Data Science: Big Mart Sales Prediction Using Machine Learning With Python

Team Name :- DIGITAL MART DJ'S

Team Members

- Harsh Vardhan Mishra - 2201220100067
- Abhinav Verma - 2201220100005
- Gyanendra Verma - 2201220100064
- Anurag Pandey - 2201220100031
- Abhishek Dube - 2201220100006

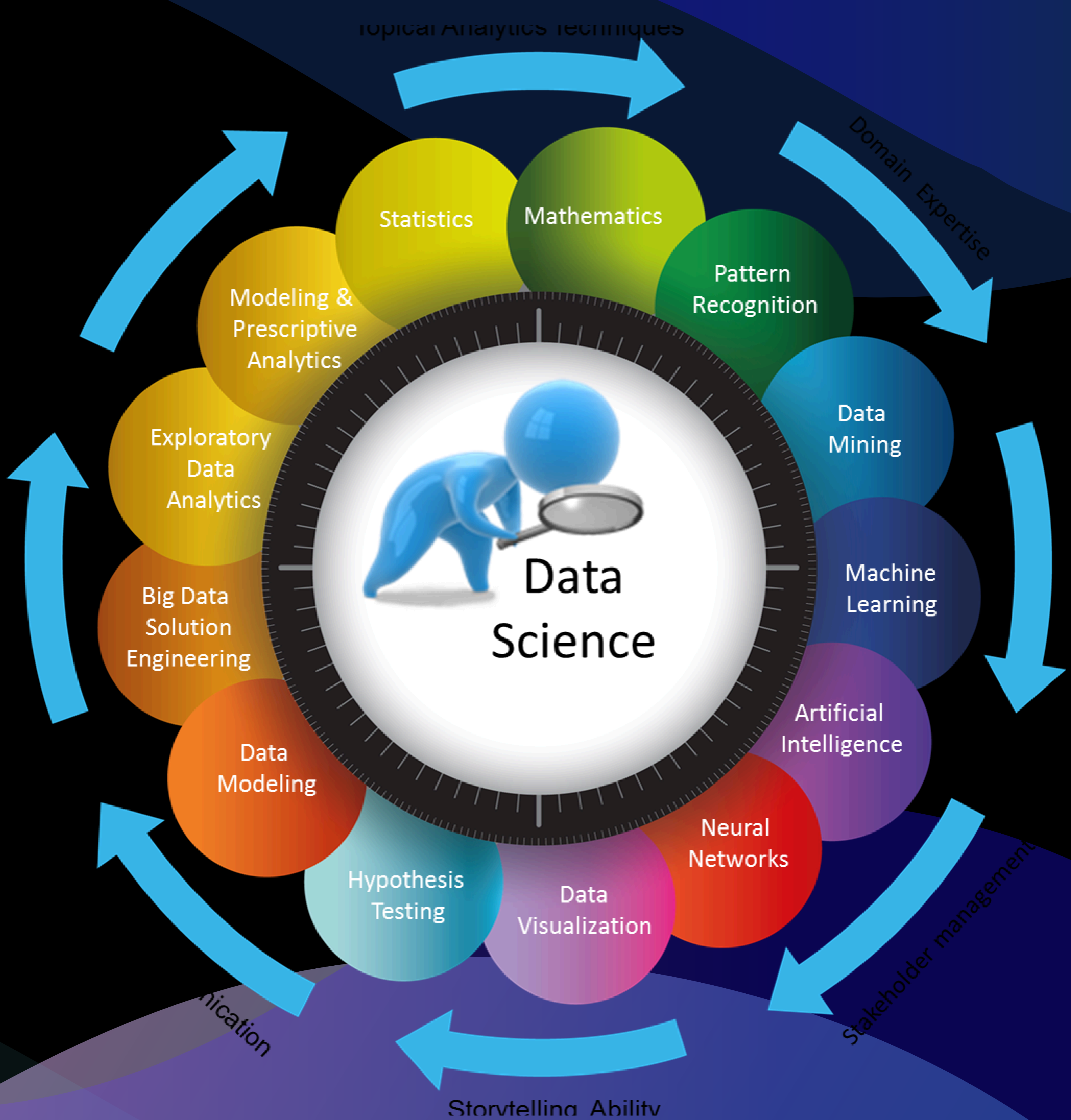
Data science is the transformation of data into actionable knowledge.

CONTENTS

- Introduction
- Data Analytics
- Machine Learning
- Key Algorithms in Machine Learning
- Big Mart Sales Prediction
- Conclusion

INTRODUCTION

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



Data Science Project

Problem Definition

Data Collection

Data Cleaning

Exploratory Data Analysis (EDA)

Data Modeling

Data Visualization

Introduction to Machine Learning

Machine Learning (ML) is a subset of artificial intelligence that focuses on building systems that learn from data, identify patterns, and make decisions with minimal human intervention.

Machine Learning Process

Problem Definition: Clearly define the problem you want to solve.

Data Collection: Gather relevant data from various sources.

Data Preprocessing: Cleaning and transforming data for model training.

Types of Machine Learning

Supervised Learning: Learning from labeled data to make predictions.

Unsupervised Learning: Finding hidden patterns in unlabeled data.

Reinforcement Learning: Learning through trial and error to maximize cumulative reward.

Key Algorithms in Machine Learning

Supervised Learning

Linear Regression: Used for predicting continuous values.

