STATISTICAL ANALYSIS OF PERFORMANCE OF THERMAL POWER STATIONS

A Project Report Submitted in Partial Fulfillment of the
Requirements for the Degree of
M.Sc. Statistics
(With Specialization in Industrial Statistics)

Submitted by

Mr. Pawar Prasanna Deepak (387441)
Ms. Sapkal Komal Ramesh (387445)
Ms. Patil Diksha Dilip (387429)

Under the Guidance of
Asst. Prof. MANOJ C. PATIL

in the

Department of Statistics, School of Mathematical Sciences,
Kavayitri Bahinabai Chaudhari North Maharashtra
University, Jalgaon-425001.

(Academic Year : 2022-2023)

DEPARTMENT OF STATISTICS
KAVAYITRI BAHINABAI CHAUDHARI
NORTH MAHARASHTRA UNIVERSITY, JALGAON

CERTIFICATE

This is to certify that **Ms. Patil Diksha Dilip, Mr. Pawar Prasanna Deepak** and **Ms. Sapkal Komal Ramesh** are the student of M.Sc. Statistics (with specialization in Industrial Statistics) at Department of Statistics, School of Mathematical Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon have successfully completed their project entitled “Statistical Analysis of Performance of Thermal Power Stations” under my guidance and supervision during the academic year 2022-2023.

Place :- Jalgaon
Date :-

Asst. Prof. Manoj C. Patil
(Project Guide)
Department Of Statistics
Kavayitri Bahinabai Chaudhari
North Maharashtra University,
Jalgaon.

Acknowledgements

On the completion of this project we must acknowledge from the core of our heart to Dr. R. L. Shinde, Head of the Department of Statistics, School of Mathematical Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon for seeking us the desire permission for this project.

We take this opportunity to express our sense of gratitude to our project guide Asst. Prof. Manoj C. Patil for his valuable guidance, immense support, motivation and encouragement to which we could complete our project work successfully. We would like to express deepest appreciation towards Mr. Sachin Kavitke (Chief Executive Engineer, BTPS) whose valuable guidance supported us in completing this project.

We thanks to our parents, friends and classmates to give us moral support. We are also thankful to all for directly or indirectly helping us for our project work.

Mr. Pawar Prasanna Deepak (Seat No. 387441)

Ms. Sapkal Komal Ramesh (Seat No. 387445)

Ms. Patil Diksha Dilip (Seat No. 387429)

Contents

Chapter 1

Introduction

1.1 Introduction

Thermal Power Stations (TPS) play a crucial role in meeting the energy demands of modern society. These power plants utilize the combustion of fossil fuels, such as coal, oil, or natural gas, to generate electricity through the conversion of thermal energy into mechanical energy and then into electrical energy. The performance of thermal power stations is of utmost importance as it directly impacts the efficiency, reliability, and environmental sustainability of power generation.

Statistical analysis of performance data provides a valuable tool for evaluating and understanding the operational characteristics of thermal power stations. By analyzing large datasets containing information on various operational parameters, such as power output, fuel consumption, heat rate, emissions, and maintenance records, we can gain insights into the factors influencing power station performance.

The primary objectives of this statistical analysis are to identify patterns, trends, and relationships within the performance data, as well as to assess the overall efficiency and reliability of the thermal power stations under consideration. Additionally, this analysis aims to identify potential areas for improvement, operational optimization, and cost reduction, leading to enhanced power generation and environmental sustainability.

The statistical analysis will involve employing various techniques, including descriptive statistics, regression analysis, time series analysis, and

hypothesis testing. In this project report, we will present a detailed statistical analysis of the performance data from a selected set of thermal power stations. We will describe the data collection process, explore the data using appropriate statistical techniques, and interpret the results to gain insights into the factors influencing power station performance. Based on these findings, we will provide recommendations and implications for improving the operational efficiency and sustainability of thermal power stations in the future.

1.2 Motivation

Conducting a statistical analysis on performance data of thermal power stations provides an opportunity to apply the statistical techniques and methodologies learned during the M.Sc. program to real-world data. It allows us to gain hands-on experience in analyzing complex datasets and solving real-world problems using statistical tools. The analysis of performance data plays a significant role in the energy sector, as it helps improve the efficiency, reliability, and sustainability of thermal power stations. By conducting this analysis, we can contribute to the field by providing valuable insights and recommendations for optimizing power generation, reducing fuel consumption, and minimizing environmental impact.

1.3 Objectives of the Project

- **Performance Evaluation :** To assess the overall performance of thermal power stations by analyzing key performance indicators (KPIs) such as power output, heat rate, fuel consumption, emissions, and availability. Identify variations in performance across different power stations and over time.
- **Identify Factors Influencing Performance :** To identify the factors that significantly influence the performance of thermal power stations. This may include factors such as ambient temperature, fuel quality, maintenance practices, operational parameters, and design charac-

teristics. Determine the relationships between these factors and the performance indicators.

- **Trend Analysis :** To analyze performance data over time to identify trends and patterns. Determine if there are any significant changes or improvements in performance over the years. Explore the presence of seasonality, long-term trends, or cyclicality in performance data.
- **Benchmarking :** To compare the performance of different thermal power stations within a region or across different regions. Establish benchmarking metrics to evaluate and compare the efficiency and reliability of power stations. Identify best practices and areas for improvement based on performance benchmarks.

By achieving these objectives, a statistical analysis on performance data of thermal power stations can facilitate better understanding, optimization, and management of these crucial energy assets, leading to improved performance, efficiency, and sustainability in power generation.

1.4 Limitations of Study

1. **Data Quality and Availability :** The quality and availability of the data is significantly impacting the statistical analysis. Inaccurate, incomplete, or inconsistent data is leading to biased or unreliable results. The available performance data is not include all relevant variables that can influence the power station's performance. Some important factors, such as maintenance procedures, operator experience, and grid conditions, are not be adequately captured in the available data.
2. **Causality vs. Correlation :** The absence of these variables are limiting the analysis and the ability to fully understand the performance drivers. Statistical analysis can identify correlations between different variables, but it may not establish causality. While certain factors may be correlated with the performance indicators, determining the causal relationships requires further investigation and potentially experimental design.
3. **External Factors :** Thermal power station performance can be

influenced by external factors beyond the control of the plant operators, such as changes in fuel prices, government policies, or environmental regulations. These external factors can introduce additional complexities into the analysis and may confound the interpretation of the results.

4. **Assumptions and Simplifications :** Statistical analysis often relies on certain assumptions and simplifications to make the analysis manageable. These assumptions may not fully capture the complexity of the power station operations, and the simplifications can introduce uncertainties and limitations into the analysis.

5. **Complex Interactions :** The performance of thermal power stations is influenced by the interactions among multiple factors, which can be challenging to capture in statistical models. Complex relationships and interactions may exist, requiring advanced modeling techniques or alternative analytical approaches beyond traditional statistical methods.

6. **Technological Advancements :** The performance data used for analysis may be outdated, and the power station technology and operational practices may have evolved since the data was collected. Changes in technology, equipment upgrades, or operational improvements may render the analysis less relevant to current power station operations.

1.5 Thermal Power Plant Working

◇ **Thermal Electricity**

- **Traditional thermal power plants :** also called combustion power plants, they operate with energy produced by a steam boiler fuelled by coal, natural gas, heating oil, as well as by biomass. The steam activates a turbine which, in turn, drives an alternator to produce electricity.
- **Combined cycle gas power plants (or steam-gas turbine plants) :** These combine a gas turbine and a traditional thermal plant to generate electricity. Unlike conventional thermal power plants, the residual energy of the gases is used for another cycle. This is one of the reasons

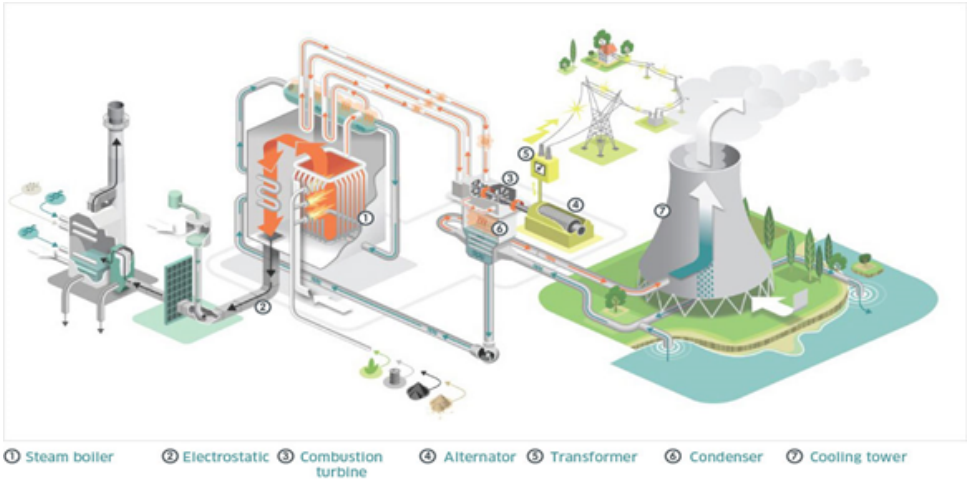

Figure 1.1: Thermal Power Plant

for which these kind of plants are more efficient (by 56%).

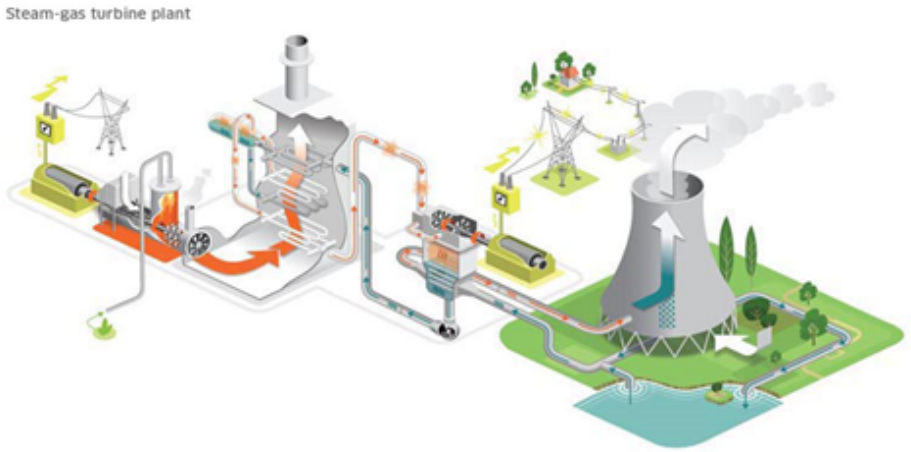

Figure 1.2: Steam-Gas Turbine Plant

Initially, gas is injected into the combustion turbine. It generates steam, which is then supplied to another turbine. The combustion turbine and steam turbine work in tandem to turn one or more alternators, which produce electricity.

- **Recovery of blast furnace gas :** Electricity production can also be obtained by recovering and recycling gases from iron and steel production (blast-furnace gas, coking plant gas, steel plant gas), using a traditional boiler (a comparable technology to traditional thermal power plants) or in a heat recovery boiler in a combined cycle gas plant.

- **Gas turbines and turbojets :** Mostly used to supplement the electricity production of other thermal plants, gas turbine and turbojet units can take over very rapidly in the event of a failure of other power plants or of unexpected peaks in consumption.

