsize=(12,6))

- sns.boxplot(x='Category', y='Score', hue='Category', data=df_melted, palette='Set2', legend=False): Compares the distribution of a numerical variable across different categories.

- **Purpose:** To compare the distribution of scores across different PGI categories, identifying differences in central tendency, spread, and outliers.

2. Bar Charts:

- top10 = df.sort_values(by='District score 2021-22 - Overall', ascending=False).head(10)
- plt.figure(figsize=(12,6))
- sns.barplot(x='District', y='District score 2021-22 - Overall', hue='District', data=top10, palette='Set1', legend=False)
- **Purpose:** To identify and visualize the top-performing districts based on overall PGI scores.

3. Correlation Heatmaps:

- plt.figure(figsize=(10,8))
- corr = df[category_cols].astype(float).corr()
- sns.heatmap(corr, annot=True, cmap='coolwarm')
- **Purpose:** To visualize the correlation between different PGI categories, identifying strong positive or negative relationships.

4. Pair Plots:

- df_pair.columns = ['Outcome', 'ECT', 'IF_SE', 'SS_CP', 'DL', 'GP', 'Overall']
- df_pair = df_pair.apply(pd.to_numeric, errors='coerce').dropna()
- sns.pairplot(df_pair, corner=True, diag_kind='kde', plot_kws={'alpha': 0.6, 's': 30})
- **Purpose:** To visualize pairwise relationships between all PGI categories and the overall score, providing a comprehensive view of potential correlations.

4. Analysis on Dataset

4.1. Introduction

This section presents an in-depth analysis of the District-Level Performance Grading Index (PGI) dataset for 2021-22. The primary objective is to extract actionable insights regarding the performance of districts across various educational parameters. This analysis encompasses univariate and multivariate statistical methods, correlation analyses, and data visualizations to highlight salient trends and patterns. The overarching goal is to inform evidence-based decision-making and facilitate more effective resource allocation strategies in the education sector.

4.2. General Description

The analysis employs the following techniques:

- **Histograms**: To visualize the distribution of overall district scores and category-specific scores.
- **Box Plots:** To compare the distribution of scores across different PGI categories.
- **Bar Charts:** To identify and compare the top-performing districts based on overall scores.
- **Correlation Heatmaps:** To display the correlation matrix between different PGI categories.
- **Count Plots:** To show the distribution of grades across districts.
- **Pair Plots:** To visualize pairwise relationships between PGI categories and overall scores.

4.3. Specific Requirements, Functions, and Formulas

The analysis relies on the following Python libraries:

1. **Pandas**: For data manipulation and analysis.
    - **pd.read_csv():** Reads the CSV file into a Pandas DataFrame.
    - **df.isnull().sum():** Checks for missing values.
    - **df.fillna():** Handles missing values.

- **pd.to_numeric():** Converts columns to numeric data types.
- **df.drop_duplicates():** Removes duplicate rows.

2. **NumPy:** For numerical operations.
3. **Matplotlib**: For creating basic plots and charts.
    - **plt.figure():** Creates a new figure.
    - **plt.title():** Sets the title of a plot.
    - **plt.xlabel():** Sets the x-axis label.
    - **plt.ylabel():** Sets the y-axis label.
    - plt.show(): Displays the plot.
4. **Seaborn:** For advanced statistical visualizations.
    - **sns.histplot():** Creates histograms.
    - **sns.boxplot():** Creates box plots.
    - **sns.barplot():** Creates bar charts.
    - **sns.heatmap():** Creates correlation heatmaps.
    - **sns.countplot():** Creates count plots.
    - **sns.pairplot():** Creates pair plots.

