A

SYNOPSIS

of

MINOR PROJECT

on

Data Mining and Analysis on Wal-Mart Data Set

तमसो मा ज्योतिर्गमय्

Darkness to Light

Submitted by

Harshita Singh Dulawat

22EGICS046

Project Guide:
Dr. Paras Kothari

Head of Department:
Dr. Mayank Patel

Geetanjali Institute of Technical Studies, Dabok , Udaipur (Raj.)
Department of Computer Science and Engineering
October,2023

Problem Statement: Data Mining and Analysis on Wal-Mart Data Set

Brief Description:

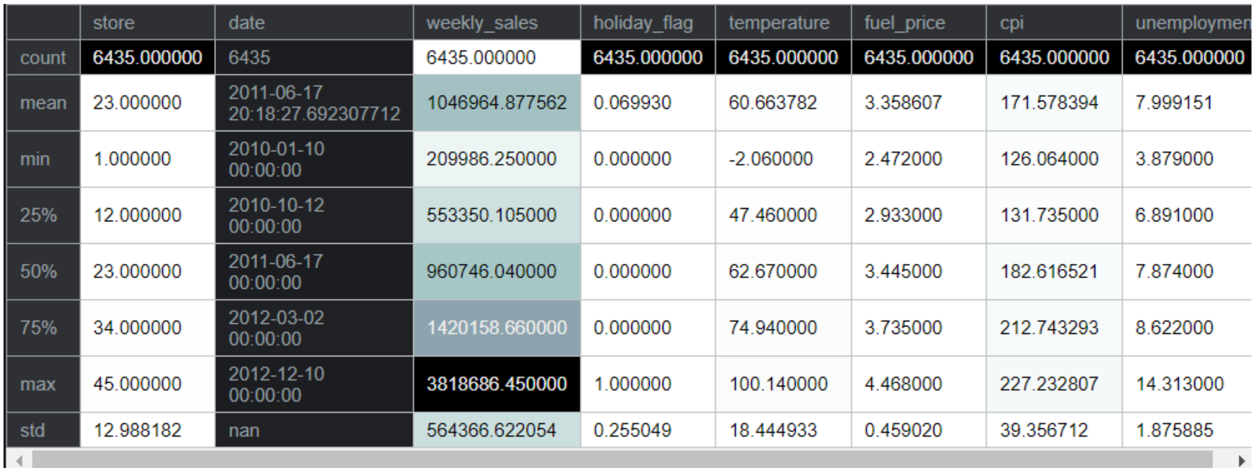
This project conducts an in-depth analysis of retail sales data, specifically focusing on Walmart store performance across different locations. Using Python for data exploration and visualization, I aim to uncover insights into sales trends, seasonal variations, and the influence of external factors like holidays, temperature, fuel prices, CPI, and unemployment rates. Through this analysis, I aim to provide actionable insights for retail management decision-making and enhance understanding of consumer behavior and market dynamics.

Summary of the Dataset: Store: Identifier for the retail store. Date: Date of sales record. Holiday_Flag: Indicator for holiday week (1) or non-holiday week (0). Temperature: Temperature in the region of the store. Fuel_Price: Fuel price in the region. CPI: Consumer Price Index.

Objective and Scope: The scope of the "Data Mining and Analysis on Wal-Mart Data Set" project includes collecting, preprocessing, and analyzing Walmart sales data to uncover sales trends, seasonal variations, and the influence of external factors like holidays, temperature, fuel prices, CPI, and unemployment rates. Using Python libraries such as Pandas, Matplotlib, scipy, calendar, and Seaborn for data manipulation and visualization, the project aims to conduct exploratory data analysis, correlation analysis, and predictive modeling. The primary objective is to provide actionable insights for retail management decision-making, enhance the understanding of consumer behavior and market dynamics, and develop reliable sales forecasts to aid inventory management and strategic planning.