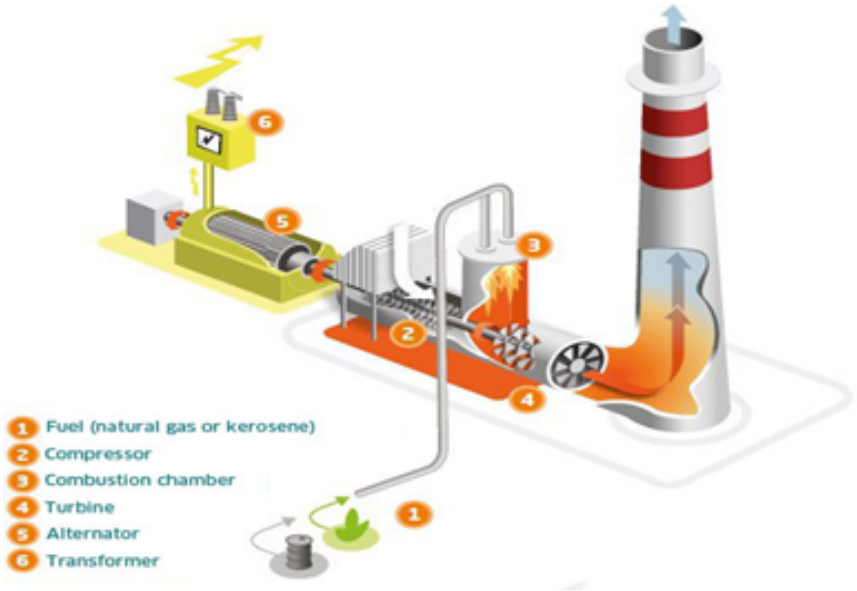

Figure 1.3: Gas turbine and turbojets

- **Cogeneration units :** These produce heat (their principal role) simultaneously with electricity (their secondary role) in a single installation and employing a single fuel. It is a highly energy-efficient solution. By recovering thermal energy normally lost in power generation, these plants are able to produce electricity and heat with efficiency of close to 90%, which is of great interest for industrial sites.
- **Co-generation :** A gas-powered generator drives an alternator that produces electricity. Heat recovered from the cooling of the motor and the combustion gases heats a water circuit thanks to heat exchangers.

◇ Working Components of Thermal Power Plant

1. Thermal Power Plant
2. Fuel Storage and Water Handling Plant
3. Water Treatment Plant
4. Steam Boiler

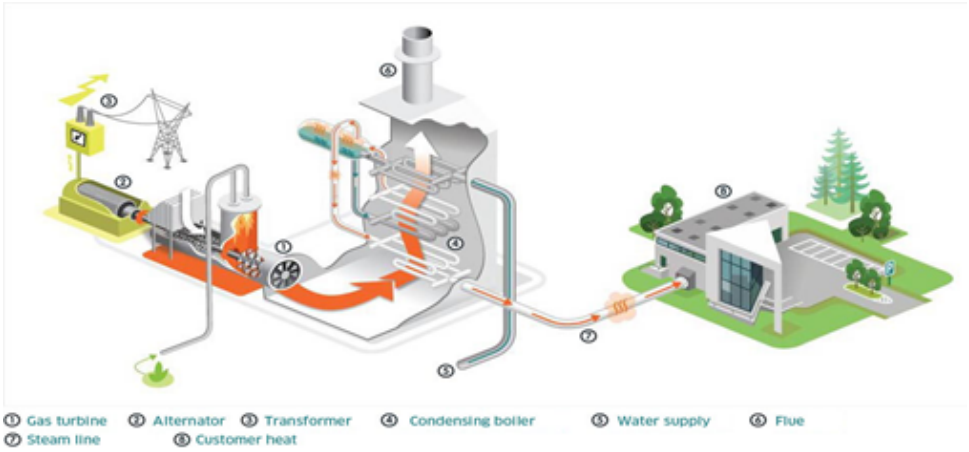

Figure 1.4: Co-generation

5. Forced Draft (FD) fan
6. Turbine

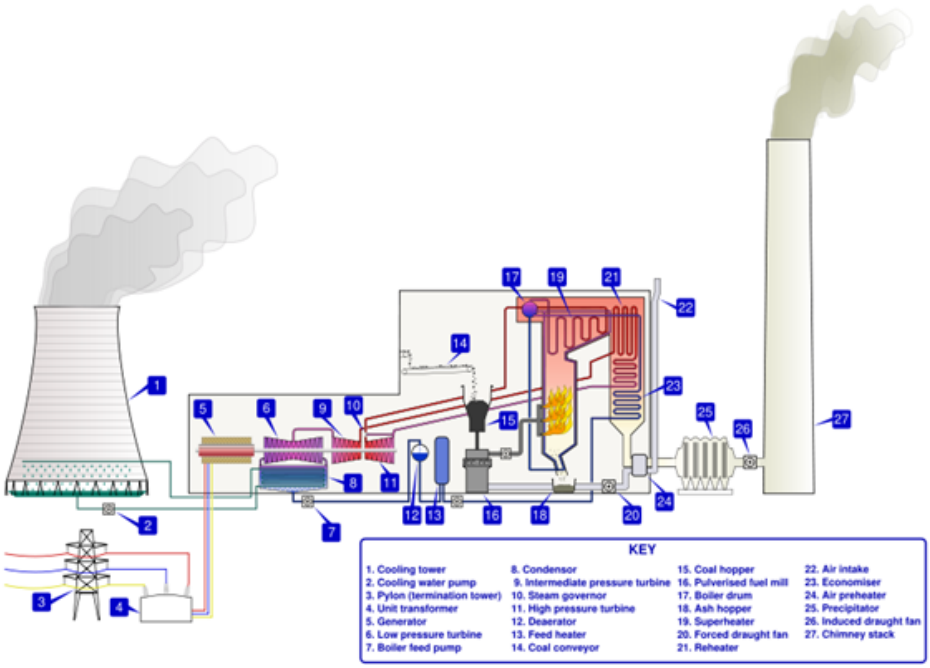

Figure 1.5: Working Components of Thermal Power Plant

1.6 Organization Profile

- **Founded :** 6 June 2005
- **Headquarters :** Bandra, Mumbai, India
- **Owner :** Government of Maharashtra

- **Organization type :** State-owned enterprise
- **Area served :** Maharashtra
- **Purpose :** Electricity generation
- **Parent organization :** Maharashtra State Electricity Board, M.S.E.B. Holding Company Limited

INSTALLED CAPACITY OF MAHAGENCO : (As on 31-12-2022) (Thermal Power Stations Only)

SR.NO.	POWER STATION	UNITS & CAPACITY (MW)	INSTALLED CAP. (MW)
1	KORADI	1 x 210 + 3 x 660	2190
2	NASIK	3 x 210	630
3	BHUSAWAL	1 x 210 + 2 x 500	1210
4	PARAS	2 x 250	500
5	PARLI	3 x 250	750
6	KHAPERKHEDA	4 x 210 + 1 x 500	1340
7	CHANDRAPUR	2 x 210 + 5 x 500	2920
Total			9540

Table 1.1: INSTALLED CAPACITY OF MAHAGENCO(2022)

1.7 Data Collection

We have collected these data from Bhusawal Thermal Power Station. It is located 8 km away from Bhusawal city of Jalgaon district in Maharashtra. The name of place where it is located is Deepnagar, which means City of Lights. The power plant is a coal based power plants of MAHAGENCO. We Identify the sources from which the required data can be obtained. This may include internal data sources such as plant management systems, operational databases, and maintenance records. Also we ensured that the necessary permissions and access rights are obtained to collect the required

data. This involved coordination with the power plant operators, data custodians, and relevant authorities responsible for providing access to the data.

1.8 Data Features

1. **M.U. Generated :** In Power Plants we generate electricity in bulk amount. It is convenient to measure electricity in LARGE unit. M.U. Generated refers to the total electrical energy generated by a thermal power plant, measured in million units (MUs) or megawatt-hours (MWh). It represents the actual output of electricity produced by the power plant during a specific period, such as a day, month, or year. M.U. Generated is a crucial parameter for assessing the power plant's performance and productivity.So, in MU i.e., Millions of Units we measure generated electricity.
   1 MU=1 million Units=10,00,000 Units
   1**Unit** = $1KWh = 1000Wh$

2. **Plant Load Factor (PLF) (CEA) :** It is the ratio of actual energy generated by a power plant to the maximum possible energy that could have been generated during a specific period, usually a year. It indicates the efficiency and utilization of the power plant. The formula for calculating PLF is as follows:

$$PLF = \frac{\text{Actual Energy Generated}}{\text{Maximum Possible Energy Generation}} \times 100$$

3. **Availability Factor (AVF) (CEA) :** It is the ratio of the actual available hours of operation of a power plant to the total hours in a specific period. It represents the plant's availability for generating electricity. The formula for calculating the Availability Factor is as follows:

$$\text{Availability Factor} = \frac{\text{Actual Output}}{\text{Maximum Possible Output}} \times 100$$

4. **Loadability (%) :** Loadability refers to the maximum load that a power plant can handle without exceeding its design limits. It is expressed as

a percentage of the power plant's maximum capacity. The formula for calculating the Loadability Factor is as follows:

$$\text{Loadability Factor} = \frac{\text{Actual Load}}{\text{Maximum Capacity}} \times 100$$

5. **PLF (%) (as per MERC regulation) :** This parameter represents the Plant Load Factor required to meet the regulations set by the Maharashtra Electricity Regulatory Commission (MERC). It specifies the minimum utilization level that the power plant must achieve.

6. **AVF (%) (as per MERC regulation) :** Similar to PLF, this parameter represents the Availability Factor required to meet the regulations set by the Maharashtra Electricity Regulatory Commission (MERC). It specifies the minimum availability level that the power plant must maintain.

7. **Oil Consumption :**

   (a) Total in KL : The total quantity of oil consumed by the power plant, measured in kiloliters.

   (b) Sp. Oil Cons. ML/Kwh : Specific Oil Consumption, which indicates the amount of oil consumed per unit of electricity generated, measured in milliliters per kilowatt-hour (ML/Kwh).

   (c) Total oil cost (Rs. Crs) : The total cost of oil consumed by the power plant, measured in crores of Indian Rupees (Rs. Crs).

   (d) Oil Cost (Paise/Kwh) : The cost of oil per unit of electricity generated, measured in paise (hundredths of a Rupee) per kilowatt-hour (Kwh).

8. **% Forced Outages (Total) :** The formula for calculating % Forced Outages is as follows:

$$\%\text{Forced Outages} = \frac{\text{Total Duration of Forced Outages}}{\text{Total Operating Time}} \times 100$$

   (a) Due to tube leakages : The percentage of forced outages caused by tube leakages in the power plant's equipment.

(b) Due to turbine & Aux. : The percentage of forced outages caused by issues related to the turbine and auxiliary systems.

(c) Any Other reasons : The percentage of forced outages caused by reasons other than tube leakages or turbine and auxiliary issues.

9. **% Planned Outages (Total) :** The formula for calculating % Planned Outages is as follows:

$$\%\text{Planned Outages} = \frac{\text{Total Duration of Planned Outages}}{\text{Total Operating Time}} \times 100$$

(a) Unit Overhaul : The percentage of planned outages for conducting unit overhauls, which involve maintenance, repair, or replacement of major components.

(b) Other Outages : The percentage of planned outages for reasons other than unit overhauls, such as scheduled maintenance or repairs.

10. **% Reserve Shut Down/Outages :**

(a) L.D./System problem : The percentage of shutdowns or outages due to load dispatch or system-related issues.

(b) Less coal receipt : The percentage of shutdowns or outages caused by inadequate coal supply to the power plant.

(c) Wet coal Problem : The percentage of shutdowns or outages caused by issues related to wet coal, which may impact its combustion efficiency.

11. **% Coal Realisation :** Percentage of coal realized or obtained compared to the required or planned coal quantity.

$$\%\text{ Coal Realisation} = \frac{\text{Coal Consumption}}{\text{Required Coal}} \times 100$$

12. **Coal Consumption (MT) :** Total consumption of coal by the power plant over a given period, measured in metric tons.

13. **Sp. Coal Consumption (Kg/Kwh) :** Specific coal consumption (in

kilograms) per unit of electricity generated (in kilowatt-hours).

$$\text{Sp. Coal Consumption} = \frac{\text{Coal Consumption } \times 1000}{\text{Net Generation}}$$

14. **G.C.V. (Kcal/kg) :** Gross Calorific Value (in kilocalories per kilogram) of the coal being used in the power plant.

15. **Heat Rate (Kcal/Kwh) :** Heat rate is a measure of the thermal efficiency of a power plant. It represents the amount of heat energy (in kilocalories) required to generate one kilowatt-hour of electricity.

$$\text{Heat Rate } = \frac{\text{Coal Consumption} \times G.C.V.}{\text{Net Generation}}$$

16. **Transit Loss in Coal (%) :** Transit Loss in Coal represents the percentage of coal lost during transportation from the source to the power plant.

17. **Sp. Raw Water consumption (Lt./Kwh) :** Specific raw water consumption (in liters) per unit of electricity generated (in kilowatt-hours).

18. **Sp. Softened Water consumption (Lt./Kwh) :** Specific softened water consumption (in liters) per unit of electricity generated (in kilowatt-hours).