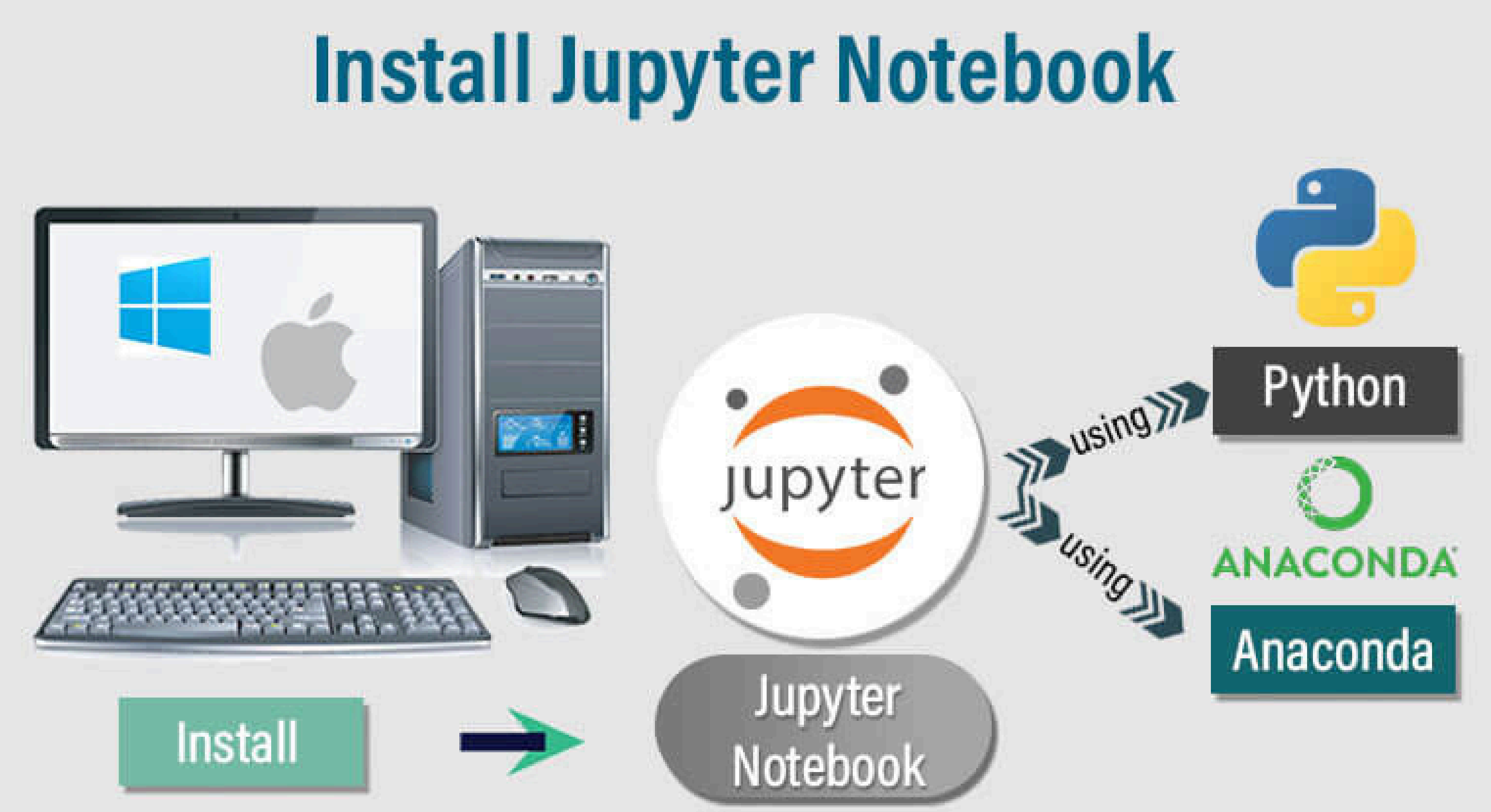
Decision Trees: A flowchart-like structure for classification and regression.

Support Vector Machine (SVM): A classification technique that finds the hyperplane maximizing the margin between classes.

Jupyter Notebook Installation:-

Steps :-

- Install Python
- Install Jupyter using pip (pip install notebook)
- Launch via terminal/command prompt.



