Data Analytics Internship

A Project Report submitted to the

GLOBAL NEXT CONSULTING INDIA PVT LTD

(Six – Week Internship Program)

By

Yatham Tejeswar Reddy

Under the Supervision of

Dr. Anuradha Gupta
(Project Director)

Submitted To:

Global Next Consulting India Pvt. Ltd.

Duration of Internship:

23-March-2026 to 6 -May -2026

23-March-2026

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report, **"Data Analyst Internship (GNCIPL)"**, submitted as per the requirements for the Data Analyst/ Business Analyst/ Data Science role, This is the result of original work carried out by me under the guidance of **Ms. Anuradha Gupta** during the time period from March – May (2026).

I further declare that this report represents authentic record of my own work and does not contain any falsely fabricated ideas, data, facts or sources. I also declare that I have adhered to all principles of academic honesty and integrity and that this report has not been submitted, either in part or in full, to any other institute, university, or organization for the award of any degree, diploma, or certification.

Yatham Tejeswar Reddy

CERTIFICATE

This is to certify that the project report entitled **"Data Analyst Internship Report"** has been carried out by **Yatham Tejeswar Reddy,** a Fresher in job search and improve skill in Data analyst & Data Science role with the Past Experience in Data Analyst around one and half year. This work was carried out under the guidance of **Ms. Anuradha Gupta** from March 2025 to May 2025. It is further certified that this work has not been submitted to any other university or institution for the award of any other degree, diploma or certificate.

Ms. Anuradha Gupta
Program Director
GNCIPL

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude to all those who contributed to the successful completion of this project report.

I would like to express my sincere gratitude to my supervisor, Ms. Anuradha Gupta, for their invaluable guidance, encouragement, and constructive suggestions throughout the course of this work. Their expertise and constant support played a crucial role in the successful completion of this report.

I am also thankful to the staff of Global Next Consulting India Private Limited for providing the necessary resources, facilities and their assistance, without which this study would not have been possible.

Finally, I would also like to acknowledge my peers and teachers whose support and discussions have been helpful in the completion of this report.

ABSTRACT

This report summarizes my six-week internship as a Data Analyst Intern at Global Next Consulting India Pvt. Ltd., Noida. The internship was structured into six Projects In these five minor projects as per each tool and one major project, aimed at developing practical skills in data handling, statistical analysis, and visualization.

The internship projects as a whole strengthened my technical skills in Python, Machine Learning (ML), Excel, and data visualization (Tableau & Power BI), while also improving my Presentation, analytical thinking and problem-solving approach. The work highlights how data analytics can uncover meaningful insights to support informed decision-making in domains such as public health, environment, and business.

INDEX

Chapter 1- Introduction

1.1 Company's Profile

Global Next Consulting India Private Limited (GNCIPL), headquartered in Greater Noida, Uttar Pradesh, is a cybersecurity-focused consulting firm dedicated to helping organizations protect their digital assets, data, and reputation. As threats evolve in today's digital world, GNCIPL offers proactive, customized solutions rather than reactive fixes. The company serves clients in diverse sectors including finance, healthcare, manufacturing, and technology, providing services like threat detection, risk assessment, incident response, compliance consulting, and 24/7 monitoring. GNCIPL's core values are integrity, innovation, customer-centricity, excellence, and collaboration - ensuring that technical solutions align with clients' specific needs and long-term goals.

Contact Details

Location- B5,402 P4 PHi2, CGEWHO TOWER, GREATER NOIDA 201310
Contact Numbers- 0120-4001768, +91-9315504902. +91-7666141260
Mail- hr@gncipl.com

1.2 Objectives of Internship

During my six-week internship at GNCIPL as a Data Analyst Intern, the main objectives were:

- To gain hands-on experience in data analytics tools and techniques, especially using Python (Google Colab, Jupyter Notebook), R, ETL Process and Microsoft Excel.

- To work on real-world datasets and deliver meaningful insights, visualizations, and dashboard reports.

- To learn data preprocessing, cleaning, transformation, and applying formulas and classification logic.

- To enhance analytical thinking, effective communication, and presentation skills through weekly minor projects and a major end project.

Chapter 2 – Projects

2.1 Customer Churn Analysis (Week 1)

2.1.1 Introduction

Data analysis plays a crucial role in understanding patterns, trends, and insights from raw data. In today's data-driven world, organizations rely on data analytics to make informed decisions and improve operational efficiency.

This project focuses on performing **Exploratory Data Analysis (EDA)** using Python to analyse structured datasets. The primary goal is to clean, process, and visualize the data in order to extract meaningful insights.

The project utilizes powerful Python libraries such as **pandas, numpy, matplotlib, and seaborn** to handle data preprocessing and visualization tasks. The dataset includes multiple attributes such as time-based data, numerical values, and categorical variables, which are analysed to identify trends and patterns.

2.1.2 Objectives

➤ **Primary Objectives:**

•To load and preprocess the dataset for analysis.
•To clean and handle missing or inconsistent data.
•To perform exploratory data analysis to understand data patterns.
•To create visualizations that clearly represent trends and relationships.
• To derive actionable insights from the dataset.

➤ Specific Analytical Goals:

•Analyse trends over time using line charts.
•Compare different categories using bar plots.
•Identify correlations between variables.
•Detect outliers and unusual patterns.
• Visualize distributions using histograms and box plots.

2.1.3 Methodology:

a) Dataset Preparation

¬ Loaded dataset using pandas (CSV format).

¬ Checked dataset structure using:

- .head()
- .info()
- .describe()

¬ Handled missing values using:

- Dropping null values
- Filling values using mean/median

¬ Converted data types where necessary (date, numeric, categorical)

b) Data Cleaning and Transformation

¬Removed duplicate records.

¬ Renamed columns for better readability.

¬ Converted date columns into proper datetime format.

¬ Created new derived columns where required.

---

c) Analysis Techniques

Performed statistical analysis using:

- Mean, median, standard deviation

Grouped data using:

- groupby()

Aggregated values for comparison:

- Sum, average, count

d) Data Visualization

Visualizations were created using **matplotlib** and **seaborn**:

• Line charts → Trend analysis
• Bar charts → Category comparison
• Heatmaps → Correlation analysis
• Histograms → Distribution analysis
• Box plots → Outlier detection

2.1.4 Results and Insights

a) Trend Analysis

• Observed patterns in data over time.
• Identified increasing/decreasing trends.

---

b) Category-wise Insights

• Certain categories showed higher values compared to others.
• Variation across groups indicates different behaviour patterns.

---

c) Correlation Insights

• Strong relationships identified between some variables.
• Helps in understanding dependent and independent factors.

---

d) Distribution Analysis

• Data distribution shows presence of skewness in some variables.
• Outliers were detected in specific columns.

2.1.5 Recommendations

¬ Improve data collection quality to reduce missing values.
¬ Focus on key variables that show strong impact on results.
¬ Use advanced techniques like:

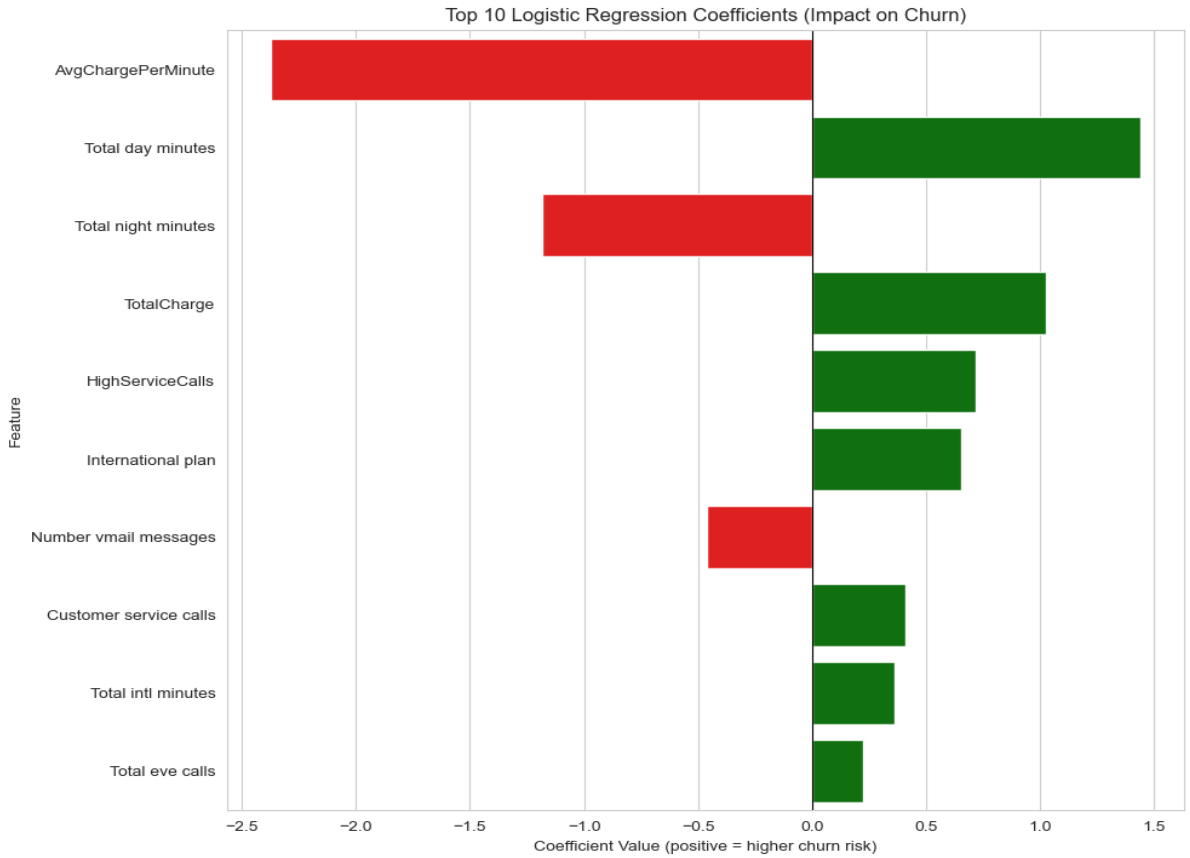
- Machine Learning models for prediction
- Forecasting future trends

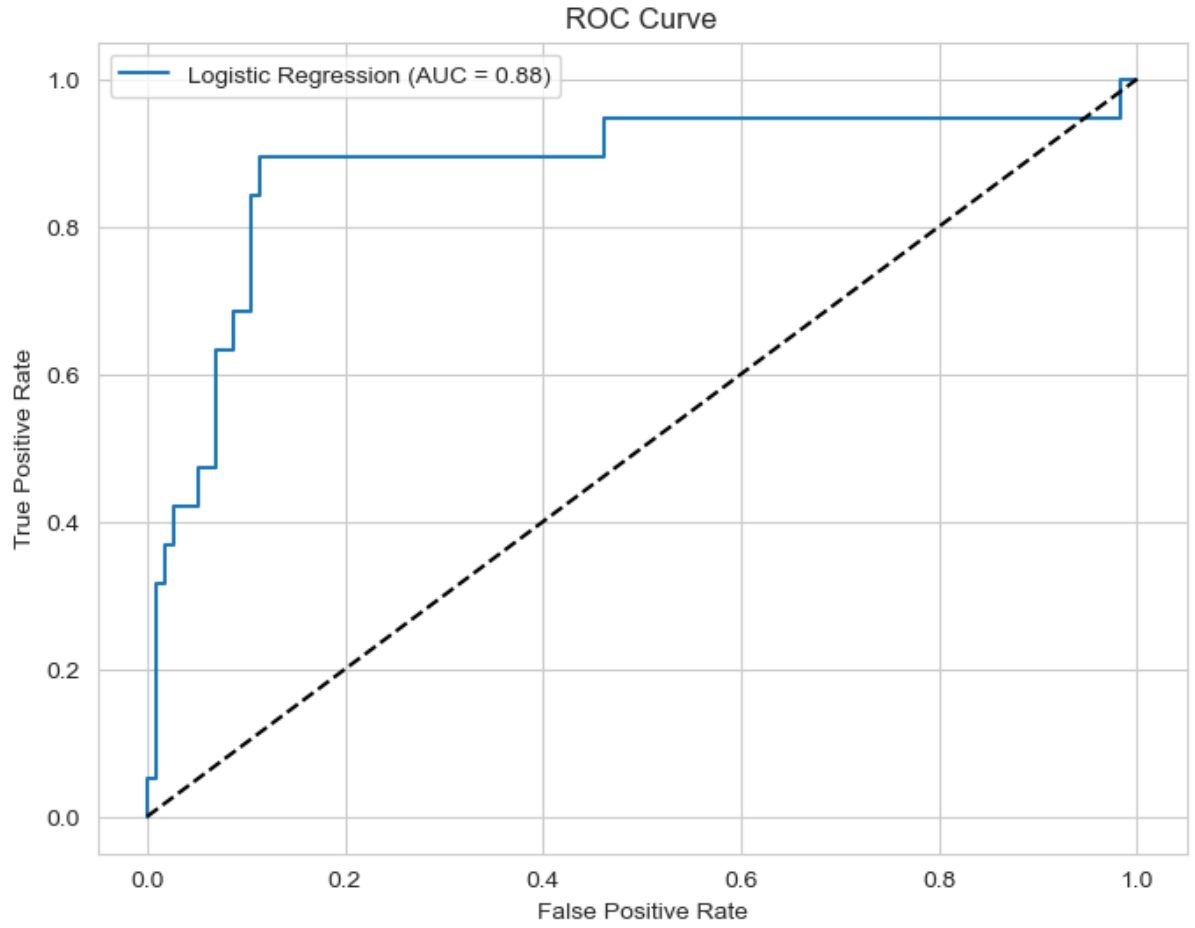
¬ Build dashboards using tools like Power BI or Tableau for better visualization.

