Exploratory Data Analysis on the Capacity Utilization for Power Generation in India

CP 318: Data Science for Smart City Applications, August-Nov Semester 2023, Project 2

Paritosh Tiwari

1 Introduction

The advancement of India's power sector is critical to the nation's growth and development, and smart grid management is a cornerstone of this progress. The Grid Operation and Distribution Wing, specifically through its Operation Performance Monitoring Division, plays a crucial role in this domain by meticulously tracking and optimizing the use of power generation capacities. The dataset in focus for this project, incorporating comprehensive daily generation reports, provides a window into the functioning of power generation sources and their capacity utilization. This data is pivotal for scrutinizing the effectiveness of energy distribution and identifying opportunities for enhancements in the power grid.

This project's problem statement centres on the necessity to analyze the operational efficiency of power generation in the context of smart city development. The challenge lies in parsing through the daily generation data to identify trends, performance metrics, and capacity utilization rates across various types of power generation. Through data pre-processing, descriptive analytics, clustering, and dimensionality reduction, the project seeks to reveal insights that could lead to more informed decisions in power management, contributing to the reliability and sustainability of smart city infrastructures in India.

2 Dataset Description and Processing

The dataset essentially contains daily records of power generation, via various methods, across India, from the year 2017 to 2023. It is a summation of all sources of power generation of a particular type, throughtout the country. The dataset is provided by the Central Electricity Authority, Ministry of Power, Government of India, and hosted on the National Data and Analytics Platform (NDAP). The dataset tracks and analyses power generation through various performance metrics and indicators. The generation capacity is measured in Mega Watts (MW). Table 5 provides necessary details of the attributes in the dataset. We could not track down any relevant studies conducted using this dataset. Hence, to the best of our knowledge, this exploratory data analysis is the first of its kind.

Redundant Information. There were several columns in the dataset relaying the same information of day, month and year. These columns were dropped: *SourceCalendarDay, SourceYear, SourceMonth, YearCode, Year, MonthCode, Month, CalendarDayCode*. We kept the one most relevant, i.e. *Calendar Day*, which contains the day, month and year, converted it into "datetime" format for the convenience of time-series analysis, and made it the first column for organisational relevance. The *Country* column was dropped as well since the report is generated for India.

Organisation. We sorted the dataset in ascending order according to the *Calender Day* column. This aggregated all power generation records, from different sources, for a particular day. We also find out that the records exist from the 1^{st} of September, 2017 to the 2^{nd} of November, 2023.

Missing/Null Values. We encountered a lot of missing/null values in the dataset. Handling this issue required a deeper understanding of the attributes. We **note** here that Thermal Coal Lignite as a source of power was renamed and split into Thermal Coal and Thermal Lignite. Lignite is a type of coal. The numbers recorded for these new names did not add up as expected, and therefore, we did not attempt

to aggregate them. We began by addressing columns with few missing values, suspecting them to be data entry errors or omissions. Analysis confirmed this; these entries, showing almost complete data absence, were dropped from the dataset. The columns *Gross sum of maximum power generated* and *Net sum of maximum power generated* are pivotal in assessing power generation efficiency and output. With their relatively few missing values, we employed the monthly mean for imputation. High missing values in *New capacity under stabilization or completion of balance works* indicate its specificity, likely pertaining to new or significantly upgraded plants. This suggests infrequent reporting, leading to fewer applicable cases. Likewise, significant missing values in *Capacity under long outage* hint at the rarity of such events, with the absence of data often signifying uninterrupted operations. In these cases, filling missing values with 0 was deemed appropriate. A consistent pattern of missing values in *Capacity of power generation from prime movers (pm), fuel oil (fo), other sources, and Total capacity* suggested simultaneous collection or reporting. Missing data in one typically meant the same for others, likely due to shared reporting channels or documentation. Further analysis revealed that these missing data points predominantly concerned the *Thermal Diesel* power type, which appeared intermittently throughout the month. Consequently, we opted to impute these missing values using the monthly mean of the available data.