Python Basics : _

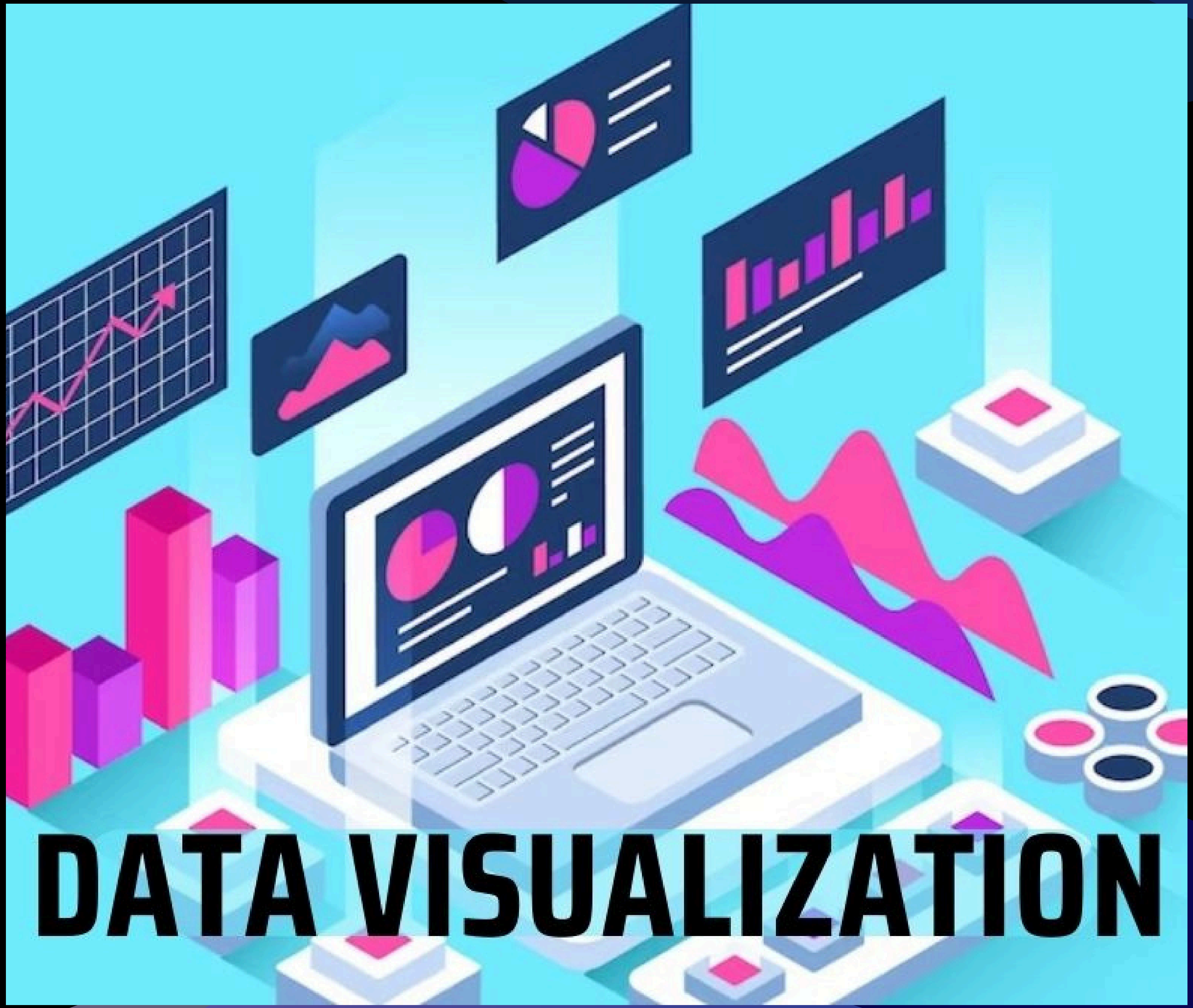
- **Operators**: Arithmetic, comparison, logical operators.
- **Data Types**: Integer, float, string, list, tuple, dictionary.
- **Conditional Statements & Loops**: If-else, for loop, while loop.
- **Functions**: Define reusable blocks of code with def keyword.

Introduction To Data Visualisation :-

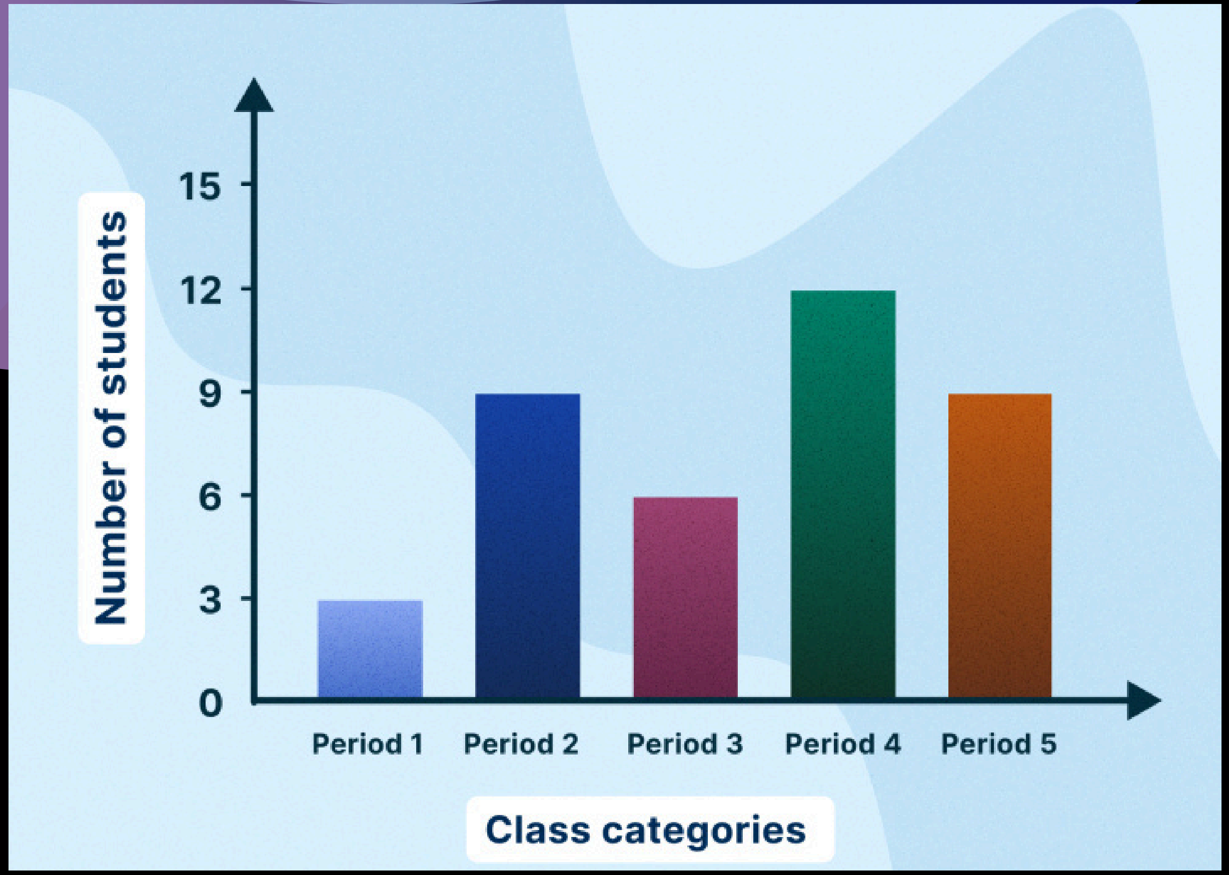
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization provides an easy way to representing and analysing details like understand trends, outliers, and patterns presented in data.