19. **DM water consumption (Lt./Kwh) :** Specific demineralized (DM) water consumption (in liters) per unit of electricity generated (in kilowatt-hours).

20. **Sp. Lub.Oil consumption (ml/Kwh) :** Specific lubricating oil consumption (in milliliters) per unit of electricity generated (in kilowatt-hours).

21. **M.U. Lost :** M.U. Lost represents the electricity generation lost or not achieved due to various reasons.

    (a) Due to less coal receipt : Percentage of lost electricity generation due to insufficient coal receipt.

    (b) Wet Coal and choking Problem: Percentage of lost electricity generation due to issues with wet coal and equipment choking. L.D. restriction: Percentage of lost electricity generation due to load dispatch restrictions.

22. **Aux. consumption :** Aux. consumption represents the electricity consumption by auxiliary equipment and systems in the power plant.

23. **% Aux. consumption :** Percentage of electricity consumed by auxiliary equipment and systems compared to the total electricity generated.

$$\% \text{ Aux. consumption} = \frac{\text{Aux. consumption}}{\text{Net Generation}} \times 100$$

24. **Net Generation :** It represents the actual electricity generated by the thermal power plant after deducting the auxiliary consumption. It is measured in megawatt-hours (MWh).
Formula for Net Generation:

$$\text{Net Generation} = \text{M.U. Generated} - \text{Aux. consumption}$$

These parameters helps to evaluate the operational efficiency, resource utilization, and overall performance of a thermal power plant. The formulas mentioned provide a quantitative measure of these parameters based on various inputs and outputs of the power plant.

◇ **Abbreviates :**

We have created these abbreviations for performing analysis, and it was complicated to use their full forms during the analysis process. Using abbreviations in the analysis helps streamline the process by condensing lengthy terms or phrases into shorter representations. This saves time and effort, allowing analysts to focus on the core aspects of their work. However, relying solely on abbreviations can sometimes be challenging, as it requires a clear understanding and familiarity with the full forms they represent. While performing the analysis, the use of abbreviations facilitates efficient communication and documentation, ensuring clarity and conciseness in the analytical reports or discussions.

Sr.No.	**Parameters**	**Abbreviates**
1	M.U. Generated	MUG
2	Plant Load Factor (CEA)	PLFC
3	Availability Factor (CEA)	AFC
4	Loadability(%)	L
5	PLF(%)(as per MERC regulation)	PLFM
6	AVF(%)(as per MERC regulation)	AVFM
7	Oil Consumption	Oil Cons
	A) Total in KL	Total
	B)Sp.Oil Cons. ML/Kwh.	SOC
	C) Total oil cost (Rs. Crs)	TOC
	D) Oil Cost (Paise/Kwh)	OC
8	% Forced Outages(Total)	FO
	A] Due to tube leakages	FO1
	B] Due to turbine & Aux.	FO2
	C] Any Other reasons	FO3
9	% Planned Outages(Total)	PO
	A] Unit Overhaul	UO
	B] Other Outages	OO
10	% Reserve Shut Down/Outages	RSD
	A] L.D./System problem	RSD1
	B] Less coal receipt	RSD2
	C] Wet coal Problem	RSD3
	D] Water shortage problem	RSD4
11	% Coal Realisation	CR
12	Coal Consumption (MT)	CC
13	Sp. Coal Consumption (Kg/Kwh)	SCC
14	G.C.V. (Kcal/kg)	GCV
15	Heat Rate (Kcal/Kwh)	HR
16	Transit Loss in Coal(%)	TL
17	Sp. Raw Water consumption (Lt./Kwh)	SRWC
18	Sp. Softened Water consumption(Lt./ Kwh)	SFWC
19	Sp.DM water consumption (Lt./ Kwh)	SDWC
20	Sp. Lub.Oil consumption (ml/Kwh)	SLOC
21	M.U.Lost	MUL
	1] Due to less coal receipt	R1
	1a] Economic shutdown	R11
	2] Wet Coal & choking Problem	R2
	3] L.D.restriction	R3
	4] Exessive Rain	R4
22	Aux.consumption	AC
23	% Aux.consumption	PAC
24	Net Generation (MU)	NG

Table 1.2: Thermal Performance Features Abbreviates

Chapter 2

EDA of Performance Data

In Statistics, Exploratory Data Analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.

2.1 Summary Statistics of Performance Parameters

- **Performance Data of Thermal Power Stations**

	MW	Station	Months	MUG	PLFC	AFC	L	PLFM	AVFM
0	420	Bhusawal	2011-04-30	237.404	78.506614	95.97	81.803286	84.14	84.14
1	420	Bhusawal	2011-05-31	213.522	68.331413	86.72	78.795449	70.60	70.60
2	420	Bhusawal	2011-06-30	227.244	75.146825	94.37	79.629994	76.96	76.96
3	420	Bhusawal	2011-07-31	170.155	54.453085	73.88	73.704771	56.68	56.68
4	420	Bhusawal	2011-08-31	138.150	44.210829	76.64	57.686364	46.23	46.23

- This is the data of 7 thermal power stations namely;

  'Bhusawal', 'Chandrapur', 'Khaperkheda', 'Koradi','Nashik','Paras', 'Parli'.

	Oil Cons	Total	SOC	TOC	OC	FO	FO1	FO2	FO3	PO	UO
0	Oil Cons	583.0	2.455729	1.59	6.697444	4.04	1.04	0.0	3.00	0.00	0.00
1	Oil Cons	687.0	3.217467	1.91	8.945214	13.28	13.14	0.0	0.14	0.00	0.00
2	Oil Cons	627.0	2.759149	1.70	7.480946	5.63	5.39	0.0	0.24	0.00	0.00
3	Oil Cons	811.0	4.766243	3.25	19.100232	14.16	11.79	0.0	2.37	11.96	0.00
4	Oil Cons	2302.0	16.663047	8.93	64.639884	7.59	5.24	0.0	2.35	15.08	15.08

	OO	RSD	RSD1	RSD2	RSD3	RSD4	CR	CC	SCC	GCV	HR
0	0.00	0.00	0.0	0.0	0.00	0.0	81.80	206895.0	0.871489	3101.0	2725.0
1	0.00	0.00	0.0	0.0	0.00	0.0	64.28	187896.0	0.879984	3099.0	2756.0
2	0.00	0.00	0.0	0.0	0.00	0.0	53.76	199974.0	0.879997	3070.0	2729.0
3	11.96	0.00	0.0	0.0	0.00	0.0	85.84	149739.0	0.880015	3050.0	2730.0
4	0.00	0.69	0.0	0.0	0.69	0.0	79.31	129633.0	0.938350	3023.0	3000.0

	SFWC	SDWC	SLOC	MUL	R1	R11	R2	R3	R4	AC	PAC	NG
0	2.20	0.08	0.019	0.000	0.0	0.0	0.000	0.0	0.0	24.041	10.126620	213.363
1	2.38	0.11	0.032	0.000	0.0	0.0	0.000	0.0	0.0	23.096	10.816684	190.426
2	2.13	0.09	0.027	0.000	0.0	0.0	0.000	0.0	0.0	24.248	10.670469	202.996
3	2.48	0.14	0.038	12.955	0.0	0.0	12.955	0.0	0.0	19.951	11.725192	150.204
4	2.68	0.17	0.042	52.728	0.0	0.0	52.728	0.0	0.0	19.434	14.067318	118.716

Table 2.1: First 5 rows of performance data

- It has 433 rows and 45 columns.
- The variables ('MW', 'Station', 'Months', 'MUG', 'PLFC', 'AFC', 'L', 'PLFM', 'AVFM','Oil Cons', 'Total', 'SOC', 'TOC', 'OC', 'FO', 'FO1', 'FO2', 'FO3','PO', 'UO', 'OO', 'RSD', 'RSD1', 'RSD2', 'RSD3', 'RSD4', 'CR', 'CC','SCC', 'GCV', 'HR', 'TL', 'SRWC', 'SFWC', 'SDWC', 'SLOC', 'MUL', 'R1"R11', 'R2', 'R3', 'R4', 'AC', 'PAC', 'NG') are of float64 datatype.

- **Descriptive Statistics of Performance Features**

	MW	MUG	PLFC	AFC	L	PLFM
count	433.000000	433.000000	433.000000	433.000000	433.000000	433.000000
mean	895.473441	380.950114	59.490836	78.312044	72.783848	62.389630
std	578.380017	283.068825	20.831523	22.771752	16.951568	22.556629
min	420.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	500.000000	201.560000	45.694124	68.679000	67.909885	47.235044
50%	630.000000	301.069000	63.094086	86.720000	76.126171	65.740000
75%	1130.000000	451.103000	75.531021	94.370000	82.223554	79.640000
max	2340.000000	1364.110000	94.218083	100.000000	97.621272	99.980000

	AVFM	Total	SOC	TOC	OC	FO
count	433.000000	433.000000	433.000000	433.000000	433.000000	433.000000
mean	66.073304	1072.835187	3.718848	4.605989	16.054633	6.825814
std	23.533123	1500.293354	5.781719	6.488092	24.962011	7.514304
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	50.100000	220.000000	0.641135	0.990000	2.864815	2.079861
50%	69.520000	498.000000	1.458517	2.170000	6.463251	4.920000
75%	85.220000	1154.000000	3.819722	5.060000	16.559508	9.000909
max	100.000000	9446.000000	35.772633	39.880000	159.985655	61.830000

	FO1	FO2	FO3	PO	UO	OO
count	433.000000	433.000000	433.000000	433.000000	433.000000	433.000000
mean	2.893074	0.246129	3.686611	6.556560	4.555819	2.000741
std	4.346701	1.426098	6.281575	12.446547	11.135056	6.689703
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.240000	0.000000	0.000000	0.000000
50%	1.810000	0.000000	1.430000	0.000000	0.000000	0.000000
75%	4.550531	0.000000	4.175373	7.320265	0.000000	0.000000
max	45.917000	18.354643	61.830000	82.000000	82.000000	36.920000

	RSD	RSD1	RSD2	RSD3	RSD4	CR
count	433.000000	433.000000	433.000000	433.000000	433.000000	433.000000
mean	8.306957	0.633059	3.470053	0.258302	3.833118	64.154907
std	20.465978	4.505409	10.128799	1.592675	17.495948	35.868517
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	49.050000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	70.110000
75%	5.330866	0.000000	0.000000	0.000000	0.000000	85.960000
max	100.000000	65.040214	66.990000	19.730000	100.000000	309.000000

	CC	SCC	GCV	HR	TL	SRWC
count	4.330000e+02	433.000000	433.000000	433.000000	433.000000	433.000000
mean	3.167061e+05	2.995923	2976.183434	2601.486365	25.424520	5.593895
std	2.392383e+05	22.449133	643.794659	469.433833	254.665613	8.262752
min	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.788210e+05	0.800881	2894.211713	2545.957663	0.040000	3.508831
50%	2.421590e+05	0.844991	3089.861916	2667.000000	0.660000	4.573874
75%	3.741660e+05	0.908312	3252.000000	2751.000000	0.820000	6.140000
max	1.174486e+06	254.966234	4158.000000	3429.000000	2672.139055	160.343345

	SFWC	SDWC	SLOC	MUL	R1	R11
count	433.000000	433.000000	433.000000	433.000000	433.000000	433.000000
mean	2.154536	0.150617	0.040750	83.947434	20.013294	7.704226
std	2.023749	0.439849	0.196901	152.828438	46.040501	36.551048
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.687000	0.069000	0.017449	4.450000	0.000000	0.000000
50%	2.137000	0.093000	0.024000	25.645000	0.000000	0.000000
75%	2.510000	0.134000	0.036000	91.505000	11.802000	0.000000
max	29.428044	7.196067	4.030389	840.720000	280.359800	312.480000

	R2	R3	R4	AC	PAC	NG
count	433.000000	433.000000	433.000000	433.000000	433.000000	433.000000
mean	13.242813	14.116913	28.870189	39.398070	11.158358	342.675283
std	40.427808	31.493549	135.221523	24.631403	5.727865	258.616318
min	0.000000	0.000000	0.000000	1.818240	0.000000	-3.357000
25%	0.000000	0.000000	0.000000	22.789000	9.697880	176.608000
50%	0.000000	1.848000	0.000000	33.153000	10.909755	271.457000
75%	3.304000	12.331000	0.000000	49.745000	12.271053	403.097000
max	387.824000	300.304000	840.720000	125.025800	88.803664	1246.397000

Table 2.2: Descriptive Statistics of Performance Features

- Now we have calculated correlation among all parameters is as follows and visualize it using heatmap plot:
- Lets visualize the positively correlated variables using heatmap plot figure 2.2.

