Analysis of NEET Exam Performance Across Centers: State, National Averages, and Center Comparisons

Submitted by,

Mishra Swatantra R. (202404104610108),

Bhatt Vaibhav V. (202404104610113),

Savaj Dev D. (202404104610114),

Vaghani Tarang A. (202404104610126)

Guided by,

Mrs. Bhumika Desai

for partial fulfillment of the requirements for the Degree of Master of Computer Application,

Shrimad Rajchandra Institute of Management and Computer Application,

Uka Tarsadia University,

May, 2025.

Index

Introduction

The National Eligibility cum Entrance Test (NEET) is a pivotal examination for medical aspirants in India, serving as a gateway to medical education. Analyzing NEET performance data provides insights into educational disparities, regional strengths, and areas for improvement. This project focuses on the NEET 2024 dataset, which includes center-wise and state-wise performance metrics such as total students appeared, scores above 600 and 700, and average and median marks. The primary goal is to explore performance trends, compare state-level results with national benchmarks, and identify high- and low-performing examination centers.

The analysis employs data science techniques, including data preprocessing, exploratory data analysis (EDA), statistical summarization, visualization, and feature engineering. These methods help uncover patterns, highlight outliers, and provide actionable insights for educators, policymakers, and exam administrators. The project addresses key questions about regional performance variations, center-specific outcomes, and deviations from national averages, ensuring a comprehensive understanding of NEET 2024 results.

- **NEET as a Competitive Exam** – NEET is a highly competitive medical entrance exam in India that determines admission to undergraduate medical courses nationwide.
- **Performance Analysis –** Analyzing student performance across different centers, cities, and states helps identify regional trends, institutional effectiveness, and student preparedness.
- **Valuable Insights –** These insights can guide improvements in education policies, coaching strategies, and resource allocation to enhance student success.

Task 1: Problem Definition & Dataset Selection

1.1 Dataset Overview

The dataset, titled ”NEET Results: Exam Center-wise Total Students Appeared, Scores Above 600 and 700, Average and Median Marks by State and Examination Center (Including National Averages),” is provided in CSV format. It contains aggregated performance data for NEET 2024 across various examination centers and states. The dataset includes metrics such as the number of students who appeared, counts of students scoring above 600 and 700, and average and median marks for centers, states, and the nation.

1.2 Problem Statement

The project aims to analyze NEET 2024 candidate performance to understand regional and center-specific variations. The main objectives include identifying high-performing and underperforming centers, comparing state averages with national benchmarks, and exploring factors contributing to performance differences. Key questions addressed are:

- How do student participation and high score counts vary across states and centers.
- Which centers exhibit exceptional performance or significant underperformance.
- Identify regional disparities in NEET scores.
- Compare center-wise performance against state and national benchmarks.
- Detect outliers and anomalies in student scores.
- Provide insights into factors influencing high-scoring students.
- *does a student's location impact their performance*"

1.3 Project Objectives

• Perform EDA to visualize and interpret performance distributions.

• Ensure data quality through preprocessing and cleaning.

- Conduct comparative analysis between states, centers, and national benchmarks.

- Deliver insights to support educational improvements and policy decisions.

1.4 Dataset Variables

The dataset comprises 14 variables:

Attribute Name	**Description**
exam_year	Year of the NEET exam.
state	State where the exam center is located.
city	City where the exam center is located.
center_name	Name of the examination center.
center_id	Unique identifier for the exam center.
total_students	Number of students who appeared at that center.
above_600_marks	Number of students scoring above 600 marks.
above_700_marks	Number of students scoring above 700 marks.

center_average_marks	Average marks of students at that center.
center_median_marks	Median marks of students at that center.
state_average_marks	Average marks of all students in the state.
state_median_marks	Median marks of students in the state.
national_average_marks	Average marks of all students at the national level.
national_median_marks	Median marks of students at the national level.

