### STATISTICAL ANALYSIS ON PERFORMANCE DATA OF THERMAL POWER STATIONS

A Project Report Submitted in Partial Fulfilment of the Requirements for the Degree of

M.Sc. Statistics

(With Specialization in Industrial Statistics)



Submitted by,

###### **Ms. Patil Diksha Dilip**

###### **Ms. Sapkal Komal Ramesh**

###### **Mr. Pawar Prasanna Deepak**

Under the guidance of

## **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)**

**CERTIFICATE**

This is to certify that **Ms. Patil Diksha Dilip, Ms. Sapkal Komal Ramesh and Mr. Pawar Prasanna Deepak** 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 on Performance Data of Thermal Power Stations”** under my guidance and supervision dur ing the academic year 2022-2023.

###### Prof. Manoj C. Patil

##### (Project Guide)

**ACKNOWLEDGEMENT**

###### 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 Mr. Manoj C. Patil for his valuable guidance,immense support, motivation and encouragement to which we could com- plete our project work successfully.

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

##### Ms. Patil Diksha Dilip (387429)

##### Ms. Sapkal Komal Ramesh (387445)

##### Mr. Pawar Prasanna Deepak(387441)

##### CONTENT

##### Chapters Page no.

##### 

##### Introduction 5-18

##### Introduction 5

##### Motivation 6

##### Objectives of the Project 7

##### Limitations of the Study 8

##### Working of Thermal Power Plant

##### Organization Profile

##### Data Collection

##### Data Features

##### Exploratory Data Analysis

##### Data Importing and Manipulations

##### Data Summary

##### Data Visualization

##### Statistical Techniques

##### Factor Analysis

##### Multiple Linear Regression

##### Time Series Analysis

##### Canonical Correlation Analysis

##### Repeated Measures Anova

##### Non-Parametric Tests

##### Conclusions based on Statisitcal Techniques

##### Factor Analysis

##### Multiple Linear Regression

##### Time Series Analysis

##### Canonical Correlation Analysis

##### Repeated Measures Anova

##### Non-Parametric Tests

##### References

##### Appendix

##### Chapter 1

##### Introduction

##### Introduction

Thermal power stations 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. Descriptive statistics will provide a summary of the central tendencies, dispersion, and distributional characteristics of the performance data. Regression analysis will enable us to examine the relationships between key performance indicators and independent variables, such as ambient temperature, fuel quality, and maintenance schedules. Time series analysis will allow us to explore temporal patterns and trends in performance data, identifying seasonality, long-term trends, and potential anomalies. Hypothesis testing will be employed to evaluate the significance of observed differences or relationships and to draw reliable conclusions based on statistical evidence.

The findings of this statistical analysis will contribute to a better understanding of the performance characteristics of thermal power stations and provide valuable insights for power plant operators, engineers, and policymakers. The results can inform decision-making processes related to maintenance scheduling, fuel procurement, operational optimization, and environmental impact mitigation. Ultimately, the goal is to improve the overall efficiency, reliability, and sustainability of thermal power stations, thereby ensuring a stable and sustainable energy supply for economic and social development.

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. **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 MSc 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.

##### Objectives of the Project

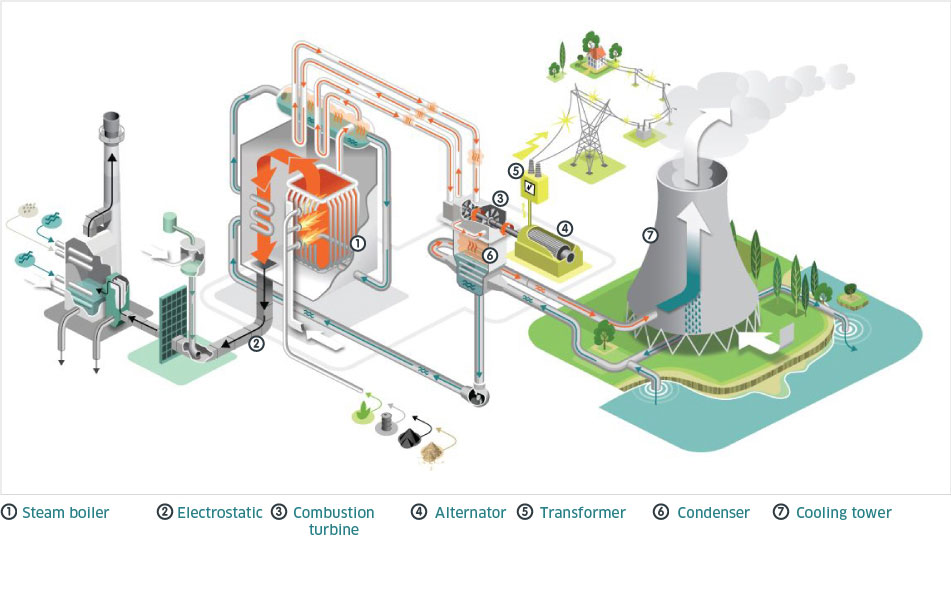
* Performance Evaluation: Assess the overall performance of thermal power stations by analyzing key performance indicators (KPIs) such as power output, heat rate, fuel consumption, emissions, and availability. Identify variations in performance across different power stations and over time.
* Identify Factors Influencing Performance: 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 characteristics. Determine the relationships between these factors and the performance indicators.
* Trend Analysis: 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: 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.
* Performance Prediction: Develop models or statistical techniques to predict the future performance of thermal power stations based on historical data and relevant factors. Forecast power generation, fuel consumption, emissions, and other performance indicators to aid in planning and decision-making processes.
* Optimization and Efficiency Improvement: Use statistical analysis to identify areas for operational optimization and efficiency improvement. Determine the most effective strategies to enhance performance, reduce fuel consumption, minimize emissions, and optimize maintenance schedules. Identify potential cost-saving measures without compromising performance.
* Environmental Impact Assessment: Evaluate the environmental impact of thermal power stations by analyzing emissions data and identifying factors that influence pollutant levels. Assess the effectiveness of emission control measures and identify opportunities for further environmental sustainability improvements.