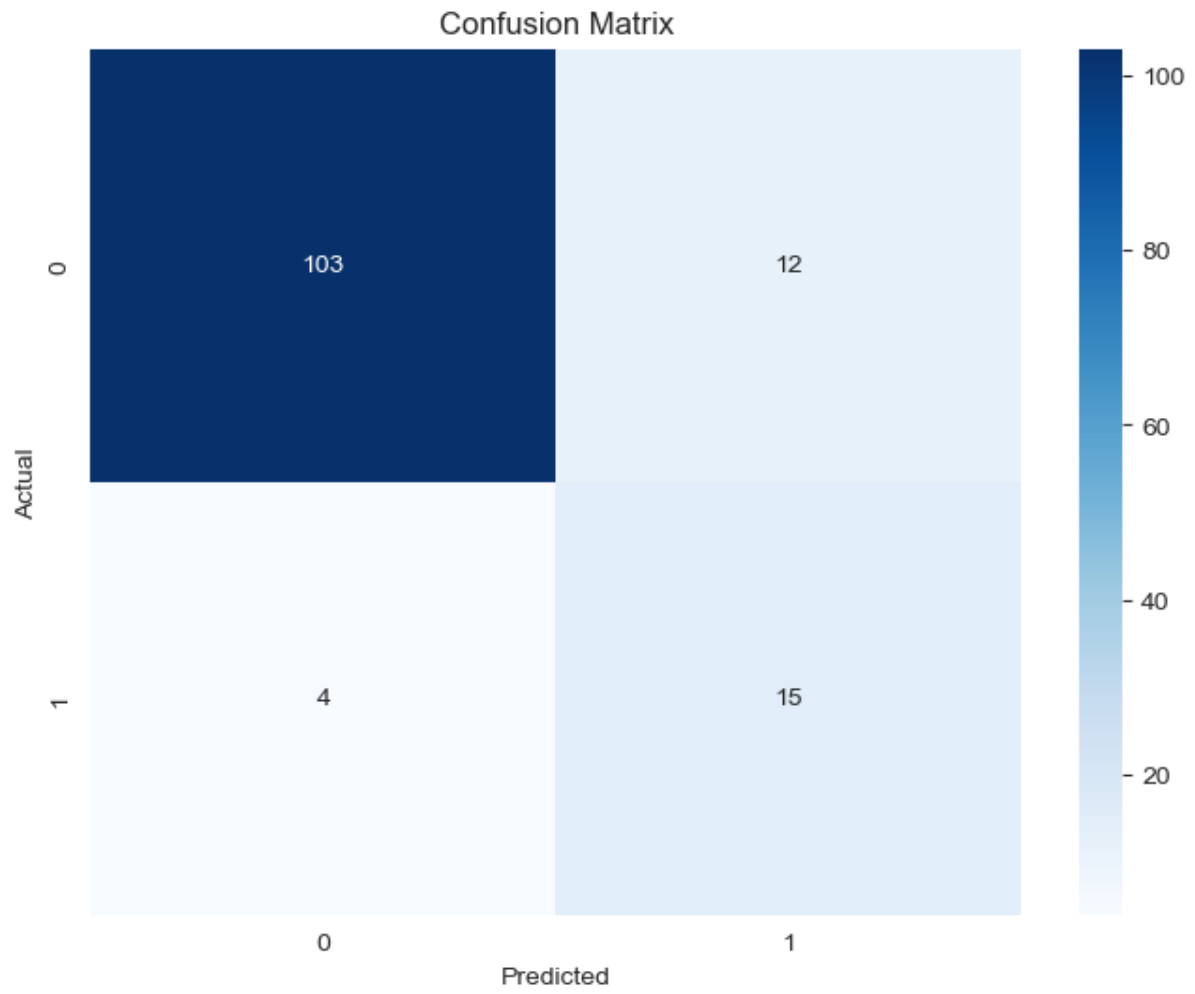
2.1.6 Conclusion

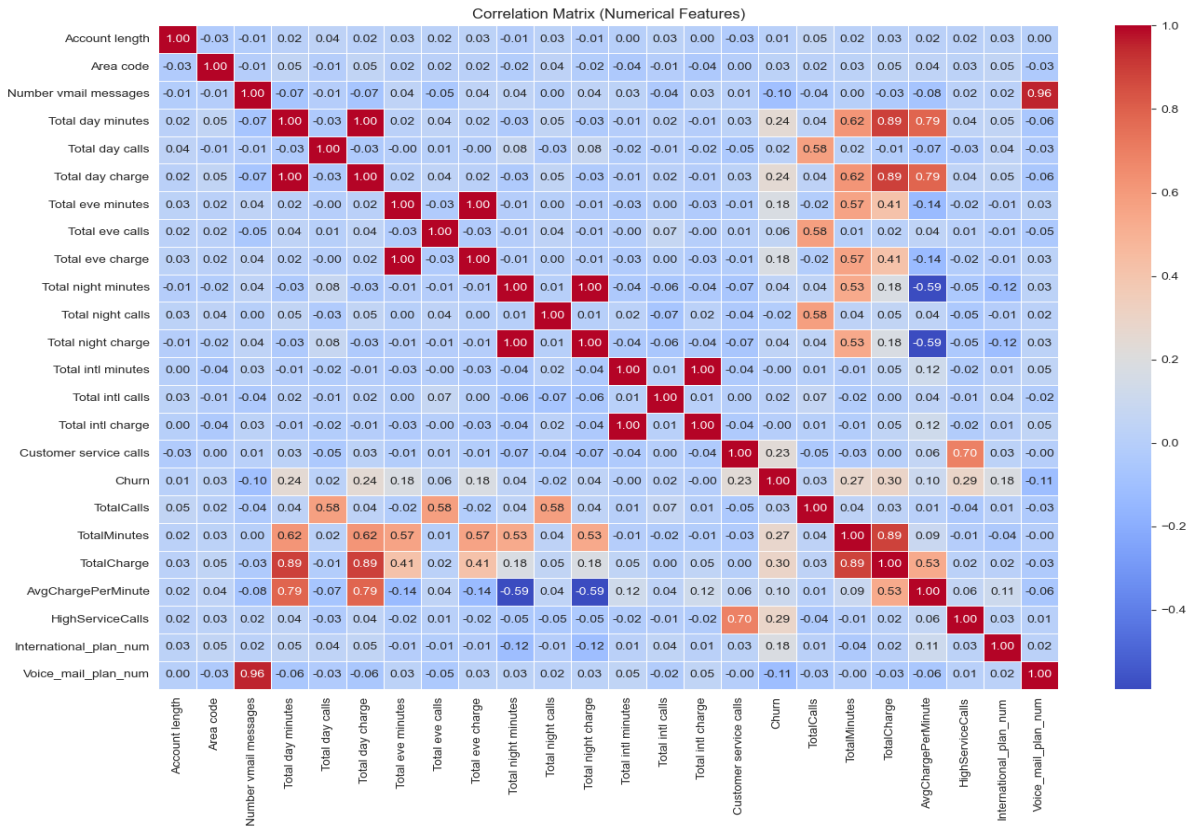
The project successfully demonstrates the use of **Exploratory Data Analysis (EDA)** techniques to understand and visualize data.

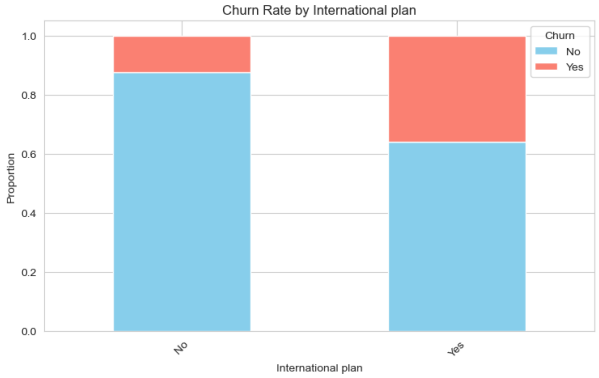
It highlights how raw data can be transformed into meaningful insights using Python tools. The findings from this project can support better decision-making and provide a strong foundation for advanced analytics such as machine learning and predictive modeling.

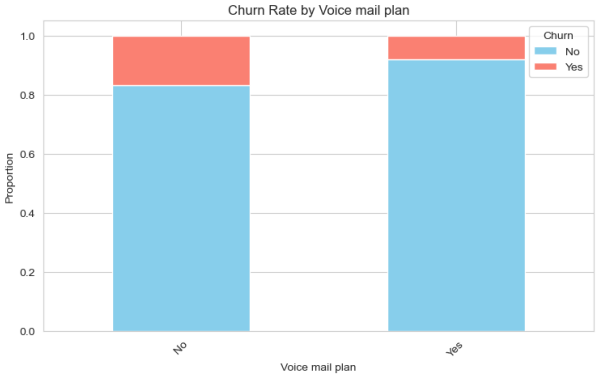

Top 10 Logistic Regression Coefficients (Impact on Churn)
AvgChargePerMinute
Total day minutes
Total night minutes
TotalCharge
HighServiceCalls
International plan
Number vmail messages
Customer service calls
Total intl minutes
Total eve calls
Feature
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Coefficient Value (positive = higher churn risk)

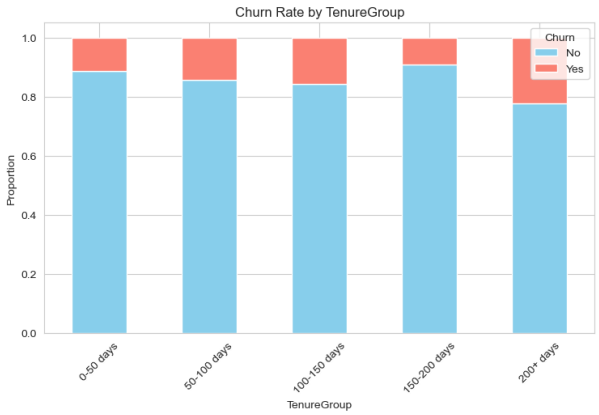

ROC Curve
Logistic Regression (AUC = 0.88)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

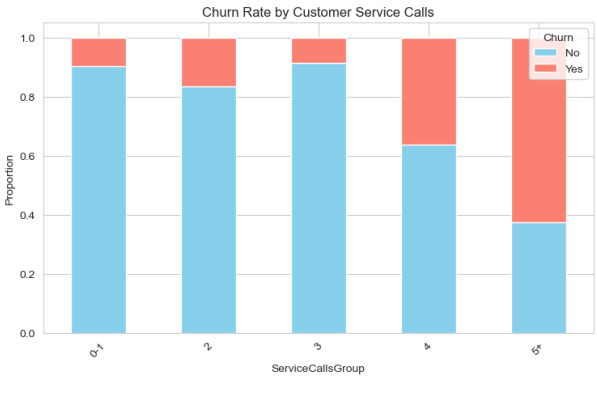

Confusion Matrix
103
12
4
15
0
1
Actual
0
1
Predicted
100
80
60
40
20

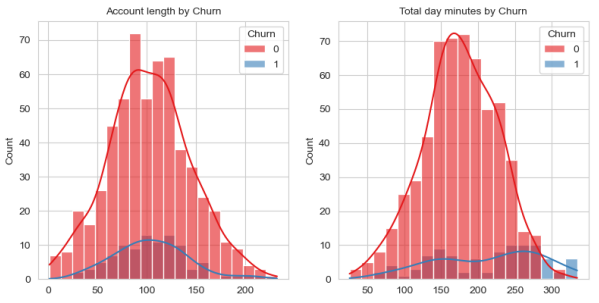

Correlation Matrix (Numerical Features)
Account length 1.00 -0.03 -0.01 0.02 0.04 0.02 0.03 0.02 0.03 -0.01 0.03 -0.01 0.00 0.03 0.00 -0.03 0.01 0.05 0.02 0.03 0.02 0.02 0.03 0.00
Area code -0.03 1.00 -0.01 0.05 -0.01 0.05 0.02 0.02 0.02 -0.02 0.04 -0.02 -0.04 -0.01 -0.04 0.00 0.03 0.02 0.03 0.05 0.04 0.03 0.05 -0.03
Number vmail messages -0.01 -0.01 1.00 -0.07 -0.01 -0.07 0.04 -0.05 0.04 0.04 0.00 0.04 0.03 -0.04 0.03 0.01 -0.10 -0.04 0.00 -0.03 -0.08 0.02 0.02 0.96
Total day minutes 0.02 0.05 -0.07 1.00 -0.03 1.00 0.02 0.04 0.02 -0.03 0.05 -0.03 -0.01 0.02 -0.01 0.03 0.24 0.04 0.62 0.89 0.79 0.04 0.05 -0.06
Total day calls 0.04 -0.01 -0.01 -0.03 1.00 -0.03 -0.00 0.01 -0.00 0.08 -0.03 0.08 -0.02 -0.01 -0.02 -0.05 0.02 0.58 0.02 -0.01 -0.07 -0.03 0.04 -0.03
Total day charge 0.02 0.05 -0.07 1.00 -0.03 1.00 0.02 0.04 0.02 -0.03 0.05 -0.03 -0.01 0.02 -0.01 0.03 0.24 0.04 0.62 0.89 0.79 0.04 0.05 -0.06
Total eve minutes 0.03 0.02 0.04 0.02 -0.00 0.02 1.00 -0.03 1.00 -0.01 0.00 -0.01 -0.03 0.00 -0.03 -0.01 0.18 -0.02 0.57 0.41 -0.14 -0.02 -0.01 0.03
Total eve calls 0.02 0.02 -0.05 0.04 0.01 0.04 -0.03 1.00 -0.03 -0.01 0.04 -0.01 -0.00 0.07 -0.00 0.01 0.06 0.58 0.01 0.02 0.04 0.01 -0.01 -0.05
Total eve charge 0.03 0.02 0.04 0.02 -0.00 0.02 1.00 -0.03 1.00 -0.01 0.00 -0.01 -0.03 0.00 -0.03 -0.01 0.18 -0.02 0.57 0.41 -0.14 -0.02 -0.01 0.03
Total night minutes -0.01 -0.02 0.04 -0.03 0.08 -0.03 -0.01 -0.01 -0.01 1.00 0.01 1.00 -0.04 -0.06 -0.04 -0.07 0.04 0.04 0.53 0.18 -0.59 -0.05 -0.12 0.03
Total night calls 0.03 0.04 0.00 0.05 -0.03 0.05 0.00 0.04 0.00 0.01 1.00 0.01 0.02 -0.07 0.02 -0.04 -0.02 0.58 0.04 0.05 0.04 -0.05 -0.01 0.02
Total night charge -0.01 -0.02 0.04 -0.03 0.08 -0.03 -0.01 -0.01 -0.01 1.00 0.01 1.00 -0.04 -0.06 -0.04 -0.07 0.04 0.04 0.53 0.18 -0.59 -0.05 -0.12 0.03
Total intl minutes 0.00 -0.04 0.03 -0.01 -0.02 -0.01 -0.03 -0.00 -0.03 -0.04 0.02 -0.04 1.00 0.01 1.00 -0.04 -0.00 0.01 -0.01 0.05 0.12 -0.02 0.01 0.05
Total intl calls 0.03 -0.01 -0.04 0.02 -0.01 0.02 0.00 0.07 0.00 -0.06 -0.07 -0.06 0.01 1.00 0.01 0.00 0.02 0.07 -0.02 0.00 0.04 -0.01 0.04 -0.02
Total intl charge 0.00 -0.04 0.03 -0.01 -0.02 -0.01 -0.03 -0.00 -0.03 -0.04 0.02 -0.04 1.00 0.01 1.00 -0.04 -0.00 0.01 -0.01 0.05 0.12 -0.02 0.01 0.05
Customer service calls -0.03 0.00 0.01 0.03 -0.05 0.03 -0.01 0.01 -0.01 -0.07 -0.04 -0.07 -0.04 0.00 -0.04 1.00 0.23 -0.05 -0.03 0.00 0.06 0.70 0.03 -0.00
Churn 0.01 0.03 -0.10 0.24 0.02 0.24 0.18 0.06 0.18 0.04 -0.02 0.04 -0.00 0.02 -0.00 0.23 1.00 0.03 0.27 0.30 0.10 0.29 0.18 -0.11
TotalCalls 0.05 0.02 -0.04 0.04 0.58 0.04 -0.02 0.58 -0.02 0.04 0.58 0.04 0.01 0.07 0.01 -0.05 0.03 1.00 0.04 0.03 0.01 -0.04 0.01 -0.03
TotalMinutes 0.02 0.03 0.00 0.62 0.02 0.62 0.57 0.01 0.57 0.53 0.04 0.53 -0.01 -0.02 -0.01 -0.03 0.27 0.04 1.00 0.89 0.09 -0.01 -0.04 -0.00
TotalCharge 0.03 0.05 -0.03 0.89 -0.01 0.89 0.41 0.02 0.41 0.18 0.05 0.18 0.05 0.00 0.05 0.00 0.30 0.03 0.89 1.00 0.53 0.02 0.02 -0.03
AvgChargePerMinute 0.02 0.04 -0.08 0.79 -0.07 0.79 -0.14 0.04 -0.14 -0.59 0.04 -0.59 0.12 0.04 0.12 0.06 0.10 0.01 0.09 0.53 1.00 0.06 0.11 -0.06
HighServiceCalls 0.02 0.03 0.02 0.04 -0.03 0.04 -0.02 0.01 -0.02 -0.05 -0.05 -0.05 -0.02 -0.01 -0.02 0.70 0.29 -0.04 -0.01 0.02 0.06 1.00 0.03 0.01
International_plan_num 0.03 0.05 0.02 0.05 0.04 0.05 -0.01 -0.01 -0.01 -0.12 -0.01 -0.12 0.01 0.04 0.01 0.03 0.18 0.01 -0.04 0.02 0.11 0.03 1.00 0.02
Voice_mail_plan_num 0.00 -0.03 0.96 -0.06 -0.03 -0.06 0.03 -0.05 0.03 0.03 0.02 0.03 0.05 -0.02 0.05 -0.00 -0.11 -0.03 -0.00 -0.03 -0.06 0.01 0.02 1.00
Account length
Area code
Number vmail messages
Total day minutes
Total day calls
Total day charge
Total eve minutes
Total eve calls
Total eve charge
Total night minutes
Total night calls
Total night charge
Total intl minutes
Total intl calls
Total intl charge
Customer service calls
Churn
TotalCalls
TotalMinutes
TotalCharge
AvgChargePerMinute
HighServiceCalls
International_plan_num
Voice_mail_plan_num
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Churn Rate by International plan
Churn
No
Yes
Proportion
International plan
No
Yes