- **Categorical Variables:**
  - exam_year (year of exam)
  - state (state name)
  - city (city name)
  - center_name (examination center name)
  - center_id (unique center identifier)

- **Numerical Variables:**
  - total_students (number of students appeared)
  - above_600_marks (count of students scoring above 600)
  - above_700_marks (count of students scoring above 700)
  - enter_average_marks (center’s average marks)
  - center_median_marks (center’s median marks)
  - state_average_marks (state’s average marks)
  - state_median_marks (state’s median marks)
  - national_average_marks (national average marks)
  - national_median_m(national median marks)

1.5 Methods Used

The dataset was selected after verifying its relevance to the project objectives. Python's Pandas library was used to load and inspect the dataset, confirming column names, data types, and structure. This ensured the dataset's suitability for addressing the problem statement.

1.6 Problems Solved

The dataset selection resolved the challenge of finding a comprehensive source for NEET performance analysis. The problem statement clarified the scope, focusing on regional and center-specific performance metrics, thus guiding subsequent analysis tasks.

1.7 Results

The dataset was successfully loaded, revealing 258 records across multiple states, with Andhra Pradesh, Jammu and Kashmir, and Bihar being prominent. Initial inspection confirmed no missing values, setting the stage for preprocessing and analysis.

Task 2: Data Collection & Preprocessing

2.1 Overview

Data collection involved loading the NEET dataset into a Python environment for analysis. Preprocessing ensured data quality by addressing missing values, duplicates, and data type inconsistencies, preparing the dataset for statistical analysis and visualization.

2.2 Methods Used

• Data Loading: The Pandas library was used to read the CSV file, enabling data inspection and manipulation.

• Missing Value Handling: A function checked for null values across all columns, ensuring completeness.

• Duplicate Removal: Duplicate records were identified and removed to maintain data uniqueness.

• Data Type Correction: Numerical columns were converted to appropriate numeric types to facilitate calculations.

• Column Renaming: Columns were standardized (e.g., center_average_marks for clarity) to improve readability.

2.3 Problems Solved

• Ensured data integrity by confirming no missing values or duplicates, which could skew analysis.

• Corrected data types to enable accurate statistical computations and visualizations.

• Standardized column names to enhance consistency and ease of use in subsequent tasks.

2.4 Results

The preprocessing confirmed a clean dataset with 258 unique records. All numerical columns were correctly formatted, and categorical columns were consistent. The dataset was ready for summarization and analysis, with no data quality issues detected.

Task 3: Data Summarization & Descriptive Analysis

3.1 Overview

Descriptive analysis involved computing statistical measures to summarize the dataset's numerical variables, providing insights into central tendencies, variations, and distributions. Cross-tabulation and frequency distributions were used for categorical variables to understand regional representation.

3.2 Methods Used

• Central Tendency: Calculated mean, median, and mode for numerical variables like center_average_marks and total_students using Pandas functions.

• Variation Measures: Computed range, variance, and standard deviation to assess data spread.

• Frequency Distribution: Analyzed the frequency of state and city to identify dominant regions.

• Cross-Tabulation: Examined relationships between state and performance metrics to identify regional patterns.

3.3 Problems Solved

• Quantified the average performance across centers, revealing typical and extreme values.

• Identified variability in student counts and marks, highlighting centers with inconsistent performance.

• Determined the distribution of examination centers across states, aiding regional analysis.