Methodology:

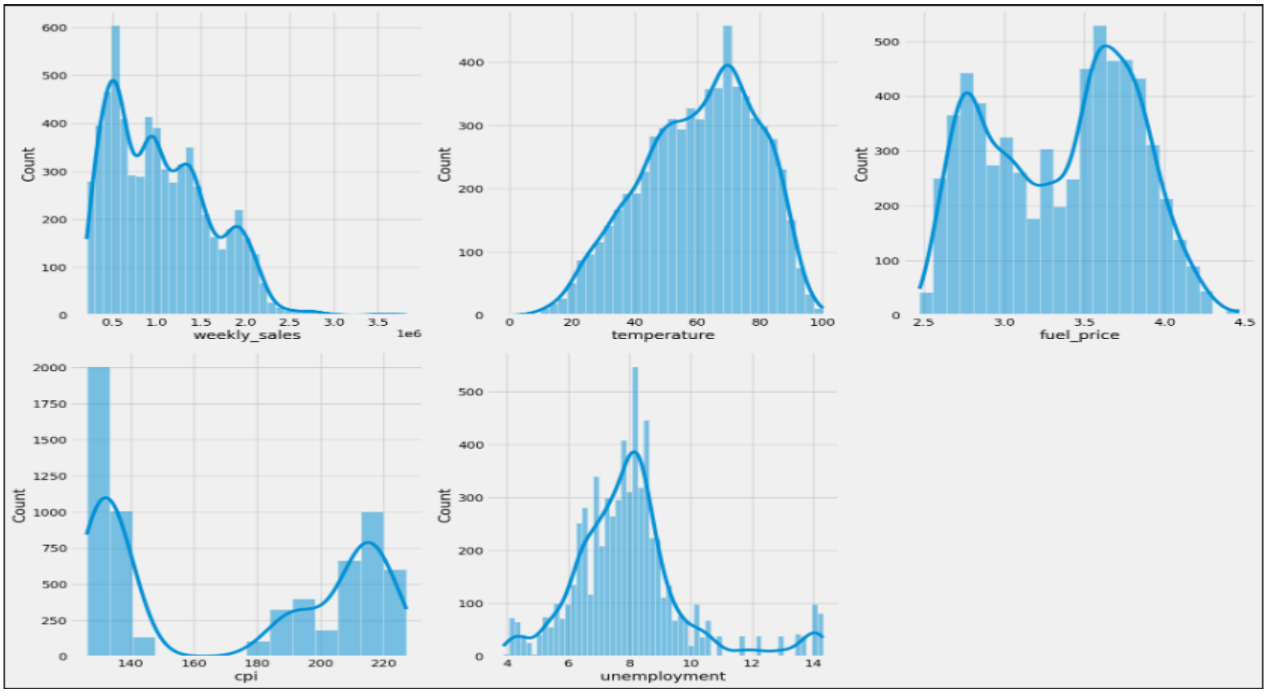
1. Data Collection and Preprocessing:
   - Load the Walmart sales dataset into a Python environment using Pandas.
   - Clean the dataset by handling missing values, removing duplicates, and correcting inconsistencies.
   - Ensure proper data types for each column and create new variables if necessary (e.g., extracting month and year from the Date column).



	store	date	weekly_sales	holiday_flag	temperature	fuel_price	cpi	unemploymen
count	6435.000000	6435	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
mean	23.000000	2011-06-17 20:18:27.692307712	1046964.877562	0.069930	60.663782	3.358607	171.578394	7.999151
min	1.000000	2010-01-10 00:00:00	209986.250000	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	12.000000	2010-10-12 00:00:00	553350.105000	0.000000	47.460000	2.933000	131.735000	6.891000
50%	23.000000	2011-06-17 00:00:00	960746.040000	0.000000	62.670000	3.445000	182.616521	7.874000
75%	34.000000	2012-03-02 00:00:00	1420158.660000	0.000000	74.940000	3.735000	212.743293	8.622000
max	45.000000	2012-12-10 00:00:00	3818686.450000	1.000000	100.140000	4.468000	227.232807	14.313000
std	12.988182	nan	564366.622054	0.255049	18.444933	0.459020	39.356712	1.875885

