INFOSYS SPRINGBOARD INTERNSHIP

Data Vista: Sales Data Analysis And Visualization

Topic : Big Mart Sales Data Analysis

Name : N. Karthik

Email id : karthik0505n@gmail.com

KEY LEARNINGS

WHAT IS DATA SCIENCE ?

Data Science is a interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

WHAT IS AI ?

Artificial Intelligence is the creation of computer systems capable of performing tasks that typically require human intelligence.

AI stimulates human intelligence in machines for tasks like learning and problem solving.

WHAT IS MACHINE LEARNING ?

Machine Learning is teaching computers to learn patterns from data and make predictions or decisions without explicit programming. It works well with structured data and smaller datasets

WHAT IS DEEP LEARNING ?

Deep Learning is a more advanced type of ML that uses neural networks to process large amounts of raw data. It’s great for tasks like image recognition, natural language processing, and speech translation.

KEY LEARNINGS

WHAT IS RNN ? TYPES ? ISSUES IN IT.

A Recurrent Neural Network (RNN) is a type of neural network designed for sequential data, where the output from previous steps is fed back into the model to influence future steps

Types :
One-One, One- Many, Many-One, Many-Many

Issues :
Short Memory, While training loop more possibility of chances for data vanishing.

LSTM (Long Short Term Memory):

Built upon RNN, To solve all the issues with RNN. LSTM was built.

Memory Larger, Process and store large sequence data therefore model accuracy increases

LLM (Large Language Model):

Pretrained model on vast amount of data. GPT,BERT are the examples. It uses Encoder and Decoder architecture, Transformer Architecture, Self Attention mechanisim.

LIBRARIES LEARNED DURING THIS INTERNSHIP :

1. Data Operations – NUMPY

   Learned how to perform Linear Regression, Logistic Regression, Euclidean distance, PCA, K-Means, Navie Bayes, SVM.

2. Data Processing – PANDAS

   Used for manipulating, transforming and analysing the data.

3. Data Visualization – MATPLOTLIB

   Used for visualizing the data for better understanding.

(line plot, scatter plot, bar plot, histogram, pie chart, combination of line and bar plot, line styles, logarithmic scale, plotting with annotations).

LIBRARIES LEARNED DURING THIS INTERNSHIP :

4. Data Visualization – SEABORN

Used for creating visually appealing, statistically focused plots with minimal code.

5. Deep learning Image Processing – OPENCV

Digitizing and Executing images with the help of vision framework/libraries.

6. Deep Learning Video Processing – OPENCV

Digitizing and Executing videos with the help of vision framework/libraries.

GITHUB :

I have learned how to effectively use GitHub for pushing documents, editing files, and updating repositories. I also gained experience in creating and managing README files to provide essential project information. This knowledge has helped streamline collaborative development and version management.

PROJECT WORK : BIG MART SALES DATA ANALYSIS

The project aims to analyze sales data from Big Mart to uncover key insights, extract meaningful information, and interpret findings to guide stakeholders in making data-driven decisions.

Problem Statement :

Big Mart is a chain of retail outlets aiming to maximize its sales and profitability. The management has collected detailed data from 2013 on 1,559 products sold across 10 different stores in various cities. The dataset includes attributes related to the products, the outlets, and their sales performance.

The key objectives of this project are:

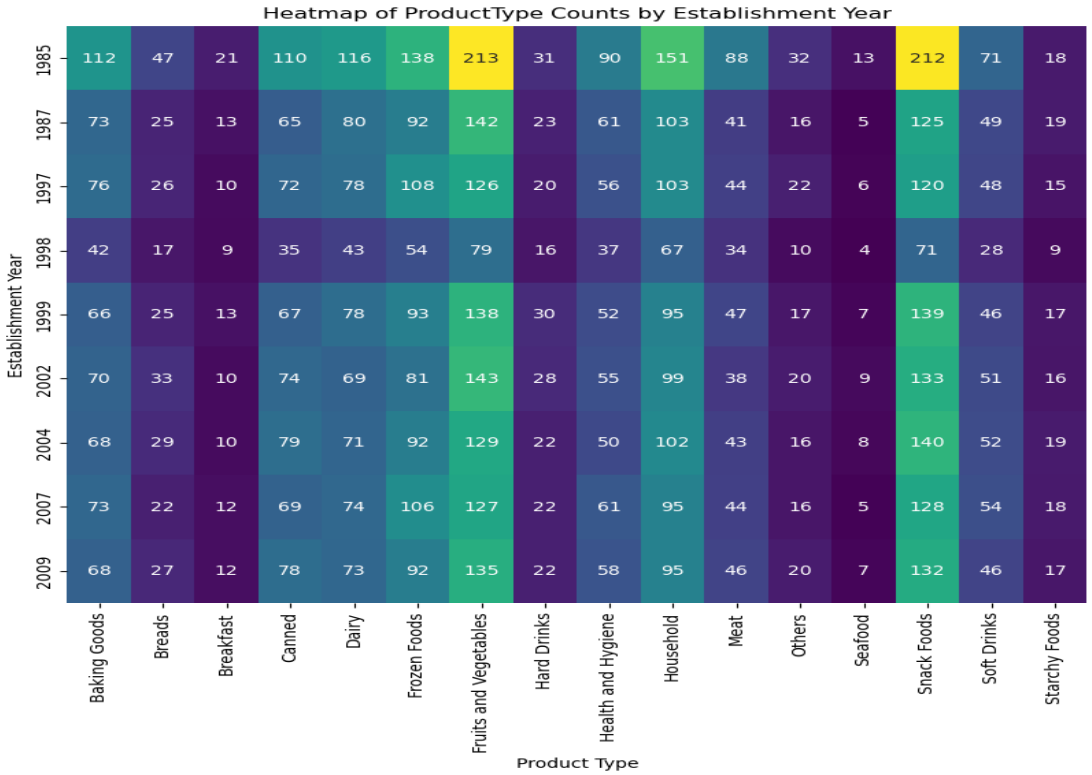
1. To **analyze sales trends and patterns** across different products and stores.
2. To **build a predictive model** that estimates the outlet sales based on its characteristics and store attributes.
3. To **Cluster the Outlet** based on 'Weight', 'ProductVisibility', 'MRP'.
4. To **identify key factors** influencing sales performance to assist in strategic decision-making.

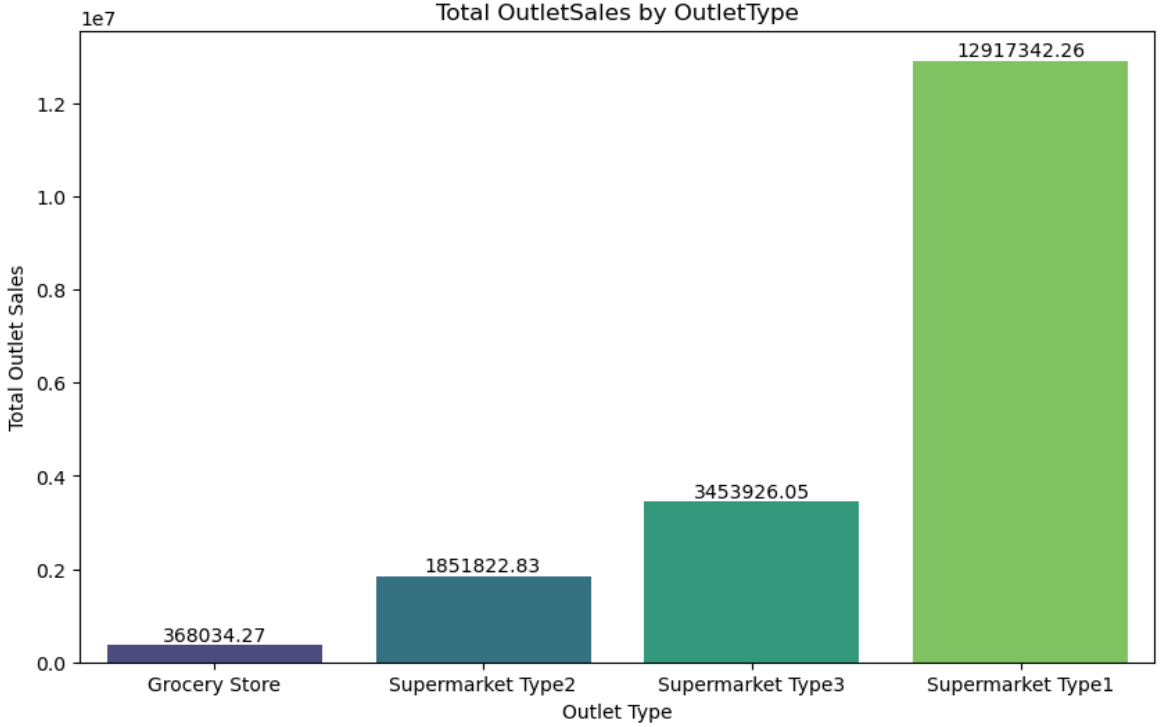
DATASET :