Figure 2.1: Correlation between all parameters

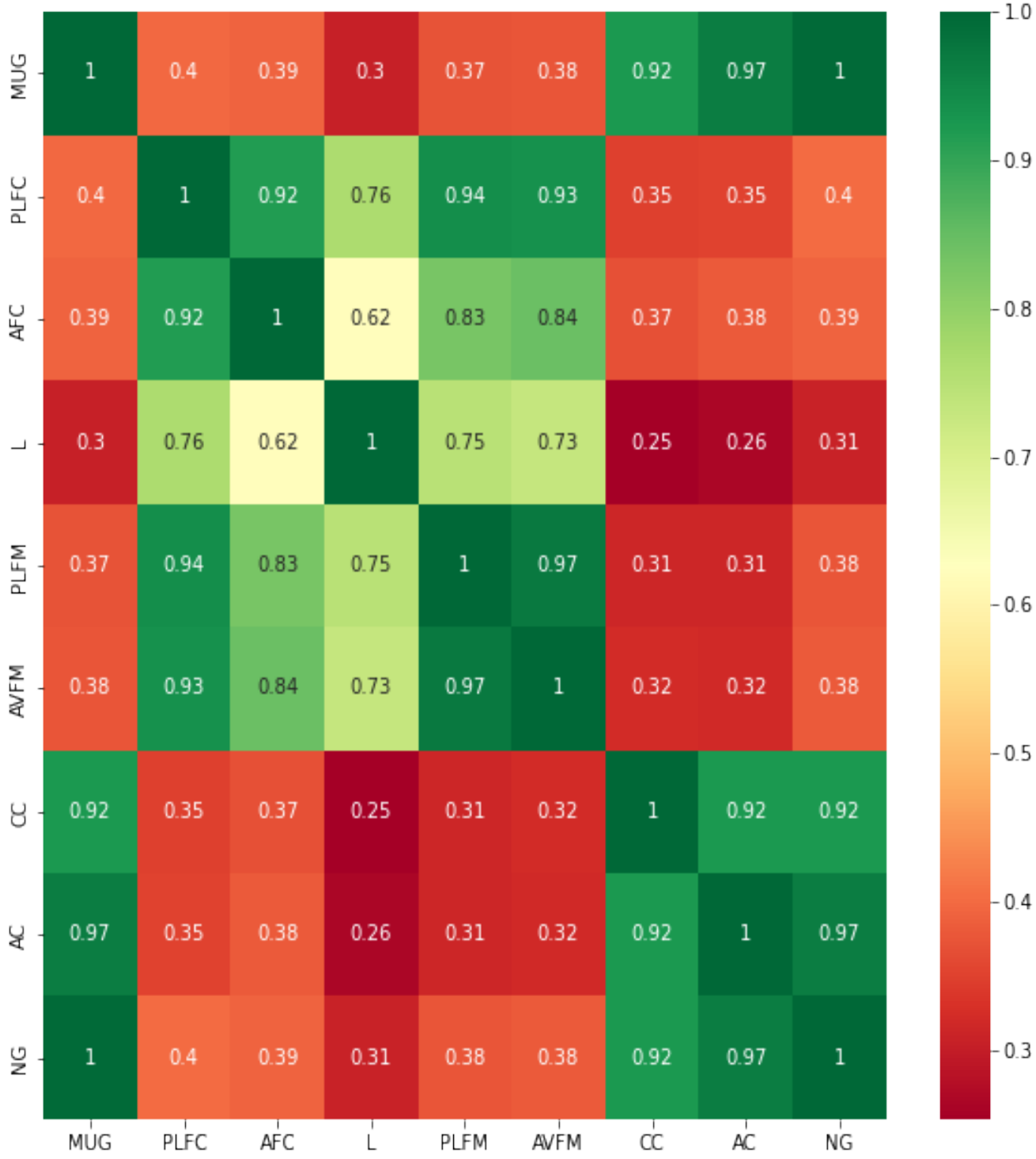

Figure 2.2: Positively Correlated Variables

◇ Calculating Average Net Generation

If we take average net generation of all seven stations from April 2011 to May 2016 and plot it on pie chart then we can observe that Chandrapur has maximum electricity generation after that Khaperkheda, Nashik, Parli, Paras, Koradi, Bhusawal resp. Note that yearly data collection starts from April to March.

Stations	Avg NG 2011-16	Avg NG 2011-12	Avg NG 2012-13	Avg NG 2013-14	Avg NG 2014-15	Avg NG 2015-16
Chandrapur	738.75839	997.70221	339.4146	732.91102	955.51784	923.92916
Khaperkheda	522.75451	439.55027	412.46141	518.15483	620.426	616.45424
Nashik	314.03070	307.92088	323.30121	309.92891	308.48083	324.64866
Parli	293.16595	343.2905	388.57914	188.66931	340.43777	104.22601
Paras	231.91440	212.58470	222.05783	253.95575	226.75166	251.0511
Koradi	175.08226	240.878	174.6277	164.65008	169.23041	122.41668
Bhusawal	132.65424	170.93658	144.769	130.99891	61.390916	120.63372

Table 2.3: Average Net Generation for all Stations

In above table 2.3 we have calculated Average Net Generation for each year separately and arrange it with respective stations.