2. Exploratory Data Analysis (EDA):

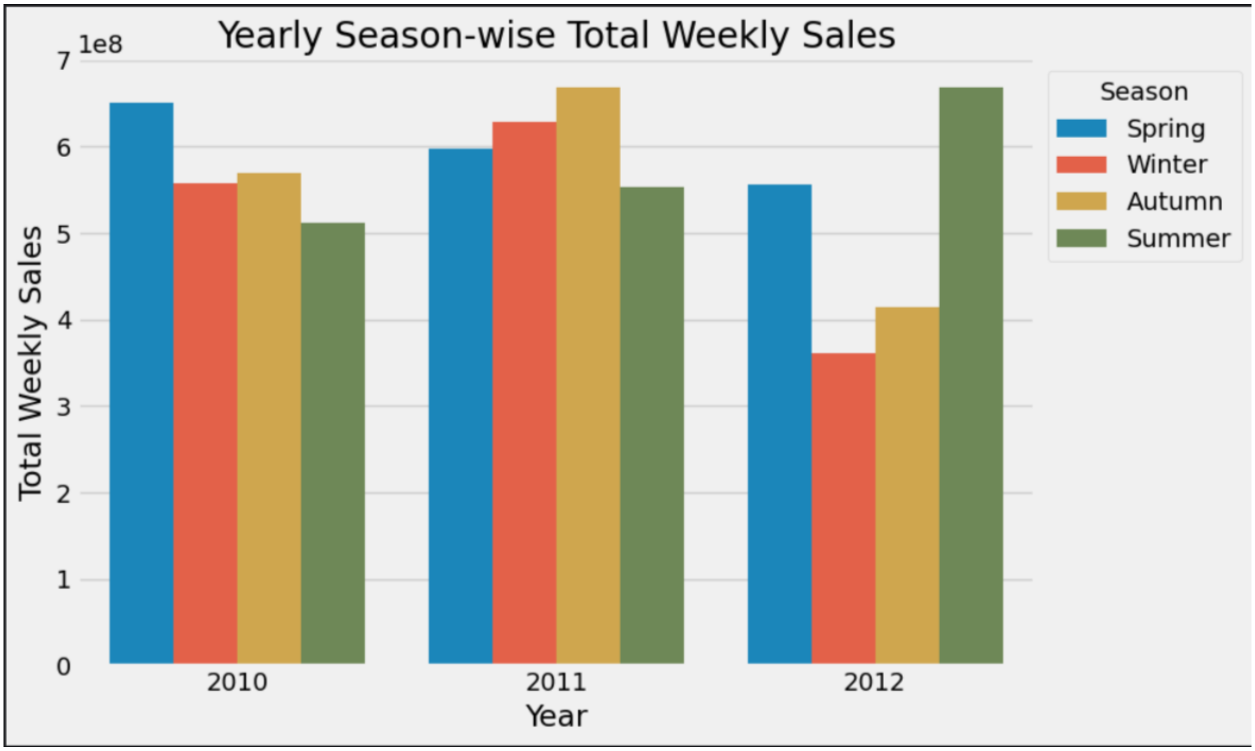
- Generate descriptive statistics for key variables (mean, median, standard deviation, etc.).
- Create visualizations such as histograms, box plots, and scatter plots using Matplotlib and Seaborn to visualize data distributions and identify patterns.



3. Time Series Analysis:

- Plot sales trends over time to identify long-term trends and seasonal variations.