Out[38]:

	ProductID	Weight	FatContent	ProductVisibility	ProductType	MRP	OutletID	EstablishmentYear	OutletSize	LocationType	OutletType	OutletSales
0	FDA15	9.300	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
1	DRC01	5.920	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
2	FDN15	17.500	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
3	FDX07	19.200	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	Grocery Store	732.3800
4	NCD19	8.930	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
5	FDP36	10.395	Regular	0.000000	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
6	FDO10	13.650	Regular	0.012741	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
7	FDP10	NaN	Low Fat	0.127470	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636
8	FDH17	16.200	Regular	0.016687	Frozen Foods	96.9726	OUT045	2002	NaN	Tier 2	Supermarket Type1	1076.5986
9	FDU28	19.200	Regular	0.094450	Frozen Foods	187.8214	OUT017	2007	NaN	Tier 2	Supermarket Type1	4710.5350

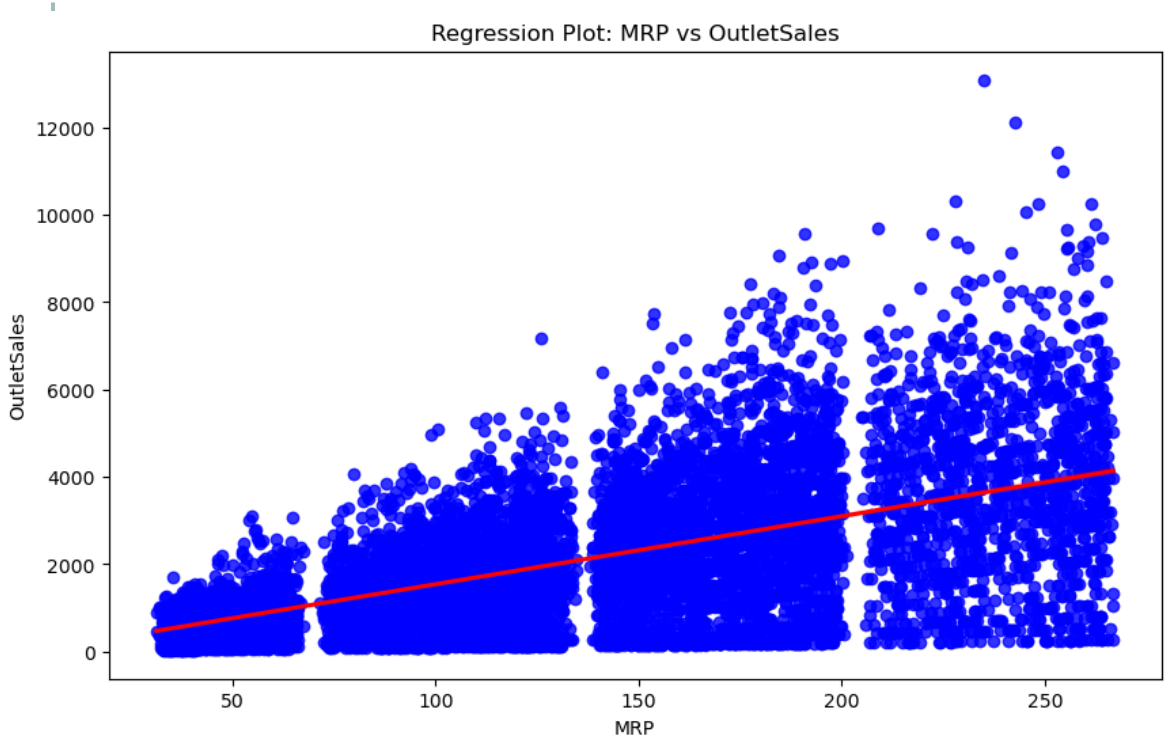
Exploratory Data Analysis :