◇ Plotting Average Net Generation

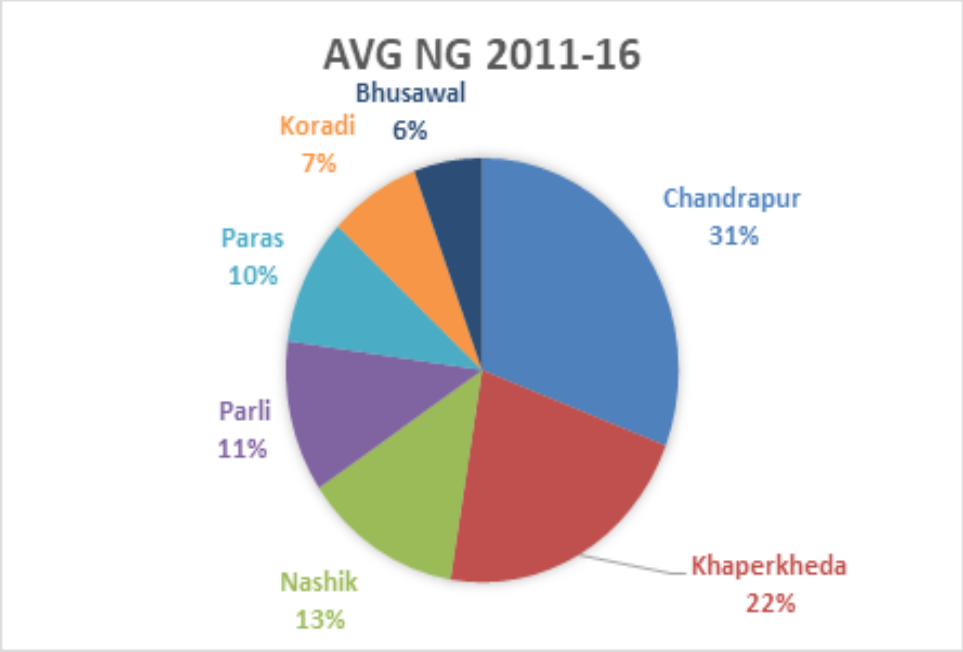

Figure 2.3: Average Net Generation for 2011-16

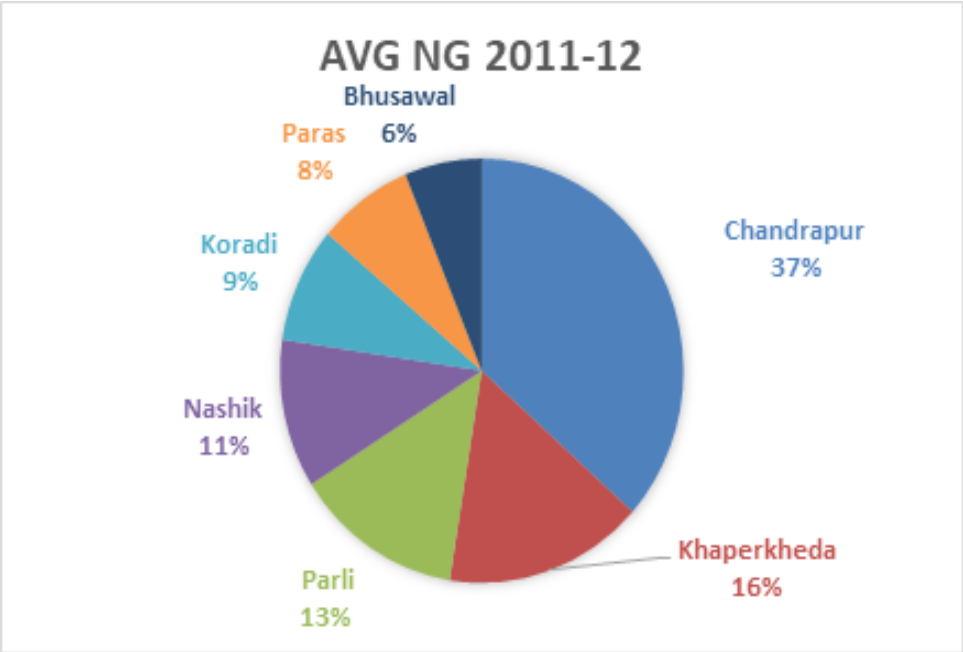

Figure 2.4: Average Net Generation for 2011-12

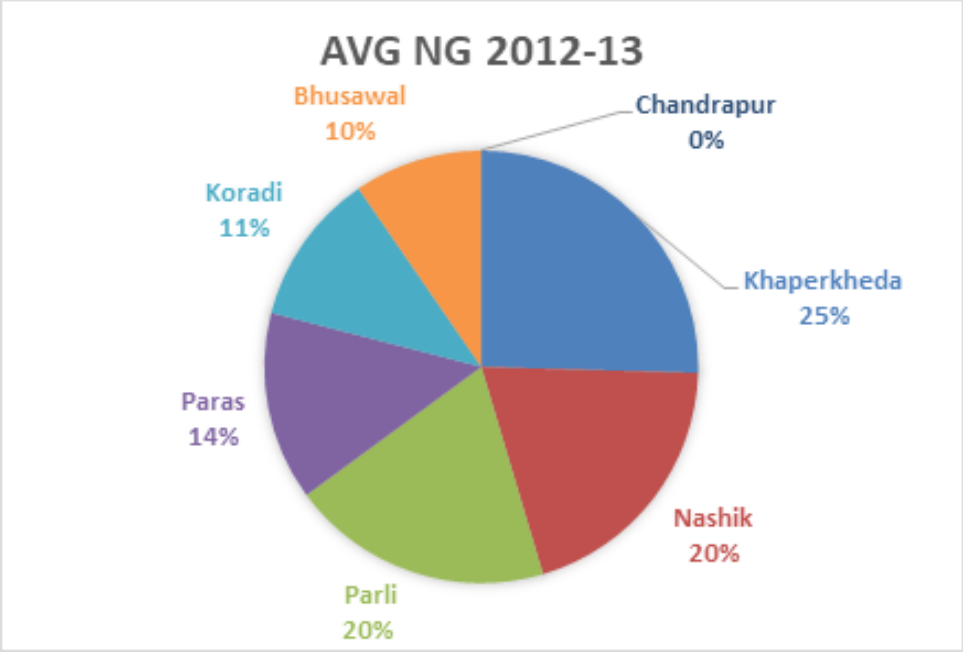

Figure 2.5: Average Net Generation for 2012-13

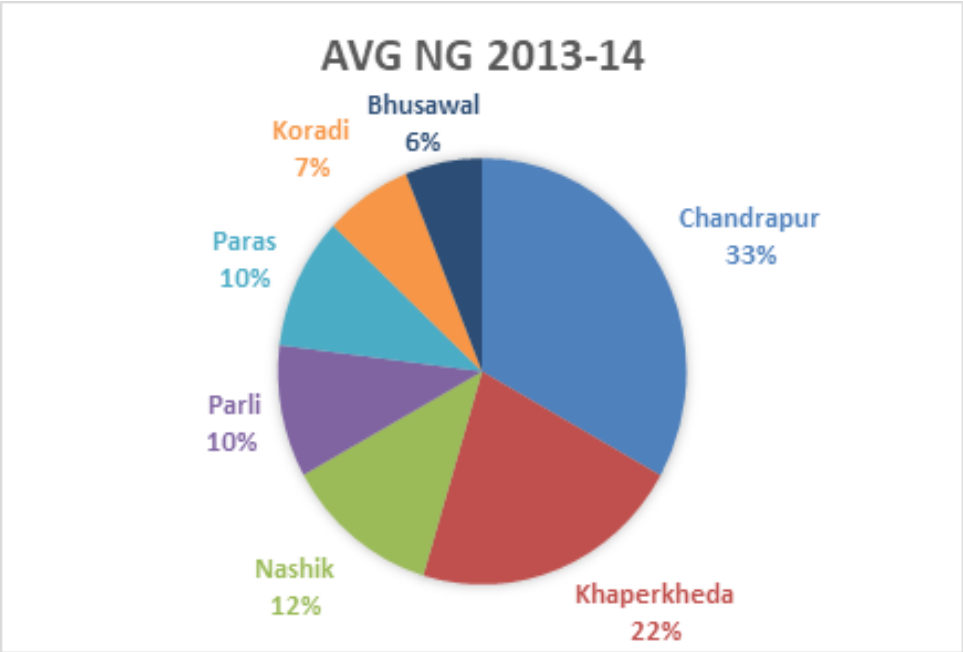

Figure 2.6: Average Net Generation for 2013-14

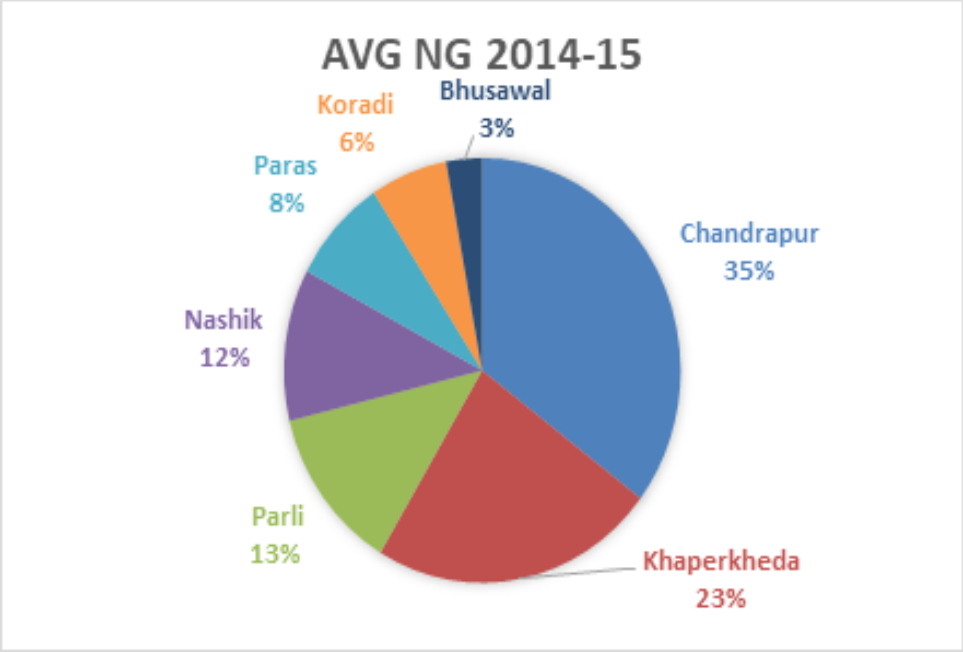

Figure 2.7: Average Net Generation for 2014-15

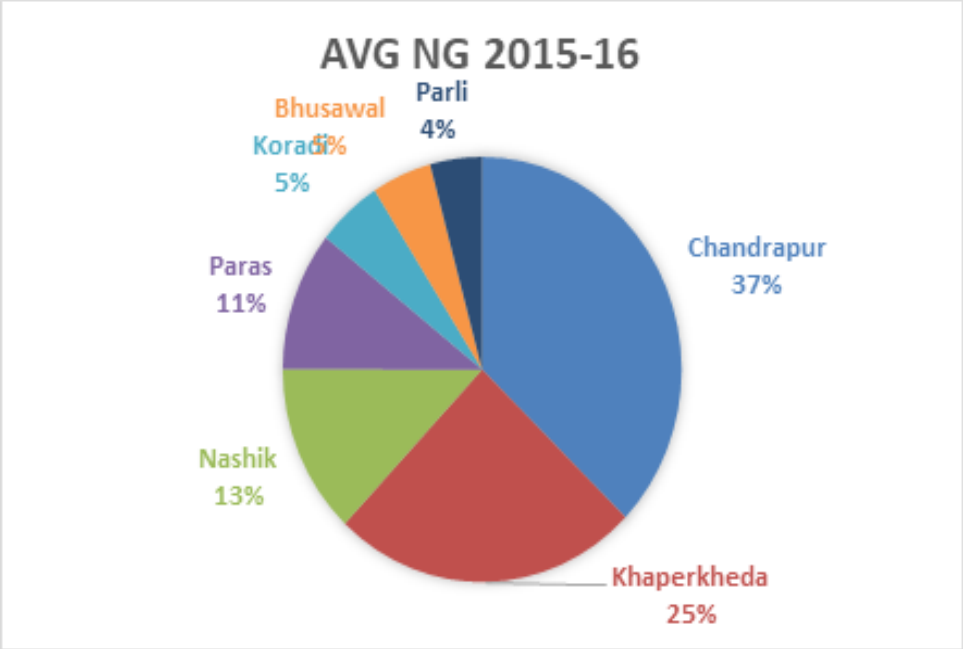

Figure 2.8: Average Net Generation for 2015-16

2.2 Efficiency of Thermal Power Plants

For calculating efficiency firstly we need to find possible generation in 24 hrs for every plant. Using below formula we can calculate possible generation in 24 hrs as shown in following table.

$$\text{Possible Generation in 24 hrs (MU)} = 24 \times \frac{(\text{Plant(MW)})}{1000}$$

Using above formula we can calculate possible generation in 24 hrs as shown in following table.

	Power Plant Capacity	**Generation in 24 hrs (MU)**	
Stations	**Plant (MW) 2011-13-16**	**in 24 hrs(2011-13)**	**in 24 hrs(2013-16)**
Chandrapur	2340	56.16	56.16
Khaperkheda	1340 & 840	20.16	32.16
Parli	1130	27.12	27.12
Nashik	630	15.12	15.12
Koradi	620	14.88	14.88
Paras	500	12	12
Bhusawal	420	10.08	10.08

Table 2.4: Possible generation in 24Hrs

Using above table values we can calculate possible generation for 1 year

(i.e. April to March). Note that 2012 and 2016 are leap years so both have 366 days.

Now possible generation in 1 year (i.e. April to March) can be calculated as :

Possible Generation in 1 year (MU)= 365 days × Possible Generation in 24 hrs

Using above formula we can calculate possible generation for 1 year as shown in following table.

	Power Plant Capacity	**Possible Generation(MU)**				
Stations	**Plant (MW) 2011-13-16**	**2011-12**	**2012-13**	**2013-14**	**2014-15**	**2015-16**
Chandrapur	2340	20554.56	20498.4	20498.4	20498.4	20554.56
Khaperkheda	1340 & 840	9925.92	9898.8	11738.4	11738.4	11770.56
Parli	1130	7378.56	7358.4	9898.8	9898.8	9925.92
Nashik	630	5533.92	5518.8	5518.8	5518.8	5533.92
Koradi	620	5446.08	5431.2	5431.2	5431.2	5446.08
Paras	500	4392	4380	4380	4380	4392
Bhusawal	420	3689.28	3679.2	3679.2	3679.2	3689.28

Table 2.5: Possible Generation (MU) for 1 year

Now Actual total generation in 1 year is given as follows:

	Power Plant Capacity	**Actual Generation(MU)**				
Stations	**Plant (MW) 2011-13-16**	**2011-12**	**2012-13**	**2013-14**	**2014-15**	**2015-16**
Chandrapur	2340	2062.086	12209.83	0	9918.073	11385.84
Khaperkheda	1340 & 840	12209.83	5330.859	4813.777	6453.282	7513.382
Parli	1130	5330.859	4499.896	3690.874	3007.971	4066.334
Nashik	630	2951.544	3781.66	3777.767	3678.265	3730.103
Koradi	620	3781.66	2951.544	2159.782	2071.034	2028.721
Paras	500	2551.016	2551.016	2670.064	2998.624	2598.618
Bhusawal	420	4499.896	2062.086	1810.74	1826.904	939.716

Table 2.6: Actual Generation (MU) for 1 year

$$\text{Efficiency (\%)} = \frac{\text{(Actual Generation (MU))}}{\text{(Possible Generation (MU))}} \times 100$$

By calculating efficiency we have obesrved that Khaperkheda Thermal Power Station has most efficient power plant during the year 2011 to 2013 after that Nashik Thermal Power Station has second most efficient power

	Power Plant Capacity	**Efficinecy %**				
Stations	**Plant (MW) 2011-13-16**	**2011-12**	**2012-13**	**2013-14**	**2014-15**	**2015-16**
Chandrapur	2340	59.40	0.00	48.38	55.55	49.45
Khaperkheda	1340 & 840	72.25	65.42	54.98	64.01	57.61
Parli	1130	45.33	37.29	30.39	41.08	11.55
Nashik	630	68.34	68.45	66.65	67.59	63.76
Koradi	620	54.20	39.77	38.13	37.35	26.49
Paras	500	58.08	60.96	68.46	59.33	66.07
Bhusawal	420	55.89	49.22	49.65	25.54	35.97

Table 2.7: Efficiency %

plant during the year 2014 to 2015. Each year efficiency is changing due to external and internal factors affecting performance of all Thermal Power Stations.

Chapter 3

Statistical Techniques for Analysis

3.1 Factor Analysis

Factor Analysis (FA) is an exploratory data analysis method used to search influential underlying factors or latent variables from a set of observed variables. It helps in data interpretations by reducing the number of variables. It extracts maximum common variance from all variables and puts them into a common score.
Factor analysis is widely utilized in market research, advertising, psychology, finance, and operation research. Market researchers use factor analysis to identify price-sensitive customers, identify brand features that influence consumer choice, and helps in understanding channel selection criteria for the distribution channel.

Factor analysis is a linear statistical model. It is used to explain the variance among the observed variable and condense a set of the observed variable into the unobserved variable called factors. Observed variables are modeled as a linear combination of factors and error terms. Factor or latent variable is associated with multiple observed variables, who have common patterns of responses. Each factor explains a particular amount of variance in the observed variables. It helps in data interpretations by reducing the number of variables.

Factor analysis is a method for investigating whether a number of variables of interest $X_1, X_2, \ldots\ldots, X_l$, are linearly related to a smaller number of unobservable factors $F_1, F_2, \ldots\ldots\ldots, F_k$.

Assumptions :

1. There are no outliers in data.
2. Sample size should be greater than the factor.
3. There should not be perfect multicollinearity.
4. There should not be homoscedasticity between the variables.

3.1.1 Types of Factor Analysis

- **Exploratory Factor Analysis :** It is the most popular factor analysis approach among social and management researchers. Its basic assumption is that any observed variable is directly associated with any factor.
- **Confirmatory Factor Analysis (CFA) :** Its basic assumption is that each factor is associated with a particular set of observed variables. CFA confirms what is expected on the basic.

3.1.2 Factor Analysis Working

The primary objective of factor analysis is to reduce the number of observed variables and find unobservable variables. These unobserved variables help the market researcher to conclude the survey. This conversion of the observed variables to unobserved variables can be achieved in two steps:

- **Factor Extraction :** In this step, the number of factors and approach for extraction selected using variance partitioning methods such as principal components analysis and common factor analysis.
- **Factor Rotation :** In this step, rotation tries to convert factors into uncorrelated factors —the main goal of this step to improve the overall interpretability. There are lots of rotation methods that are available such as: Varimax rotation method, Quartimax rotation method, and Promax rotation method.