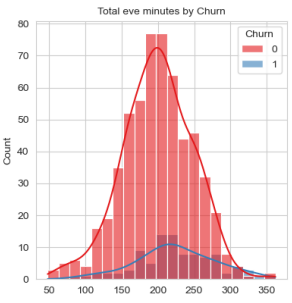

Churn Rate by Voice mail plan
Churn
No
Yes
Proportion
Voice mail plan
No
Yes

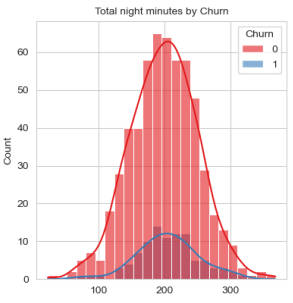

Churn Rate by TenureGroup
Churn
No
Yes
Proportion
TenureGroup
0-50 days
50-100 days
100-150 days
150-200 days
200+ days

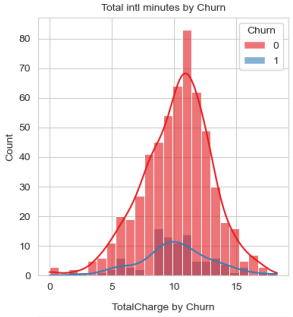

Churn Rate by Customer Service Calls
Churn
No
Yes
Proportion
ServiceCallsGroup
0-1
2
3
4
5+

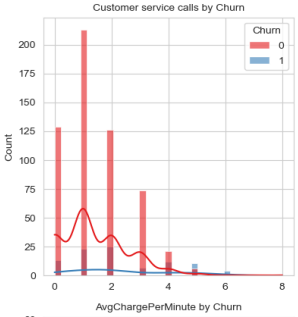

Account length by Churn
Churn
0
1
Count

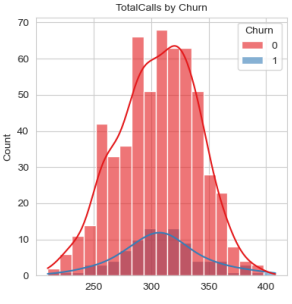
Total day minutes by Churn
Churn
0
1
Count

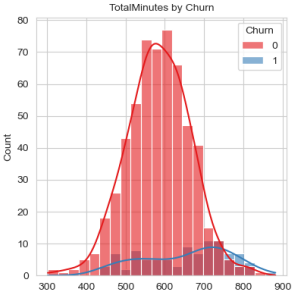

Total eve minutes by Churn
Churn
0
1
Count

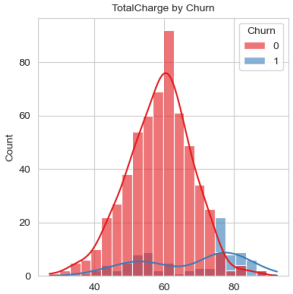

Total night minutes by Churn
Churn
0
1
Count

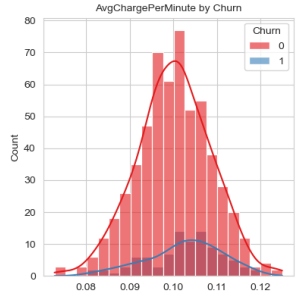

Total intl minutes by Churn
Churn
0
1
Count

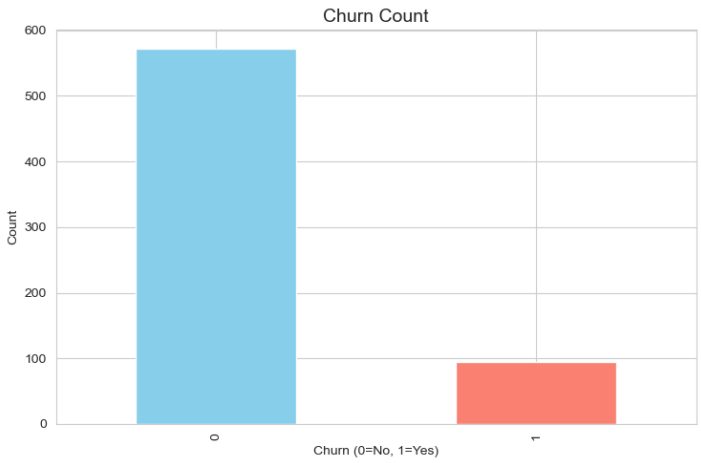

Customer service calls by Churn
Churn
0
1
Count

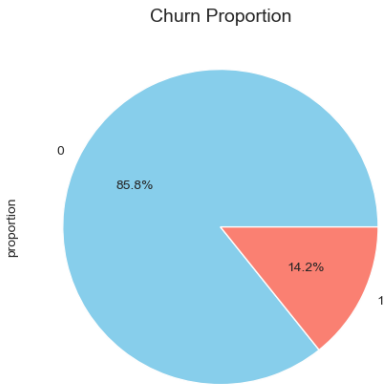

TotalCalls by Churn
Churn
0
1
Count


TotalMinutes by Churn
Churn
0
1
Count


TotalCharge by Churn
Churn
0
1
Count


AvgChargePerMinute by Churn
Churn
0
1
Count





2.1.6 Conclusion

The project successfully demonstrates the use of **Exploratory Data Analysis (EDA)** techniques to understand and visualize data.

It highlights how raw data can be transformed into meaningful insights using Python tools. The findings from this project can support better decision-making and provide a strong foundation for advanced analytics such as machine learning and predictive modelling.

2.2 Climate Change Trends Analysis (Week 2)

2.2.1 Introduction

Climate change is one of the most pressing global challenges, significantly impacting environmental stability, ecosystems, and human life. Rising global temperatures, increasing carbon emissions, and changing atmospheric conditions have led to long-term climatic variations.

This project, **Climate Change Trends Analysis**, focuses on analysing historical climate data to identify trends, patterns, and relationships among key environmental indicators. The dataset includes parameters such as temperature, carbon dioxide (CO_2) emissions, humidity, and other atmospheric variables recorded over multiple years.

Using data analysis and visualization techniques, the project aims to understand how climate variables have changed over time and how they are interrelated. Tools such as Python and visualization libraries were used to uncover insights and present findings effectively.

2.2.2 Objectives

➤ Primary Objectives

- To analyse historical climate data for identifying long-term trends.
- To clean and preprocess environmental datasets for accurate analysis.
- To study the relationship between temperature and CO_2 emissions.
- To visualize climate trends using graphs and charts.
- To derive insights that highlight the impact of climate change

➤ Specific Analytical Goals

- Analyse temperature variations over time.
- Study CO_2 emission trends across years.
- Identify correlations between environmental variables.
- Detect seasonal and regional variations.
- Visualize distributions and patterns in climate data.

2.2.3 Methodology

a) Dataset Preparation

¬ Loaded the dataset containing climate indicators such as temperature, CO_2 emissions, and humidity.
¬ Checked dataset structure using:

- .head(), .info(), .describe()

¬ Handled missing values:

- Removed null values
- Applied mean/median imputation where required

¬ Converted data types:

- Date/time columns into proper format
- Numerical fields standardized

---

b) Data Cleaning and Transformation
¬ Removed duplicate entries.
¬ Renamed columns for clarity.
¬ Extracted time-based features:

- **Year, Month**

¬ Created derived columns where required for analysis.
c) Analysis Techniques
¬ Performed statistical analysis:

- **Mean, median, standard deviation**

¬ Used grouping techniques:

- **groupby() for year-wise analysis**

¬ Performed correlation analysis to identify relationships between variables.

---

d) Data Visualization
Visualizations were created using Python libraries:
• Line charts → Temperature trends over time
• Bar charts → Category comparisons
• Histograms → Distribution analysis
• Heatmaps → Correlation analysis

2.2.4 Results and Insights
a) Temperature Trends
• A steady increase in temperature over the years was observed.
• Indicates long-term global warming patterns.

---

b) CO_2 Emission Trends

• CO_2 emissions show a continuous upward trend.

• Reflects the impact of industrialization and human activities.

---

c) Correlation Insights

• Strong positive correlation between CO_2 emissions and temperature.

• Confirms the role of greenhouse gases in climate change.

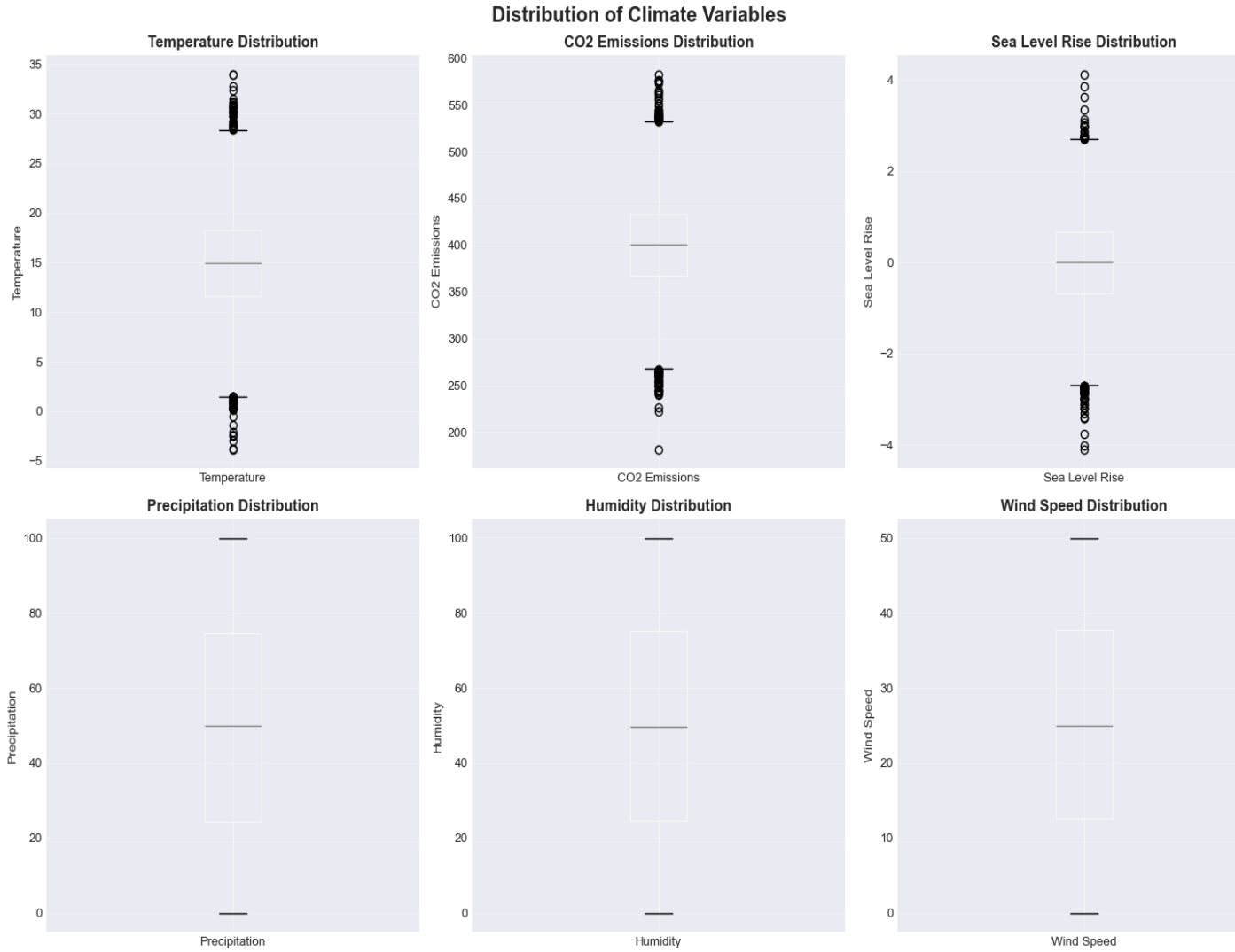
---

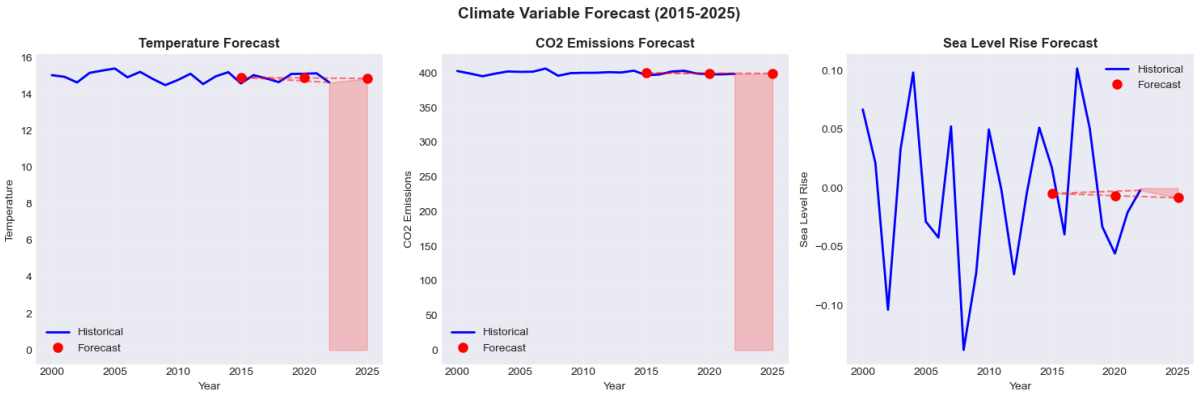
d) Variability and Patterns

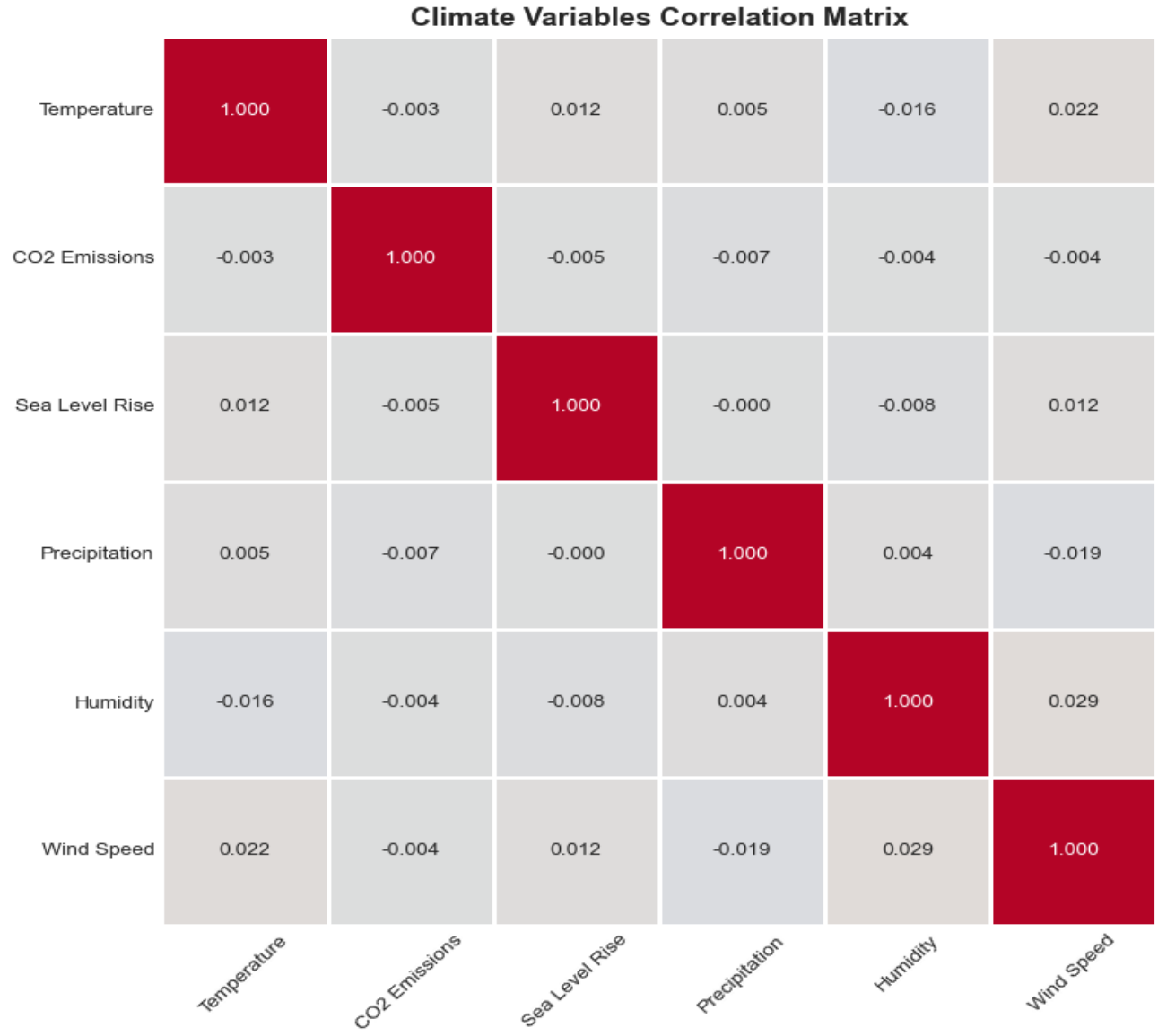
• Seasonal and yearly fluctuations observed.