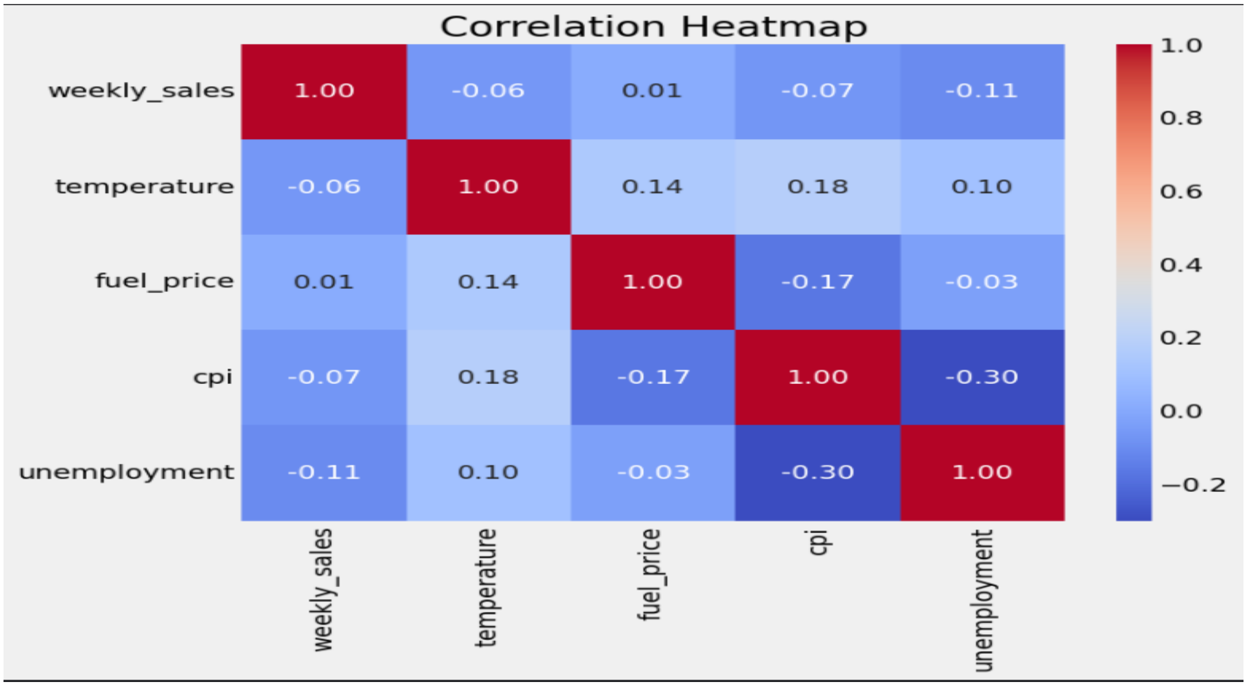
- Analyze the impact of holidays on sales by comparing holiday weeks with non-holiday weeks.



4. Correlation Analysis:

- Investigate relationships between sales and external factors like temperature, fuel prices, CPI, and unemployment rates.

- Use statistical techniques to quantify the strength and direction of these relationships.