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.

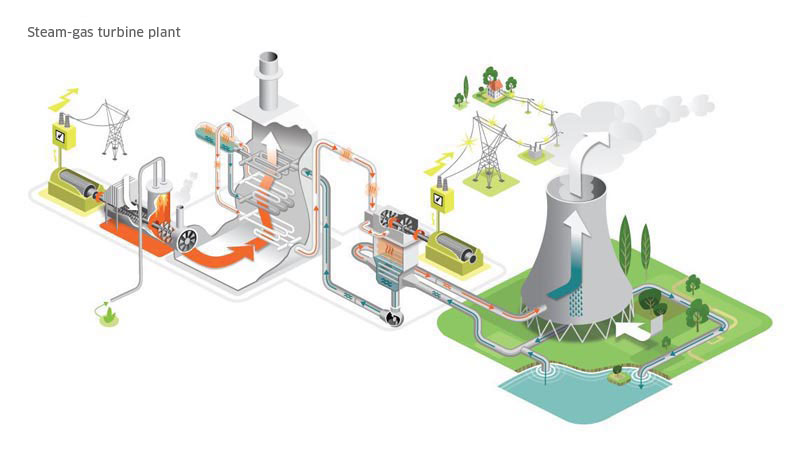
* 1. **Limitations of Study**

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. **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.

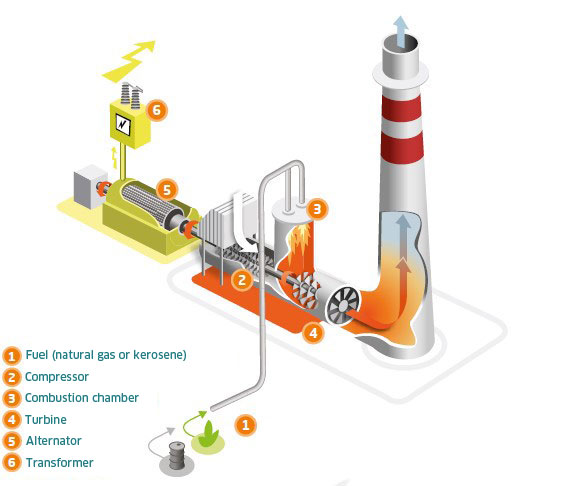


* **Combined cycle gas power plants (or steam-gas turbine plants)**: These combine a gas turbine and a traditional thermal plant to generate electricity. Unlike conventional thermal power plants, the residual energy of the gases is used for another cycle. This is one of the reasons for which these kind of plants are more efficient (by 56%), also meaning that they produce lower CO2 emissions than conventional plants.