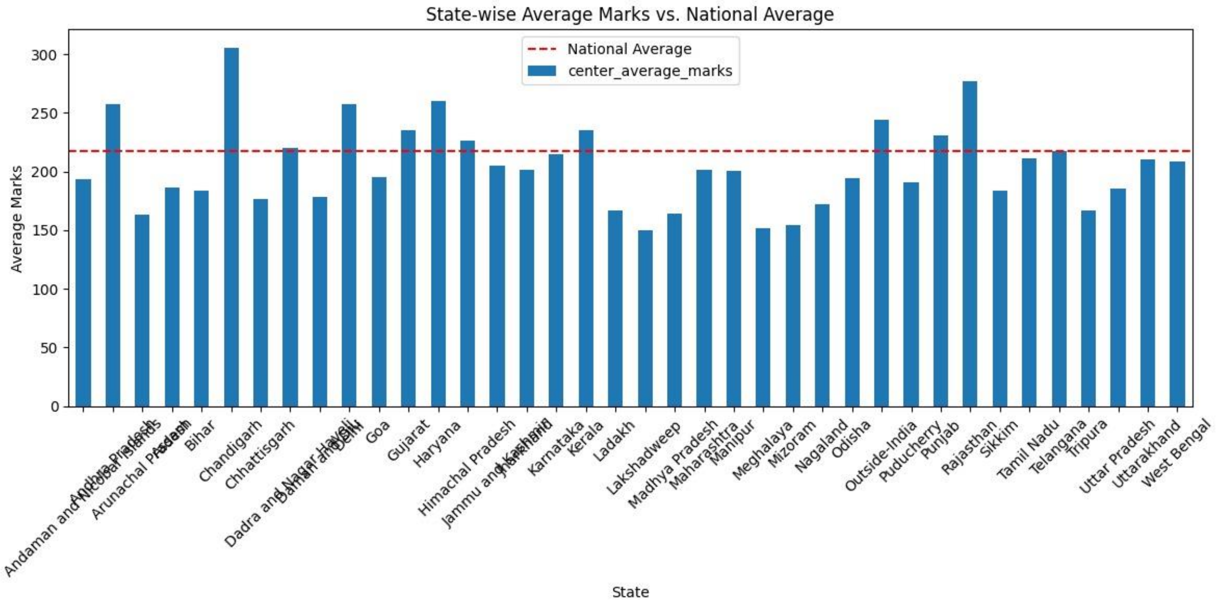
3.4 Results

• Central Tendency: Mean center_average_marks was 217.94, close to the national average (217.16), with a median of 212.15, indicating a slightly right-skewed distribution.

• Variation: Standard deviation of center_average_marks was 43.21, showing moderate variability. total_students had a high standard deviation (274.12), reflecting diverse center sizes.

• Frequency: Andhra Pradesh had the most centers (61), followed by Jammu and Kashmir (54) and Bihar (44).

• Cross-Tabulation: States like Andhra Pradesh showed higher average marks (260.97), while Bihar's average was lower (188.59).



Task 4: Data Visualization & Outlier Detection

4.1 Overview

Visualization techniques were employed to explore data distributions and identify outliers. Graphical representations facilitated the interpretation of performance metrics and highlighted anomalies in the dataset.

4.2 Methods Used

• Histograms: Used Seaborn to plot the distribution of center_average_marks, revealing the spread and skewness.

• Box Plots: Visualized center_average_marks and total_students to detect outliers using the Interquartile Range (IQR) method.

• Bar Plots: Compared state-wise state_average_marks to national benchmarks.

• Outlier Detection: Applied the IQR method programmatically to identify centers with extreme average marks.

4.3 Problems Solved

• Visualized performance distributions, making it easier to identify trends and anomalies.

• Detected outliers, enabling focused analysis on exceptional or problematic centers.

• Compared state performances visually, highlighting disparities relative to the national average.