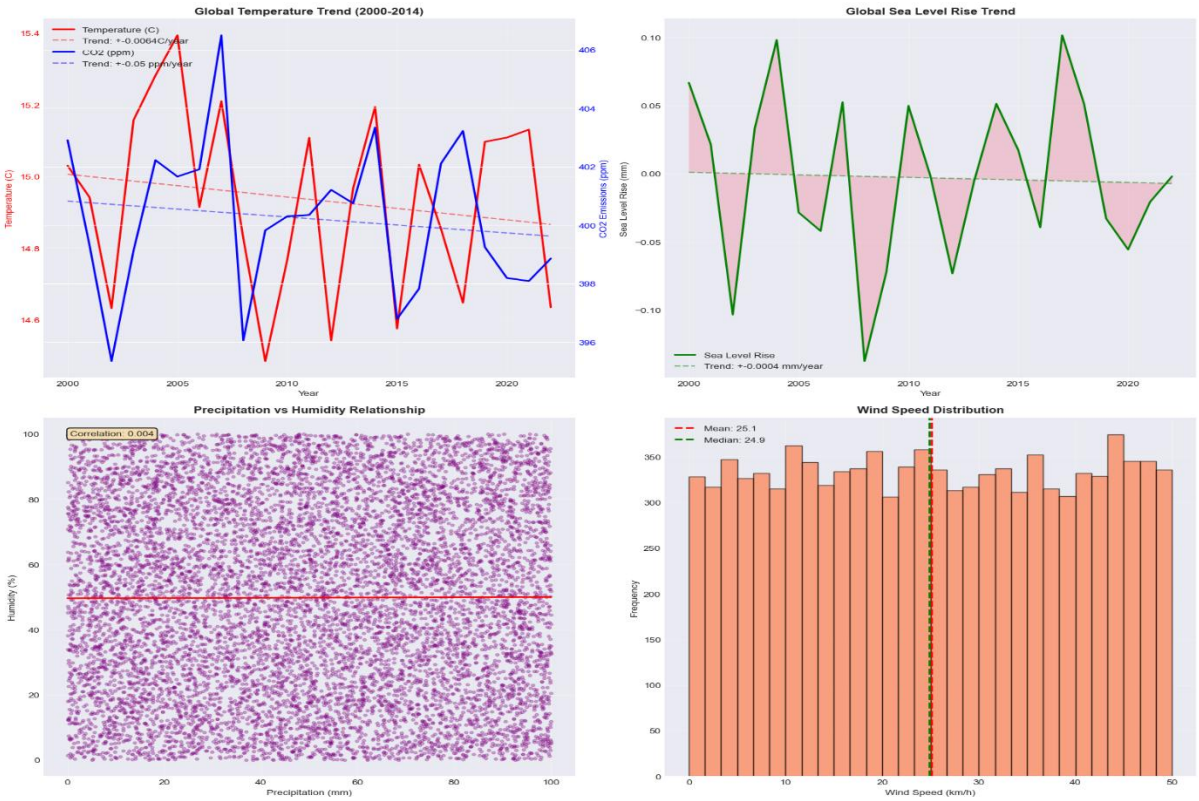
• Long-term trend shows increasing environmental instability.

Output:




Climate Variable Forecast (2015-2025)
Temperature Forecast
Temperature
Year
Historical
Forecast
CO2 Emissions Forecast
CO2 Emissions
Year
Historical
Forecast
Sea Level Rise Forecast
Sea Level Rise
Year
Historical
Forecast


Climate Variables Correlation Matrix
Temperature
CO2 Emissions
Sea Level Rise
Precipitation
Humidity
Wind Speed
1.000 -0.003 0.012 0.005 -0.016 0.022
-0.003 1.000 -0.005 -0.007 -0.004 -0.004
0.012 -0.005 1.000 -0.000 -0.008 0.012
0.005 -0.007 -0.000 1.000 0.004 -0.019
-0.016 -0.004 -0.008 0.004 1.000 0.029
0.022 -0.004 0.012 -0.019 0.029 1.000



2.2.5 Conclusion

The Climate Change Trends Analysis project demonstrates how data analysis techniques can be used to study environmental changes over time. The results highlight a strong relationship between rising CO_2 emissions and increasing temperatures, emphasizing the impact of human activities on climate change. The project underscores the importance of data-driven insights in supporting sustainable development and environmental policy decisions.

2.3 SMART CROP YIELD PREDICTION SYSTEM

(Week – 3)

2.3.1 Introduction

Agriculture plays a vital role in India's economy by contributing significantly to employment, food security, and rural development. However, agricultural productivity is highly influenced by environmental conditions such as rainfall, fertilizer usage, pesticide application, and seasonal climate variations. Farmers often face uncertainty regarding crop production, which affects financial planning, market decisions, and resource allocation.

This project, *Smart Crop Yield Prediction System*, focuses on developing an AI-powered prediction system capable of forecasting crop yield using historical agricultural and environmental data. The system combines machine learning techniques with interactive data visualization to provide accurate and meaningful yield predictions.

The project utilizes a structured dataset containing 19,689 records of 55 crop varieties across multiple Indian states and cultivation seasons. Important parameters such as annual rainfall, fertilizer usage, pesticide consumption, cultivated area, crop type, state, and season were used for analysis and prediction.

The system was developed using Python, SQLite, Scikit-learn, and Streamlit. A Random Forest Regression model was trained to identify complex relationships between agricultural inputs and crop productivity. An interactive web-based dashboard was also created to visualize trends, state-wise performance, crop rankings, and environmental impacts on yield.

The project demonstrates how machine learning and data analytics can support precision agriculture and assist farmers in making informed decisions for better crop management and production planning.

2.3.2 Objectives

Primary Objectives

- **To develop a machine learning-based crop yield prediction system using agricultural and environmental parameters.**
- **To analyse historical crop production data across Indian states and seasons.**
- **To build an interactive web application for real-time crop yield prediction.**
- **To identify relationships between rainfall, fertilizer usage, pesticide consumption, and crop productivity.**
- **To generate analytical dashboards and visual insights for agricultural decision-making.**

Specific Analytical Goals

- **To create a structured relational database using crop production datasets.**
- **To preprocess and clean agricultural data for accurate prediction.**
- **To apply feature engineering and categorical encoding techniques.**
- **To train and evaluate a Random Forest Regression model for yield prediction.**
- **To visualize state-wise, crop-wise, and year-wise agricultural trends.**
- **To analyse environmental impacts on crop yield using interactive charts and heatmaps.**

2.3.3 Methodology

a) Dataset Preparation and Data Ingestion

The dataset used in this project contains historical crop production records from various Indian states across multiple cultivation seasons.

Dataset Features Included:

- **Crop Name**
- **Crop Year**
- **State**
- **Season**
- **Area (hectares)**
- **Production (tonnes)**
- **Annual Rainfall (mm)**
- **Fertilizer Usage (kg/ha)**
- **Pesticide Usage (kg/ha)**
- **Yield (tonnes/hectare)**

Data Preparation Steps:

- **Imported dataset using Pandas.**
- **Standardized column names and formats.**
- **Converted yield values into numeric format.**
- **Removed null and invalid records.**
- **Cleaned inconsistent text values and duplicate entries.**
- **Inserted cleaned records into SQLite database.**

---

b) Database Architecture

A relational SQLite database was created to store and manage agricultural records efficiently.

Database Tables Created:

- **crop_data → Stores complete crop records.**
- **yearly_summary → Average yield and rainfall by year.**
- **state_summary → State-wise average crop performance.**
- **crop_summary → Crop-wise production statistics.**

Benefits of Database Structure:

- **Faster querying and analysis.**
- **Improved dashboard performance.**
- **Better organization of agricultural records.**

---

c) Feature Engineering and Preprocessing

The prediction model uses both numerical and categorical variables.

Numerical Features:

- **Crop Year**
- **Area**
- **Annual Rainfall**
- **Fertilizer**
- **Pesticide**

Categorical Features:

- **Crop**
- **State**
- **Season**

Preprocessing Techniques Used:

- **Standard Scaler for numerical normalization.**
- **One Hot Encoding for categorical variables.**
- **Column Transformer pipeline for combined preprocessing.**
- **Feature transformation using Scikit-learn Pipeline.**

---

d) Model Training and Evaluation

A Random Forest Regression model was used due to its ability to handle non-linear agricultural relationships.

Training Process:

- **Split dataset into 80% training and 20% testing data.**
- **Used Random Forest Regressor with 100 decision trees.**
- **Enabled parallel processing for faster computation.**
- **Trained model using Scikit-learn Pipeline.**

Evaluation Metrics:

- **R^2 Score**
- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Error (MAE)**

Model Performance:

- **Training Accuracy (R^2): 0.94 – 0.97**
- **Testing Accuracy (R^2): 0.85 – 0.91**

The model demonstrated strong predictive performance with minimal overfitting.

---

e) Web Application Development

The prediction system was deployed using Streamlit.

Features of Web Application:

- **Interactive sidebar for user inputs.**
- **Real-time crop yield prediction.**
- **Animated agricultural user interface.**
- **Multi-tab analytics dashboard.**
- **Responsive layout for desktop and mobile devices.**

Dashboard Visualizations:

- **Year-wise Yield Trends**
- **State-wise Yield Comparison**
- **Crop Ranking Analysis**
- **Rainfall vs Yield Scatter Plot**
- **Fertilizer Impact Analysis**
- **Correlation Heatmaps**

2.3.4 Results and Insights

i) Dataset Analysis

- **Total records processed: 19,689**
- **Total crop varieties analysed: 55**
- **Multiple Indian states and seasons included.**
- **Historical data ranged across several cultivation years.**

---

ii) State-wise Yield Performance

- **Goa recorded the highest average crop yield.**
- **States with better rainfall and irrigation facilities showed stronger productivity.**
- **Plantation crops contributed significantly to higher yield averages.**

---

iii) Crop-wise Yield Analysis

- **Coconut recorded the highest yield among all crops.**
- **Rice and Sugarcane showed strong positive response to rainfall.**
- **Millets and Groundnut maintained stable yields under lower rainfall conditions.**

---

iv) Rainfall Impact on Crop Yield

- **Moderate-to-high rainfall positively influenced crop productivity.**
- **Excessive or insufficient rainfall reduced yield stability.**
- **Different crop categories responded differently to rainfall conditions.**

---

v) Fertilizer and Pesticide Analysis

- **Fertilizer usage showed positive correlation with crop yield.**
- **Yield improvements gradually reduced at excessive fertilizer application levels.**
- **Balanced input utilization produced optimal results.**

---

vi) Machine Learning Model Insights

- **Random Forest successfully captured non-linear agricultural relationships.**
- **The model generalized well on unseen data.**
- **Minimal difference between training and testing accuracy confirmed reduced overfitting.**

---

vii) Dashboard Insights

The analytics dashboard provided:

- **State-wise agricultural comparisons.**
- **Historical yield trend analysis.**
- **Environmental impact visualization.**
- **Crop performance rankings.**
- **Interactive filtering and exploration.**

2.3.5 Technologies Used

Programming & Development

- **Python 3.x**
- **Streamlit**

Database & Data Handling

- **SQLite3**
- **Pandas**
- **NumPy**

Machine Learning

- **Scikit-learn**
- **Random Forest Regression**

Visualization Tools

- **Plotly Express**
- **Plotly Graph Objects**

Model Serialization

- **Pickle**

2.3.6 Challenges and Limitations

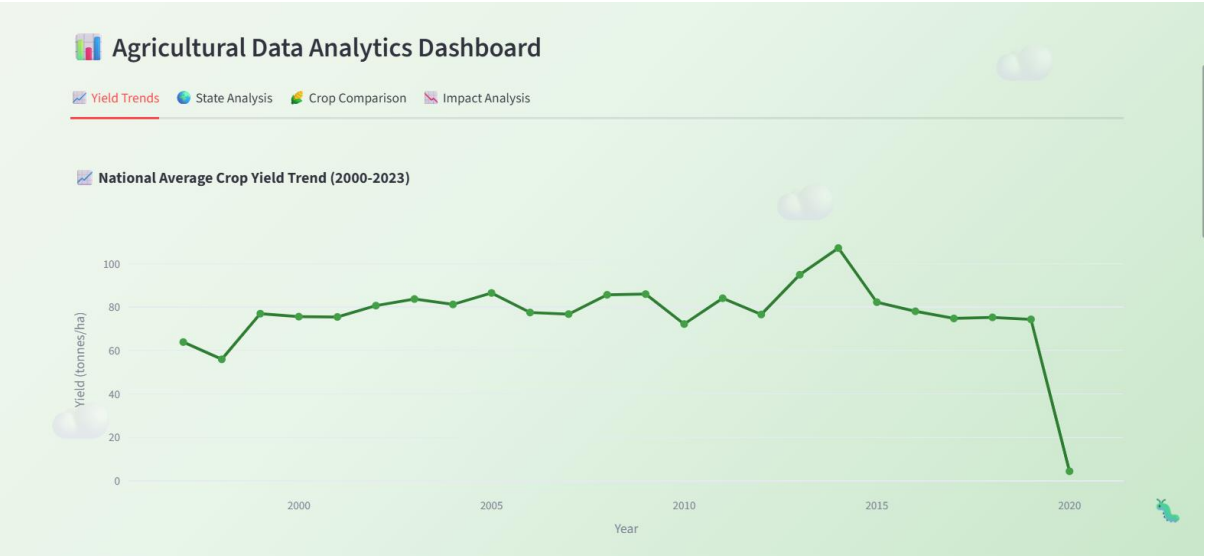
Challenges Faced

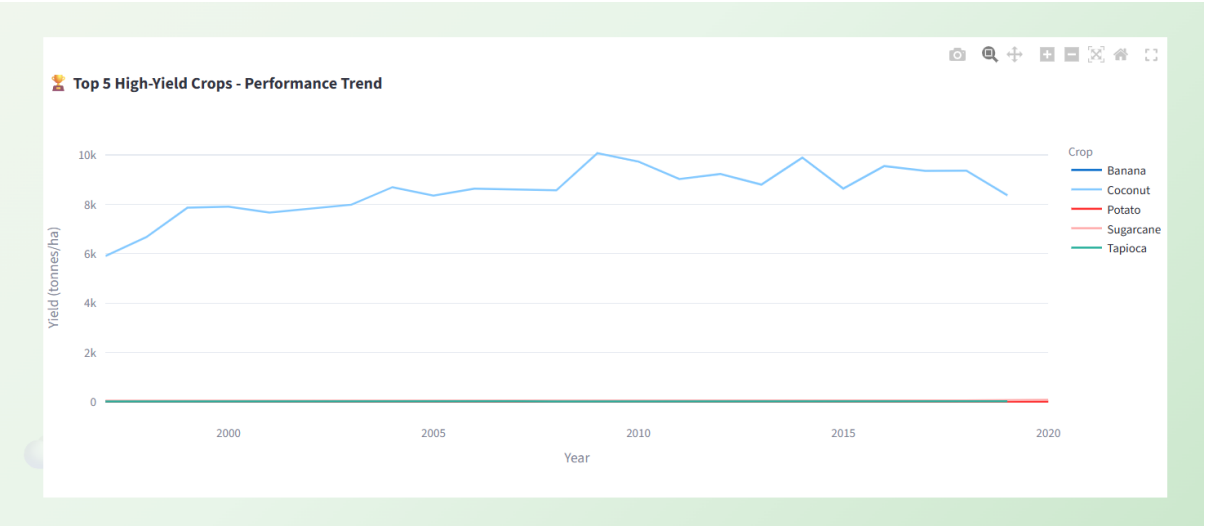
- **Highly skewed agricultural yield distributions.**
- **Large variation between plantation and cereal crop yields.**
- **Missing soil health and irrigation data.**
- **Data inconsistency across regions and seasons.**