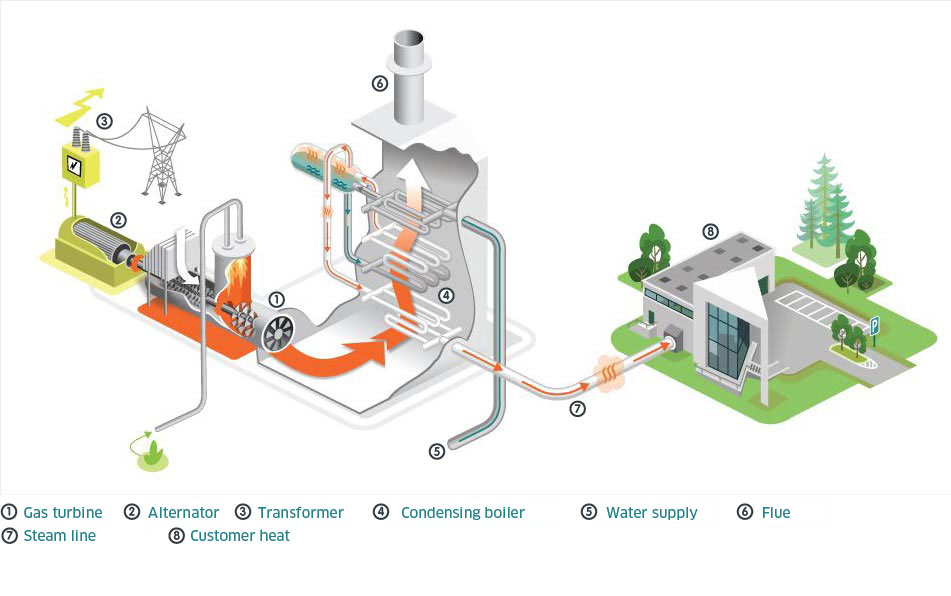
Initially, gas is injected into the combustion turbine. It generates steam, which is then supplied to another turbine. The combustion turbine and steam turbine work in tandem to turn one or more alternators, which produce electricity.

* **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.



**Gas turbine and turbojets:** The compressor draws in air, compresses it and injects it into the combustion chamber. Natural gas (gas turbine) or kerosene (turbojet) is injected into the chamber to be burned. The hot combustion gases rotate the turbine, which drives an alternator to produce electricity.

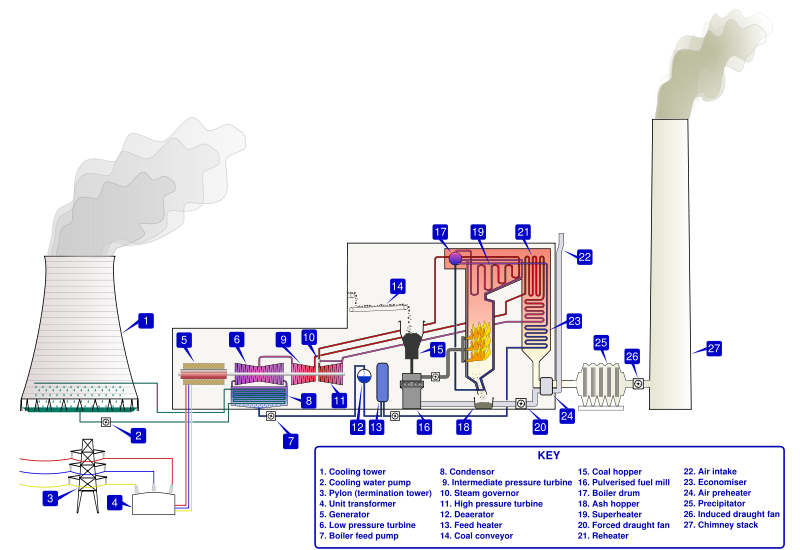
* **Cogeneration units:** these produce heat (their principal role) simultaneously with electricity (their secondary role) in a single installation and employing a single fuel. It is a highly energy-efficient solution. By recovering thermal energy normally lost in power generation, these plants are able to produce electricity and heat with efficiency of close to 90%, which is of great interest for industrial sites.



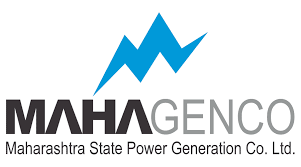
**Co-generation:** A gas-powered generator drives an alternator that produces electricity. Heat recovered from the cooling of the motor and the combustion gases heats a water circuit thanks to heat exchangers.

* **Working Components of Thermal Power Plant**

1. Thermal Power Plant
2. Fuel Storage and Water Handling Plant
3. Water Treatment Plant
4. Steam Boiler
5. **Forced Draft (FD) fan**
6. Turbine