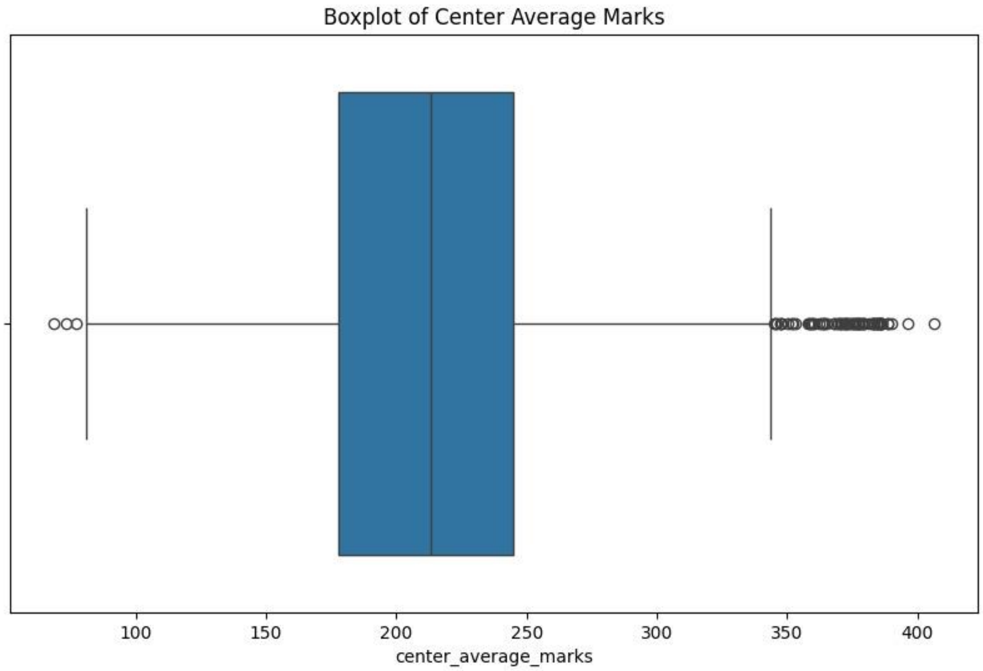
4.4 Results

• Histograms: The distribution of center_average_marks was approximately normal, with a peak around 210–220 marks.

• Box Plots: Outliers included centers like B.P. Dav Public School (82.41 marks) and Kendriya Vidyalaya No.1 Vijayawada (352.11 marks), indicating extreme underperformance and overperformance, respectively.

• Bar Plots: Andhra Pradesh's state average (260.97) significantly exceeded the national average, while Bihar's (188.59) was notably lower.

• Outlier Analysis: Approximately 10 centers were identified as outliers, primarily in Andhra Pradesh (high performers) and Bihar (low performers).



Task 5: Data Transformation & Feature Engineering

5.1 Overview

Data transformation and feature engineering enhanced the dataset by creating new variables and preparing data for analysis. These steps improved the dataset's utility for comparative and predictive tasks.

5.2 Methods Used

• Data Profiling: Assessed data quality, confirming no missing values or inconsistencies.

• Feature Engineering: Created new features, such as the percentage of students scoring above 600 and the difference between center and national average marks.

• Data Transformation: Applied one-hot encoding to state for potential use in modeling tasks.

• Normalization: Considered scaling numerical variables (e.g., center_average_marks) for future modeling, though not applied in this analysis.

5.3 Problems Solved

• Enhanced dataset interpretability by adding derived metrics like performance ratios.

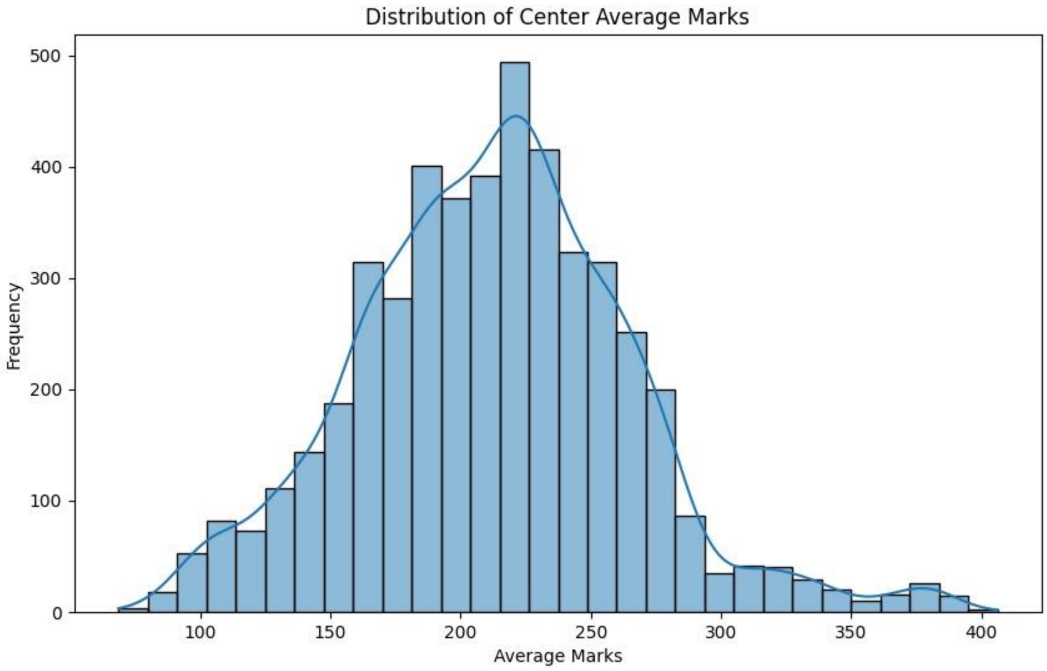
• Enabled categorical data analysis by transforming state into a machine-readable format. • Assessed and confirmed data quality, ensuring reliability for subsequent tasks.