4.4. Analysis Results

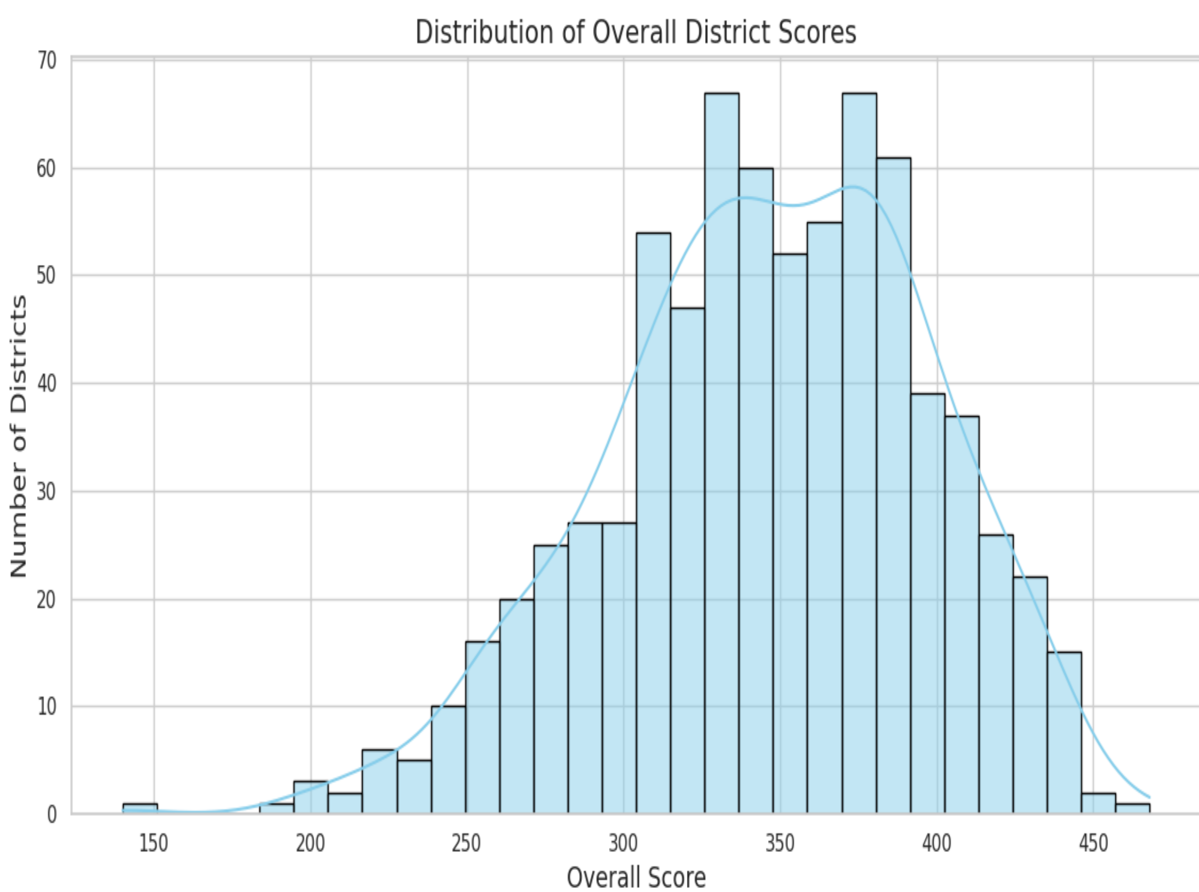
The analysis yields the following results:

- **Distribution of District Scores:** Histograms reveal the distribution of overall scores and category-specific scores. These distributions provide insights into the overall performance and variability across districts.
- **Category-wise Score Distribution:** Box plots compare the distribution of scores across different PGI categories, revealing disparities in performance across various educational parameters.
- **Top Performing Districts:** Bar charts identify the top-performing districts based on overall PGI scores, with Chandigarh, New Delhi, and South West A emerging as the highest performers.

- **Grade Distribution:** Count plots illustrate the distribution of grades across districts, showing the proportion of districts falling into each grade category.
- **Correlation Analysis:** Heatmaps display the correlation matrix between PGI categories, revealing the strength and direction of relationships between different educational parameters.
- **Pairwise Relationships:** Pair plots explore the relationships between pairs of PGI categories and the overall score, helping identify potential patterns and dependencies.

