**INVERTIS UNIVERSITY**

Bareilly, U.P. (India)

**Department of Computer Applications**

**BCA 4th Semester** **[2024-25]**

**CSED (Centre for Skill and Entrepreneurship Development)**

**Project Report**

**Subject Name:** Machine Learning

**Code:**IIOT4

**PROJECT NAME:** SALES ANALYTICS

**DCS MENTOR:** Mr. MAYANK SINGH

**TEAM :** 4

**TEAM LEADER :** ABHISHEK KUMAR

**TEAM FORMATION**

# All Team Members -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr.**  **No.** | **Student ID** | **Roll No.** | **Student Name** | **Status** |
| 1 | 0 | 0 | ABHISHEK KUMAR **(Leader)** | Active |
| 2 | 0 | 0 | ABHISHEK MAHESHWARI | Active |
| 3 | 0 | 0 | VAISHNAVI GUPTA | Active |
| 4 | 0 | 0 | RIDHIMA MAHESHWARI | Active |
| 5 | 0 | 0 | VANSH GUPTA | Active |
| 6 | 0 | 0 | SAMINA NOOREEN | Active |
| 7 | 0 | 0 | SONI | Active |
| 8 | 0 | 0 | MOHD SHEEBAN KHAN | Active |
| 9 | 0 | 0 | ANKIT KUMAR | Active |
| 10 | 0 | 0 | ANKIT KUMAR | Active |
| 11 | 0 | 0 | SANJAY MISHRA | Active |
| 12 | 0 | 0 | ARPIT GUPTA | Active |
| 13 | 0 | 0 | ARYAN SRIVASTAVA | Active |
| 14 | 0 | 0 | ANUBHAV GANGWAR | Active |

**Table of Contents:**

|  |  |
| --- | --- |
| **Sr. No.** | Table of Contents |
| **1** | Introduction |
| **2** | Problem Statement & Objectives |
| **3** | Tools and Technologies Used |
| **4** | Dataset Description |
| **5** | Data Preprocessing |
| **6** | Exploratory Data Analysis (EDA) |
| **7** | Feature Engineering |
| **8** | Correlation & Multicollinearity |
| **9** | Model Building |
| **10** | Model Evaluation |
| **11** | Final Result Visualization |
| **12** | Insights and Conclusion |
| **13** | Future Work |
| **14** | References |

1. **Introduction** Retail businesses generate vast amounts of sales data every day. Predicting future sales accurately can help them optimize stock, staffing, and marketing efforts. Machine Learning (ML) techniques enable automated and intelligent forecasting based on historical trends. This project focuses on using these techniques to drive meaningful retail strategies.

1. **Problem Statement & Objectives** The objective of this project is to predict weekly retail sales using historical data. It includes cleaning, exploring, and modelling data to build a predictive system. We focus on data-driven insights to support retail decision-making.

Goals:

* + Understand key drivers of sales
  + Clean and preprocess data
  + Compare various ML algorithms
  + Choose the best performing model

# 3. Tools and Technologies Used

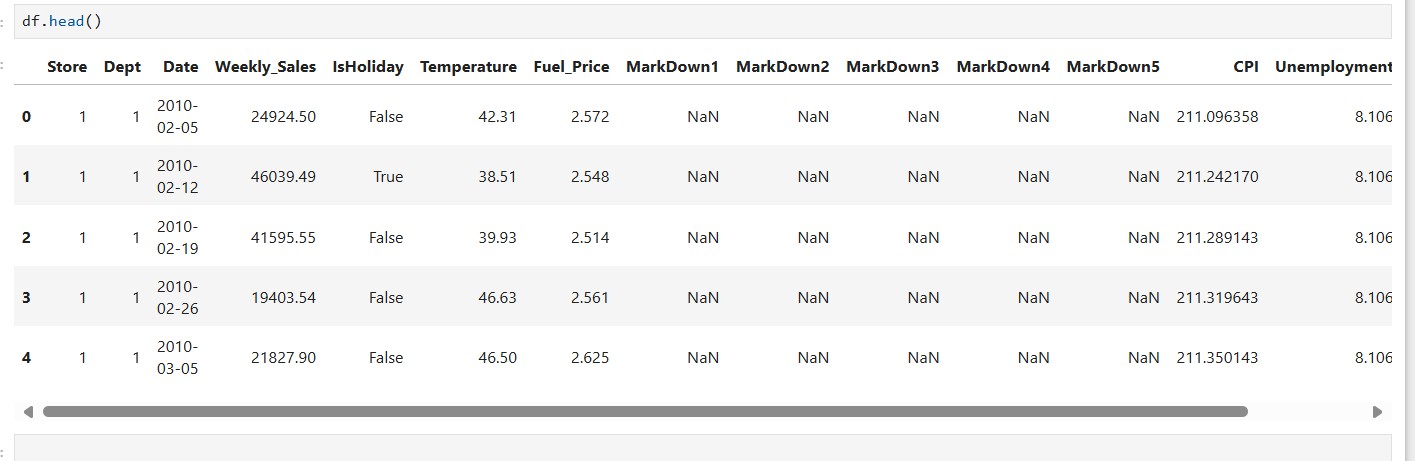
* Programming Language: Python 3.x
* IDE: Jupyter Notebook / VS Code
* Libraries: Pandas, NumPy, Seaborn, Scikit-learn, Matplotlib



# 4. Dataset Description

* **features.csv**: Store-level features like temperature, fuel price, promotions
* **sales.csv**: Weekly sales data by department and store
* **stores.csv**: Store type, size

The datasets collectively provide both internal sales metrics and external conditions like holidays or fuel prices that influence demand.

