DATA ANALYSIS USING PYTHON

FINAL PROJECT

SUBMITTED TO : SIR MOOSA

Data Set:

First of all, we discuss data set. The Data set is about **Walmart Stores.** The data set contains6435 rows and 8 columns.
We read the data set by using df=pd.read_csv("The path of file")

The column names are

- Store
- Date
- Weekly_Sales
- Holiday_Flag
- Temperature
- Fuel_Price
- CPI (Consumer Price Index)
- Unemployment

Task A: Data Cleaning

- **Handle Missing Values and Duplicate Rows**

There is no missing Values and no Duplicate Rows that are found in Walmart Stores data set.

- **Fix Inconsistencies**

There is no inconsistencies tha are found in data set. So, We can convert the date column into Date format

Task B: Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the process of looking at data to understand its main features, patterns, and relationships using graphs and statistics. It helps you find out what the data is telling you before doing deeper analysis.

- **Revies first rows and explain Data types**

  We can use the methods of
  df.head() and df.dtypes

- **Data Distribution: Visualize the distribution of numerical columns (e.g., histograms or boxplots).**

We can make histogram or box blot of numerical coumns like Weekly Sales, CPI etc.

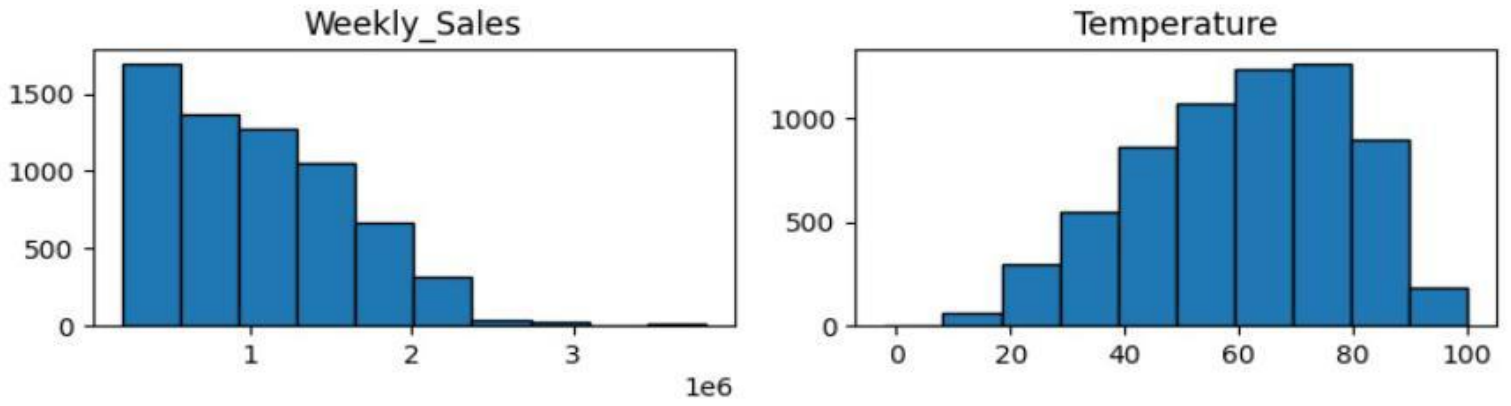
Histograms

A **histogram** is a type of bar graph that shows how often different ranges of values appear in a dataset. It helps you see the distribution and spread of the data, like whether it's mostly low, high, or centered.

• Histograms of Weekly_Sales and Temperature

•**Weekly sales** Represents ranges of weekly sales amounts (in dollars). Represents the number of times (frequency) weekly sales fell into each range (bin). For example: If the first bar reaches 1,600, that means about 1,600 weeks had sales between 0and 500,000.

•**Temperature Histogram** X-axis (bottom) shows temperature ranges (like 0–20, 20–40, 40–60, etc.). Y-axis (side) shows how many times the temperature fell into each range.