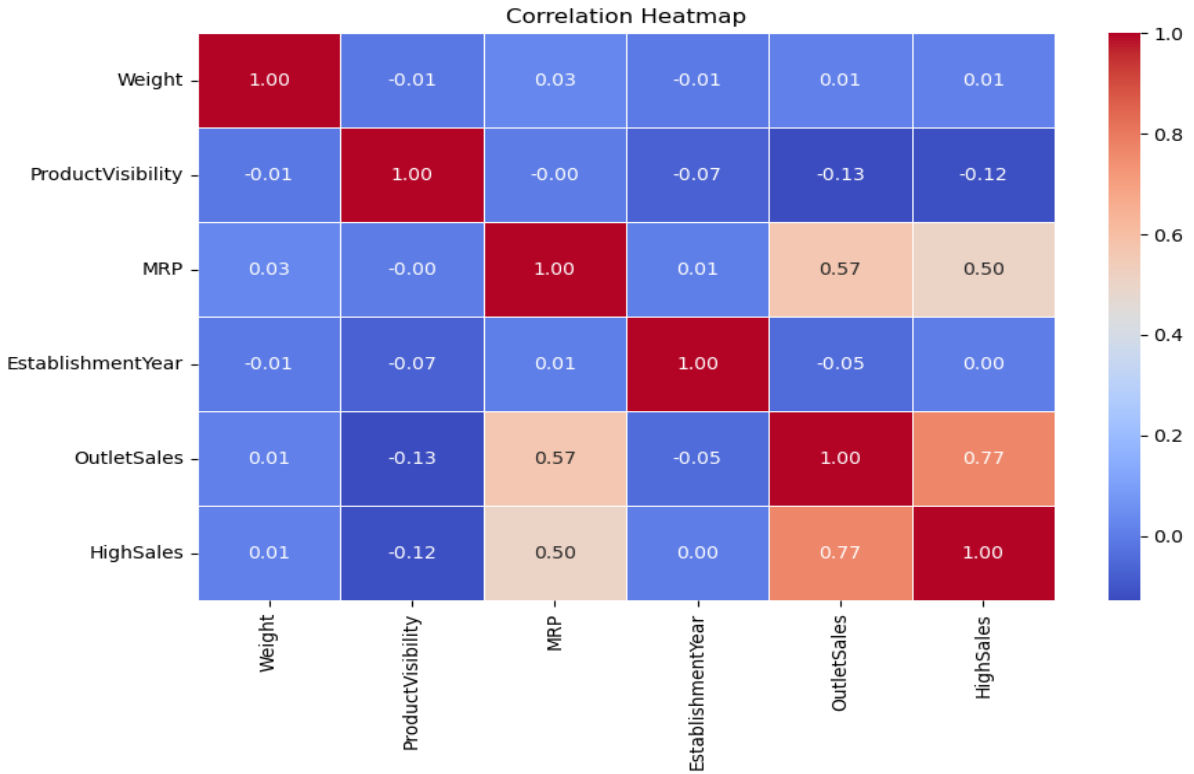


This heatmap shows the count of each product type according to the establishment year. ***Where Fruits and Vegetables have highest count***

This figure shows the Total OutletSales By each of the outletType, Where ***Supermarket Type1 has the Highest outlet sales***.