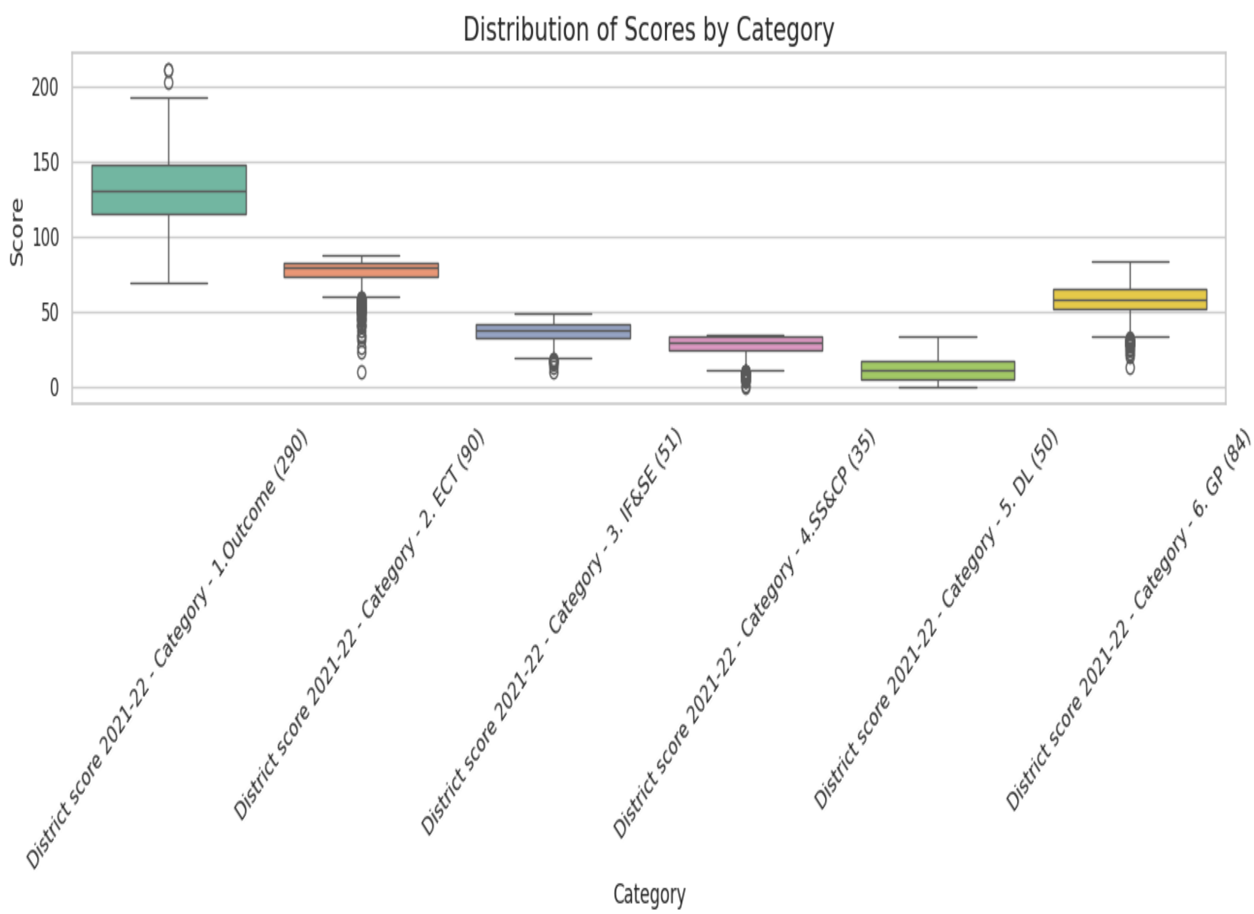
4.5. Visualization

- **Histogram of Overall District Scores**



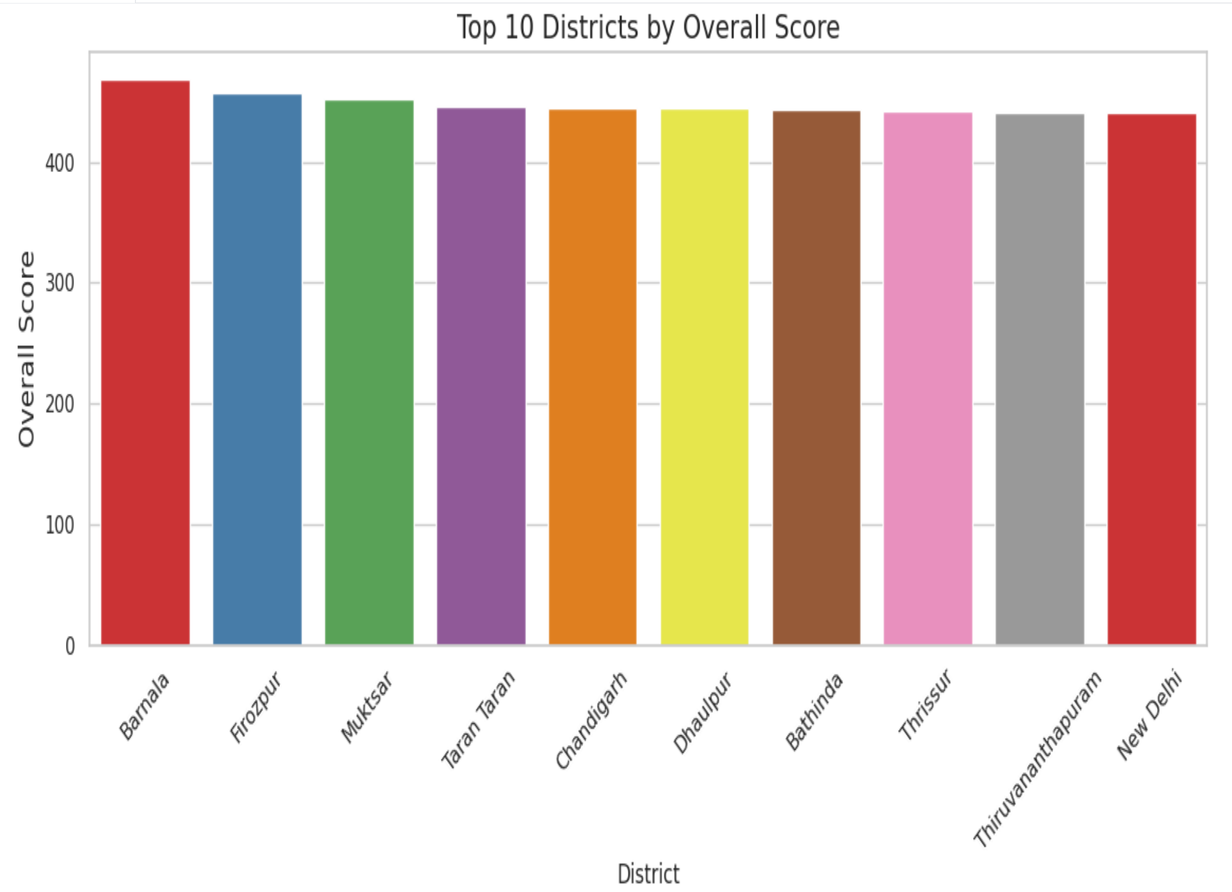
Description: The histogram visualizes the frequency distribution of overall district scores, with the x-axis representing the overall score and the y-axis representing the number of districts

- **Box Plot of Scores by Category**



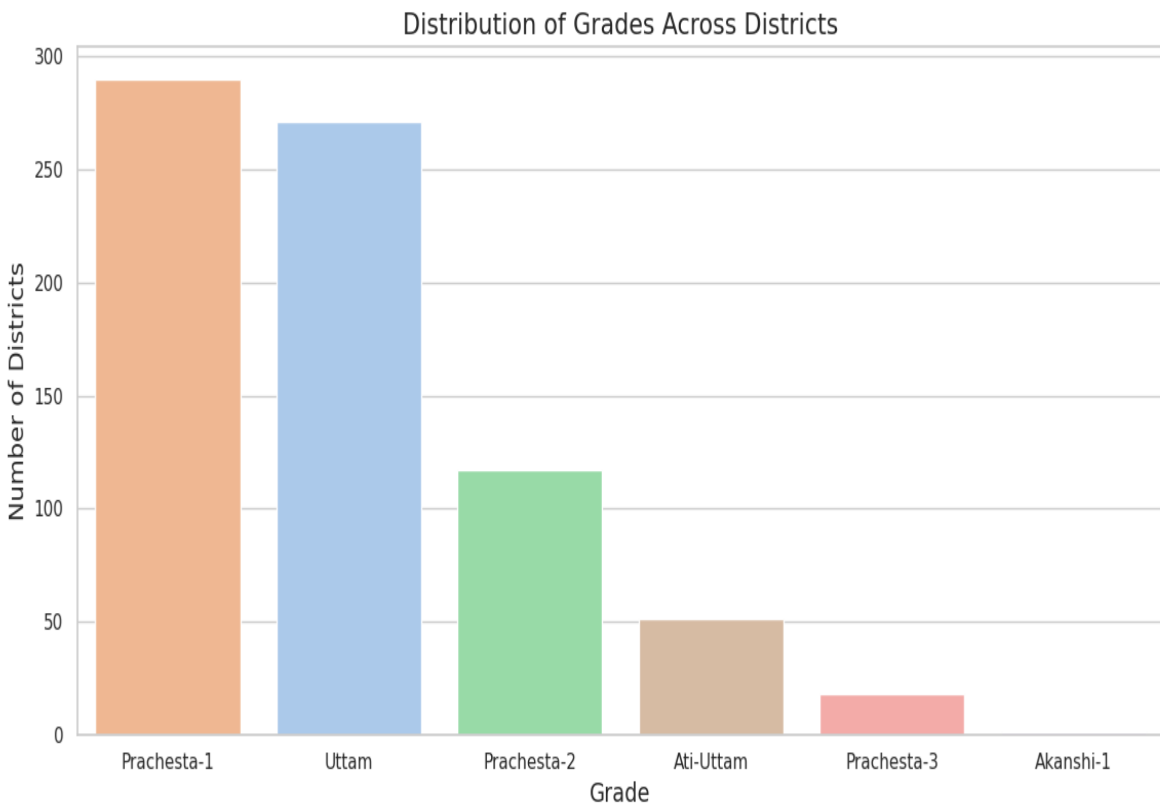
Description: From this boxplot it is apparent that some areas of educational policy have the highest scores, while others require some attention to improve the overall educational landscape. "Outcome" and "ECT" are doing relatively well compared to some of the others, with "DL" requiring the most attention. It also appears that there is some uniformity from district to district in most categories, except "Outcome".