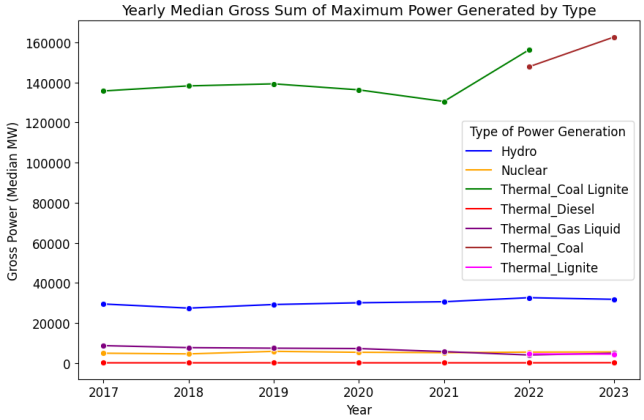
3 Descriptive Analysis

In this section, we discuss insights as a result of statistics and visualizations of the dataset. Table 1 shows some basic statistical values pulled from the dataset. It reveals a right-skewed distribution in monitored capacity, operational capacity, and both gross and net maximum power generated, with means significantly higher than medians and large standard deviations, indicating that a few large plants disproportionately influence these averages. The capacity under long outage is relatively low on average, suggesting that only a small portion of the total capacity is typically affected by long outages, despite noticeable variability. The maximum power output as a percentage of capacity on line and the percentage of monitored maximum power output both display significant variability. The former suggests varied plant efficiency levels, with some plants potentially operating above nominal capacity, while the latter highlights diverse performance against monitored expectations due to factors like operational efficiency and external influences.

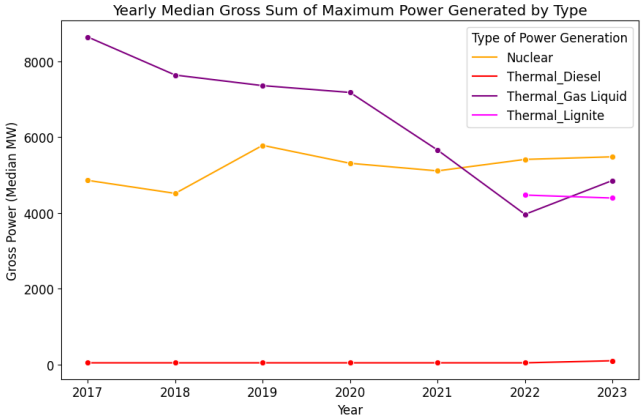
Statistic	Mean	Median	Std. Dev.	Min	Max
Monitored capacity	52425.23	24869.21	73619.23	0.00	210739.50
Capacity under long outage	3361.07	1037.00	3666.18	0.00	10810.00
Operational capacity	37196.23	7689.14	53902.97	0.00	180910.00
Gross sum of maximum power generated	36040.20	6496.99	53516.84	0.00	179008.63
Net sum of maximum power generated	33329.15	6267.70	48637.16	0.00	162897.86
Maximum power output as percentage of capacity on line	66.43	83.70	32.96	0.00	159.79
Percentage of monitored maximum power output	45.83	58.46	26.07	0.00	94.52

Table 1: Overall Statistics for the Dataset. All values are in Mega Watts (MW).

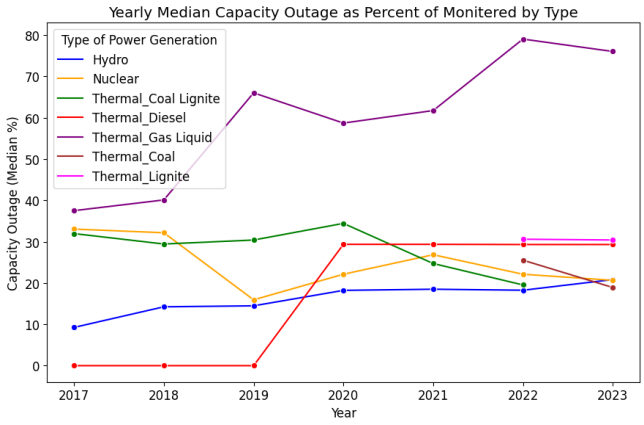
Figure 1 shows various statistical visualisations. Figure 1b is a zoomed-in version of the lower line plots from Figure 1a, since they were not easily distinguishable. Figure 1a shows significant domination of thermal power derived from coal. The extremely low contribution of Diesel power reveals it only used a backup source of power, in case of failure of other sources. From Figure 1c we see high fluctuations in outages from gas and diesel sources. This might suggest they are only required when the demand for power increases beyond expected limits. The box plots in Figure 1d are a visualisations of the information in Table 1. The significant variability in thermal power from coal in the years 2021 and 2022 point to changes in infrastructure and undesired higher dependence on coal as a source of power.