* 1. **Organization Profile**



* [**Founded**](https://www.google.com/search?client=firefox-b-d&q=maharashtra+state+power+generation+company+founded&stick=H4sIAAAAAAAAAC3LMQ7CMAwAwAkJMTDwAs8sKbCgfiayEtNGaezIcanKc_rSIsR2yx3Pl5MrrnvEZ35311sOvcvVT4ujiYJpCslWLzogpw9aEu5fMnOk6NuISnE73At-hW00RWiGRlBlIYWBmPR3IEipyCv87w71VXyndgAAAA&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQ6BMoAHoECFkQAg)**:** 6 June 2005
* [**Headquarters**](https://www.google.com/search?client=firefox-b-d&q=maharashtra+state+power+generation+company+headquarters&stick=H4sIAAAAAAAAAE3LPQrCQBAG0EoQCwtPMLXNBCyUXCZ8bIbsstkfZyeGeBxPGkjl69_5ertw4u4xvuKnu3N0Pcc6zCvLLM40uGDbUHRCDl9YKLn3gvG9QE20_U7PBA9F86agZjChWlZRmiSLHoNcSRV5o_-5AxHlh9B5AAAA&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQ6BMoAHoECFQQAg)**:** Bandra, Mumbai, India
* [**Owner**](https://www.google.com/search?client=firefox-b-d&q=maharashtra+state+power+generation+company+owner&stick=H4sIAAAAAAAAAB3LMQrDMAxA0akQOnTICTR3kSFLyGWMcERiHFtGFjXpcXrShmx_-G94jU_M6KZ1Th_3xhQWTNUfHfngYBpDtNOLblTilyxKWaQXVt92Ul5_D5fpKmq7KUEzMoYqnRU2vrZbQJBcqZxwyz8M_ws5cgAAAA&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQ6BMoAHoECFMQAg)**:** [Government of Maharashtra](https://www.google.com/search?client=firefox-b-d&q=Government+of+Maharashtra&si=AMnBZoEofOODruSEFWFjdccePwMH96ZlZt3bOiKSR9t4pqlu2DS-KM317HPSySOM__b_dHXscp_oR6Zdvo7YVLNmnz_Mq0uZIfiSUDkt-xkAlLFV5FYi7O961V6YAfIyAt8Rft_BYdW4WeOQ_O0y2IDPD0ViJh5KtlwsleCMvAb-UzxboOVb9XQAajPO35ChQMQCUe5fAKjAQOgAItlzoG3teTRijiDE8g%3D%3D&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQmxMoAXoECFMQAw)
* [**Organization type**](https://www.google.com/search?client=firefox-b-d&q=maharashtra+state+power+generation+company+organization+type&stick=H4sIAAAAAAAAAOPgE-LSz9U3ME6xyC4z0NLLTrbSzy6IzynXT81JTS4pykzOLKmMzy9KT8zLrEosyczPsyqpLEiNL85ILEpNWcRqk5sIZCUWZ5QUJSoUlySWpCoU5JenFimkp-alFoE1KCTn5xYk5lUqIJuiADIFANtxdHZ9AAAA&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQ6BMoAHoECFAQAg)**:** [State-owned enterprise](https://www.google.com/search?client=firefox-b-d&q=State-owned+enterprise&si=AMnBZoEofOODruSEFWFjdccePwMH96ZlZt3bOiKSR9t4pqlu2Mwt0yzLvj4upKj6FR-5evrlgCUbJlBy0U5kjADqBq7FuLcPW79EbbIYaBVcRu0YEZum5z3V8y-spblAgr8Hqod7oQUe5-IUvyQW57iIqLRogCGhKdQ4xuqYS16kQwF-iHYO6LMgoY9rxARcMiSkcKPf0qvjhFE1fjH6R4b8n8hD7ArtEw%3D%3D&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQmxMoAXoECFAQAw)
* [**Area served**](https://www.google.com/search?client=firefox-b-d&q=maharashtra+state+power+generation+company+area+served&stick=H4sIAAAAAAAAACXLTQrCMBCG4ZUgLlx4glm7SaFUpJcJQ_LRhjQ_TIaWehxPatXd-y6e8_V2Mcl0vX_GtbsP0Y0mVrtsBgucSnBBd1tk4hxerKHkkQVsG2SFt20-zr9Pj8RHcZtVmJqygmrZIDQhQ36OXEmV805fT3__AUEaEth-AAAA&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQ6BMoAHoECFIQAg)**:** [Maharashtra](https://www.google.com/search?client=firefox-b-d&q=Maharashtra&si=AMnBZoFk_ppfOKgdccwTD_PVhdkg37dbl-p8zEtOPijkCaIHMmMiB1saJRbyoXBSNY-KathwrBZS2wyFcQWB9tXq7ivOamnHLVp3QPGDuXBchy3O_gjsScRaOXLYySsyM3c8AZqJ5YLWHlM9O-sHe7rz5kQ-HJkfg35b0R0JmcSZBGeULxAri7_fKE67mlRb5Zja_J2JDGcY&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQmxMoAXoECFIQAw)
* [**Purpose**](https://www.google.com/search?client=firefox-b-d&q=maharashtra+state+power+generation+company+purpose&stick=H4sIAAAAAAAAAC3LwQqDMAwA0NNg7LDDviBnL5V5Gf6MhBq01DYhzSz6OX6pIN7fe74_L5dc242_uLZNE33vogxLdbSQNw0-2DawTpjDjhY49_JX4ULH45twRsUymyIUQyMQrqQwUSa9MHhOgnmDO50I6dPNbwAAAA&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQ6BMoAHoECFgQAg)**:** Electricity generation
* [**Parent organization**](https://www.google.com/search?client=firefox-b-d&q=maharashtra+state+power+generation+company+parent+organization&stick=H4sIAAAAAAAAAOPgE-LSz9U3ME6xyC4z0LLNTrbSzy6IzynXT81JTS4pykzOLKmMzy9KT8zLrEosyczPsypILErNK0ERiy_OAAqmLGK1y00EshKLM0qKEhWKSxJLUhUK8stTixTSU_NSi8BqFZLzcwsS8yoVIOYoIJsDAEmEe62OAAAA&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQ6BMoAHoECFEQAg)**:** [Maharashtra State Electricity Board](https://www.google.com/search?client=firefox-b-d&q=Maharashtra+State+Electricity+Board&si=AMnBZoEofOODruSEFWFjdccePwMH96ZlZt3bOiKSR9t4pqlu2FLzHUO5_g89GPgF2HB7-NUsulS4QgCBGTpuNFI6hgfA0D8ZvIdWKNymQaCsYa5-SOzhy9SthcgpWFPFuFLGiruY0SK7Tzl0H51JgYnleDeaAfLfd_8WWczC4nRkTS7HTFcnDUZWPwpskoDPLVzrE9RNs6g8vQQ2XFCavx6NpcoahBRFWkn-RtVrXHgeQyx4hRCZ_8OXxENLAY3QA2deD06BpLSjGlDI60vlqPgEetJOTBJtHizc45TRILOnEhiUrR2GelqFyuZqi5gTSUs6XxpIVREro-ronzcI775NSDTHbye1bw%3D%3D&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQmxMoAXoECFEQAw), [M.S.E.B. Holding Company Limited](https://www.google.com/search?client=firefox-b-d&q=M.S.E.B.+Holding+Company+Limited&si=AMnBZoEZ8aFftZu792frFYrnK9KQYGXRL3UTeDeHB9-uc0sfFVLz-cxWm2mqQlWs0jlPoQPacSJ-gWFfQ6KpNM6pONNXQeUCxHMutHSX84n307_H0rtvaQSpH9eKLSz1ZhoNo27b_uRJL9CpkQdYGAiN6KrIGHDI_zA_IV7ELM-qpKEErCVOjJxmbY0MHFEDnnRLEfDnPvGyaBpR5VLBhtCOOdOaehB4jGs6ZTrsvi3-vqjtO6r1wDx3d7hoLyyNYqKu_anjii6BlPSLnXWy3Sp_5uQPyEkCfmMmy6Zhc8fjGEBzV1Bn_QfhOS1wxgXmEzMFsRcL2psRmkcnLtM4dN1dCGZswCGYGA%3D%3D&sa=X&ved=2ahUKEwjM1dPtyIT_AhVcplYBHYdvC4cQmxMoAnoECFEQBA)