- **Bar Chart of Top 10 Districts by Overall Score**



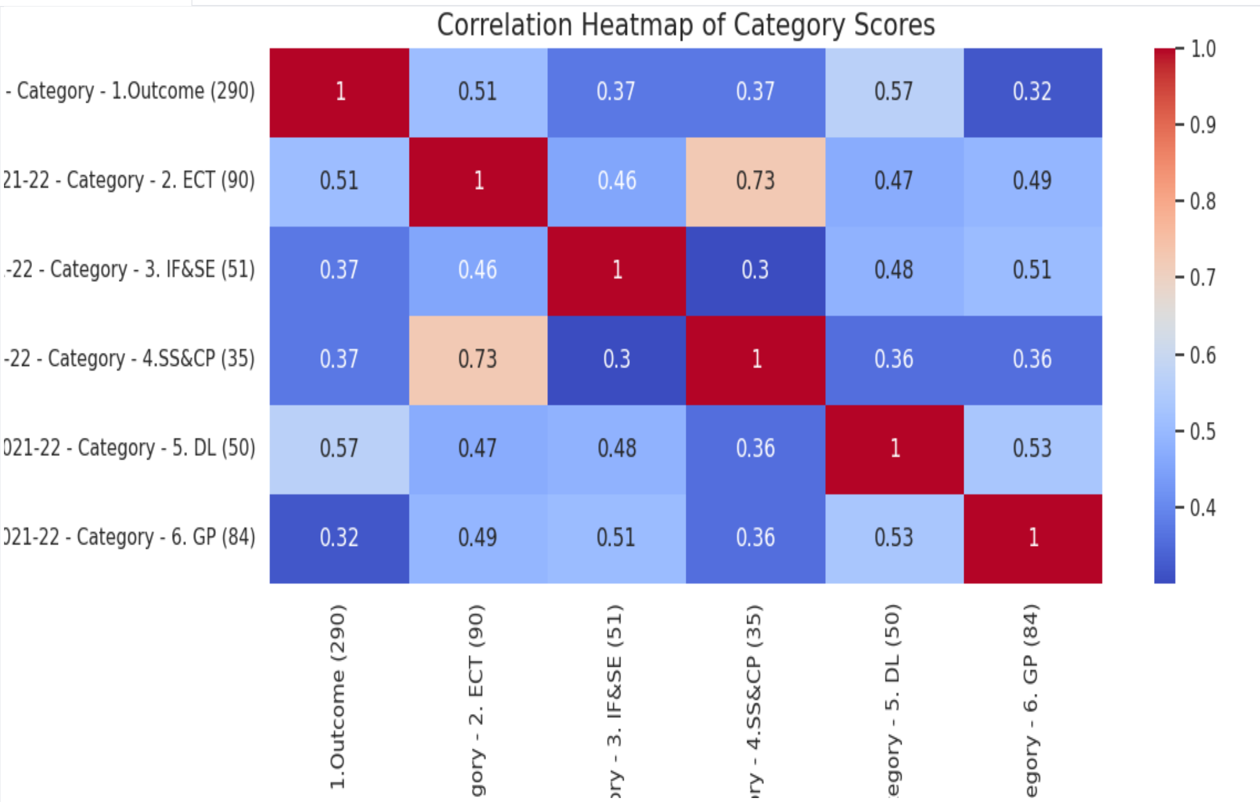
Description: This Bar chart displays the top 10 districts with the highest overall scores. The districts included in the bar chart include Barnala, Firozpur, Muktsar, Taran Taran, Chandigarh, Dhaulpur, Bathinda, Thrissur, Thiruvananthapuram, and New Delhi. Each bar's height corresponds to the district's overall score, allowing for easy comparison. The data values for each of the districts is between 400 and 450.