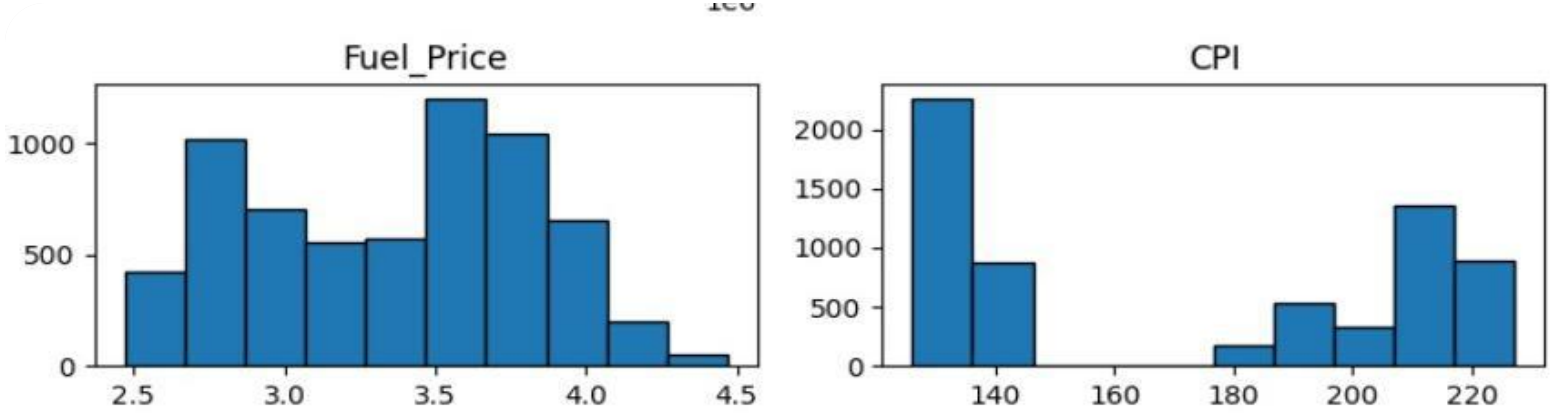
• Historgams of Fuel_Price and CPI

•Fuel Price :

•X-axis shows fuel price ranges (like 2.5to4.5 per gallon). Y-axis shows how many times fuel prices were in those ranges.

•CPI :

X-axis shows CPI values (Consumer Price Index), which measures the cost of goods over time. Y-axis shows how many times each CPI range occurred. This shape is called bimodal (it has two peaks)