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 |

* 1. **Data Collection**

We have collected these data from **Bhusawal Thermal Power Station.** It is located 8 km away from [Bhusawal](https://en.wikipedia.org/wiki/Bhusawal) city of [Jalgaon district](https://en.wikipedia.org/wiki/Jalgaon_district) in [Maharashtra](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Maharashtra_State_Power_Generation_Company_Limited).

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. **Data Features**
  2. **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

* 1. **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 = (Actual Energy Generated / Maximum Possible Energy Generation) x 100

* 1. **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:

Availability Factor = (Actual Output / Maximum Possible Output) x 100

* 1. **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:

Loadability Factor = (Actual Load / Maximum Capacity) x 100

* 1. **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.
  2. **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.
  3. **Oil Consumption:**

1. Total in KL: The total quantity of oil consumed by the power plant, measured in kiloliters.
2. 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).
3. Total oil cost (Rs. Crs): The total cost of oil consumed by the power plant, measured in crores of Indian Rupees (Rs. Crs).
4. Oil Cost (Paise/Kwh): The cost of oil per unit of electricity generated, measured in paise (hundredths of a Rupee) per kilowatt-hour (Kwh).
   1. **% Forced Outages (Total):** The formula for calculating % Forced Outages is as follows:

% Forced Outages = (Total Duration of Forced Outages / Total Operating Time) x 100

1. Due to tube leakages: The percentage of forced outages caused by tube leakages in the power plant's equipment.
2. Due to turbine & Aux.: The percentage of forced outages caused by issues related to the turbine and auxiliary systems.
3. Any Other reasons: The percentage of forced outages caused by reasons other than tube leakages or turbine and auxiliary issues.
   1. **% Planned Outages (Total):** The formula for calculating % Planned Outages is as follows:

% Planned Outages = (Total Duration of Planned Outages / Total Operating Time) x 100

1. Unit Overhaul: The percentage of planned outages for conducting unit overhauls, which involve maintenance, repair, or replacement of major components.
2. Other Outages: The percentage of planned outages for reasons other than unit overhauls, such as scheduled maintenance or repairs.
   1. **% Reserve Shut Down/Outages:**
3. L.D./System problem: The percentage of shutdowns or outages due to load dispatch or system-related issues.
4. Less coal receipt: The percentage of shutdowns or outages caused by inadequate coal supply to the power plant.
5. Wet coal Problem: The percentage of shutdowns or outages caused by issues related to wet coal, which may impact its combustion efficiency.
   1. **% Coal Realisation:** Percentage of coal realized or obtained compared to the required or planned coal quantity.