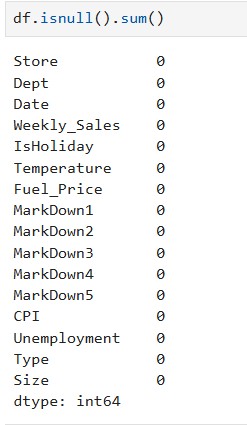
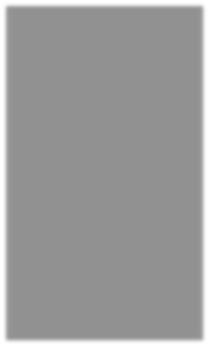
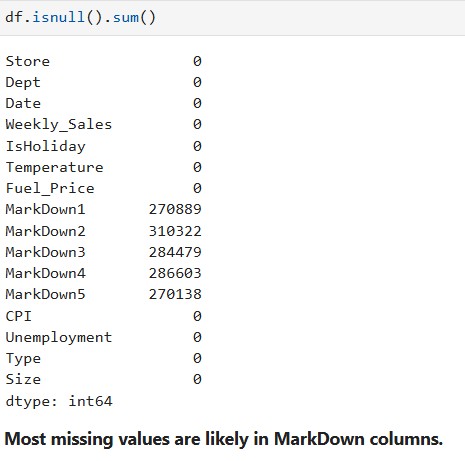
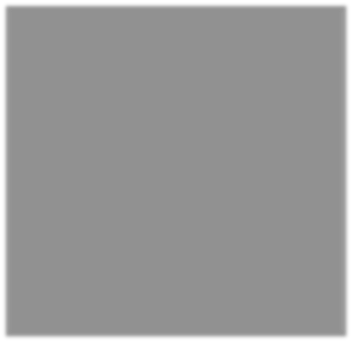


# 5. Data Preprocessing

* Converted 'Date' field into datetime format
* Handled missing values using mean/median or dropping
* Encoded categorical variables using label encoding
* Scaled features like temperature and fuel price

Data preprocessing improves the quality and compatibility of data for ML models, reducing chances of error or bias.

Before & After =

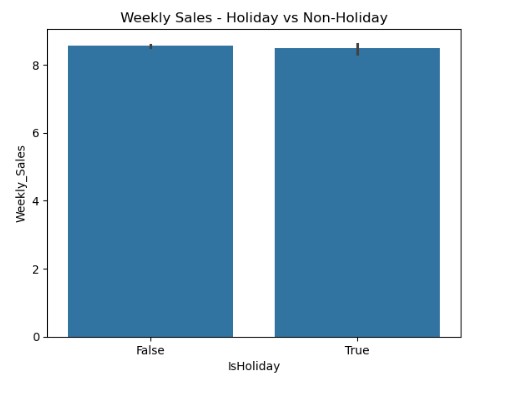


# 6. Exploratory Data Analysis (EDA)

* Sales trend comparison between stores
* Impact of holidays on weekly sales

EDA helps identify hidden patterns, anomalies, and trends through visualization and statistical summaries.

* Reveals seasonal patterns and anomalies in data.
* Helps visualize performance across store types.



# 7. Feature Engineering

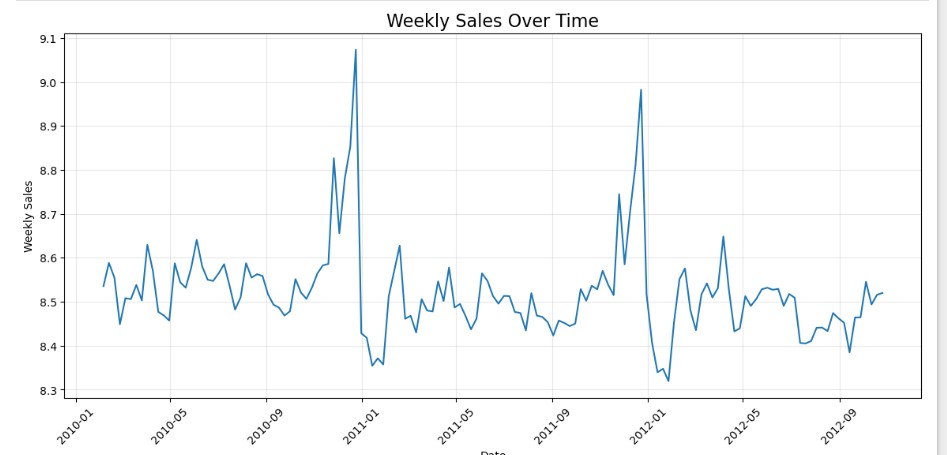
* Extracted Month, Year, Week from date
* Created binary flag for 'IsHoliday'

Feature engineering enhances the predictive power of the dataset by creating new informative features from existing ones.