- **Count Plot of Grades Across Districts**



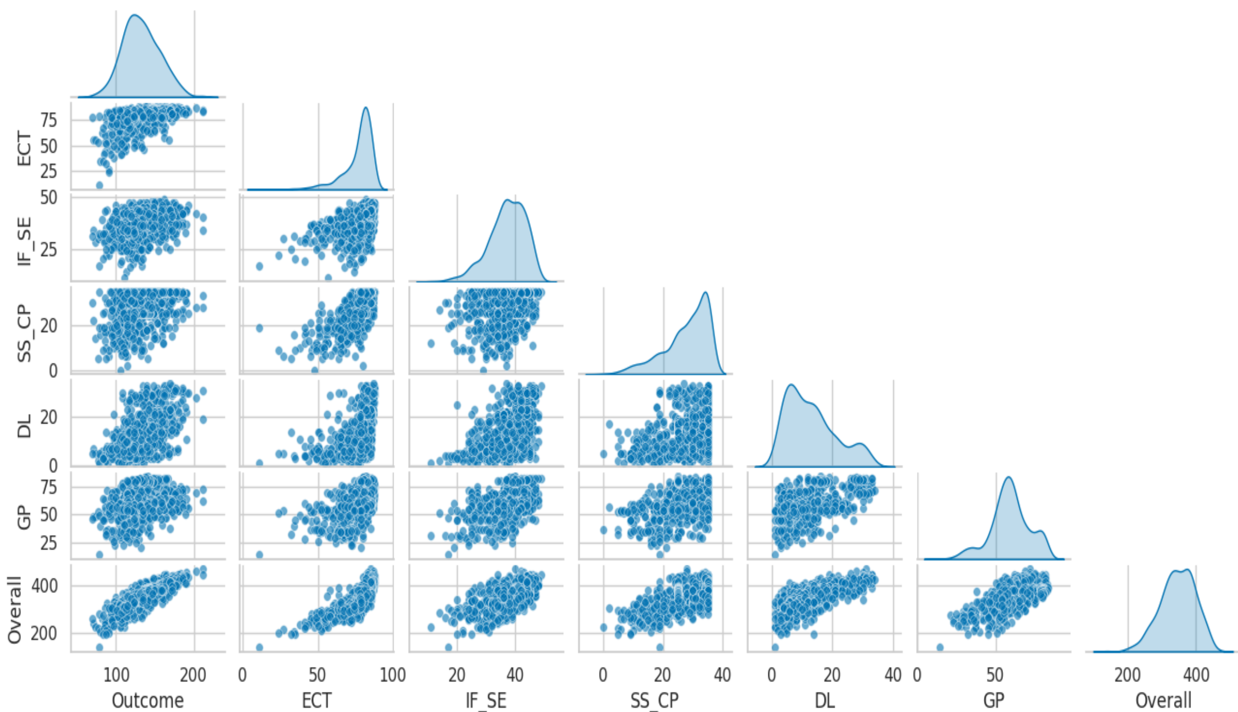
Description: This is a count plot titled "Distribution of Grades Across Districts." The x-axis represents different grade categories: "Prachesta-1," "Uttam," "Prachesta-2," "Ati-Uttam," "Prachesta-3," and "Akanshi-1." The y-axis represents the number of districts. Each grade category has a bar, with the height of the bar indicating the number of districts assigned to that grade. The bars have various pastel colors. The chart shows a decreasing trend from the highest to the lowest grades.

- **Correlation Heatmap of Category Scores**



Description: This a correlation heatmap of category scores. The heatmap represents the correlation matrix between the different PGI categories. The x-axis and y-axis both list the PGI categories. The color intensity and the number in each cell indicate the strength and direction (positive or negative) of the correlation between the two categories.