Formula: % Coal Realisation = (Coal Consumption / Required Coal) \* 100

* 1. **Coal Consumption (MT):** Total consumption of coal by the power plant over a given period, measured in metric tons.
  2. **Sp. Coal Consumption (Kg/Kwh):** Specific coal consumption (in kilograms) per unit of electricity generated (in kilowatt-hours).

Formula: Sp. Coal Consumption = (Coal Consumption \* 1000) / (Net Generation)

* 1. **G.C.V. (Kcal/kg):** Gross Calorific Value (in kilocalories per kilogram) of the coal being used in the power plant.
  2. **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.

Formula: Heat Rate = (Coal Consumption \* G.C.V.) / Net Generation

* 1. **Transit Loss in Coal (%):** Transit Loss in Coal represents the percentage of coal lost during transportation from the source to the power plant.
  2. **Sp. Raw Water consumption (Lt./Kwh):** Specific raw water consumption (in liters) per unit of electricity generated (in kilowatt-hours).
  3. **Sp. Softened Water consumption (Lt./Kwh): Specific** softened water consumption (in liters) per unit of electricity generated (in kilowatt-hours).
  4. **DM water consumption (Lt./Kwh**): Specific demineralized (DM) water consumption (in liters) per unit of electricity generated (in kilowatt-hours).
  5. **Sp. Lub.Oil consumption (ml/Kwh):** Specific lubricating oil consumption (in milliliters) per unit of electricity generated (in kilowatt-hours).
  6. **M.U. Lost:** M.U. Lost represents the electricity generation lost or not achieved due to various reasons.

1. Due to less coal receipt: Percentage of lost electricity generation due to insufficient coal receipt.
2. Wet Coal & 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.
   1. **Aux. consumption:** Aux. consumption represents the electricity consumption by auxiliary equipment and systems in the power plant.
   2. **% Aux. consumption:** Percentage of electricity consumed by auxiliary equipment and systems compared to the total electricity generated.

Formula: % Aux. consumption = (Aux. consumption / Net Generation) \* 100

* 1. **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: Net Generation = M.U. Generated - Aux. consumption

These parameters help 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 |

**Chapter 2**

**Exploratory Data Analysis**

In statistics, exploratory data analysis 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 encour- age 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 hy- pothesis testing, and handling missing values and making transformations of variables as needed.

In this chapter, we first import the dataset and make suitable transformations(changes) required for the further analysis. After that we Data described and visualised to make suitable conclusions.

**2.1 Data Importing and manipulation**

**2.2 Data Summary and Visualization**

|  |  |
| --- | --- |
| **Stations** | **Avg NG 2011-16** |
| Chandrapur | 740.288 |
| Khaperkheda | 523.598 |
| Nashik | 313.497 |
| Parli | 278.162 |
| Paras | 232.544 |
| Koradi | 180.572 |
| Bhusawal | 135.024 |

|  |  |
| --- | --- |
| **Stations** | **Avg NG 2011-12** |
| Chandrapur | 1017.486 |
| Khaperkheda | 444.238 |
| Parli | 374.991 |
| Nashik | 315.138 |
| Koradi | 245.962 |
| Paras | 212.585 |
| Bhusawal | 171.841 |

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 2012-13** |
| Khaperkheda | 401.148 |
| Nashik | 314.814 |
| Parli | 307.573 |
| Paras | 222.505 |
| Koradi | 179.982 |
| Bhusawal | 150.895 |
| Chandrapur | 0.000 |

|  |  |
| --- | --- |
| **Stations** | **Avg NG 2014-15** |
| Chandrapur | 948.820 |
| Khaperkheda | 626.115 |
| Parli | 338.861 |
| Nashik | 310.842 |
| Paras | 216.552 |
| Koradi | 169.060 |
| Bhusawal | 78.310 |

|  |  |
| --- | --- |
| **Stations** | **Avg NG 2013-14** |
| Chandrapur | 826.506 |
| Khaperkheda | 537.774 |
| Nashik | 306.522 |
| Parli | 250.664 |
| Paras | 249.885 |
| Koradi | 172.586 |
| Bhusawal | 152.242 |

|  |  |
| --- | --- |
| **Stations** | **Avg NG 2015-16** |
| Chandrapur | 923.929 |
| Khaperkheda | 616.454 |
| Nashik | 320.778 |
| Paras | 263.798 |
| Koradi | 131.151 |
| Bhusawal | 120.63 |
| Parli | 104.226 |

Above we have calculated Average Net Generation for each year separately and arrange it in decreasing order with respective stations.