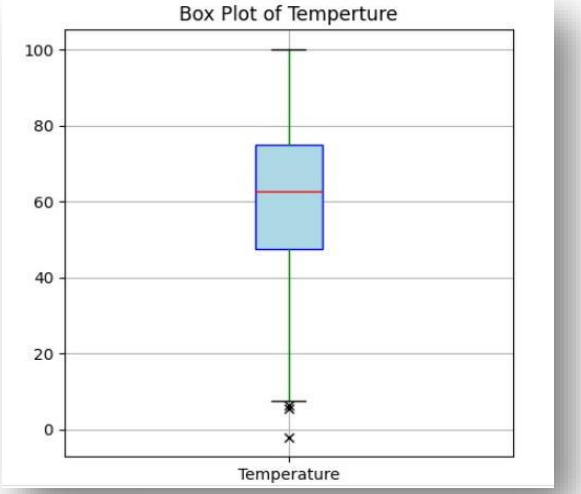


- **Box plot**

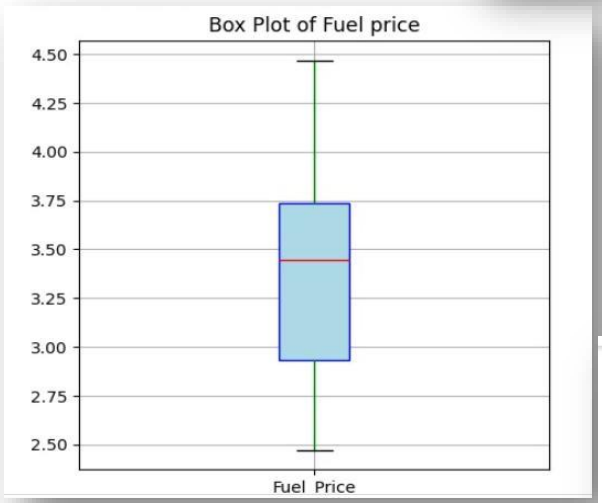
A boxplot is a graphical summary of a dataset that displays its median, quartiles, and potential outliers. It shows the distribution through a box (interquartile range) and "whiskers" that extend to the minimum and maximum values within a defined range.

Weekly Sales: The median weekly sales appear to be around 2.0, with the majority of the data falling between approximately 1.0 and 3.0. There are several outliers above 3.0, indicating weeks with unusually high sales.

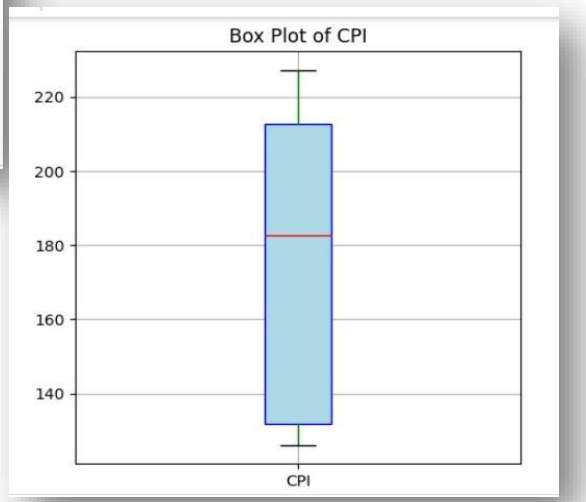
Temperature: The median temperature is roughly 70, with most of the data ranging from about 40 to 90. There's one outlier near 0.



Fuel Price: The median fuel price is around 3.50, with the data spread between approximately 2.75 and 4.25.



CPI (Consumer Price Index): The median CPI is about 190, and the data ranges from 150 to 220.



- **Relationships Between Variables: Explore correlations or relationships between variables using plots.**