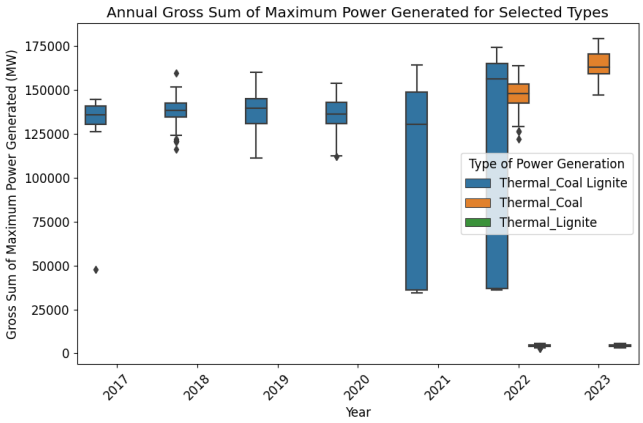
(a) Yearly Median Gross Sum of Maximum Power Generated by Type



(b) Yearly Median Gross Sum of Maximum Power Generated by Type (zoomed)



(c) Yearly Median Capacity Outage as Percent of Monitered by Type



(d) Annual Gross Sum of Maximum Power Generated for Selected Types

Figure 1: Descriptive Analysis

4 Clustering and Dimension Reduction Analysis

A clustering analysis approach for the type of dataset we are dealing with is not really applicable, since most of the valuable insights can be obtained via the descriptive analysis. What we can do however is reverse verify our dataset distribution. We do this by applying a clustering algorithm and then check to see if the clusters verify any type of class assignment that might be applicable to our dataset. In this case, we can verify if the clusters assignments agree with the *Type of Power Generation* column. This approach is often referred to as a *cross-tabulation* or *contingency table* analysis in the context of clustering. It involves examining the relationship between the categorical variable (in your case, the type of power generation) and the cluster labels assigned through clustering.

Specifications. Our analysis employed *K-Means* and *DBSCAN*, two diverse clustering methods. K-Means, a centroid-based approach, forms K clusters by iteratively assigning points to the nearest centroid until stable, ideal for spherical clusters with an estimable count. DBSCAN, focusing on density, excels in complex-shaped datasets and outlier identification, grouping points in high-density areas without predefining cluster numbers. We clustered using attributes like *Capacity under long outage* and *Operational capacity of power generation*, chosen for their uniqueness. Dimensionality reduction was achieved through *PCA*, retaining significant data variance in fewer variables. With two principal components, we effectively visualized distinct clusters and outliers in our dataset (see Figure 3).

K-Means. We first use k-means iteratively to determine the optimal value of k using the *elbow method*. Figure 2 shows that the optimal value turns out to be $k = 3$. The *WCSS* (Within-Cluster Sum of Squares) score measures cluster compactness, reflecting the closeness of data points to their cluster centroids by

summing squared distances within clusters. The elbow method selects an optimal k, balancing minimized WCSS with a reasonable cluster count.

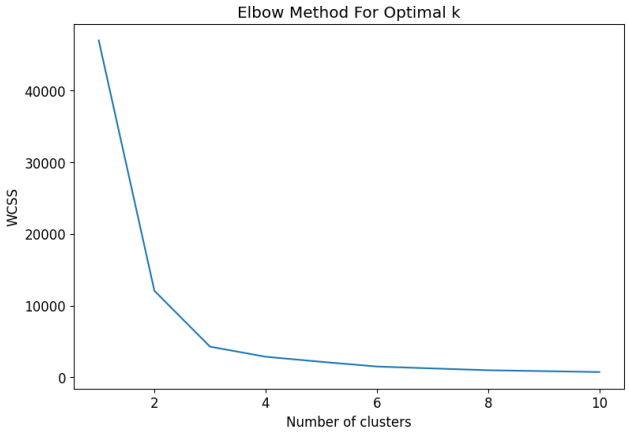


Figure 2: Elbow method for optimal 'k'.

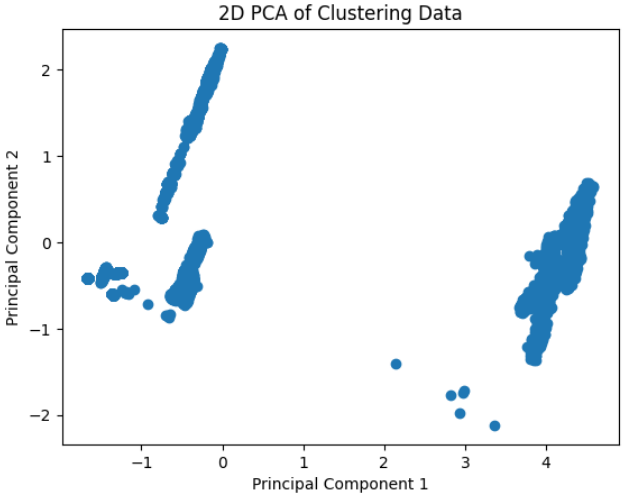


Figure 3: PCA applied to the clustering data.