Limitations

- **Model does not include real-time weather forecasting.**
- **Soil nutrient information was unavailable.**
- **Temporal sequence modelling was limited.**
- **Crop variety-level differentiation was not included.**

Results :





2.3.7 Conclusion

The Smart Crop Yield Prediction System demonstrates the practical application of machine learning in agriculture for accurate crop yield forecasting. By integrating environmental, agronomic, and geographical data, the system successfully predicts crop productivity using a Random Forest Regression model.

The project achieved high predictive accuracy while also providing meaningful analytical insights through an interactive web dashboard. The system enables better agricultural planning, resource optimization, and data-driven decision-making for farmers and policymakers.

The modular architecture of the project allows future integration of real-time weather APIs, soil health records, satellite data, and advanced deep learning techniques to further improve prediction accuracy and precision agriculture capabilities.

2.4 E-Commerce Sales Dashboard – ETL PIPELINE AND DATA VISUALIZATION (Week – 4)

2.4.1 Introduction

Data engineering plays a crucial role in modern analytics and business intelligence systems. Before meaningful analysis and visualization can be performed, raw data must be extracted, cleaned, transformed, and loaded into a structured format suitable for analytical processing. This project focuses on implementing a complete ETL (Extract, Transform, Load) pipeline using the Sample Superstore retail dataset to analyse sales performance, customer behaviour, product profitability, and regional business trends.

The project processes approximately 9,994 transactional records containing 21 columns related to e-commerce and retail sales operations across the United States from 2014 to 2017. The dataset includes information such as customer segments, product categories, order dates, sales, profit, discount rates, and shipping details.

The ETL pipeline was implemented entirely in Python using Jupyter Notebook. Pandas was used for data extraction, transformation, and cleaning operations, while Matplotlib, Seaborn, and Plotly were used for static and interactive visualizations. The cleaned dataset was exported into a CSV format compatible with Power BI and Tableau for further dashboard development.

The project demonstrates how structured ETL workflows and effective visualizations can convert raw retail transaction data into actionable business intelligence insights that support operational and strategic decision-making.

2.4.2 Objectives

Primary Objectives

- **To design and implement a complete ETL pipeline for retail transaction data.**
- **To clean, preprocess, and transform raw sales data for analytical usage.**
- **To create meaningful derived metrics such as profit margin and shipping duration.**

- **To generate multiple business intelligence visualizations using Python.**
- **To prepare a cleaned dataset compatible with Power BI and Tableau.**
  Specific Analytical Goals
- **To analyse revenue and profit trends over time.**
- **To evaluate category-wise and sub-category-wise profitability.**
- **To identify top-performing products and customer segments.**
- **To examine the relationship between discounts and profitability.**
- **To analyse regional sales performance across US states.**
- **To identify loss-making transactions and operational inefficiencies.**

2.4.3 Methodology

a) Dataset Description

The dataset used for this project is the *Sample Superstore Dataset*, a synthetic retail transaction dataset commonly used in business analytics and visualization projects.

Dataset Features Included:

- **Order ID**
- **Order Date**
- **Ship Date**
- **Customer ID**
- **Customer Segment**
- **State**
- **Region**
- **Product Category**
- **Sub-Category**
- **Product Name**
- **Sales**
- **Quantity**
- **Discount**
- **Profit**
- **Ship Mode**

Dataset Overview:

- **Total Records: 9,994**
- **Total Columns: 21**
- **Time Period: 2014 – 2017**
- **Product Categories:**
  - **Furniture**

- **Office Supplies**
- **Technology**

---

b) Extract Phase

The extraction phase focuses on locating and importing the raw CSV dataset into the Python environment.

Extraction Steps:

- **Implemented automatic file path discovery.**
- **Searched multiple system directories for dataset location.**
- **Loaded dataset using pd.read_csv().**
- **Used Latin-1 encoding for special character handling.**
- **Verified successful extraction through row and column validation.**

Libraries Used:

- **Pandas**
- **OS Module**

---

c) Transform Phase

The transformation phase involved multiple cleaning, preprocessing, and feature engineering operations.

Data Cleaning Operations:

- **Removed duplicate rows using drop_duplicates().**
- **Removed null values in critical financial columns.**
- **Validated structural consistency of dataset.**

Date Processing:

- **Converted Order Date and Ship Date into datetime format.**
- **Extracted:**
  - **Year**
  - **Month**

Feature Engineering:

Created additional analytical columns:

- **Profit Margin**
- **Days to Ship**
- **Loss Making Flag**

Derived Metrics:

Profit Margin Formula:

$$\textbf{Profit Margin} = \frac{\textbf{Profit}}{\textbf{Sales}}$$

Shipping Duration:

$$\text{Days to Ship} = \text{Ship Date} - \text{Order Date}$$

Benefits of Transformation:

- **Improved analytical capability.**
- **Enhanced dashboard performance.**
- **Better business intelligence insights.**

---

d) Load Phase

The transformed dataset was exported into a new cleaned CSV file.

Loading Steps:

- **Saved cleaned dataset as cleaned_superstore.csv.**
- **Removed unnecessary index columns.**
- **Ensured compatibility with Power BI and Tableau.**

Final Dataset Summary:

- **Total Records After Cleaning: 9,994**
- **Total Columns After Feature Engineering: 26**
- **Missing Values: 0**
- **Duplicate Rows: 0**

---

e) Data Quality Report

A data quality validation report was generated after ETL execution.

Data Quality Metrics:

- **Total Revenue: $2,297,201**
- **Total Profit: $286,397**
- **Average Profit Margin: 12.5%**
- **Unique Orders Count**
- **Total Loss-Making Orders**
- **Overall Loss Rate Percentage**

---

f) Visualization Design

Ten analytical visualizations were generated to analyse different business dimensions.

Visualizations Created:

1. Monthly Revenue and Profit Trend

- **Dual-axis line chart comparing revenue and profit trends over time.**

2. Category-wise Revenue and Profit

- **Grouped bar chart for Furniture, Technology, and Office Supplies.**

3. Top 10 Products by Revenue

- **Horizontal bar chart showing highest-selling products.**

4. Customer Segment Analysis

- **Revenue, order count, and profit comparison by customer segment.**

5. Profit Margin by Sub-Category

- **Visual identification of profitable and loss-making sub-categories.**

6. Loss Rate Analysis

- **Monthly and category-level loss trend analysis.**

7. Geographic Revenue Distribution

- **Interactive US state-wise choropleth sales map.**

8. Yearly Performance Summary

- **Year-wise revenue and profit comparison table.**

9. Discount vs Profit Analysis

- **Scatter plot showing impact of discounts on profitability.**

10. Executive Summary Dashboard

- **Multi-panel business intelligence dashboard.**

2.4.4 Results and Insights

i) Overall Business Performance

- **Revenue increased steadily from 2014 to 2017.**
- **Total business growth reached approximately 51%.**
- **Profit growth outpaced revenue growth, indicating operational improvements.**

---

ii) Year-wise Revenue and Profit Trends

Revenue Growth:

- **2014 Revenue: $484,247**
- **2017 Revenue: $733,215**

Profit Growth:

- **2014 Profit: $49,544**
- **2017 Profit: $93,439**

Insight:

The organization demonstrated continuous business expansion and profitability growth over the four-year period.

---

iii) Product Category Analysis

Technology:

- **Highest revenue-generating category.**

- **Strong profit margins.**
  Furniture:
- **Lower profit margins despite high sales.**
- **Several sub-categories consistently recorded losses.**
  Office Supplies:
- **Stable profitability and moderate sales.**

---

iv) Sub-Category Profitability
Loss-Making Sub-Categories:

- **Tables**
- **Bookcases**
  Insight:
  Negative profit margins suggest:
- **Excessive discounting**
- **Pricing inefficiencies**
- **High operational costs**

---

v) Discount Impact Analysis
Key Finding:
Orders with discounts above 40% were mostly loss-making.
Insight:
Aggressive discounting significantly reduced profitability despite increased sales volume.

KPI	Value	Notes
Total Revenue	$2,297,201	4-year cumulative (2014–2017)
Total Profit	$286,397	Net across all transactions
Avg Profit Margin	12.5%	Mean across all orders
Unique Orders	4,009	Distinct order IDs
Loss-Making Orders	~1,871	Approx. 18.7% of all orders
Best Year (Revenue)	2017 — $733,215	51% above 2014 baseline

---

vi) Customer Segment Analysis
Customer Segments:

- **Consumer**
- **Corporate**

- **Home Office**
  Segment Contribution:
- **Consumer Segment: ~50% of total sales**
- **Corporate Segment: ~31%**
- **Home Office Segment: ~19%**
  Insight:
  Consumer customers generated the highest overall business revenue.

---

vii) Geographic Sales Analysis
Top Revenue States:

- **California**
- **New York**
- **Texas**
  Insight:
  High-population and economically active states generated maximum sales revenue.

---

viii) Dashboard Insights
The executive dashboard provided:

- **Revenue trend monitoring**
- **Profitability diagnostics**
- **Geographic business analysis**
- **Customer behaviour insights**
- **Product performance comparison**

2.4.5 Technologies Used
Programming Environment

- **Python 3.x**
- **Jupyter Notebook**
  Data Processing
- **Pandas**
- **NumPy**
  Visualization Libraries
- **Matplotlib**
- **Seaborn**
- **Plotly Express**
  File Handling
- **OS Module**

Business Intelligence Compatibility

- **Power BI**
- **Tableau**

2.4.6 Challenges and Limitations

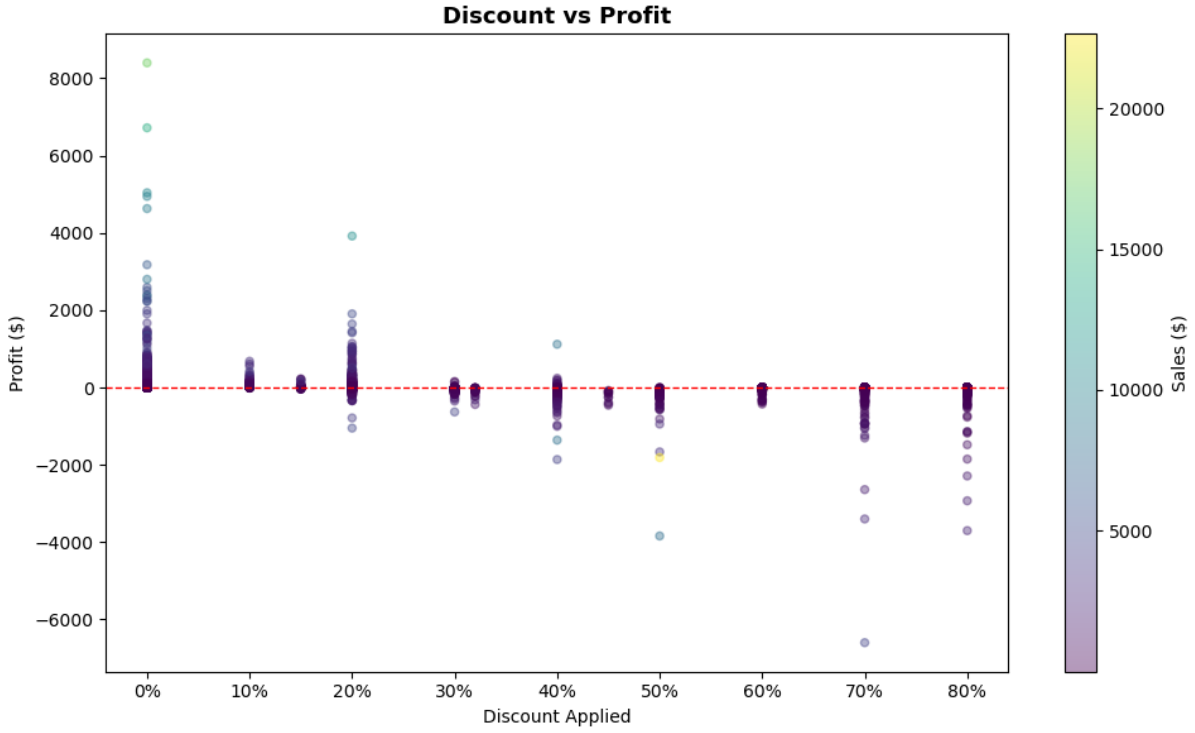
Challenges Faced

- **Complex dataset file path discovery.**
- **Dual-axis chart readability management.**
- **Large number of transactional records requiring optimized processing.**

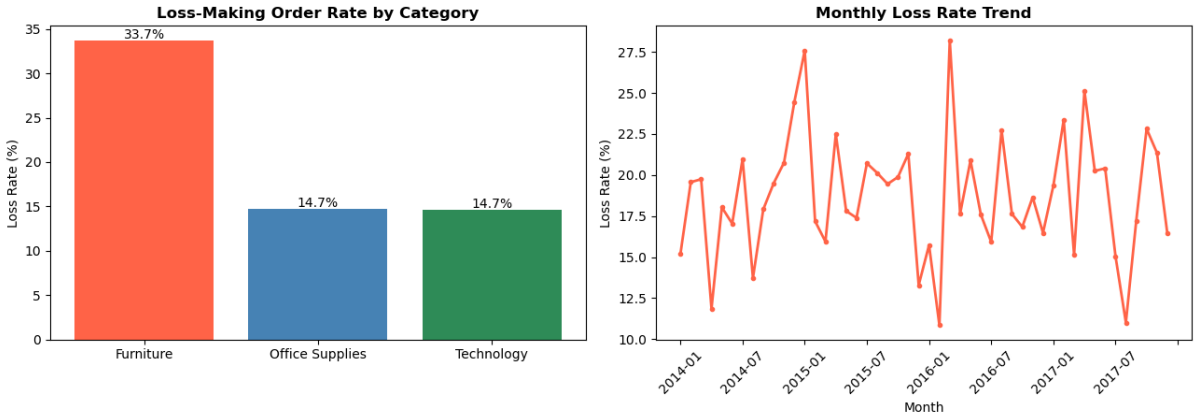
Limitations

- **Static CSV dataset without live database integration.**
- **No incremental data loading mechanism.**
- **Limited automated schema validation.**
- **No real-time sales data connectivity.**