We can explore correlation between between variables by using Heat map

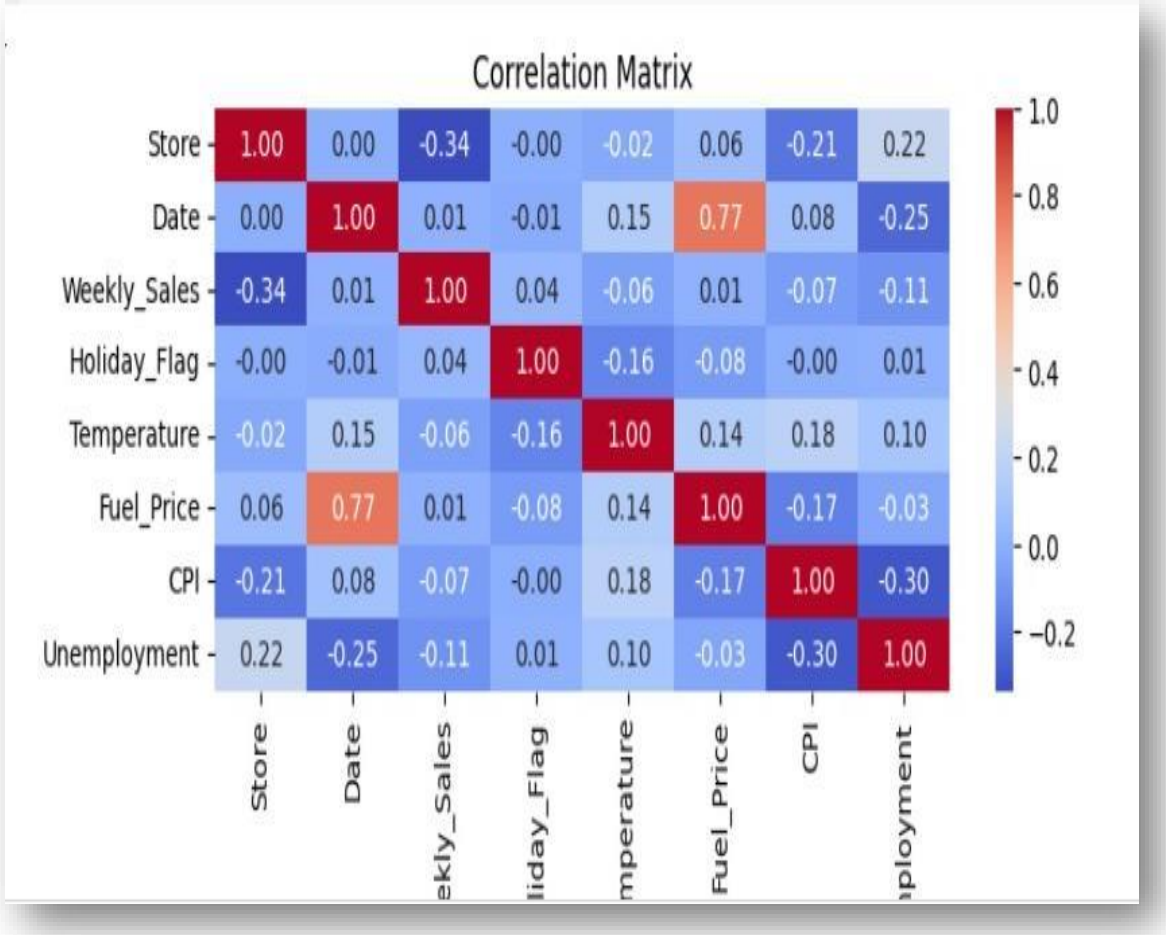
Store vs. Store: The correlation is 1.00 because it's comparing the same variable to itself.

***Date vs. Fuel Price** : The correlation is 0.77, suggesting a positive correlation between the date and fuel price. Over the period analyzed, fuel prices may have generally increased.

Weekly Sales vs. Store: The correlation is -0.34, suggesting a weak negative correlation between weekly sales and the store.

CPI vs. Unemployment: The correlation is -0.30, indicating a weak negative correlation between CPI and unemployment.

Other Correlations: The matrix provides a comprehensive view of the relationships between all pairs of variables, allowing for the identification of potential dependencies and patterns



Task C: Data Preprocessing for Modeling

- **Feature Engineering: Create new features if relevant**

Feature engineering is the process of creating, transforming, or selecting variables (features) to improve a machine learning model's performance. It involves techniques like encoding, scaling, and deriving new features from raw data.

- **Scaling/Normalization: Scale or normalize numerical data if necessary.**

Normalization is a data preprocessing technique that scales features to a common range, typically [0, 1], without distorting differences in the ranges of values. It helps improve the performance and training stability of machine learning models.

.