The results after running k-means with $k = 3$, with and without using PCA on the data are tabulated in Table 2. As the dimension of our clustering dataset is 4, we choose to avoid plotting the results for the raw k-means run. Visualisation of the k-means results with PCA is in Figure 4.

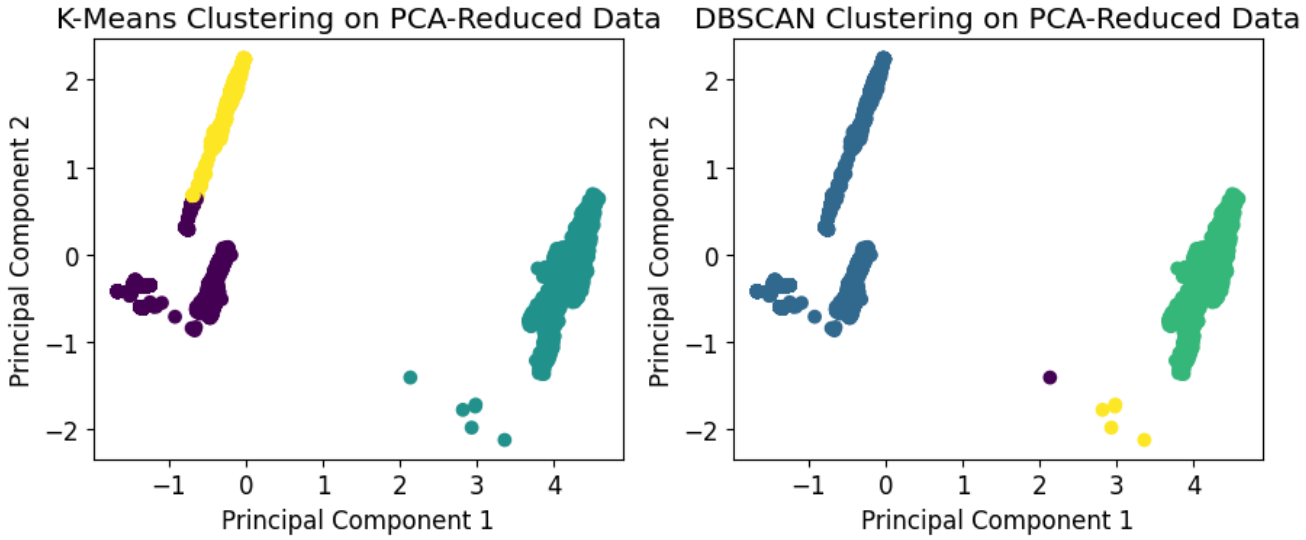


Figure 4: Clustering results after PCA.

K-Means Analysis. In our analysis, the clustering outcomes remained consistent before and after applying PCA 2, suggesting that the dimensionality reduction effectively preserved the variance critical to clustering 4. This observation is evident in the grouping of power types, with *Thermal Coal* and *Thermal Gas Liquid* forming distinct clusters, reflecting their unique characteristics in India's power landscape. The line plots (Figures 1a and 1b) further corroborate these findings. Notably, the application of PCA led to reduced execution times for k-means clustering (Table 4), demonstrating the efficiency gains from dimensionality reduction.

DBSCAN. As stated previously, this algorithm does not need to number of clusters it is supposed to be searching for. However, it takes in two other parameters: ϵ, which is the maximum distance between two samples for one to be considered as in the neighborhood of the other; and the number of samples in a neighborhood for a point to be considered as a core point. In this case, the DBSCAN algorithm automatically identifies the number of relevant clusters to be 3, with another class reserved for outliers/noise.

Type of Power	K-Means			K-Means after PCA		
	Cluster 0	Cluster 1	Cluster 2	Cluster 0	Cluster 1	Cluster 2
Hydro	2177	0	0	2177	0	0
Nuclear	2172	0	0	2172	0	0
Thermal_Coal	0	514	0	0	514	0
Thermal_Coal Lignite	363	1663	0	363	1663	0
Thermal_Diesel	2171	0	0	2171	0	0
Thermal_Gas Liquid	164	0	2013	164	0	2013
Thermal_Lignite	514	0	0	514	0	0

Table 2: Comparison of K-Means clustering results

The results after running DBSCAN, with and without using PCA on the clustering data are tabulated and visualised in Table 3 and Figure 4 respectively.