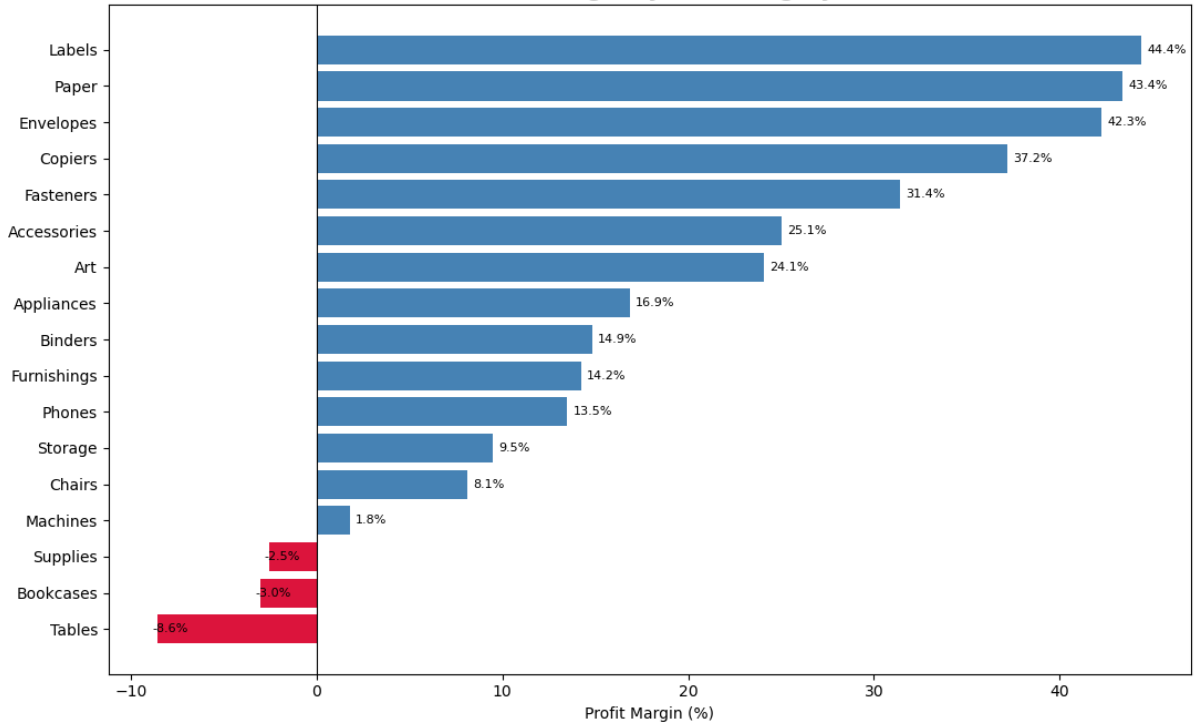
Results :



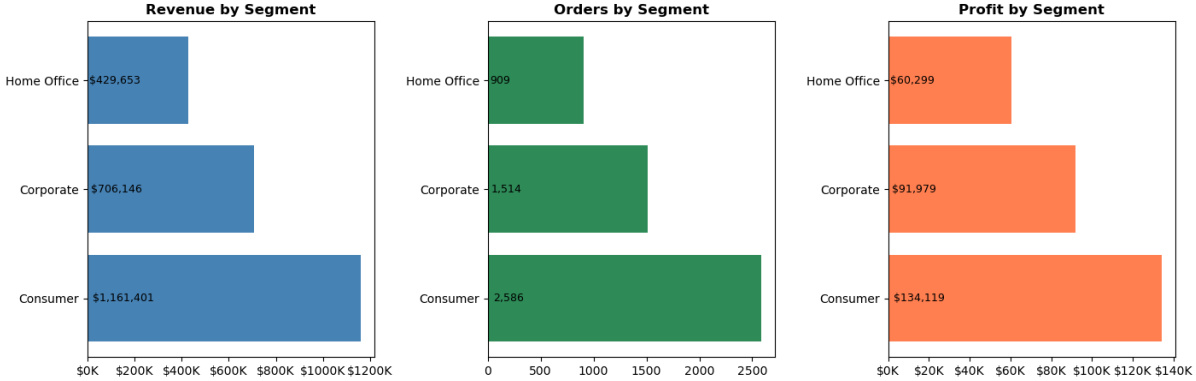
Loss Rate Analysis

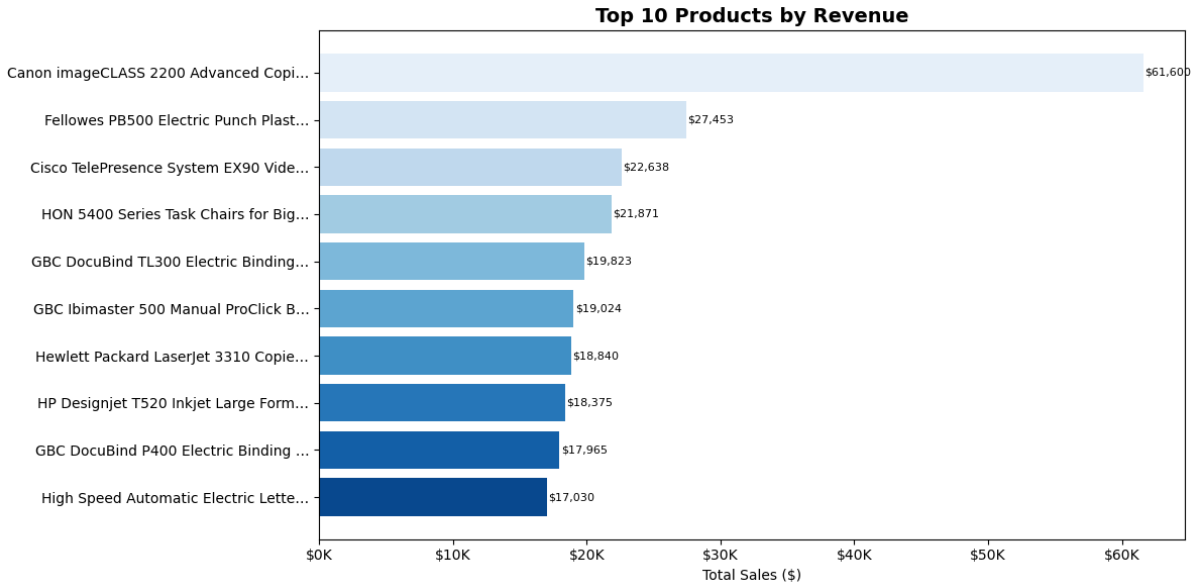


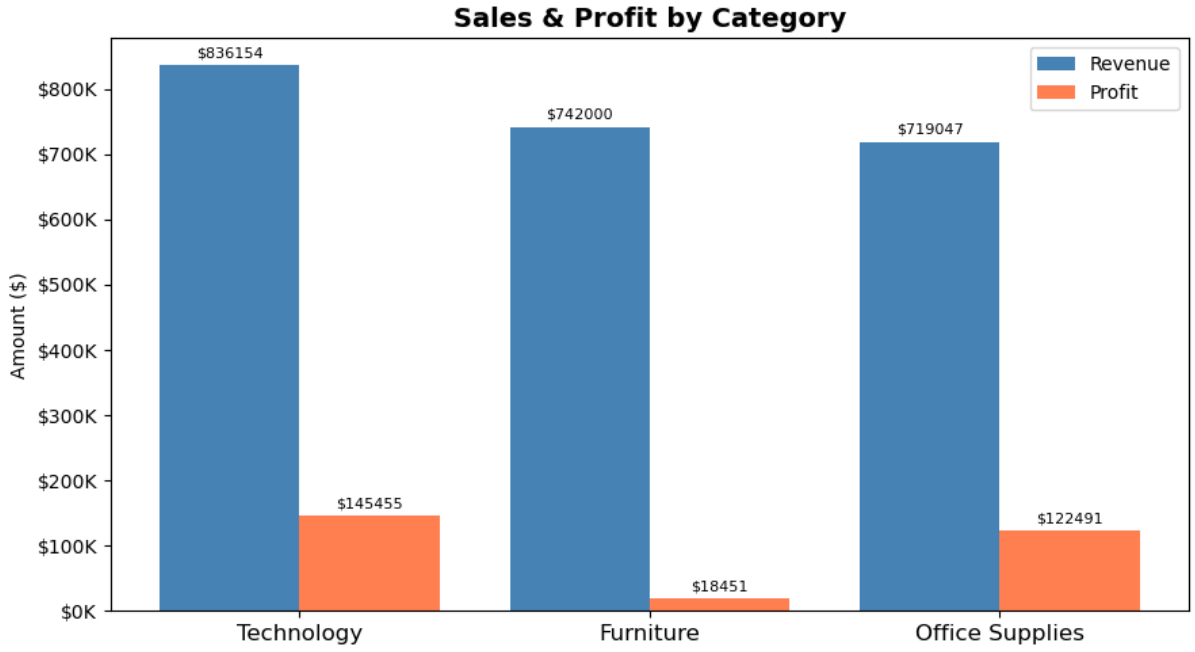
Profit Margin by Sub-Category

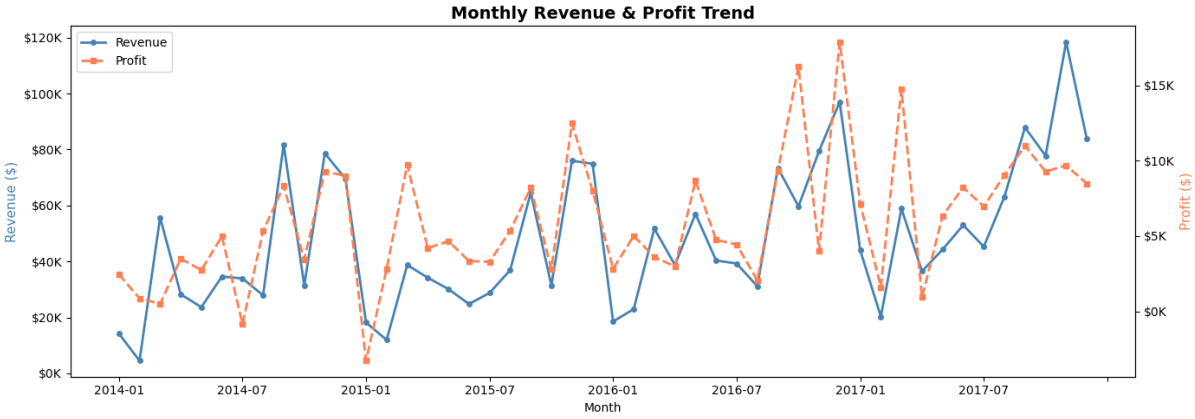


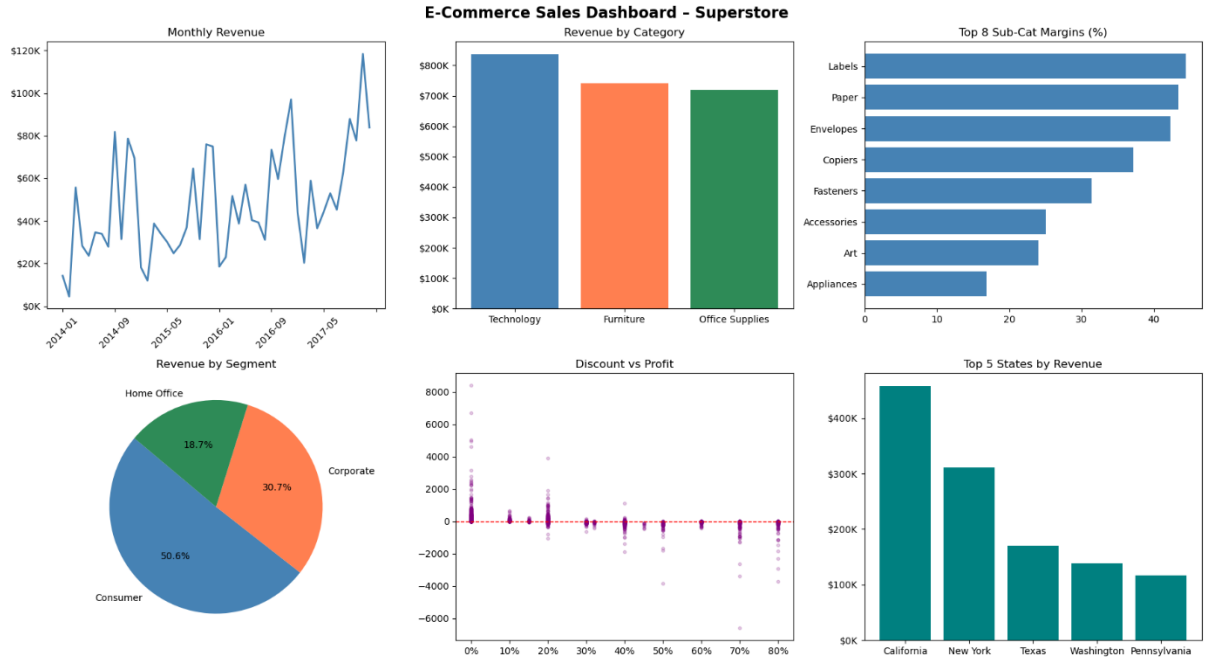
Sales Funnel by Customer Segment

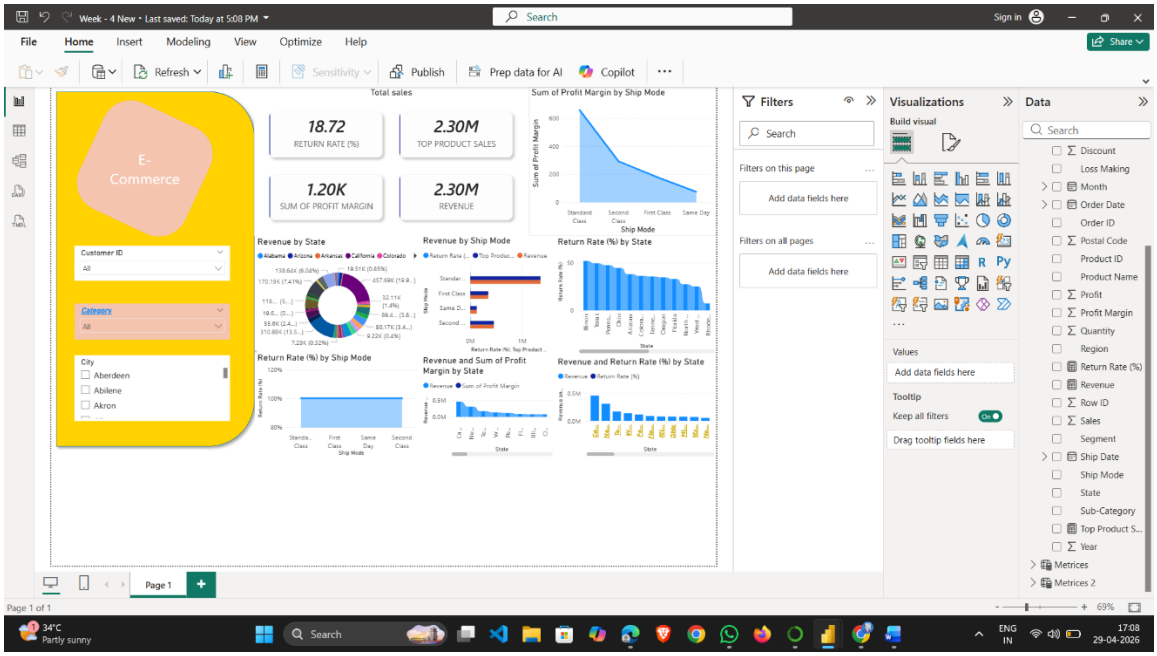



Top 10 Products by Revenue
Canon imageCLASS 2200 Advanced Copi...
Fellowes PB500 Electric Punch Plast...
Cisco TelePresence System EX90 Vide...
HON 5400 Series Task Chairs for Big...
GBC DocuBind TL300 Electric Binding...
GBC Ibimaster 500 Manual ProClick B...
Hewlett Packard LaserJet 3310 Copie...
HP Designjet T520 Inkjet Large Form...
GBC DocuBind P400 Electric Binding ...
High Speed Automatic Electric Lette...
$61,600
$27,453
$22,638
$21,871
$19,823
$19,024
$18,840
$18,375
$17,965
$17,030
$0K
$10K
$20K
$30K
$40K
$50K
$60K
Total Sales ($)


Sales & Profit by Category
Revenue
Profit
Amount ($)
$0K
$100K
$200K
$300K
$400K
$500K
$600K
$700K
$800K
$836154
$145455
$742000
$18451
$719047
$122491
Technology
Furniture
Office Supplies


Monthly Revenue & Profit Trend
Revenue
Profit
Revenue ($)
Profit ($)
$0K
$20K
$40K
$60K
$80K
$100K
$120K
$5K
$10K
$15K
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
Month





2.4.7 Conclusion

This project successfully demonstrates the implementation of a complete ETL pipeline for retail transaction analysis using Python. The pipeline effectively extracted, cleaned, transformed, and loaded raw transactional data into a structured analytical format suitable for business intelligence applications.