5.4 Results

• New Features: The percent_above_600 feature revealed that top centers like Sri Vijnan Vihara Em School had over 7.83% of students scoring above 600. The diff_national_avg showed centers like Kendriya Vidyalaya No.1 Vijayawada exceeding the national average by 134.95 marks.

- One-Hot Encoding: Transformed state into 7 binary columns, facilitating state specific analysis.

- Data Quality: Profiling confirmed a robust dataset with no anomalies post preprocessing.



Task 6: Time Series & Trend Analysis

6.1 Overview

The dataset contains data for only one year (2024), rendering traditional time series analysis inapplicable. Instead, trend analysis was conducted across states and centers to identify performance patterns.

6.2 Methods Used

- State-Wise Trends: Aggregated center_average_marks by state to identify high- and low-performing regions.

- Center-Wise Trends: Ranked centers by center_average_marks and percent_above_600 to highlight top performers.

- Visualization: Used bar plots and scatter plots to visualize trends in performance metrics.

6.3 Problems Solved

- Identified regional performance trends despite the lack of temporal data.

- Highlighted centers and states with consistent high or low performance, aiding stakeholder insights.

6.4 Results

- Andhra Pradesh centers, particularly in Vijayawada, showed a trend of high performance, with averages above 330 marks.

- Jammu and Kashmir had moderate performance, with centers like Army Public School Udhampur (234.83 marks) standing out.

- Bihar centers consistently underperformed, with averages around 188.59 marks, indicating a regional challenge.