* **Average Net Generation with Pie Charts**
* **Efficiency**

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.

**Possible Generation in 24 hrs (MU) =24 ×**

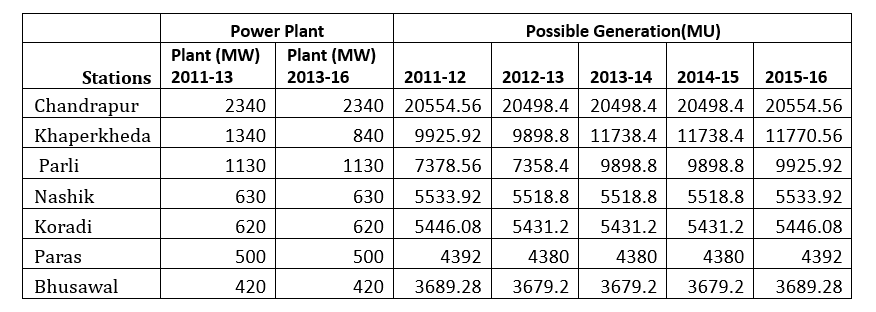
Using below table values we can calculate possible generation for 1 year (i.e. April to March). Note that 2012 and 2016 are leap years so both have 366 days.

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

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.



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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Power Plant** | | **Actual Generation(MU)** | | | | |
| **Stations** | **Plant (MW) 2011-13** | **Plant (MW) 2013-16** | **2011-12** | **2012-13** | **2013-14** | **2014-15** | **2015-16** |
| Chandrapur | 2340 | 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 | 1130 | 5330.859 | 4499.896 | 3690.874 | 3007.971 | 4066.334 |
| Nashik | 630 | 630 | 2951.544 | 3781.66 | 3777.767 | 3678.265 | 3730.103 |
| Koradi | 620 | 620 | 3781.66 | 2951.544 | 2159.782 | 2071.034 | 2028.721 |
| Paras | 500 | 500 | 2551.016 | 2551.016 | 2670.064 | 2998.624 | 2598.618 |
| Bhusawal | 420 | 420 | 4499.896 | 2062.086 | 1810.74 | 1826.904 | 939.716 |

**Efficiency (%)= × 100**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Power Plant** | | **Efficinecy** | | | | |
| **Stations** | **Plant (MW) 2011-13** | **Plant (MW) 2013-16** | **2011-12** | **2012-13** | **2013-14** | **2014-15** | **2015-16** |
| Chandrapur | 2340 | 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 | 1130 | 45.33% | 37.29% | 30.39% | 41.08% | 11.55% |
| Nashik | 630 | 630 | 68.34% | 68.45% | 66.65% | 67.59% | 63.76% |
| Koradi | 620 | 620 | 54.20% | 39.77% | 38.13% | 37.35% | 26.49% |
| Paras | 500 | 500 | 58.08% | 60.96% | 68.46% | 59.33% | 66.07% |
| Bhusawal | 420 | 420 | 55.89% | 49.22% | 49.65% | 25.54% | 35.97% |

**Chapter 3**

##### Statistical Techniques

**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.

## **3.1.1 Factor Analysis**

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 X1, X2,……., Xl, are linearly related to a smaller number of unobservable factors F1, F2,..……, Fk.

Source: This image is recreated from an image that I found in factor analysis notes. The image gives a full view of factor analysis.

**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.

## **3.1.2 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.

## **3.1.3 How does Factor Analysis Work?**

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.

## **3.1.4 Terminology**

**What is a 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.

**What are the 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.

**What is Eigenvalues?**

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

**What are 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.

**What is 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.

**3.1.5 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 (H0) 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.

## **3.1.6 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.

## **3.1.7 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 Multiple Linear Regression

Multiple linear regression is a statistical technique used to analyze the relationship between multiple independent variables and a dependent variable. It is an extension of simple linear regression, where only one independent variable is considered.

In multiple linear regression, the goal is to determine how the independent variables, denoted as X₁, X₂, X₃, ..., Xₚ, collectively affect the dependent variable, denoted as Y. The equation for multiple linear regression can be expressed as:

**Y = β₀ + β₁X₁ + β₂X₂ + β₃X₃ + ... + βₚXₚ + ɛ**

where:

* Y represents the dependent variable or the variable being predicted.
* β₀, β₁, β₂, ..., βₚ are the regression coefficients or the parameters that represent the relationship between each independent variable and the dependent variable.
* X₁, X₂, X₃, ..., Xₚ are the independent variables or predictors.
* ɛ represents the error term, which accounts for the random variation or unexplained factors in the relationship between the independent variables and the dependent variable.