The generated visualizations provided valuable insights into sales growth, category performance, customer behaviour, discount impact, and geographic market trends. The analysis identified important business patterns such as the profitability risks associated with excessive discounting and the consistently loss-making furniture sub-categories.

The cleaned dataset's compatibility with Power BI and Tableau further demonstrates the importance of upstream data preprocessing in modern analytics workflows. Overall, the project highlights how ETL pipelines and data visualization techniques can support data-driven retail decision-making and operational optimization.

2.5 FRAUD TRANSACTION DETECTION (Week – 5)

2.5.1 Introduction

Financial fraud has become one of the major challenges in the modern digital economy. Credit card fraud, in particular, causes significant financial losses to banks, financial institutions, and customers worldwide. With the rapid growth of online transactions and digital payment systems, detecting fraudulent activities in real time has become increasingly important.

This project focuses on developing and evaluating machine learning models for detecting fraudulent credit card transactions using the Kaggle Credit Card Fraud Dataset. The project follows a complete data science workflow including Exploratory Data Analysis (EDA), data preprocessing, class imbalance handling, machine learning model development, and performance evaluation.

The dataset contains highly imbalanced transaction records where fraudulent transactions represent only a very small percentage of total transactions. To overcome this challenge, preprocessing techniques such as feature scaling and SMOTE (Synthetic Minority Oversampling Technique) were applied.

Multiple classification models were implemented and compared, including:

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **Neural Network**

The models were evaluated using metrics suitable for imbalanced classification problems such as Precision, Recall, F1-Score, AUC-ROC, and Precision-Recall Curve. The project demonstrates how machine learning techniques can effectively identify fraudulent financial transactions while minimizing false positives and improving fraud detection efficiency.

2.5.2 Objectives

Primary Objectives

- **To develop machine learning models capable of detecting fraudulent credit card transactions.**
- **To analyse transaction patterns and identify characteristics of fraudulent behaviour.**
- **To handle severe class imbalance using preprocessing techniques.**
- **To compare the performance of multiple classification algorithms.**
- **To evaluate models using fraud-specific performance metrics.**

Specific Analytical Goals

- **To perform exploratory data analysis on transaction records.**
- **To standardize and preprocess transaction features.**
- **To apply SMOTE for balancing minority fraud samples.**
- **To train and evaluate Logistic Regression, Decision Tree, Random Forest, and Neural Network models.**
- **To analyse Precision-Recall and ROC performance.**
- **To identify the most effective model for real-world fraud detection deployment.**

2.5.3 Methodology

a) Dataset Description

The dataset used in this project is the *Credit Card Fraud Detection Dataset* available on Kaggle.

Dataset Overview:

- **Total Transactions: 284,807**
- **Fraudulent Transactions: 492**
- **Fraud Percentage: 0.172%**

Dataset Features:

- **V1 – V28 (PCA-transformed anonymized features)**
- **Transaction Amount**
- **Transaction Time**
- **Target Variable: Class**
  - **0 → Legitimate Transaction**
  - **1 → Fraudulent Transaction**

The dataset is highly imbalanced, making fraud detection a challenging classification problem.

---

b) Exploratory Data Analysis (EDA)

EDA was performed to understand transaction behaviour and fraud distribution.

EDA Techniques Used:

- **Class distribution analysis**
- **Transaction amount distribution**
- **Time-series fraud analysis**
- **Correlation heatmaps**
- **Boxplots and violin plots**

Key Observations:

- **Fraud cases formed a very small portion of the dataset.**
- **Fraudulent transactions mostly occurred at lower transaction amounts.**
- **Several PCA components showed strong separation between fraud and non-fraud transactions.**
- **No missing values were present in the dataset.**

---

c) Data Preprocessing

A structured preprocessing pipeline was implemented before model training.

Preprocessing Steps:

Feature Scaling

- **Amount and Time features were standardized using StandardScaler.**

Train-Test Split

- **Dataset split into:**
  - **80% Training Data**
  - **20% Testing Data**
- **Stratified splitting preserved fraud class ratio.**

Class Imbalance Handling

- **Applied SMOTE (Synthetic Minority Oversampling Technique) on training data.**
- **Generated synthetic fraud samples to balance the dataset.**

Benefits of SMOTE:

- **Improved fraud representation.**
- **Increased model sensitivity toward fraud cases.**
- **Reduced majority-class bias.**

---

d) Model Development

Four machine learning models were developed and compared.

---

i) Logistic Regression

Features:

- **Baseline classification model.**
- **Used L2 regularization.**
- **Applied balanced class weighting.**

Advantages:

- **Simple and interpretable.**
- **Fast training and prediction.**

---

ii) Decision Tree Classifier

Features:

- **Used Gini impurity criterion.**
- **Recursive feature partitioning.**

Hyperparameter Tuning:

- **Maximum depth**
- **Minimum samples split**
- **Minimum samples leaf**

Advantages:

- **Easy interpretability.**
- **Non-linear decision boundaries.**

---

iii) Random Forest Classifier

Features:

- **Ensemble model with 100 decision trees.**
- **Bootstrapped sampling and feature subsampling.**

Advantages:

- **Reduced overfitting.**
- **Strong performance on high-dimensional data.**
- **Feature importance analysis.**

---

iv) Neural Network

Architecture:

- **Input Layer**
- **Two Hidden Layers with ReLU activation**
- **Dropout layers (0.3)**
- **Sigmoid output layer**

Training Details:

- **Binary Cross Entropy loss function**
- **Adam optimizer**

- **Early stopping mechanism**

Advantages:

- **Deep feature learning.**
- **Better non-linear pattern detection.**

2.5.4 Evaluation Metrics

Since the dataset is highly imbalanced, accuracy alone was not considered reliable.

Metrics Used:

Precision

Measures correctness of predicted fraud transactions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

Measures ability to identify actual fraud cases.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score

Harmonic mean of Precision and Recall.

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

AUC-ROC

Measures overall classification performance across thresholds.

Precision-Recall Curve

More informative for rare fraud detection problems.

2.5.5 Results and Analysis

i) Logistic Regression Results

Performance:

- **AUC-ROC: ~0.97**
- **Recall: ~0.91**

- **Precision: ~0.85**
- **F1-Score: ~0.88**

Insight:
Logistic Regression provided a strong baseline but struggled with complex non-linear fraud patterns.

---

ii) Decision Tree Results
Performance:

- **AUC-ROC: ~0.94**
- **Recall: ~0.89**
- **Precision: ~0.82**

Insight:
Decision Trees showed moderate performance but suffered from slight overfitting despite tuning.

---

iii) Random Forest Results
Performance:

- **AUC-ROC: ~0.985**
- **Recall: ~0.934**
- **Precision: ~0.91**
- **F1-Score: ~0.92**

Key Features Identified:

- **V14**
- **V10**
- **V4**
- **V12**

Insight:
Random Forest achieved the best overall performance with high fraud detection capability and reduced false positives.

---

iv) Neural Network Results
Performance:

- **AUC-ROC: ~0.982**
- **Recall: ~0.928**
- **Precision: ~0.90**
- **F1-Score: ~0.914**

Insight:

Neural Networks performed similarly to Random Forest and demonstrated strong deep learning capability on fraud data.

---

v) Comparative Model Analysis

Best Performing Model:

- **Random Forest Classifier**

Reasons:

- **Strong recall performance**
- **Lower false positive generation**
- **Better generalization capability**
- **Reduced overfitting**

Operational Advantage:

Random Forest maintained high precision even at high recall thresholds, making it suitable for real-world fraud detection systems.

2.5.6 Discussion

The project highlights several important aspects of fraud detection using machine learning.

Key Findings:

- **Accuracy is not a reliable metric for highly imbalanced datasets.**
- **SMOTE significantly improved fraud detection performance.**
- **Ensemble methods outperform individual classifiers.**
- **Random Forest provides an effective balance between accuracy and interpretability.**
- **Neural Networks require careful regularization to avoid overfitting.**

Practical Implications:

- **High recall is critical because missed fraud cases directly cause financial losses.**
- **Excessive false positives increase operational review costs.**
- **Ensemble models are highly suitable for real-time fraud monitoring systems.**

2.5.7 Technologies Used

Programming & Development

- **Python 3.x**
- **Jupyter Notebook**

Data Processing

- **Pandas**

- **NumPy**

Machine Learning Libraries

- **Scikit-learn**
- **TensorFlow**
- **Keras**

Data Visualization

- **Matplotlib**
- **Seaborn**

Imbalance Handling

- **SMOTE (Imbalanced-learn)**

2.5.8 Challenges and Limitations

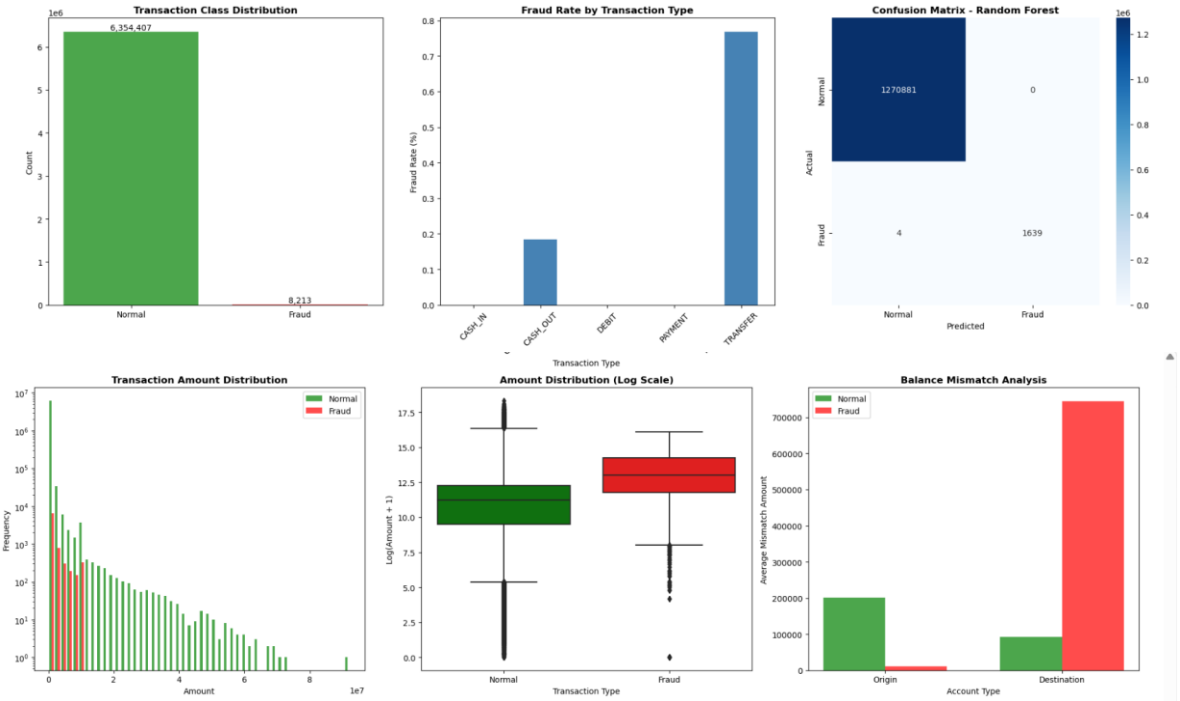
Challenges Faced

- **Severe class imbalance in fraud data.**
- **Difficulty in optimizing recall and precision simultaneously.**
- **Overfitting risk in complex models.**
- **PCA-transformed features reduced interpretability.**

Limitations

- **Dataset lacked real-world behavioural transaction features.**
- **No temporal sequence modelling was included.**
- **Static dataset without real-time streaming transactions.**
- **Fraud patterns may evolve over time, causing concept drift.**

Results:



2.5.9 Conclusion

This project successfully demonstrated the application of machine learning techniques for fraudulent credit card transaction detection. Through data preprocessing, SMOTE-based balancing, and model evaluation using imbalance-aware metrics, the study identified Random Forest as the best-performing model for fraud detection.

The Random Forest classifier achieved high recall and precision while effectively reducing false positives, making it highly suitable for practical deployment in financial fraud monitoring systems. Neural Networks also demonstrated strong performance and highlighted the potential of deep learning in fraud analytics.

The project reinforces the importance of proper preprocessing, imbalance handling, and evaluation metric selection in fraud detection problems.

Future improvements may include:

- **Real-time fraud monitoring systems**
- **Sequential deep learning models**
- **Explainable AI techniques**
- **Graph-based fraud detection**
- **Federated learning approaches**

Overall, the project demonstrates how machine learning can significantly improve financial fraud prevention and transaction security in modern digital banking systems.

2.6 ENERGY CONSUMPTION PATTERN ANALYSIS(Week – 6)

2.6.1 Introduction

The rapid growth of smart grid systems and connected metering infrastructure has resulted in large volumes of energy consumption data being generated across residential and commercial sectors. Understanding these consumption patterns is essential for improving energy efficiency, designing demand-side management programs, and developing targeted energy-saving strategies.

This project focuses on analysing residential energy consumption patterns using the ACORN (A Classification Of Residential Neighbourhoods) demographic dataset. The ACORN classification system categorizes UK households into different socioeconomic groups based on housing, finance, economy, transport, and population characteristics.

The project implements a complete analytical pipeline that includes:

- **Data loading and preprocessing**
- **Feature extraction**
- **Consumption matrix generation**
- **K-Means clustering**
- **Statistical anomaly detection**
- **Peak consumption analysis**
- **Visualization dashboard creation**
- **Energy-saving recommendation generation**

The dataset contains 826 records across 17 ACORN demographic segments (ACORN-A to ACORN-Q). A total of 177 features related to housing, finance, economy, and transport categories were analysed to identify consumption behaviour patterns and anomalous energy usage.

The project demonstrates how demographic segmentation combined with machine learning and statistical analysis can support targeted energy management and efficiency planning.

2.6.2 Objectives

Primary Objectives

- **To analyse residential energy consumption behaviour using ACORN demographic data.**

- **To identify high-consumption demographic segments using clustering techniques.**
- **To detect anomalous consumption patterns using statistical methods.**
- **To generate energy-saving recommendations for different household groups.**
- **To visualize energy consumption trends and cluster structures.**

Specific Analytical Goals

- **To extract energy-related demographic and housing features.**
- **To construct a standardized consumption matrix for analysis.**
- **To apply K-Means clustering for identifying consumption groups.**
- **To detect outliers using Z-score and IQR methods.**
- **To perform peak consumption analysis.**
- **To create an interactive visualization dashboard for analytical insights.**

2.6.3 Methodology

a) Dataset and Features

The dataset used in this project is the *ACORN Details Dataset*, which contains demographic and socioeconomic information for different UK household categories.