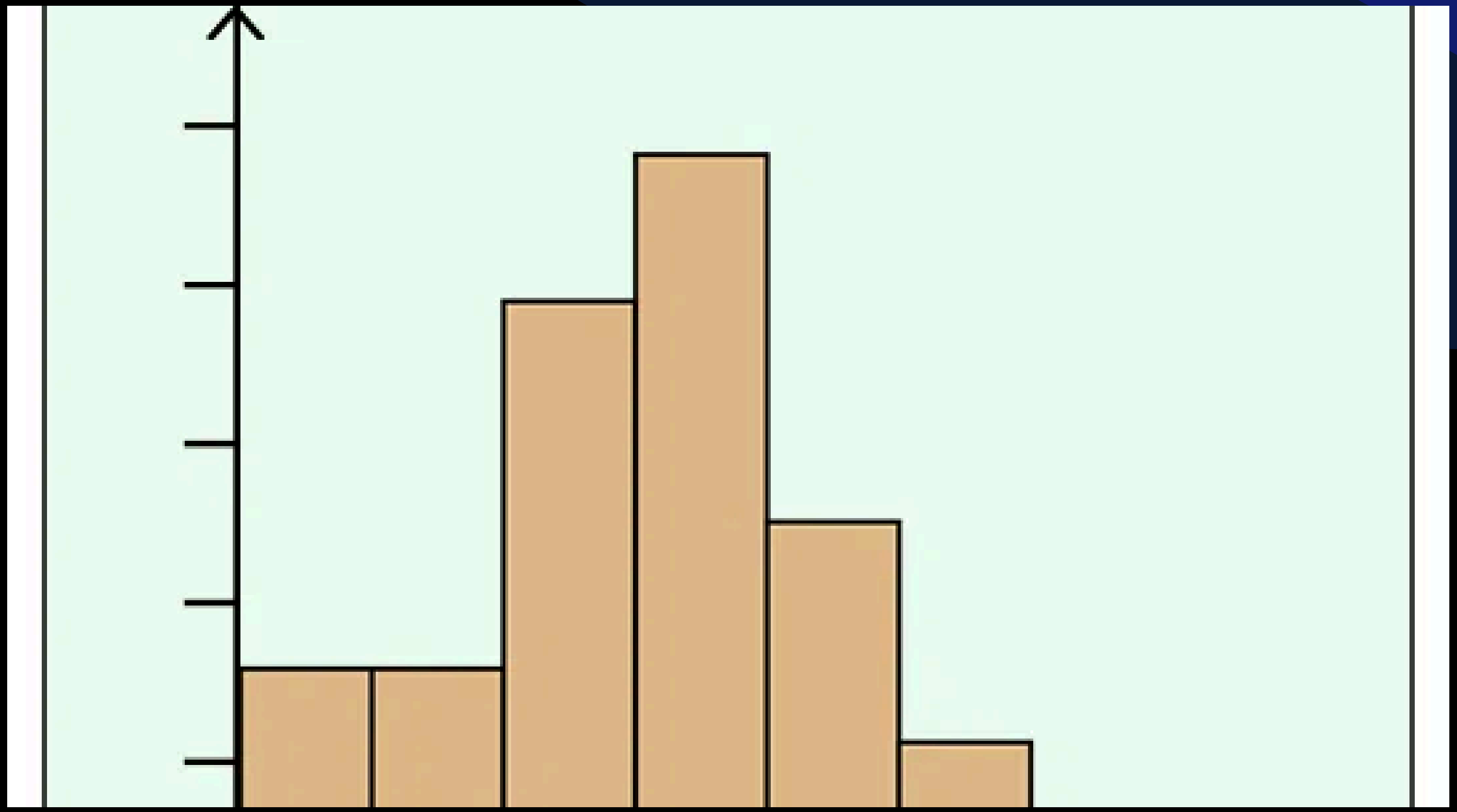
Some of the types of Data Visualization:

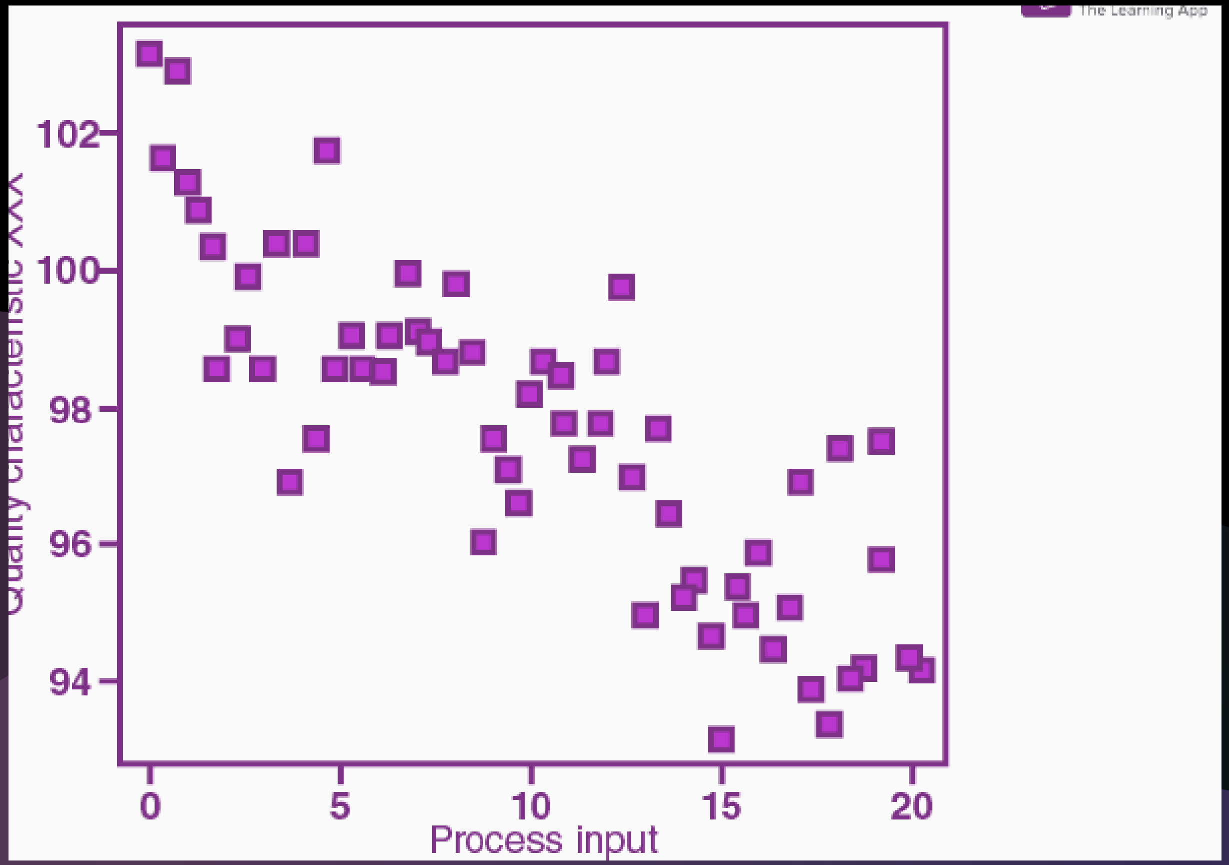
- **Bar Chart:** Useful for comparing different categories.
- **Scatter Plot:** Displays relationships between two variables.
- **Histograps:** Show the distribution of a dataset