◇ **Terminology**

- **Factor**

  A factor is a latent variable which describes the association among the number of observed variables. The maximum number of factors are equal to a number of observed variables. Every factor explains a certain variance in observed variables. The factors with the lowest amount of variance were dropped. Factors are also known as latent variables or hidden variables or unobserved variables or Hypothetical variables.

- **Factor Loadings**

  The factor loading is a matrix which shows the relationship of each variable to the underlying factor. It shows the correlation coefficient for observed variable and factor. It shows the variance explained by the observed variables.

- **Eigenvalues**

  Eigenvalues represent variance explained each factor from the total variance. It is also known as characteristic roots.

- **Communalities**

  Commonalities are the sum of the squared loadings for each variable. It represents the common variance. It ranges from 0-1 and value close to 1 represents more variance.

- **Factor Rotation**

  Rotation is a tool for better interpretation of factor analysis. Rotation can be orthogonal or oblique. It re-distributed the commonalities with a clear pattern of loadings.

3.1.3 Steps Involved In Factor Analysis

The various steps involved in factor analysis are

- Bartlett's Test of Sphericity and KMO Test
- Determining the number of factors

- Interpreting the factors

  Make sure that you have removed the outliers, standard scaled the data and also the features have to be numeric.

- **Bartlett's Test Of Sphericity**

  Bartlett's test checks whether the correlation is present in the given data. It tests the null hypothesis (H_0) that the correlation matrix is an Identical matrix. The identical matrix consists of all the diagonal elements as 1. So, the null hypothesis assumes that no correlation is present among the variables.

  We want to reject this null hypothesis because factor analysis aims at explaining the common variance i.e. the variation due to correlation among the variables. If the p test statistic value is less than 0.05, we can decide that the correlation is not an Identical matrix i.e. correlation is present among the variables with a 95% confidence level.

- **Kaiser-Meyer-Olkin (Kmo) Test**

  KMO Test measures the proportion of variance that might be a common variance among the variables. Larger proportions are expected as it represents more correlation is present among the variables thereby giving way for the application of dimensionality reduction techniques such as Factor Analysis. KMO score is always between 0 to 1 and values more than 0.6 are much appreciated. We can also say it as a measure of how suited our data is for factor analysis.

3.1.4 Choosing the Number of Factors

Kaiser criterion is an analytical approach, which is based on the more significant proportion of variance explained by factor will be selected. The eigenvalue is a good criterion for determining the number of factors. Generally, an eigenvalue greater than 1 will be considered as selection criteria for the feature.

The graphical approach is based on the visual representation of factors' eigenvalues also called scree plot. This scree plot helps us to determine the

number of factors where the curve makes an elbow.

3.1.5 Pros and Cons of Factor Analysis

Factor analysis explores large dataset and finds interlinked associations. It reduces the observed variables into a few unobserved variables or identifies the groups of inter-related variables, which help the market researchers to compress the market situations and find the hidden relationship among consumer taste, preference, and cultural influence. Also, It helps in improve questionnaire in for future surveys. Factors make for more natural data interpretation.

Results of factor analysis are controversial. Its interpretations can be debatable because more than one interpretation can be made of the same data factors. After factor identification and naming of factors requires domain knowledge.

3.2 Canonical Correlation Analysis

◇ **What is Canonical Correlation Analysis?**

Canonical correlation analysis (CCA) is a statistical technique to derive the relationship between two sets of variables. One way to understand the CCA, is using the concept of multiple regression. In multiple regression, the relationship between one single dependent variable and a set of independent variables are investigated.

In CCA, we extend the multiple regression concept to more than one dependent variable. In some applications, we confront with more than one dependent variable which are inter-correlated, so it is not sensible to ignore dependency.

$$V_i = X_1, X_2, X_3, ..., X_p$$

$$U_i = Y_1, Y_2, Y_3, ..., Y_p$$

The aim of CCA is finding the relationship between two lumped variables in a way that the correlation between these two is maximum.

- **Canonical Variables :** There are several linear combinations of variables, but the aim is to pick only those linear functions which best express the correlations between the two variable sets. These linear functions are called the **Canonical Variables.**

- **Canonical Correlations :** The correlations between corresponding pairs of canonical variables are called **Canonical Correlations.**

The basic idea behind CCA is finding a linear combination of Y_s which has the maximum correlation with linear combination of X_s . Say

$$U_i = a_1 Y_1 + a_2 Y_2 + a_3 Y_3 + ... + a_q Y_p$$

$$V_i = b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_p X_p$$

For every choice of weights, the value pair of U and V. Then the correlation between U and V can be obtained. There are k=min (q, p) number of variants, each corresponds to different set of weights, which gives us k number of different correlation. These correlations are proved to be the square root of the eigenvalue of the product of two matrices given by equation.

$$P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$$P_{\mathbf{Y}} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$$

Where:

1. X is a matrix with n rows and p columns
2. Y is a matrix with n rows and q columns.
3. The n rows denote the number of samples observed and p or q is the number of features of X and Y, respectively.

- **Eigenvalue and Eigenvectors :** The eigenvalue and eigenvectors give us the canonical correlation and corresponding conical variables U and V. As mentioned, there are k number of different canonical variables. The first canonical variables are the most important one as the corresponding correlation is the maximum correlation between others.

3.3 Repeated Measures ANOVA

3.3.1 What is Repeated Measures ANOVA?

Repeated measures ANOVA is a statistical technique used to compare the mean scores of a dependent variable measured repeatedly over time or under different conditions. It is also known as "within-subjects ANOVA" or "ANOVA with repeated measures." In this technique, the same group of subjects is tested multiple times, or the same subjects are exposed to different treatments or conditions, and their responses are measured.

3.3.2 The Purpose of Repeated Measures ANOVA

The primary purpose of repeated measures ANOVA is to determine whether there are significant differences between the mean scores of the dependent variable across different levels of the independent variable(s) or over time. It is used to test the null hypothesis that there are no significant differences between the means of the dependent variable, and the alternative hypothesis that at least one mean is different from the others.

3.3.3 Assumptions of Repeated Measures ANOVA

1. **Normality :** The distribution of the dependent variable should be approximately normal.
2. **Sphericity :** The variances of the differences between all possible pairs of the dependent variable should be equal.
3. **Homogeneity of variances :** The variances of the dependent variable should be equal across different levels of the independent variable.

3.3.4 Conducting Repeated Measures ANOVA

- **The One-Way Repeated Measures ANOVA**

  The one-way repeated measures ANOVA is used when there is only one independent variable. The following steps are involved in conducting a one-way repeated measures ANOVA:

1. Determine the research question and hypotheses.
2. Identify the dependent variable and independent variable.
3. Determine the number of levels of the independent variable.
4. Select the appropriate statistical software.
5. Test for sphericity assumption using Mauchly's test.
6. Conduct the repeated measures ANOVA.
7. Examine the output and interpret the results.
8. Check the post-hoc tests to determine which conditions differ significantly.

- **The Two-Way Repeated Measures ANOVA**

  The two-way repeated measures ANOVA is used when there are two independent variables. The following steps are involved in conducting a two-way repeated measures ANOVA:

1. Determine the research question and hypotheses.
2. Identify the dependent variable and independent variables.
3. Determine the number of levels of the independent variable.
4. Select the appropriate statistical software.
5. Test for sphericity assumption using Mauchly's test.
6. Conduct the repeated measures ANOVA.
7. Examine the output and interpret the results.
8. Check the post-hoc tests to determine which conditions differ significantly.

3.3.5 Tukey's (HSD) Test of Post Hoc Test

In statistical analysis, when comparing the means of three or more groups, it's essential to determine which groups differ significantly from each other. The ANOVA test helps determine if there is a significant difference

among the group means, but it doesn't indicate which specific groups differ significantly.

- **Tukey's (HSD) Test:**
    1. The Tukey's HSD test is a post hoc test that addresses the issue of multiple comparisons by conducting pairwise comparisons between all groups.
    2. It calculates the difference between the means of each pair of groups and compares it to the standard error of the differences.
    3. The standard error of the differences takes into account the variability within the groups and the sample size.
- **HSD Value:**
    1. The HSD value is the critical value used for determining the significance of the differences between the group means.
    2. It is based on the studentized range distribution and depends on the number of groups and the total number of observations.
    3. The HSD value represents the minimum difference between the means that is considered statistically significant.

Chapter 4

Analysis and Conclusions

4.1 Analyzing Performance Features of TPS using "Factor Analysis"

4.1.1 Adequacy Test for Analyzing Performance Features

- **Bartlett's Test**

  Here we used Bartlett's test to check whether the correlation is present in the given performance data of thermal power stations. It tests the null hypothesis (H_0) that the correlation matrix is an Identical matrix. The identical matrix consists of all the diagonal elements as 1. So, the null hypothesis assumes that no correlation is present among the variables.

  We want to reject this null hypothesis because factor analysis aims at explaining the common variance i.e. the variation due to correlation among the variables. If the p test statistic value is less than 0.05, we can decide that the correlation is not an Identical matrix i.e. correlation is present among the variables with a 95% confidence level.

  We have calculated Chi Square Value and p value which are (28257.28733605932, < 0.00001). In this Bartlett's test, the p-value is 0. The test was statistically significant, indicating that the observed correlation matrix is not an identity matrix for performance features.

- **Kaiser-Meyer-Olkin Test**

  After performing Kaiser-Meyer-Olkin Test on all performance features we can see that data has an overall proportion of variance of 0.69. It shows that our data has more correlation and dimensionality reduction techniques such as the factor analysis can be applied.

4.1.2 Choosing Number Of Factors For Performance Features

To choose number of factors we have plotted Scree plot as follows

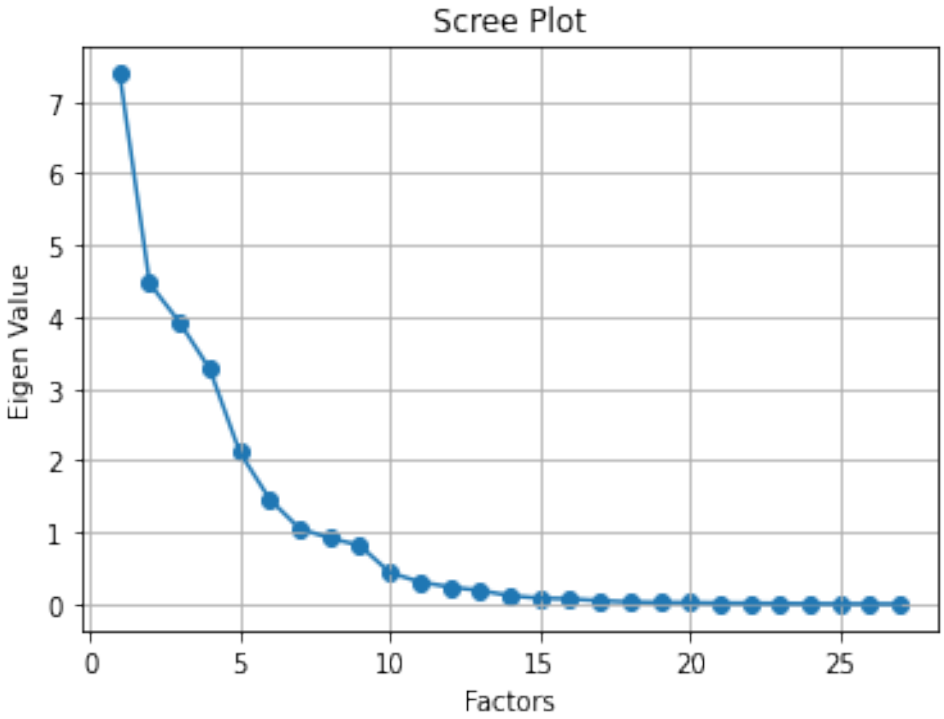

Figure 4.1: Scree plot

From the graph, we can see that the eigenvalues drop below 1 from the 1st factor. So, the optimal number of factors is 6. Create an optimal number of factors i.e. 6 in our case. Then, we have to interpret the factors by making use of loadings, variance, and commonalities.