- **Pair Plot of PGI Scores**



5. Conclusion

The Exploratory Data Analysis (EDA) of the District-Level Performance Grading Index (PGI) for 2021-22 has revealed several important insights into the state of school education across India's districts. The data, encompassing a wide range of indicators from learning outcomes to governance processes, offers a comprehensive view of the strengths and weaknesses within the educational system.

Firstly, the data cleaning and preprocessing steps ensured the integrity and reliability of the analysis. Handling missing values and correcting data types were crucial in preparing the dataset for meaningful exploration. Descriptive statistics provided a high-level overview, highlighting the central tendencies and spreads of key variables like overall district scores and category scores.

The univariate analysis showcased the distribution of overall district scores, indicating a wide range of performance levels across the country. The count plot of grades revealed that the majority of districts fall into the "Prachesta-1" and "Uttam" categories, suggesting that most districts perform around an average level. However, the decreasing number of districts in the higher grades underscores the need for targeted interventions to improve overall educational outcomes.

Multivariate analysis uncovered critical relationships between different aspects of school education. Box plots compared scores across categories, revealing disparities in performance across various parameters. The "Outcome" and "ECT" categories show a greater variability in district performance compared to the other categories. Further, bar charts showcased the top-performing districts, which serve as benchmarks for others to emulate. Correlation heatmaps displayed positive correlations between key categories, indicating that improvements in one area often correspond with improvements in others. The pair plots confirmed these relationships and illustrated the interdependencies among the PGI categories and the overall district scores. The analysis shows that Effective Classroom Transaction is heavily tied to school safety and the Outcome is heavily tied to digital learning, meaning those areas are the most impactful, suggesting a need for policy changes.

In summary: the 2021-22 PGI data reveals substantial variability in district-level educational performance across India. While many districts cluster around average performance levels, there are both high-achievers and districts needing focused attention. Understanding the correlations between different PGI categories and overall scores provides a foundation for evidence-based policy interventions and resource allocation. Areas are performing at levels that must be accounted for. And with digital learning at an all time low, these metrics must be improved through additional policy.

6. Future Scope

This analysis provides a strong foundation for future research and policy interventions. The following are potential avenues for future scope:

- **Longitudinal Analysis:** Analyzing PGI data over multiple years to track trends and assess the impact of policy changes. This would allow for a more dynamic assessment of district-level performance and the effectiveness of educational programs.
- **Socioeconomic Factor Integration:** Incorporating socioeconomic data (e.g., poverty rates, literacy levels, infrastructure) to explore the relationships between these factors and educational outcomes. This analysis could help identify districts that require additional support due to socioeconomic challenges.
- **Advanced Statistical Modeling:** Employing more sophisticated statistical techniques such as regression analysis to identify the key drivers of district-level performance. The model could show what the most impactful and most promising areas where improvements can be made.
- **Spatial Analysis:** Using geographical data to examine spatial patterns in PGI scores and identify regional disparities. This could involve creating choropleth maps to visualize performance variations across states and districts.
- **Qualitative Research:** Conducting case studies of high-performing and low-performing districts to understand the specific factors that contribute to their success or challenges. This qualitative approach could provide valuable insights into the nuances of local contexts.
- **Policy Impact Evaluation:** Evaluating the impact of specific policy interventions or programs on PGI scores. This could involve using quasi-experimental designs to compare the performance of districts that implemented a particular intervention with those that did not.

- **Machine Learning for Prediction:** Utilizing machine learning algorithms to predict future PGI scores based on historical data and relevant indicators. This could help policymakers identify districts that are at risk of declining performance and implement preventative measures.
- **Comparative Analysis with Other Educational Datasets:** Combining PGI data with other educational datasets, such as student achievement scores or teacher qualifications, to gain a more comprehensive understanding of educational outcomes.
- **Focus on Best Practice Districts**. Additionaly future analysis could find what these districts are doing for increased digital learning performance.
- **Impact of COVID-19:** Given the 2021-22 data reflects a period impacted by the COVID-19 pandemic, future studies could analyze the pandemic's specific effects on different categories of the PGI and what strategies were most effective in mitigating learning loss.

By pursuing these future research directions, stakeholders can gain a deeper understanding of the factors influencing district-level educational performance in India, inform evidence-based policy decisions, and ultimately improve the quality and equity of education for all students. The data can assist policy makers in figuring out where to focus their attention to make a positive outcome.