The regression coefficients (β₀, β₁, β₂, ..., βₚ) are estimated using a method called Ordinary Least Squares (OLS), which minimizes the sum of squared residuals. The residuals represent the differences between the observed values of the dependent variable and the predicted values based on the regression equation.

To assess the overall fit of the multiple linear regression model, several statistical measures can be used, such as the coefficient of determination (R²), adjusted R², and the F-statistic. R² represents the proportion of the total variation in the dependent variable that can be explained by the independent variables. The adjusted R² adjusts for the number of predictors in the model. The F-statistic tests the overall significance of the regression model.

* **Assumptions Of Multiple Linear Regression Include:**

1. **Linearity:** The relationship between the independent variables and the dependent variable is linear.
2. **Independence:** The observations are independent of each other.
3. **Homoscedasticity:** The variances of the errors are constant across all levels of the independent variables.
4. **Multicollinearity:** The independent variables are not highly correlated with each other.
5. **Normality:** The errors are normally distributed with a mean of zero.

It is important to assess these assumptions when applying multiple linear regression and to interpret the results appropriately. Additionally, considerations such as variable selection, model diagnostics, and potential interactions between variables should be taken into account to ensure the validity and reliability of the regression model.

##### 3.3 Time Series Analysis-ARIMA Model

The ARIMA (Autoregressive Integrated Moving Average) model is a widely used time series forecasting technique that combines autoregressive (AR), differencing (I), and moving average (MA) components. It is effective in capturing the temporal dependencies and patterns in a time series data.

The ARIMA model is denoted as ARIMA(p, d, q),

where:

* p represents the order of the autoregressive component (AR), which captures the relationship between the current observation and the previous observations.
* d represents the degree of differencing (I), which is the number of times the data needs to be differenced to achieve stationarity. Stationarity implies that the mean and variance of the series remain constant over time.
* q represents the order of the moving average component (MA), which considers the weighted average of past forecast errors to predict future values.

The general equation for an ARIMA(p, d, q) model is as follows:

Xₜ = c + φ₁Xₜ₋₁ + φ₂Xₜ₋₂ + ... + φₚXₜ₋ₚ + θ₁εₜ₋₁ + θ₂εₜ₋₂ + ... + θ\_qεₜ₋\_q + εₜ

where:

* Yₜ represents the value of the time series at time t.
* c is a constant term.
* φ₁, φ₂, ..., φₚ are the autoregressive coefficients, representing the influence of past values on the current value.
* θ₁, θ₂, ..., θ\_q are the moving average coefficients, representing the influence of past forecast errors on the current value.
* εₜ represents the white noise error term at time t.

**The ARIMA model involves several steps:**

1. **Data Preparation:** Ensure the time series data is stationary or make it stationary through differencing.
2. **Identification:** Determine the order of differencing (d) by checking for trends and seasonality in the data. Identify the order of autoregressive (p) and moving average (q) components using autocorrelation and partial autocorrelation plots.
3. **Parameter Estimation:** Estimate the coefficients (φ and θ) of the ARIMA model using methods such as maximum likelihood estimation.
4. **Model Fitting:** Fit the ARIMA model to the training data using the estimated parameters.
5. **Model Evaluation:** Assess the goodness of fit and performance of the model using metrics such as mean squared error (MSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and visual inspection of residuals.
6. **Forecasting:** Use the fitted ARIMA model to make future predictions.

It's worth noting that while ARIMA models are effective for capturing the temporal dynamics in a time series, they may not account for external factors or exogenous variables. If additional variables are influential in the time series, alternative models such as ARIMAX (ARIMA with exogenous variables) or SARIMA (seasonal ARIMA) can be considered.

##### 3.4 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.

##### The aim of CCA is finding the relationship between two lumped variables in a way that the correlation between these twos 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: T****he correlations between corresponding pairs of canonical variables are called ****canonical correlations.****

##### The basic idea behind CCA is finding a linear combination of Ys which has the maximum correlation with linear combination of Xs . Say

##### 

##### 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.

##### 

##### Where

##### X is a matrix with n rows and p columns

##### Y is a matrix with n rows and q columns.

##### 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.

##### 3.5 Repeated Measures Anova

##### 3.5.1 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.

## **3.5.2 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.

## **3.5.3 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.

## **3.5.4 Conducting Repeated Measures ANOVA**

### 1. 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.

### 2. 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 each 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.

## **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:

### The Tukey's HSD test is a post hoc test that addresses the issue of multiple comparisons by conducting pairwise comparisons between all groups.

### It calculates the difference between the means of each pair of groups and compares it to the standard error of the differences.

### The standard error of the differences takes into account the variability within the groups and the sample size.

## **HSD Value:**

## The HSD value is the critical value used for determining the significance of the differences between the group means.

## It is based on the studentized range distribution and depends on the number of groups and the total number of observations.

## The HSD value represents the minimum difference between the means that is considered statistically significant.

**Chapter 4**