4.1.3 Interpreting Analyzed Performance Features

- **Loadings**

To interprete factors we have calculated loadings corresponding to analyzed featues from performance data:

Var	0	1	2	3	4	5
PLFC	**0.777837**	0.159963	-0.12461	-0.15501	0.029103	0.576233
AFC	**0.78483**	0.161725	-0.06673	0.099551	0.04774	0.525251
L	**0.845531**	0.143861	0.01663	-0.29482	0.016873	0.029216
PLFM	**0.724077**	0.150878	-0.00161	-0.21009	0.018516	0.568985
AVFM	**0.710153**	0.156971	0.009049	-0.22532	0.030334	0.593563
RSD	**-0.87659**	-0.1203	0.017829	-0.17475	-0.05484	0.025094
CR	**0.309037**	0.09176	-0.01436	-0.09714	-0.17804	0.133613
GCV	**0.803161**	0.02973	0.060542	-0.00633	-0.4086	-0.17134
HR	**0.758667**	0.085987	0.153102	0.258115	0.128605	-0.37677
MUL	**-0.8076**	0.117596	-0.03302	-0.01701	0.057876	0.148003
MW	-0.22656	**0.943292**	0.037279	0.098359	0.124229	0.009057
MUG	0.194151	**0.952201**	-0.05091	-0.00869	0.137705	0.142336
CC	0.181522	**0.936523**	-0.06601	0.050189	-0.22066	0.119034
AC	0.192748	**0.948557**	-0.04143	0.117298	0.092676	0.10677
NG	0.19181	**0.953056**	-0.05219	-0.02283	0.140727	0.146955
SRWC	-0.01612	-0.03844	**0.513088**	-0.02944	-0.11716	-0.26432
SFWC	0.058677	0.06855	**0.932372**	0.012198	0.102226	0.02276
SDWC	-0.06861	0.019311	**0.808614**	0.017514	0.525836	0.084606
SLOC	-0.04457	-0.06234	**0.782685**	0.015446	0.044498	0.132004
PAC	0.135567	-0.1336	**0.927689**	0.101303	-0.09757	-0.26245
Total	-0.02793	0.172947	0.018043	**0.960634**	-0.02234	-0.08233
SOC	-0.0556	-0.16674	0.048591	**0.811053**	-0.00867	-0.30283
TOC	-0.04276	0.194418	0.021948	**0.935225**	-0.01269	-0.09265
SCC	-0.04362	0.120634	0.057685	-0.02823	**0.984921**	0.003008
TL	-0.04764	0.118742	0.056605	-0.0305	**0.986528**	0.004001
FO	0.075201	-0.1042	-0.01893	0.105375	0.023764	**-0.31136**
PO	-0.05878	-0.05937	0.117011	0.094731	-0.02665	**-0.56208**

Table 4.1: Factor Loadings

Here we have calculated factor loadings and we can observed that factor 0 has most number of features and factor 5 has less number of features. Factor 0 has 10 features, factor 1 and 2 has 5 features, factor 3 has 3 features and factor 4 and 5 has 2 features.

- **Variance of Analyzed Performance Features**

To interpret factors we can use variance as a measure of analyzed

performance features. Also we have calculated, 'Proportional Variance', 'Cumulative Variance' with 'Variance' to interpret properly.

	0	1	2	3	4	5
Variance	5.944950	4.813312	3.345224	2.838139	2.595197	2.247860
Proportional Var	0.220183	0.178271	0.123897	0.105116	0.096118	0.083254
Cumulative Var	0.220183	0.398454	0.522351	0.627468	0.723586	0.806840

Table 4.2: Variance of Analyzed Features

- **Communalities**

  Again to interpret analyzed performance features of TPS we have calculated Communalities.

	Communalities
MW	0.967709
MUG	0.986271
PLFC	1.003064
AFC	0.934644
L	0.823950
PLFM	0.915279
AVFM	0.933047
Total	0.961111
SOC	0.782841
TOC	0.923498
FO	0.125482
PO	0.346288
RSD	0.817370
CR	0.163116
CC	0.979759
SCC	0.990658
GCV	0.845972
HR	0.831525
TL	0.993756
SRWC	0.349453
SFWC	0.888576
SDWC	0.942904
SLOC	0.638112
MUL	0.692672
AC	0.972376
PAC	0.985494
NG	0.989752

Table 4.3: Communalities of Analyzed Features

– By obtaining and interpreting analyzed features from factor analysis the most important features are 'MW', 'MUG', 'PLFC', 'AFC', 'PLFM', 'AVFM', 'SOC', 'TOC', 'CC', 'SCC', 'GCV', 'HR', 'TL', 'SFWC', 'SDWC', 'SLOC', 'AC', 'PAC', 'NG'.

4.2 Interrelations Between Performance Features Using "Canonical Correlation Analysis"

The variables used to perform the canonical correlation analysis were X, consisting of the following independent variables: MUG, PLFC, AFC, L, PLFM, AVFM, and CC. The dependent variables Y included AC and NG.

4.2.1 Covariance Matrix

First Covariance Matrix :

$$\begin{bmatrix} 1 & 0.98558904 \\ 0.98558904 & 1 \end{bmatrix}$$

The diagonal elements of the matrix represent the correlation of each variable with itself, which is always 1. Therefore, the correlation between the first column of X and itself is 1, and the same holds true for the first column of Y.

The off-diagonal elements represent the correlation between the two variables. In this case, the correlation between the first column of X and the first column of Y is 0.98558904. This value is close to 1, indicating a strong positive linear relationship between these two variables.

Second Covariance Matrix :

$$\begin{bmatrix} 1 & -0.1041016 \\ -0.1041016 & 1 \end{bmatrix}$$

The diagonal elements are 1, representing the correlation of each variable with itself. The off-diagonal element is -0.1041016, indicating a weak negative linear relationship between the second column of X and the second column of Y. This value is close to 0, suggesting a weak or no linear association between these two variables. The matrices indicate that the parameters of the thermal power station are positively correlated with each other.

4.2.2 Plotting of Canonical Variables

To visualizing the relationships between the canonical Variables we have plotted scatter plot

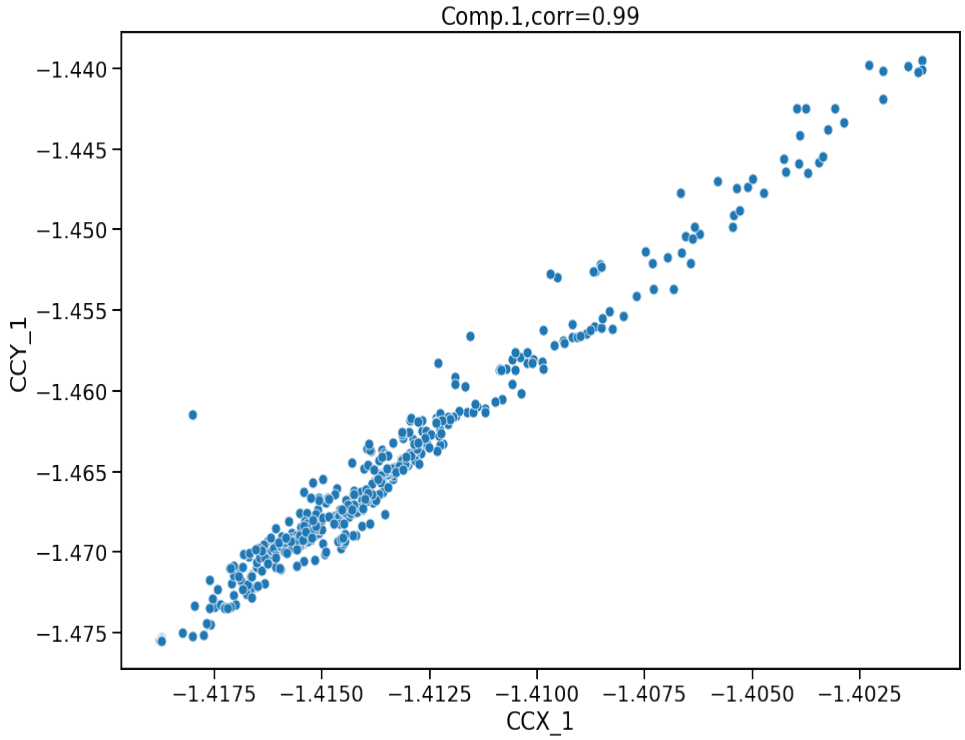

Figure 4.2: Scatter plot of Canonical Variables

From the above plot, we can observe a high correlation among the canonical variables. This suggests that the parameters of the thermal power plant are interrelated with each other.

4.3 Comparing Mean Differences of Thermal Power Stations

Using Repeated Measure ANOVA and Post Hoc Test (Tukey's HSD Test) we have comparing mean difference of thermal power stations.

4.3.1 Repeated Measure ANOVA :

The variables used for the analysis of repeated measures ANOVA were NetGeneration, PowerStation, and Month.

ANOVA				
Month	F Value	Num DF	Den DF	Pr >F
	2.772	38	228	0

Table 4.4: Repeated Measure ANOVA

- The **"F Value" is 2.772,** which represents the ratio of the between-group variability to the within-group variability. It is a measure of the significance of the differences between the groups' means.
- The **"Num DF" is 38,** indicating the number of degrees of freedom in the numerator of the F statistic.
- The **"Den DF" is 228,** representing the number of degrees of freedom in the denominator of the F statistic.
- The **"Pr > F" value is « 0.0001,** which indicates that the mean differences between the groups are statistically significant.

The output does not provide specific details about the factors or groups being compared. However, the ANOVA results are likely related to the "Month" variable, suggesting that there are significant differences in means across different months. The F value, degrees of freedom, and p-value are the same as in the mean aggregation.

Overall, the output suggests that there are significant differences in means between the groups (possibly months) being analyzed

4.3.2 Post Hoc Test (Tukey's HSD Test):

Multiple Comparison of Means - Tukey HSD, FWER=0.05						
group1	group2	meandiff	p-adj	lower	upper	reject
Bhusawal	Chandrapur	799.4124	0.001	723.4496	875.3752	True
Bhusawal	Khaperkheda	409.5634	0.001	333.6006	485.5262	True
Bhusawal	Koradi	61.8935	0.1939	-14.0693	137.8564	False
Bhusawal	Nashik	180.2642	0.001	104.3014	256.227	True
Bhusawal	Paras	98.8091	0.0026	22.8463	174.772	True
Bhusawal	Parli	186.571	0.001	110.6081	262.5338	True
Chandrapur	Khaperkheda	-389.849	0.001	-465.8118	-313.8861	True
Chandrapur	Koradi	-737.5189	0.001	-813.4817	-661.556	True
Chandrapur	Nashik	-619.1482	0.001	-695.111	-543.1854	True
Chandrapur	Paras	-700.6033	0.001	-776.5661	-624.6404	True
Chandrapur	Parli	-612.8414	0.001	-688.8043	-536.8786	True
Khaperkheda	Koradi	-347.6699	0.001	-423.6327	-271.707	True
Khaperkheda	Nashik	-229.2992	0.001	-305.262	-153.3364	True
Khaperkheda	Paras	-310.7543	0.001	-386.7171	-234.7914	True
Khaperkheda	Parli	-222.9924	0.001	-298.9553	-147.0296	True
Koradi	Nashik	118.3707	0.001	42.4078	194.3335	True
Koradi	Paras	36.9156	0.7501	-39.0472	112.8784	False
Koradi	Parli	124.6774	0.001	48.7146	200.6403	True
Nashik	Paras	-81.4551	0.0267	-157.4179	-5.4922	True
Nashik	Parli	6.3068	0.9	-69.6561	82.2696	False
Paras	Parli	87.7618	0.0122	11.799	163.7247	True

Table 4.5: Multiple Comparison of Means

1. The **group1** and **group2** columns indicate the two groups being compared.
2. **meandiff** represents the difference in means between the two groups.
3. **p-adj** is the adjusted p-value, which indicates the statistical significance of the difference between the groups. If the p-value is less than the chosen significance level (0.05 in this case), the result is considered statistically significant.
4. **lower** and **upper** represent the lower and upper bounds of the confidence interval for the difference in means.
5. The **reject** column indicates whether the null hypothesis is rejected for each pairwise comparison. If **reject is True,** it means there is a

statistically significant difference between the two groups. If **reject is False,** it means there is not enough evidence to reject the null hypothesis.