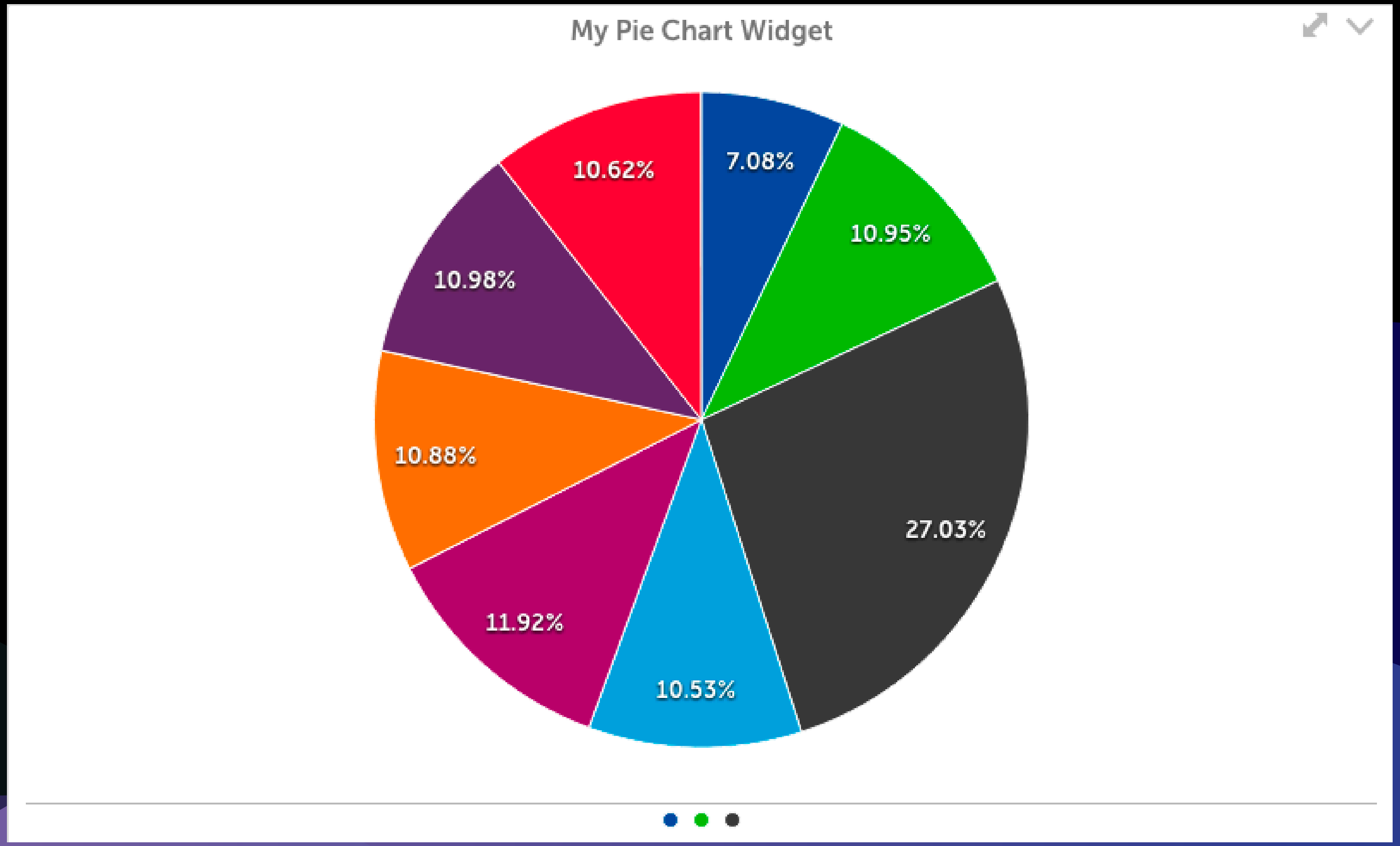


Data Visualisation Graph : -

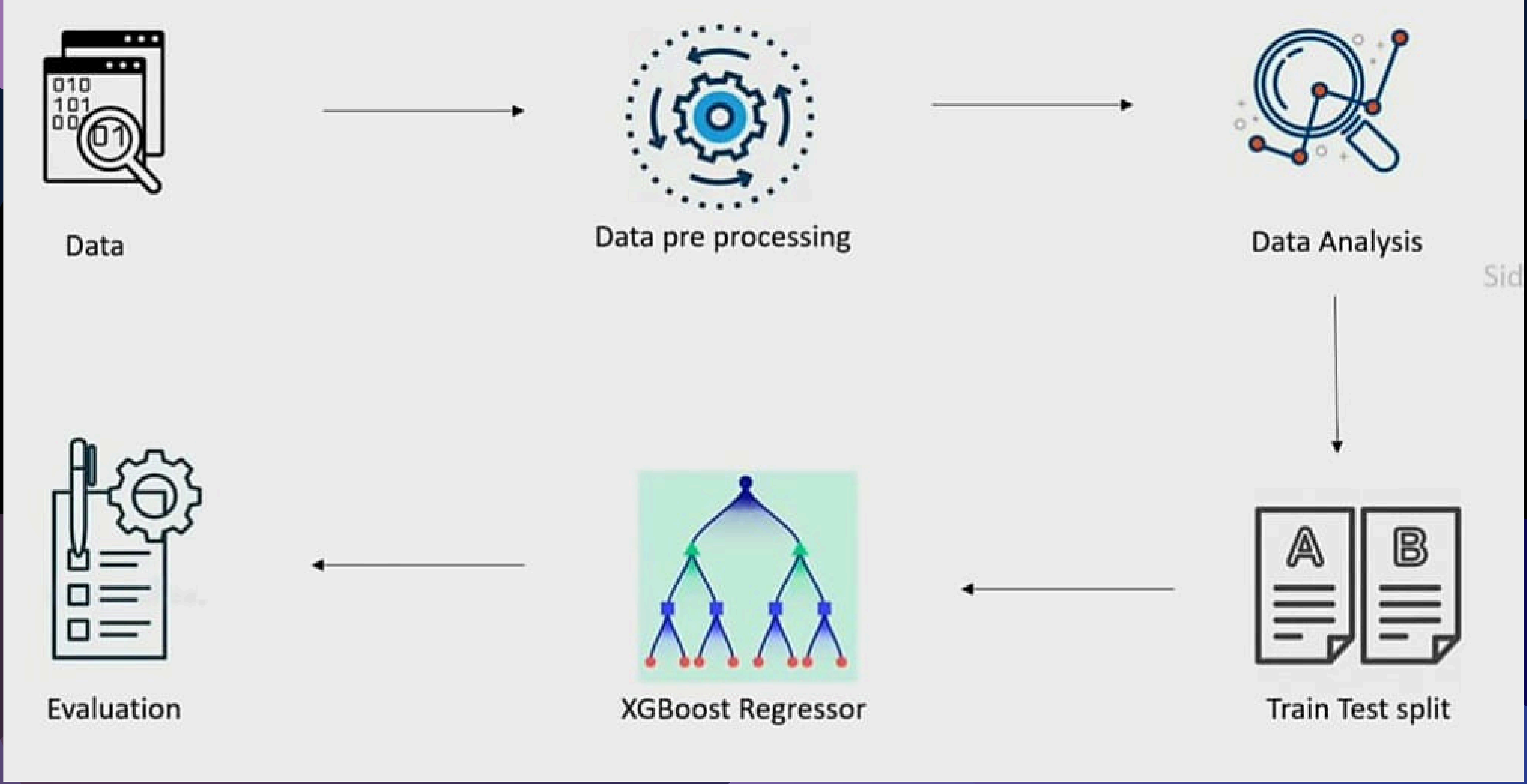






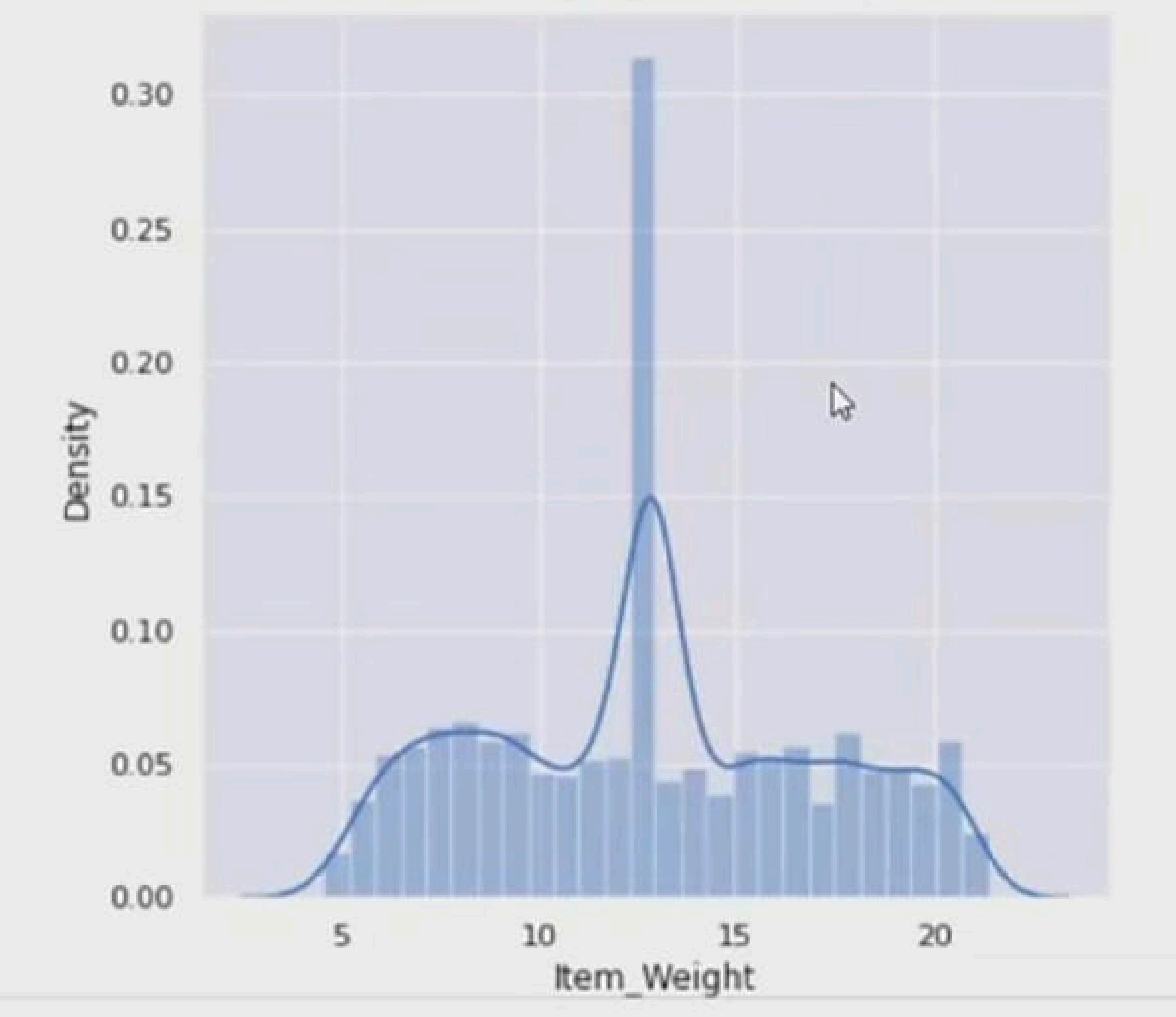