Task D: Data Visualization and Dashboarding

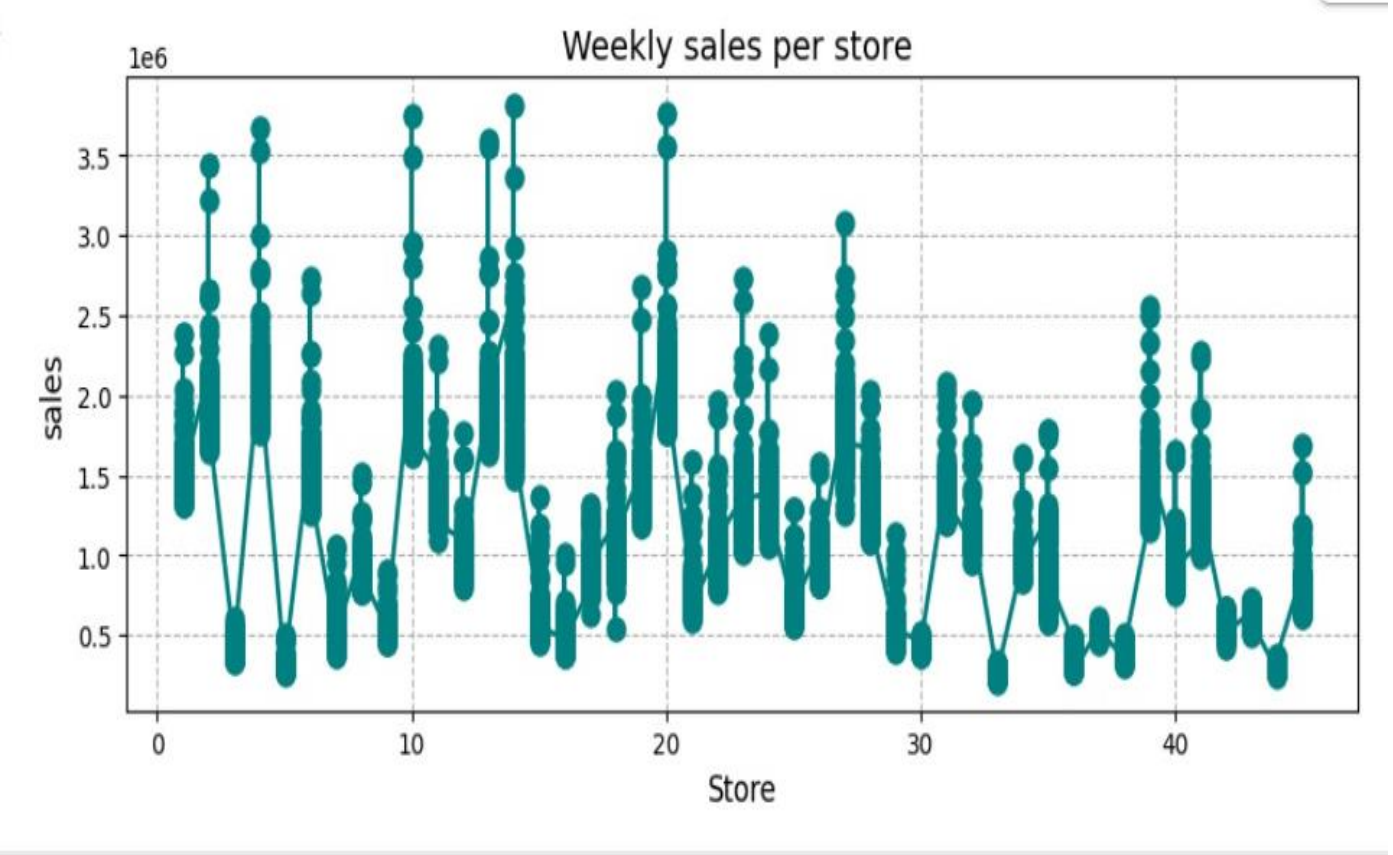
- **Static Visualizations using Matplotlib and Seaborn**
  - **Matplotlib** is a Python library used for creating static, animated, and interactive visualizations like line plots, bar charts, and histograms.

Seaborn is built on top of Matplotlib and provides a higher-level interface for making attractive and informative statistical graphics.