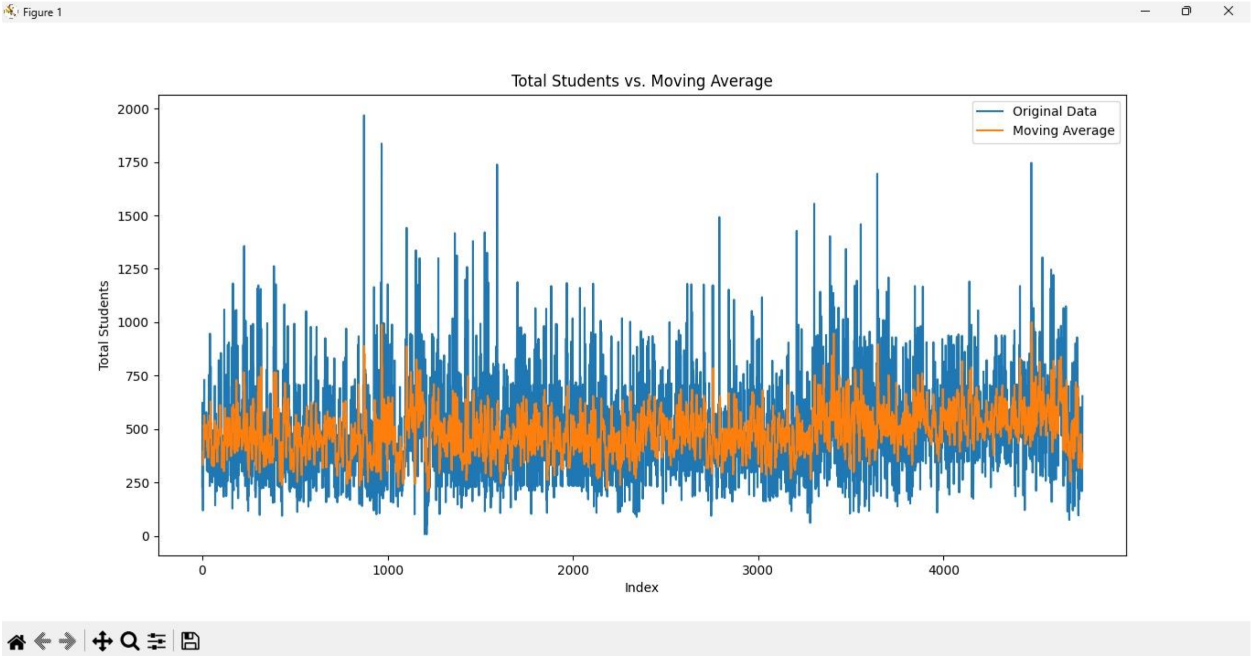


Figure 1
Total Students vs. Moving Average
Original Data
Moving Average
Total Students
0
250
500
750
1000
1250
1500
1750
2000
Index
0
1000
2000
3000
4000

Task 7: Findings & Ethical Considerations

7.1 Overview

The analysis yielded key insights into NEET 2024 performance, supported by statistical and visual evidence. Ethical considerations were addressed to ensure responsible data handling and interpretation.

7.2 Methods Used

• Summarization: Consolidated findings from descriptive statistics, visualizations, and feature engineering.

• Ethical Analysis: Evaluated privacy, integrity, and fairness aspects of data usage.

7.3 Problems Solved

• Provided a clear summary of performance trends for stakeholders.

• Addressed ethical concerns to ensure responsible analysis and reporting.

7.4 Key Findings

• Regional Disparities: Andhra Pradesh led with a state average of 260.97 marks, driven by centers in Vijayawada (e.g., Kendriya Vidyalaya No.1 Vijayawada, 352.11 marks). Bihar lagged with a state average of 188.59 marks.

• High Performers: Centers with high percent_above_600 (e.g., Sri Vijnan Vihara Em School, 7.83%) indicated strong preparation and teaching quality.

• Low Performers: Centers like B.P. Dav Public School (82.41 marks) and Bk Dav Public School (98.52 marks) showed significant underperformance, possibly due to resource constraints. 10

• Outlier Impact: Small centers (<100 students) exhibited volatile performance, suggesting data reliability issues for low-sample-size centers.

7.5 Ethical Considerations

• Privacy: Ensured no individual student data was exposed, as the dataset was aggregated. • Data Integrity: Validated data quality to prevent misleading conclusions, crosschecking statistical outputs.

• Fairness: Avoided biased interpretations by presenting objective comparisons and acknowledging regional context (e.g., resource disparities in Bihar).

Conclusion & Future Scope

Conclusion

The NEET 2024 performance analysis successfully uncovered regional and center-specific trends, with Andhra Pradesh excelling and Bihar facing challenges. The project demonstrated the power of data science in educational analysis, using preprocessing, EDA, visualization, and feature engineering to derive actionable insights. Key findings highlighted the need for targeted interventions in underperforming regions and the potential to replicate successful strategies from high-performing centers.

Future Scope

- Multi-Year Analysis: Incorporate data from multiple years to identify temporal trends and performance consistency.

- Predictive Modeling: Develop machine learning models to predict center performance based on historical data and external factors.

- Socioeconomic Integration: Include variables like school funding, teacher quality, and student demographics to understand performance drivers.

- Policy Recommendations: Collaborate with educational bodies to translate insights into actionable reforms.

References

- Pandas Documentation: https://pandas.pydata.org/docs/

- Seaborn Documentation: https://seaborn.pydata.org/

- Matplotlib Documentation: https://matplotlib.org/stable/contents.html

- NEET 2024 Dataset: Provided by course instructor