5. Predictive Modeling:

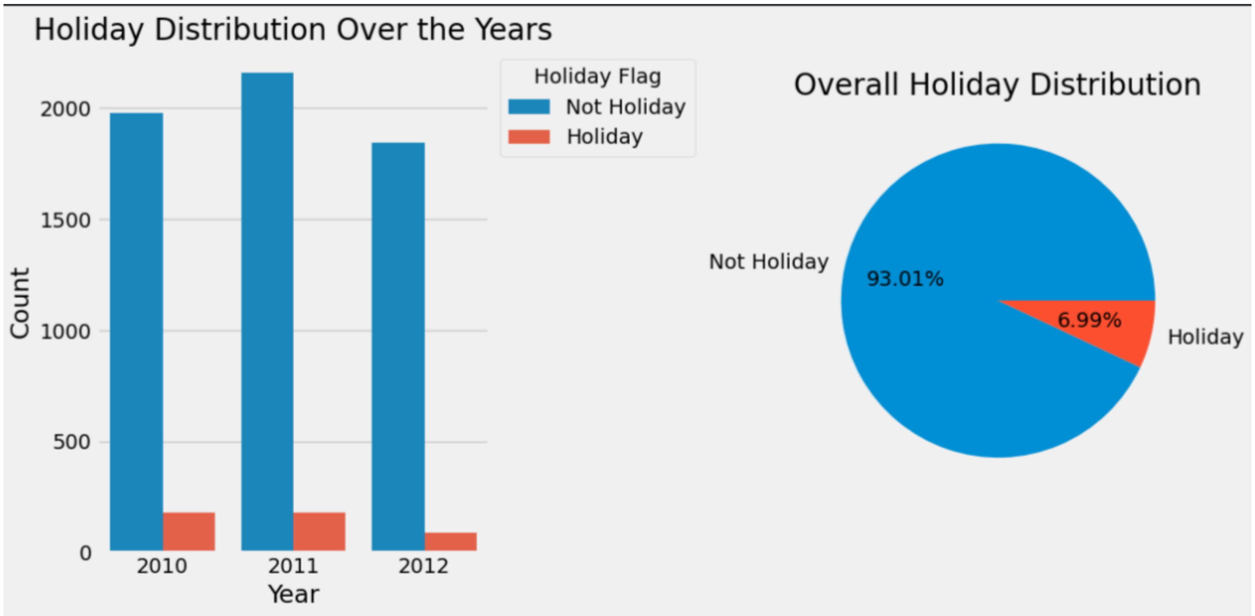
- Develop predictive models to forecast future sales based on historical data and external factors.

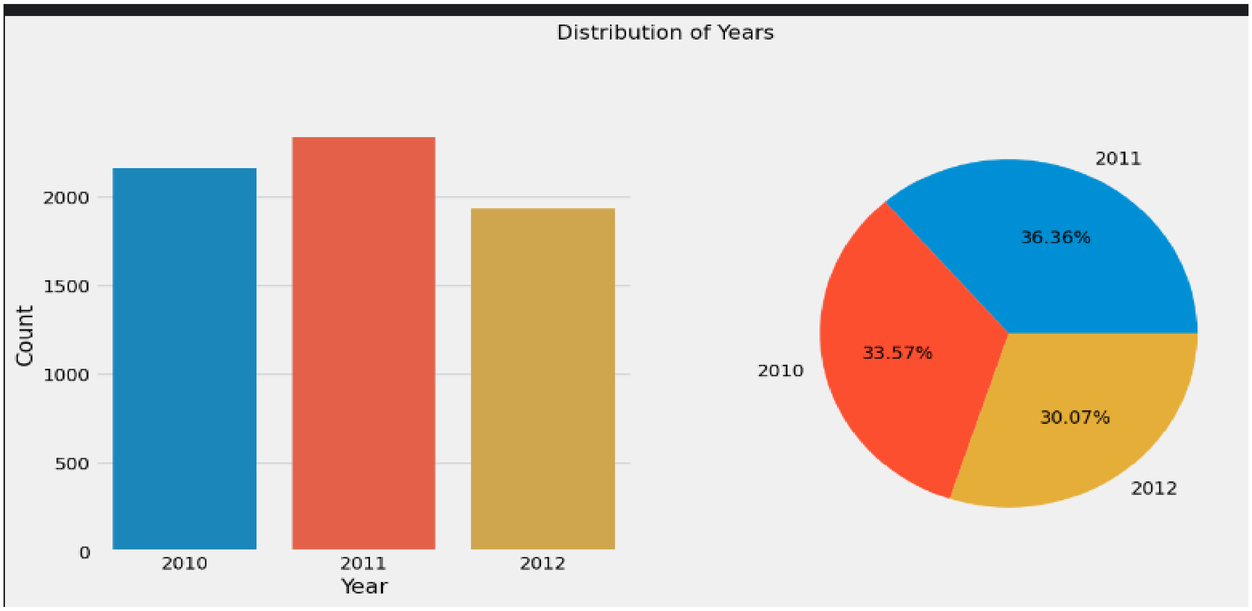
- Evaluate model performance using metrics such as RMSE or MAE and refine models for better accuracy.