We can use matplotlib to make Scatter plot

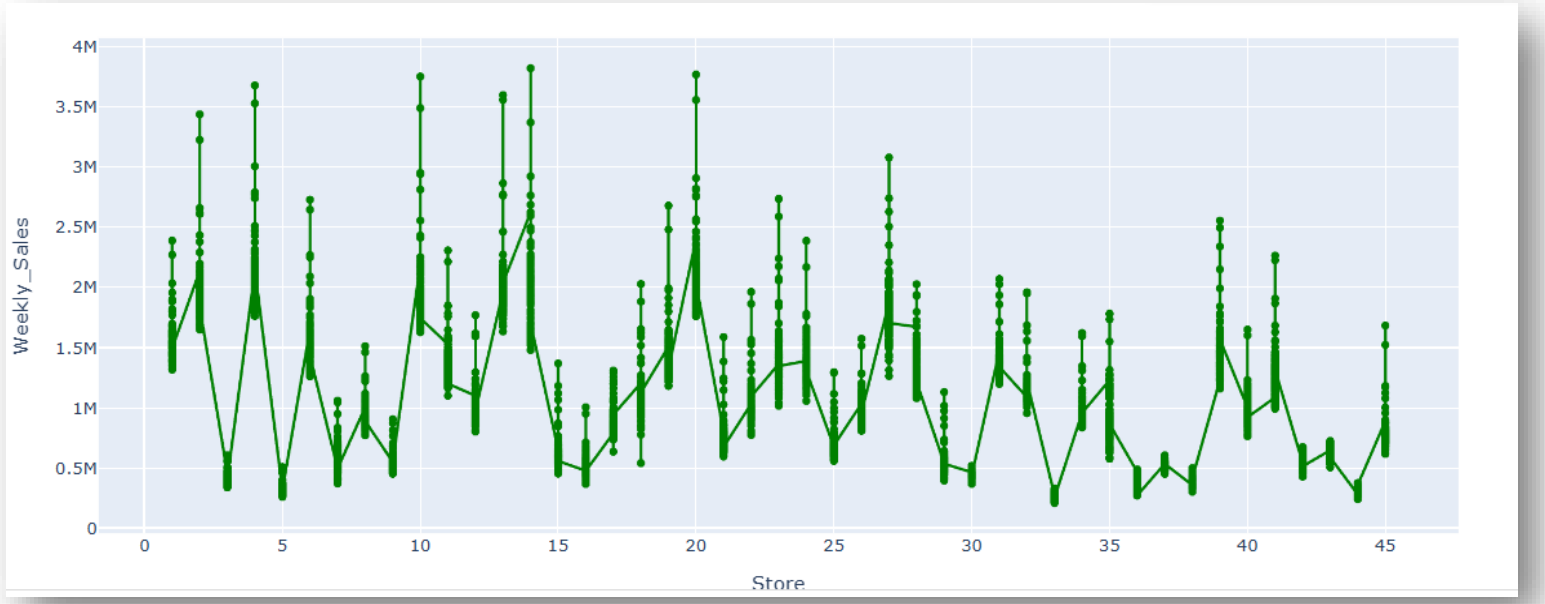
Explanation

The diagram shows weekly sales per store. The vertical axis represents sales in millions, ranging from 0 to 3.5 million. The horizontal axis represents the store number, ranging from 0 to 40. Each vertical line represents the weekly sales for a particular store. The length of the line corresponds to the sales amount, with longer lines indicating higher sales. The connected blue line appears to show the trend of sales across different stores. It shows that some stores have significantly higher sales than others and that there is considerable variation in sales performance among the stores. It is also noticeable that there are a few stores with very low sales, indicated by the short lines.



Plotly:

Plotly is a Python library for creating interactive, web-based visualizations like line charts, scatter plots, and 3D graphs. It supports rich interactivity, making it ideal for data exploration and dashboard development.



Dash Board:

A dashboard is a visual interface that displays key data and metrics in an organized, easy-to-understand format. It helps users monitor performance, identify trends, and make data-driven decisions in real time. Dashboards often include charts, graphs, and tables to summarize complex data. They are commonly used in business intelligence, analytics, and web applications.

THANK YOU