Work Flow of Our Project : -



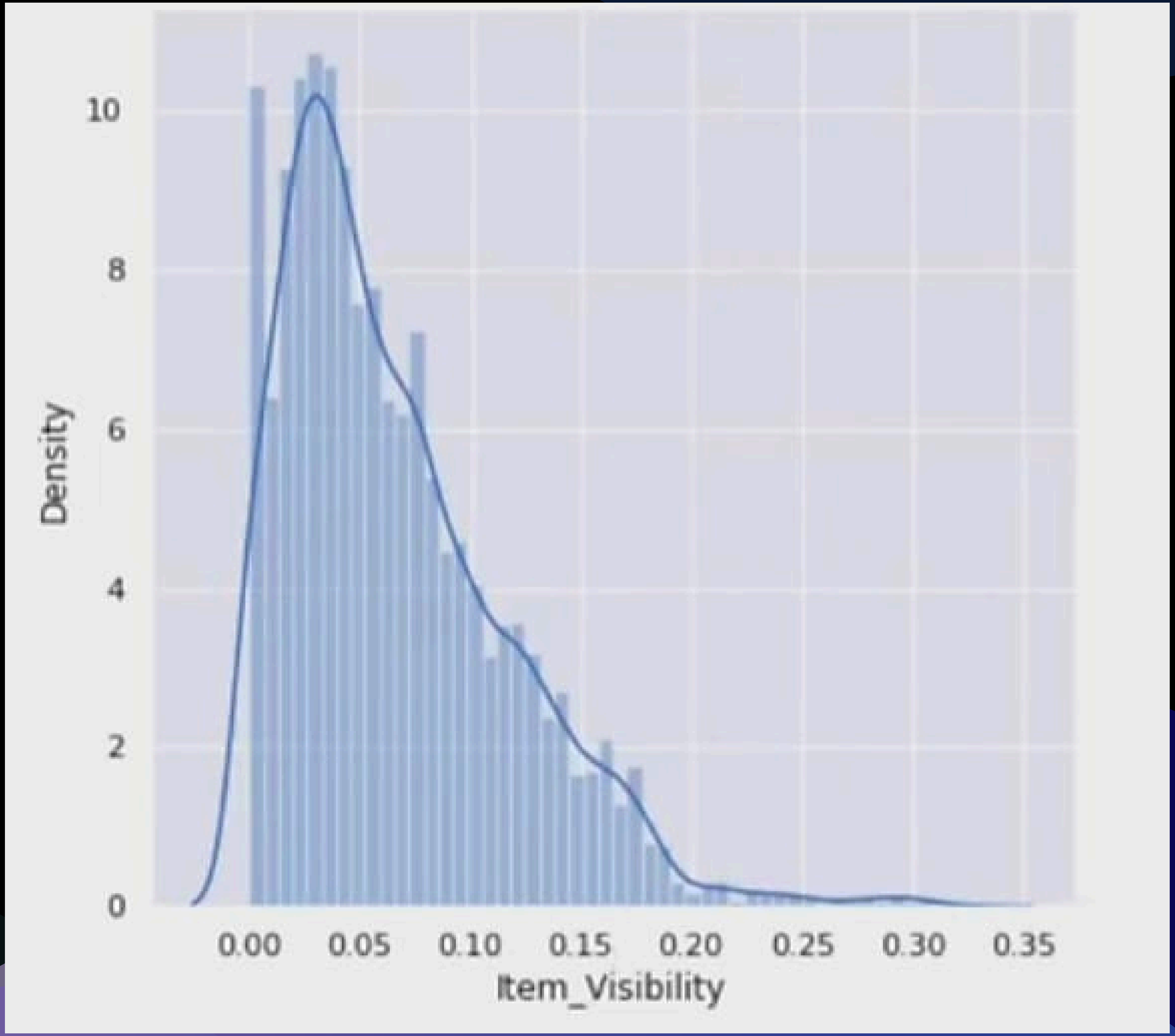
Distribution Analysis Based on Item Weigh : -