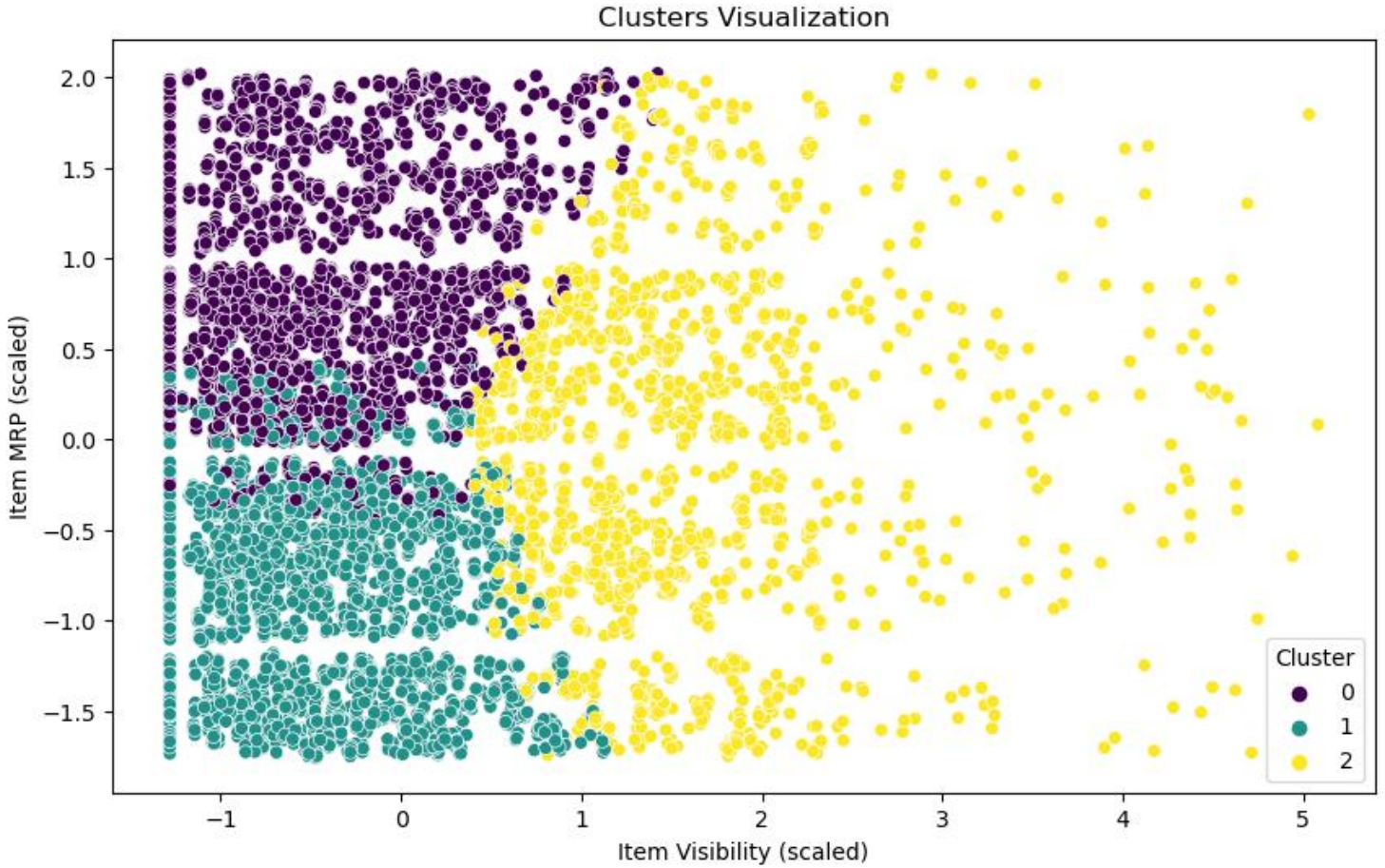


This Regression plot shows that if ***MRP*** increases then the ***OutletSales*** also increasing.



This Correlation Heatmap shows the correlation between the each variables. Where the ***MRP & OutletSales*** have (0.57), Which is the positive Correlation and MRP & HighSales have (0.50).

K Means Clustering :



Cluster Centers:

	Weight	ProductVisibility	MRP
0	13.880908	0.045785	195.263207
1	11.823541	0.043246	89.819905
2	12.888935	0.142790	136.120049

Fitted Line Equation :

```
Fitted Line Equation:
y = -52833.03 + -0.53*Weight + -265.23*ProductVisibility + 15.57*MRP + 30.72*EstablishmentYear + -9911.63*OutletSize_Mediu
m + -9948.22*OutletSize_Small + -755.73*OutletSize_Unknown + -165.09*LocationType_Tier 2 + -9610.09*LocationType_Tier 3 +
1520.79*OutletType_Supermarket Type1 + 10449.34*OutletType_Supermarket Type2 + 12913.75*OutletType_Supermarket Type3
```

Conclusion :

- **Cluster 0**: Represents products with an average weight of **13.88**, a visibility percentage of **4.58%**, and a higher MRP of **195.26**. Likely premium or high-end products.

- **Cluster 1**: Represents lighter products with a weight of **11.82**, a visibility percentage of **4.32%**, and the lowest MRP of **89.82**. Likely basic or economical items.

- **Cluster 2**: Represents moderately weighted products with a weight of **12.88**, a higher visibility percentage of **14.28%**, and a mid-range MRP of **136.12**. Likely popular or moderately priced products.

- These Clusters provide actionable insights for pricing, marketing, and store layout strategies.

- The predictive model can help forecast sales and optimize inventory across stores.

THANK YOU…