* Extracted features improve model understanding of seasonality.
* Derived metrics like sales per area help normalize store comparisons.

**8**

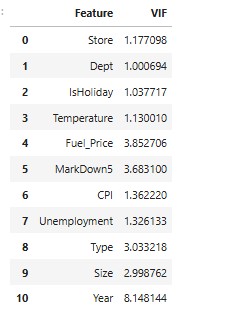
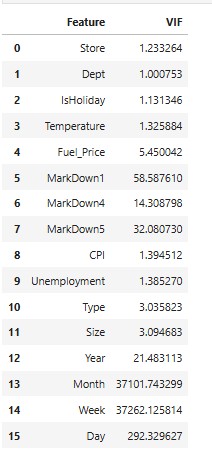
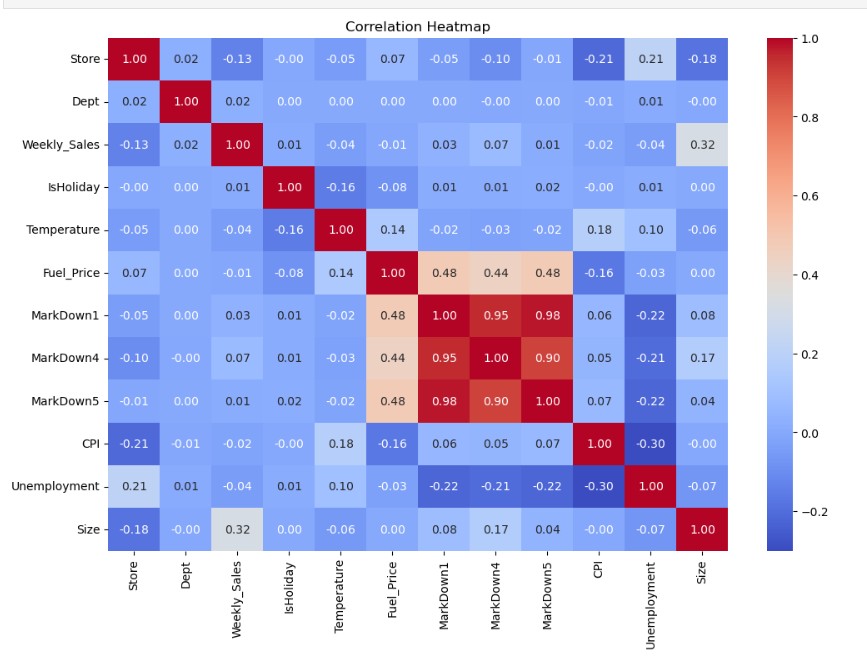
**. Correlation & Multicollinearity**



* Heatmap showing feature correlation
* Calculated Variance Inflation Factor (VIF)
* Removed highly correlated redundant features

This step helps reduce multicollinearity which can mislead model interpretation and performance.

* Reduces noise and improves model reliability.
* Helps focus on impactful variables.



1. **Model Building** Models Applied:
   * Random Forest Regressor

Different models were trained and tested to compare their accuracy and performance in sales forecasting.

* + Ensemble models often outperform basic models.

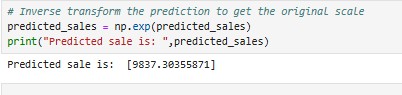
1. **Model Evaluation** Evaluation Metrics:
   * R-squared Score (R²)

Each metric provides a different perspective on how close predictions are to actual sales values.

* + R² close to 1 indicates a strong fit.

**12. Insights and Conclusion** Key Findings:

* Holidays and store types not significantly impact sales.
* Store size and promotions are strong predictors.
* Random Forest provided the most accurate results.
* Store Type A showed the highest average sales.
* Holiday periods show variable effects.
* Certain departments dominate in sales.
* Environmental features have mild correlations.
* ML can automate demand forecasting.
* Businesses can optimize staffing and marketing with forecast data. Sales Prediction=



# 13. Future Work

* Focus promotions on top departments during high seasons.
* Analyze holiday trends per department.
* Consider ML-based forecasting.
* Helps transition project to real-world deployment.
* Could support multi-store regional forecasting.

# 14. References

* Python official documentation – [https://docs.python.org](https://docs.python.org/)
* Scikit-learn official documentation – [https://scikit-learn.org](https://scikit-learn.org/)
* Pandas library – [https://pandas.pydata.org](https://pandas.pydata.org/)
* NumPy library – [https://numpy.org](https://numpy.org/)
* Matplotlib for plotting – [https://matplotlib.org](https://matplotlib.org/)
* Seaborn for data visualization – [https://seaborn.pydata.org](https://seaborn.pydata.org/)