Dataset Overview:

- **Total Records: 826**
- **Total ACORN Segments: 17**
- **ACORN Categories:**
  - **ACORN-A to ACORN-Q**

Main Categories Used:

- **Housing**
- **Finance**
- **Economy**
- **Transport**
- **Population**

Features Analysed:

- **House Size**
- **House Value**
- **Housing Expenditure**
- **Employment Status**
- **Age Distribution**
- **Economic Activity**

- **Transportation Indicators**

---

b) Data Loading and Preprocessing

The dataset was loaded using a robust encoding detection mechanism.

Encodings Attempted:

- **UTF-8**
- **Latin-1**
- **ISO-8859-1**
- **CP1252**
- **UTF-16**

Successful Encoding:

- **Latin-1**

Preprocessing Steps:

- **Numeric conversion of features**
- **Missing value handling**
- **NaN imputation**
- **Standardization of numerical data**

---

c) Feature Extraction and Consumption Matrix Construction

Energy-related features were extracted from the Housing and Finance categories.

Important Extracted Features:

- **House Value**
- **House Size**
- **Housing Expenditure**

These features were considered strong indicators of household energy consumption.

Consumption Matrix:

- **Feature Rows: 177**
- **ACORN Segment Columns: 17**

The final matrix represented demographic indicators across all ACORN household categories.

---

d) Anomaly Detection

Two statistical methods were applied for identifying anomalous consumption patterns.

---

i) Z-Score Method

The Z-score method standardizes each data point relative to the mean and standard deviation.
Formula:

$$Z = \frac{X - \mu}{\sigma}$$

Threshold Used:

- **Absolute Z-score > 2.0**

Results:

- **Total Anomalies Detected: 167**

---

ii) Interquartile Range (IQR) Method
The IQR method identifies values outside the normal quartile range.
Formula:

$$IQR = Q3 - Q1$$

Anomaly Range:

$$Q1 - 1.5(IQR), Q3 + 1.5(IQR)$$

Results:

- **Total Anomalies Detected: 202**

Overall Anomaly Rate:

- **Approximately 5.55%**

Insight:
The anomaly detection methods identified unusual demographic and consumption behaviour patterns across specific ACORN segments.

---

e) K-Means Clustering and Pattern Identification
Before clustering, the consumption matrix was standardized using StandardScaler.
Clustering Algorithm:

- **K-Means Clustering**

Optimal Cluster Selection:

- **Elbow Method**

Optimal Number of Clusters:

- **k = 2**

Cluster Details:

Cluster 1:

- **Total Features: 132**
- **Average Consumption Index: 105.63**

Cluster 2:

- **Total Features: 45**
- **Average Consumption Index: 106.93**

Insight:
Cluster 2 represented above-average consumption patterns associated with specific housing and economic characteristics.

---

f) Peak Detection and Temporal Analysis
Peak consumption analysis was performed using signal processing techniques.
Techniques Used:

- **SciPy find_peaks() function**
- **Rolling average smoothing**

Analysis Parameters:

- **Dynamic threshold:**

$$\text{Mean} + 1 \times \text{Standard Deviation}$$

Additional Demographic Analysis:

- **Age Groups:**
  - **18–24**
  - **25–34**
  - **35–49**
  - **50–64**

Economic Categories:

- **Full-Time Employee**
- **Retired**
- **Unemployed**

---

g) Visualization Dashboard
A comprehensive dashboard was developed using Matplotlib and Seaborn.
Dashboard Components:
1. Consumption Heatmap

- **Visualized consumption index values across ACORN segments.**

2. Cluster Distribution Boxplots

- **Compared consumption spread between clusters.**

3. Feature Ranking Chart

- **Ranked features by average consumption index.**

4. Segment-wise Consumption Chart

- **Compared average consumption across ACORN categories.**

5. PCA Cluster Visualization

- **Visualized clusters in reduced two-dimensional space.**

6. Rolling Average Consumption Curves

- **Highlighted smoothed consumption trends.**

2.6.4 Results and Analysis

i) Dataset and Consumption Matrix Analysis

Dataset Summary:

- **Total Records: 826**
- **Total Features Analysed: 177**
- **ACORN Segments: 17**

Consumption Statistics:

- **Mean Consumption Index: 105.96**
- **Standard Deviation: 76.14**
- **Consumption Range: 0 – 1805**

Insight:

The dataset showed significant demographic and socioeconomic diversity across household groups.

---

ii) Anomaly Detection Results

Z-Score Method:

- **Total Anomalies: 167**

IQR Method:

- **Total Anomalies: 202**

Overall Anomaly Rate:

- **5.55%**

Insight:

Anomalies were concentrated in specific demographic segments rather than uniformly distributed across all groups.

---

iii) Cluster Analysis

Cluster 1:

- **Near-average consumption patterns.**
- **Larger cluster containing most demographic indicators.**

Cluster 2:

- **Higher average consumption patterns.**
- **Greater variability and heterogeneity.**

PCA Visualization:

The PCA scatter plot confirmed clear separation between the two clusters.

---

iv) Housing and Economic Insights

Key Findings:

- **Housing characteristics strongly influenced energy consumption.**
- **Larger and higher-value houses generally consumed more energy.**
- **Economic status significantly impacted consumption behaviour.**

Demographic Factors:

- **Age distribution influenced consumption variation.**
- **Employment categories showed different energy usage patterns.**

---

v) Peak Consumption Analysis

Key Findings:

- **Certain demographic groups consistently exhibited higher peak consumption.**
- **High-consumption peaks aligned with economic and housing indicators.**

Insight:

Peak analysis can support targeted demand-response strategies and load balancing programs.

---

vi) Dashboard Insights

The dashboard successfully provided:

- **Cluster-level consumption analysis**
- **Anomaly visualization**
- **Demographic comparison**
- **Consumption trend monitoring**
- **PCA-based cluster representation**

2.6.5 Discussion

The project demonstrates the effectiveness of combining demographic segmentation with machine learning techniques for energy consumption analysis.

Important Observations:

- **K-Means clustering effectively separated average and high-consumption groups.**
- **Z-score and IQR methods provided complementary anomaly detection results.**
- **Housing and financial indicators were strong predictors of consumption behaviour.**
- **Demographic segmentation can improve energy efficiency targeting.**

Energy Efficiency Recommendations:

For High Consumption Cluster:

- **Install smart metering systems.**
- **Conduct household energy audits.**
- **Promote energy-efficient appliances.**
- **Implement time-of-use pricing systems.**
- **Encourage energy-efficient retrofitting.**

General Recommendations:

- **Personalized energy-saving suggestions.**
- **Community-level awareness programs.**
- **Behavioural change initiatives.**
- **Smart energy monitoring dashboards.**

2.6.6 Technologies Used

Programming & Development

- **Python 3.x**
- **Jupyter Notebook**

Data Processing

- **Pandas**
- **NumPy**

Machine Learning

- **Scikit-learn**
- **K-Means Clustering**
- **PCA**

Statistical Analysis

- **Z-score Analysis**
- **IQR Method**
- **SciPy**

Visualization Libraries

- **Matplotlib**
- **Seaborn**

2.6.7 Challenges and Limitations

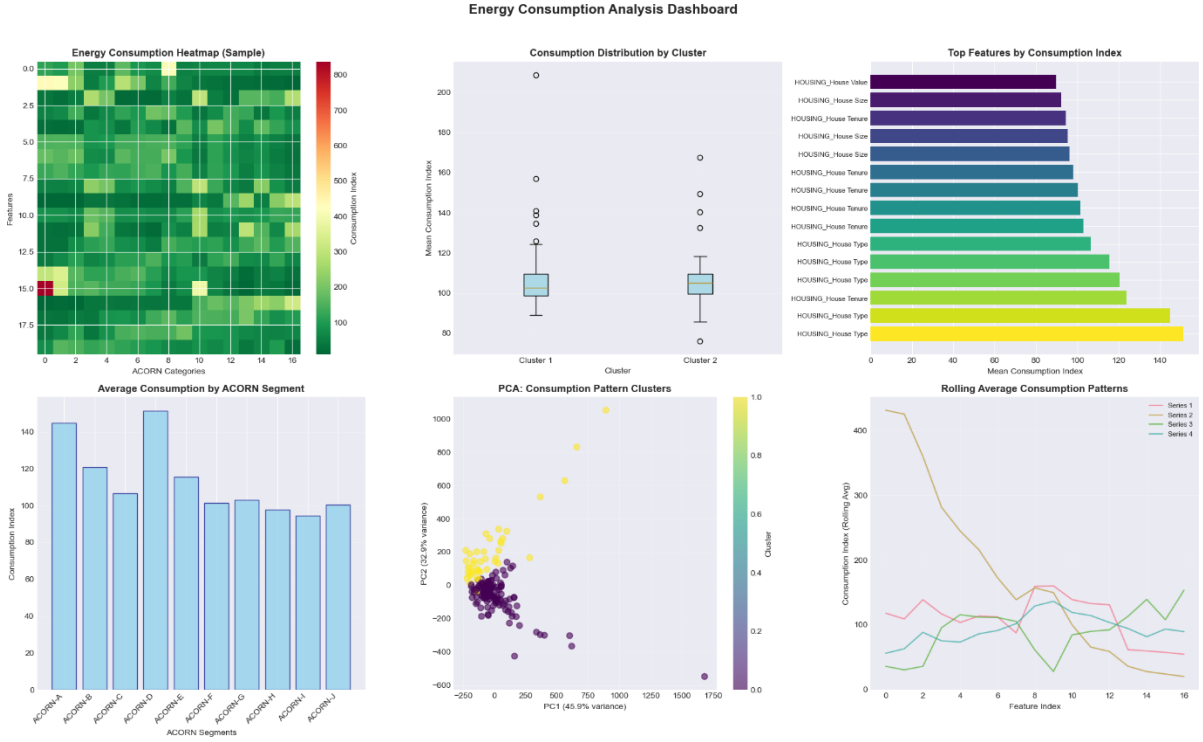
Challenges Faced

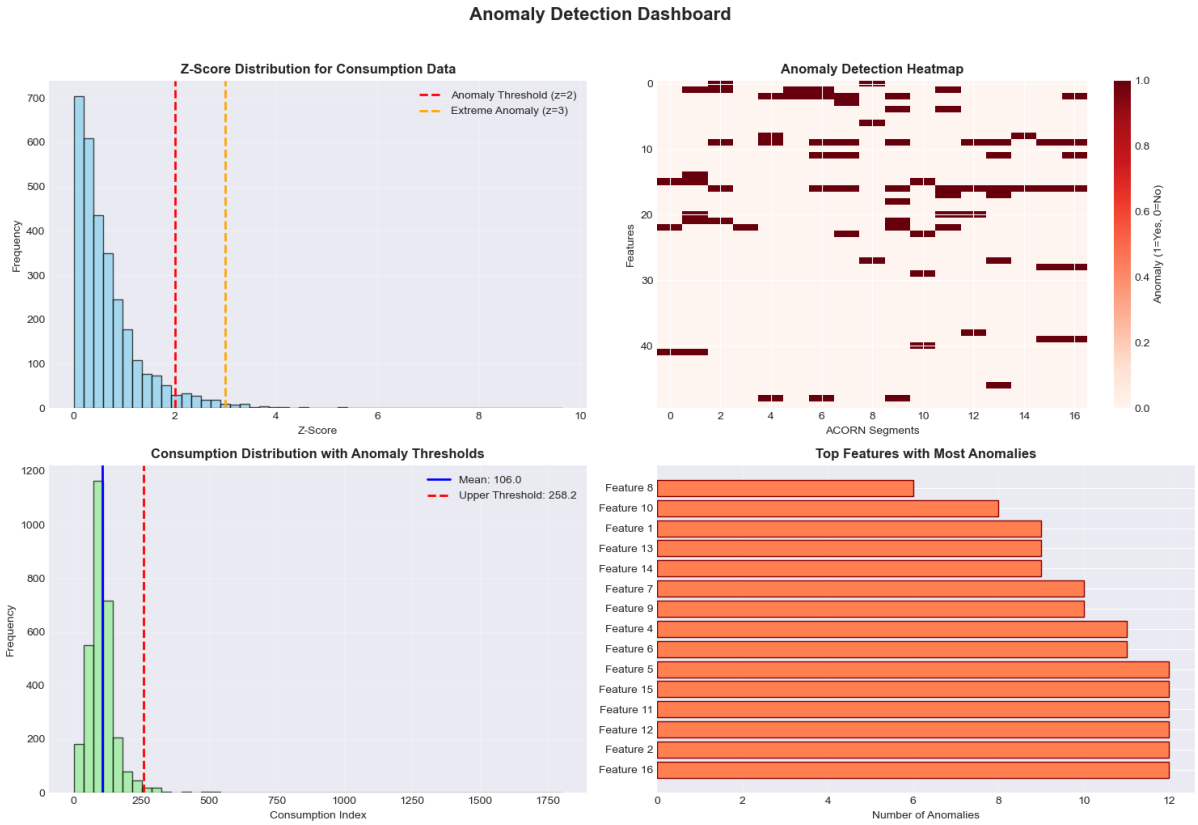
- **Handling encoding issues during data loading.**
- **Managing high-dimensional demographic data.**
- **Selecting optimal clustering parameters.**
- **Identifying meaningful anomalies in demographic indicators.**

Limitations

- **Dataset contained segment-level data rather than household-level meter readings.**
- **No real-time smart meter integration.**
- **Limited temporal energy consumption data.**
- **Lack of weather and appliance usage information.**

Results:





2.6.8 Conclusion

This project successfully demonstrated the application of machine learning and statistical analysis techniques for residential energy consumption pattern analysis using the ACORN demographic dataset.

The analytical pipeline effectively combined:

- **Feature extraction**
- **Consumption clustering**
- **Statistical anomaly detection**
- **Peak identification**
- **Visualization-based insights**

The project identified two major consumption clusters and detected significant anomalous patterns across demographic groups. Housing and economic variables emerged as strong predictors of energy consumption behaviour.

The findings highlight the potential of demographic segmentation in designing targeted energy efficiency interventions, smart grid planning, and demand-side management strategies. Future improvements may include:

- **Integration of smart meter time-series data**
- **Advanced clustering algorithms**

- **Real-time dashboard systems**
- **Weather-based energy forecasting**
- **Interactive energy planning tools**

Overall, the project demonstrates how data-driven analytics can support efficient energy management and sustainable residential energy consumption planning.

Bibilography

- **Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics, 95*(9–10), 1082–1095.**

- **CACI Ltd. (2023). *ACORN — A Classification Of Residential Neighbourhoods*. Retrieved from https://acorn.caci.co.uk**

- **Chicco, G. Napoli, R., & Piglione, F. (2006). Comparisons among clustering techniques for electricity customer classification. *IEEE Transactions on Power Systems, 21*(2), 933–940.**

- **Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. (2005). An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems, 20*(2), 596–602.**

- **Haben, S., Ward, J., Vukadinovic Greetham, D., Singleton, C., & Grindrod, P. (2016). A new error measure for forecasts of household-level, high resolution electrical energy consumption. *International Journal of Forecasting, 30*(2), 246–256.**

- **Janetzko, H., Stoffel, F., Mittelstädt, S., & Keim, D. (2014). Anomaly detection for visual analytics of power consumption data. *Computers & Graphics, 38*, 27–37.**

- **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesneau, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.**

- **Richardson, I., Thomson, M., Infield, D., & Clifford, C. (2010). Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings, 42*(10), 1878–1887.**