6. Visualization:

- Create visualizations to effectively communicate findings, trends, and patterns.

- Use tools like Matplotlib and Seaborn to generate plots, charts, and graphs for the final report.


Holiday Distribution Over the Years
Holiday Flag
Not Holiday
Holiday
2000
1500
1000
500
0
Count
2010
2011
2012
Year
Overall Holiday Distribution
Not Holiday
93.01%
6.99%
Holiday


Distribution of Years
2000
1500
1000
500
0
Count
2010
2011
2012
Year
2011
36.36%
2010
33.57%
30.07%
2012

7. Reporting:
- Compile the analysis results into a comprehensive report.
- Provide actionable insights and recommendations for retail management decision-making based on the analysis findings.

Hardware and Software Requirements:

Hardware Requirements

- **Computer**: Modern quad-core processor (Intel i5/i7 or AMD Ryzen 5/7).
- **RAM**: Minimum 8GB, recommended 16GB+.
- **Storage**: At least 256GB SSD.
- **Graphics Card**: Dedicated GPU (optional but beneficial).
- **Operating System**: Windows 10/11, macOS, or Linux.

Software Requirements

- **Python Environment**: Python 3.6+, Jupyter Notebook/Lab.
- **Python Libraries**: Pandas, NumPy, Matplotlib, Seaborn, SciPy, scikit-learn (optional).
- **Development Tools**: PyCharm, VSCode, or similar IDE; Git for version control.
- **Additional Tools**: Cloud storage (AWS S3, Google Cloud, Azure) for large datasets, Tableau/Power BI for advanced visualization (optional).

Technologies:

The project utilizes Python as the primary programming language, with Jupyter

Notebook/Lab as the development environment for interactive coding. Key libraries include Pandas for data manipulation, NumPy for numerical operations, Matplotlib and Seaborn for data visualization, and SciPy for scientific computing. scikit-learn is used for predictive modeling if needed. Development tools such as PyCharm or VSCode facilitate code writing, with Git for version control. Data is managed using local SSD storage for fast access and optionally cloud storage solutions like AWS S3, Google Cloud Storage, or Azure Blob Storage for large datasets. Advanced visualization and reporting can be achieved using tools like Tableau or Power BI if required.

Testing Techniques:

1. **Unit Testing**:

- **Description**: Test individual functions and methods to ensure they work as expected.
- **Tools**: unittest, pytest.
- **Example**: Test functions for data cleaning, feature engineering, and calculation methods.

2. **Integration Testing**:

- **Description**: Test the interaction between different modules or components of the project.
- **Tools**: pytest, unittest.
- **Example**: Ensure that the data preprocessing pipeline correctly processes the raw data and feeds it into the analysis and visualization modules.

3. **Regression Testing**:

- **Description**: Ensure that new changes or updates do not negatively impact existing functionality.
- **Tools**: pytest, unittest.

- **Example**: After adding new features or making modifications, run existing tests to confirm that everything still works as intended.

4. **Data Validation Testing**:

- **Description**: Verify the accuracy and quality of the data after each processing step.
- **Tools**: Custom scripts or data validation libraries.
- **Example**: Check for missing values, outliers, and data consistency after data cleaning and transformation.

Project Contribution:

The project on sales analysis at Walmart significantly contributes by providing a comprehensive examination of sales performance across different store locations using advanced data exploration and visualization techniques in Python. By meticulously analyzing sales trends, seasonal variations, and the impact of external factors such as holidays, temperature, fuel prices, CPI, and unemployment rates, the project uncovers critical insights that aid in understanding consumer behavior and market dynamics. These insights facilitate data-driven decision-making for retail management, enabling more effective inventory management, targeted marketing strategies, and improved overall operational efficiency. Additionally, the project enhances the ability to forecast future sales trends, ultimately contributing to better strategic planning and resource allocation.