Type of Power Generation	DBSCAN			DBSCAN after PCA		
	Cluster 0	Cluster 1	Cluster 2	Cluster 0	Cluster 1	Cluster 2
Hydro	2177	0	0	2177	0	0
Nuclear	2172	0	0	2172	0	0
Thermal_Coal	0	514	0	0	514	0
Thermal_Coal Lignite	0	1657	363	363	1657	5
Thermal_Diesel	2171	0	0	2171	0	0
Thermal_Gas Liquid	2177	0	0	2177	0	0
Thermal_Lignite	514	0	0	514	0	0

Table 3: Comparison of DBSCAN clustering results

Method	Execution Time (seconds)
K-Means	0.04289
K-Means after PCA	0.07
DBSCAN	0.58341
DBSCAN after PCA	0.63

Table 4: Execution Times of Clustering Algorithms

DBSCAN Analysis. In our analysis, DBSCAN clustering exhibited similar consistency to K-Means, with most power types like Hydro, Nuclear, Thermal Coal, Diesel, and Lignite showing distinct clustering both pre- and post-PCA. A notable shift occurred in the clustering of Thermal Coal Lignite; while initially split between two clusters, post-PCA analysis primarily grouped it in one cluster, with some outliers. This shift, as visualized in Figure 4, suggests PCA's impact on DBSCAN's perception of its density distribution, highlighting the algorithm's sensitivity to feature representation changes. The execution time for DBSCAN, as shown in Table 4, decreased post-PCA, albeit less dramatically than for K-Means. This reduction in time indicates efficiency improvements from dimensionality reduction, while the smaller drop compared to K-Means might reflect inherent algorithmic optimizations in DBSCAN.

Conclusion. Through our exploratory data analysis of the power generation dataset, we gained critical insights into the efficiency and operational characteristics of different power plants, especially when applying clustering techniques like K-Means and DBSCAN post-PCA. This analysis not only highlighted the variability and performance differences among power generation types but also underscored the importance of context-specific approaches in data handling and interpretation. The project was a valuable learning experience, enhancing our technical skills in data analytics, and deepening our understanding of the complexities inherent in energy sector data.

A Appendix - Metadata

Attribute	Description
Calendar Day	Specific day of the month and year.
Type of Power Generation	Type of power generation: Thermal Coal, Thermal Diesel, Thermal Gas Liquid, Nuclear, Hydro etc.
Monitored capacity	The total power generation capacity being monitored.
Capacity under long outage	The capacity that is not operational due to long-term outages.
New capacity under stabilization or completion of balance works	Indicates new power generation capacity that is either being stabilized or tested before being fully operational.
Capacity of power generation from prime movers (pm)	Prime movers typically refer to the main engines or turbines used in power plants, such as steam turbines, gas turbines, or diesel generators.
Capacity of power generation from fuel oil (fo)	Fuel oil is a type of liquid fuel derived from petroleum, and it is commonly used in power plants as a source of energy.
Capacity of power generation from other sources	The other sources may include alternative or renewable energy sources like solar, wind, hydroelectric, or biomass, depending on the specific power generation system.
Total capacity of power generation from all sources	Total capacity of power generation from all sources combined, Prime Movers (PM) + Fuel Oil (FO) + Other sources.
Capacity outage as percent of monitored	The percentage of the total monitored capacity that is currently out of service.
Operational capacity of power generation	Power generation capacity that is currently in operation and actively producing electricity.
Gross sum of maximum power generated	The total power output without accounting for any consumption or losses in the generation process.
Net sum of maximum power generated	The total power output after accounting for consumption or losses, indicating the actual power available to the grid.
Maximum power output as percentage of capacity on line	"Maximum Power Output" refers to the highest level of power that was generated during a given time period. "Capacity on Line" means the total capacity that is operational and available for use (not including capacity that is offline due to maintenance, outages, etc.). Thus, this percentage represents how much of the operational capacity was actually used to its maximum potential.
Percentage of monitored maximum power output	"Monitored Maximum Power Output" refers to the maximum power output that was expected or planned for a given time period. The percentage indicates how the actual maximum power output compares to this monitored or expected level. It is a measure of performance against planned or standard levels.

Table 5: Metadata for the Capacity Utilization for Power Generation dataset.