Here's a summary of the interpretations for the given output:

1. **Bhusawal and Chandrapur** have a significantly different mean electricity generation (meandiff = 799.4124).
2. **Bhusawal and Khaperkheda** have a significantly different mean electricity generation (meandiff = 409.5634).
3. There is no significant difference in mean electricity generation between **Bhusawal and Koradi** (p-value > 0.05).
4. **Bhusawal and Nashik** have a significantly different mean electricity generation (meandiff = 180.2642).
5. **Bhusawal and Paras** have a significantly different mean electricity generation (meandiff = 98.8091).
6. **Bhusawal and Parli** have a significantly different mean electricity generation (meandiff = 186.571).
7. **Chandrapur and Khaperkheda** have a significantly different mean electricity generation (meandiff = -389.849).
8. **Chandrapur and Koradi** have a significantly different mean electricity generation (meandiff = -737.5189).
9. **Chandrapur and Nashik** have a significantly different mean electricity generation (meandiff = -619.1482).
10. **Chandrapur and Paras** have a significantly different mean electricity generation (meandiff = -700.6033).
11. **Chandrapur and Parli** have a significantly different mean electricity generation (meandiff = -612.8414).
12. **Khaperkheda and Koradi** have a significantly different mean electricity generation (meandiff = -347.6699).
13. **Khaperkheda and Nashik** have a significantly different mean electricity

generation (meandiff = -229.2992).

14. **Khaperkheda and Paras** have a significantly different mean electricity generation (meandiff = -310.7543).
15. **Khaperkheda and Parli** have a significantly different mean electricity generation (meandiff = -222.9924).
16. **Koradi and Nashik** have a significantly different mean electricity generation (meandiff = 118.3707).
17. There is no significant difference in mean electricity generation between **Koradi and Paras** (p-value > 0.05).
18. **Koradi and Parli** have a significantly different mean electricity generation (meandiff = 124.6774).
19. **Nashik and Paras** have a significantly different mean electricity generation (meandiff = -81.4551).
20. There is no significant difference in mean electricity generation between **Nashik and Parli** (p-value > 0.05).
21. **Paras and Parli** have a significantly different mean electricity generation (meandiff = 87.7618).

4.3.3 Overall Conclusions

- All conclusions are based on data from the period between 2011 and 2016. It is important to note that conclusions may vary or change for more recent years.
- Factors like Forced Outages, Reserve Shut Downs and M.U.Lost affecting the performance of Thermal Power Stations.
- By obtaining and interpreting analyzed features from factor analysis, the most important features impacting the thermal power plant performance are as follows: 'MW,' 'MUG,' 'PLFC,' 'AFC,' 'PLFM,' 'AVFM,' 'SOC,' 'TOC,' 'CC,' 'SCC,' 'GCV,' 'HR,' 'TL,' 'SFWC,' 'SDWC,' 'SLOC,' 'AC,' 'PAC,' and 'NG.'
- The set of variable of Thermal Power Stations are positively correlated among the canonical variables.
- There is no significant difference in mean electricity generation between some power stations.
- By calculating efficiency we have observed that Khaperkheda Thermal Power Station has most efficient power plant during the year 2011 to 2013 after that Nashik Thermal Power Station has second most efficient power plant during the year 2014 to 2015. Each year efficiency is changing due to external and internal factors affecting performance of all Thermal Power Stations.
- We attempted to analyze performance over time to identify trends and patterns, but trend analysis is not suitable here due to its poor quality.

Suggestions

- It is needed to observe and study the parameters which impact the net electricity generation.
- Parameters which causes the problems that give rise to less generation.
- There is need of Statistician to improve electricity net generation and to deal with data properly.

References

[1] https://www.analyticsvidhya.com/blog/2020/10/dimensionality-reduction-using-factor-analysis-in-python/

[2] https://www.datacamp.com/tutorial/introduction-factor-analysis/

[3] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, 1984.

[4] S. E. Maxwell and H. D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Routledge, 2017.

Annexure

◇ Python Code for Exploratory Data Analysis of Performance Data

Data Importing and Summary

[1]:
```
# Importing  required  libraries
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

[2]:
```
# Importing dataset
df=pd.read_excel(r'C:\Users\komal ramesh sapkal\Dropbox\Project\TPPS_Perf.
 ↪xlsx')
```

[3]:
```
df.head()
```

[4]:
```
df.Station.unique()
```

[4]:
```
array(['Bhusawal', 'Chandrapur', 'Khaperkheda', 'Koradi', 'Nashik','Paras',␣
 ↪'Parli'], dtype=object)
```

[5]:
```
df.shape
```

[5]:
```
(433, 45)
```

[6]:
```
# Checking null values
df.isnull().sum()
```

[7]:
```
df.dtypes
```

[8]:
```
# Descriptive Statstics
df.describe()
```

[9]:
```
df.columns
```

[10]:
```
# Dummy variables
final_df=pd.get_dummies(final_df,drop_first=True)
```

[11]:
```
final_df.drop_duplicates()
```

[12]:
```
corr=final_df.corr()
```

Data Visualization

[13]:
```
sns.pairplot(final_df)
```

[14]:
```
corrmat=final_df.corr()
top_corr_features=corrmat.index
plt.figure(figsize=(18,18))
g=sns.heatmap(final_df[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```

[15]:
```
pos_corr = corr[corr > 0.1].dropna(how='all', axis=1).dropna(how='all',␣
 ↪axis=0)
print(pos_corr.columns.tolist())
```

[16]:
```
target_var = 'NG'
pos_corr_vars = corr[target_var][corr[target_var] > 0.3].index.tolist()
print(pos_corr_vars)
```

[17]:
```
corrmat=df1.corr()
top_corr_features=corrmat.index
plt.figure(figsize=(10,10))
g=sns.heatmap(df1[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```

◇ Python Code for Analyzing Performance Features using 'Factor Analysis'

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

[2]:
```
df=pd.read_excel(r'C:\Users\FLUTE\Dropbox\Project\TPPS_Perf.xlsx')
```

[3]:
```
df.head()
```

[4]:
```
from factor_analyzer import FactorAnalyzer
```

Adequacy Test

Bartlett's Test
Kaiser-Meyer-Olkin Test

[5]:
```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(df1)
chi_square_value, p_value
```

[5]:
```
(28257.28733605932, 0.0)
```

[6]:
```
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(df1)
```

[7]:
```
kmo_model
```

[7]:
```
0.6979285805320502
```

Choosing no. of factors

[8]:
```
# Create factor analysis object and perform factor analysis
from factor_analyzer import FactorAnalyzer
fa = FactorAnalyzer(rotation = None,impute = "drop",n_factors=df1.shape[1])
fa
```

[8]:
```
FactorAnalyzer(impute='drop', n_factors=27, rotation=None)
```

[9]:
```
fa.fit(df1)
ev,_ = fa.get_eigenvalues()
plt.scatter(range(1,df1.shape[1]+1),ev)
plt.plot(range(1,df1.shape[1]+1),ev)
plt.title('Scree Plot')
plt.xlabel('Factors')
plt.ylabel('Eigen Value')
plt.grid()
```

Loadings

[10]:
```
fa = FactorAnalyzer(n_factors=6,rotation='varimax')
fa.fit(df1)
print(pd.DataFrame(fa.loadings_,index=df1.columns))
```

Variance

[11]:
```
print(pd.DataFrame(fa.get_factor_variance(),index=['Variance','Proportional␣
 ↪Var','Cumulative Var']))
```

Communalities

[12]:
```
print(pd.DataFrame(fa.get_communalities(),index=df1.
  ↪columns,columns=['Communalities']))
```

[13]:
```
factor_scores = fa.transform(df1)
```

[14]:
```
print(factor_scores)
```

◇ Python Code for Interrelations Between Performance Features Using 'Canonical Correlation Analysis'

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Dataset

[2]:
```
df=pd.read_excel(r'C:\Users\komal ramesh sapkal\Dropbox\Project\TPPS_Perf.
  ↪xlsx')
```

[3]:
```
print(df.head().to_latex())
```

[4]:
```
data=df[['MUG','PLFC', 'AFC', 'L', 'PLFM', 'AVFM', 'CC', 'AC','NG']]
```

[6]:
```
X=data[['MUG','PLFC', 'AFC', 'L', 'PLFM', 'AVFM', 'CC', ]]
Y=data[['AC','NG']]
```

Standardize the data

[8]:
```
X_mc=(X-X.mean())/(X.std())
X_mc.head()
```

[9]:
```
Y_mc=(Y-Y.mean())/(Y.std())
Y_mc.head()
```

Fit the CCA and Transformtion

[10]:
```
from sklearn.cross_decomposition import CCA
my_cca = CCA(n_components=2)
my_cca.fit(X, Y)
X_c, Y_c = my_cca.transform(X_mc, Y_mc)
```

Canonical Variates

[12]:
```
cc_res=pd.DataFrame({"CCX_1":X_c[:,0],"CCY_1":Y_c[:,0],"CCX_2":X_c[:
 ↪,1],"CCY_2":Y_c[:,1]})
cc_res
```

[12]:

	CCX_1	CCY_1	CCX_2	CCY_2
0	-1.415159	-1.468931	-0.405678	-0.309355
1	-1.415521	-1.469414	-0.399177	-0.310484
2	-1.415353	-1.469055	-0.400666	-0.308901
3	-1.416137	-1.470483	-0.393257	-0.314727
4	-1.416915	-1.471029	-0.377586	-0.315091
..	...	...	...	...
428	-1.418720	-1.475318	-0.414565	-0.338654
429	-1.418720	-1.475424	-0.414565	-0.339862
430	-1.418720	-1.475427	-0.414565	-0.339900
431	-1.418720	-1.475459	-0.414565	-0.340259
432	-1.418720	-1.475453	-0.414565	-0.340192

[433 rows x 4 columns]

Covariance Matrix

[13]:
```
np.corrcoef(X_c[:,0],Y_c[:,0])
```

[13]:
```
array([[1.        , 0.98558904],
       [0.98558904, 1.        ]])
```

[14]:
```
np.corrcoef(X_c[:,1],Y_c[:,1])
```

[14]:
```
array([[ 1.       , -0.1041016],
       [-0.1041016,  1.       ]])
```

Plot

[17]:
```
sns.set_context("talk",font_scale=1.2)
plt.figure(figsize=(15,10))
sns.scatterplot(x="CCX_1",y="CCY_1",data=cc_res)
plt.title('Comp.1,corr=%.2f'%np.corrcoef(X_c[:,0],Y_c[:,0])[0,1])
```

[17]:
```
Text(0.5, 1.0, 'Comp.1,corr=0.99')
```

◇ Python Code for Comparing Mean Differences in Thermal Power Stations

Reapeated Measure ANOVA

[1]:
```
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import AnovaRM
```

[8]:
```
# Load data into a pandas DataFrame
df = pd.read_csv(r'C:\Users\komal ramesh sapkal\Dropbox\Project\anovarm_data.
  ↪csv')
df
```

[9]:
```
# Define the repeated measures ANOVA model
model = ols('NetGeneration ~ C(Power_Station)*C(Month)', data=df).fit()

# Compute the repeated measures ANOVA table
aov_table = AnovaRM(df, 'NetGeneration', 'Power_Station',␣
  ↪within=['Month'],aggregate_func='mean').fit()

# Print the ANOVA table
print(aov_table.summary())
```

```
                Anova
====================================
      F Value  Num DF  Den DF  Pr > F
------------------------------------
Month  2.7720 38.0000 228.0000 0.0000
====================================
```

Post Hoc Test (Tukey's HSD Test):

[2]:
```
import pandas as pd
import numpy as np
from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

[3]:
```
# Load data into a pandas DataFrame
df = pd.read_csv(r'C:\Users\komal ramesh sapkal\Dropbox\Project\anovarm_data.
  ↪csv')
df
```

[7]:
```
print(df['Power_Station'].unique())
```

```
['Bhusawal' 'Chandrapur' 'Khaperkheda' 'Koradi' 'Nashik' 'Paras' 'Parli']
```

[11]:
```
tukey = pairwise_tukeyhsd(endog=df['NetGeneration'],
                          groups=df['Power_Station'],
                          alpha=0.05)
```

[12]:
```
print(tukey)
```

Softwares and IDE Used

1. Python
2. Jupyter Notebook
3. MS Excel