Accompanying this analysis, we present a histogram overlaid with a Kernel Density Estimate (KDE) to illustrate the distribution of items based on their weight. The histogram effectively shows the frequency of items within specific weight categories, while the KDE provides a smooth curve representing the overall distribution.



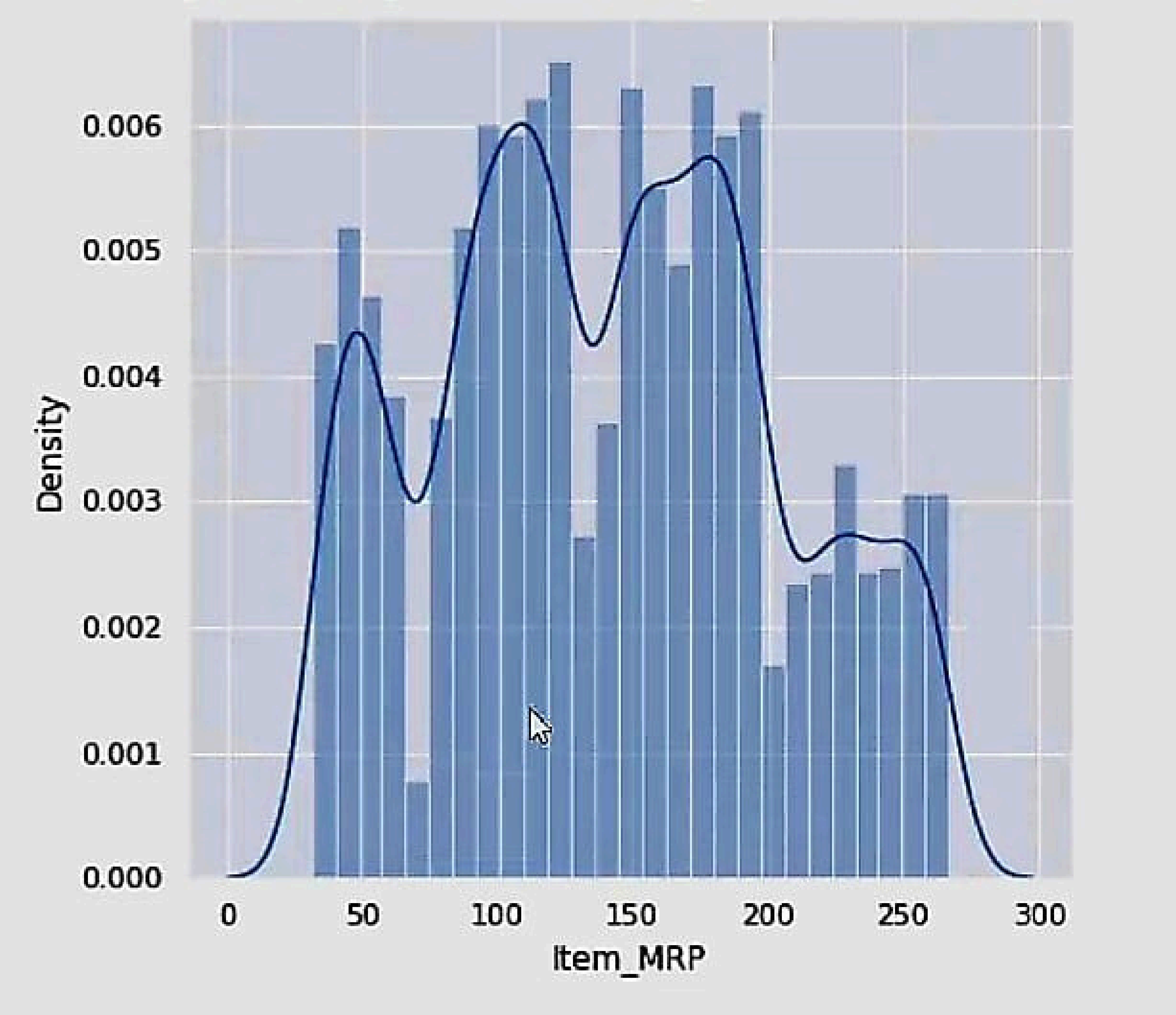
Distribution On Basis Of Item Visibility : -

Given alongside we have a histogram with a KDE (Kernel Density Estimate) that displays the distribution of Item Visibility.



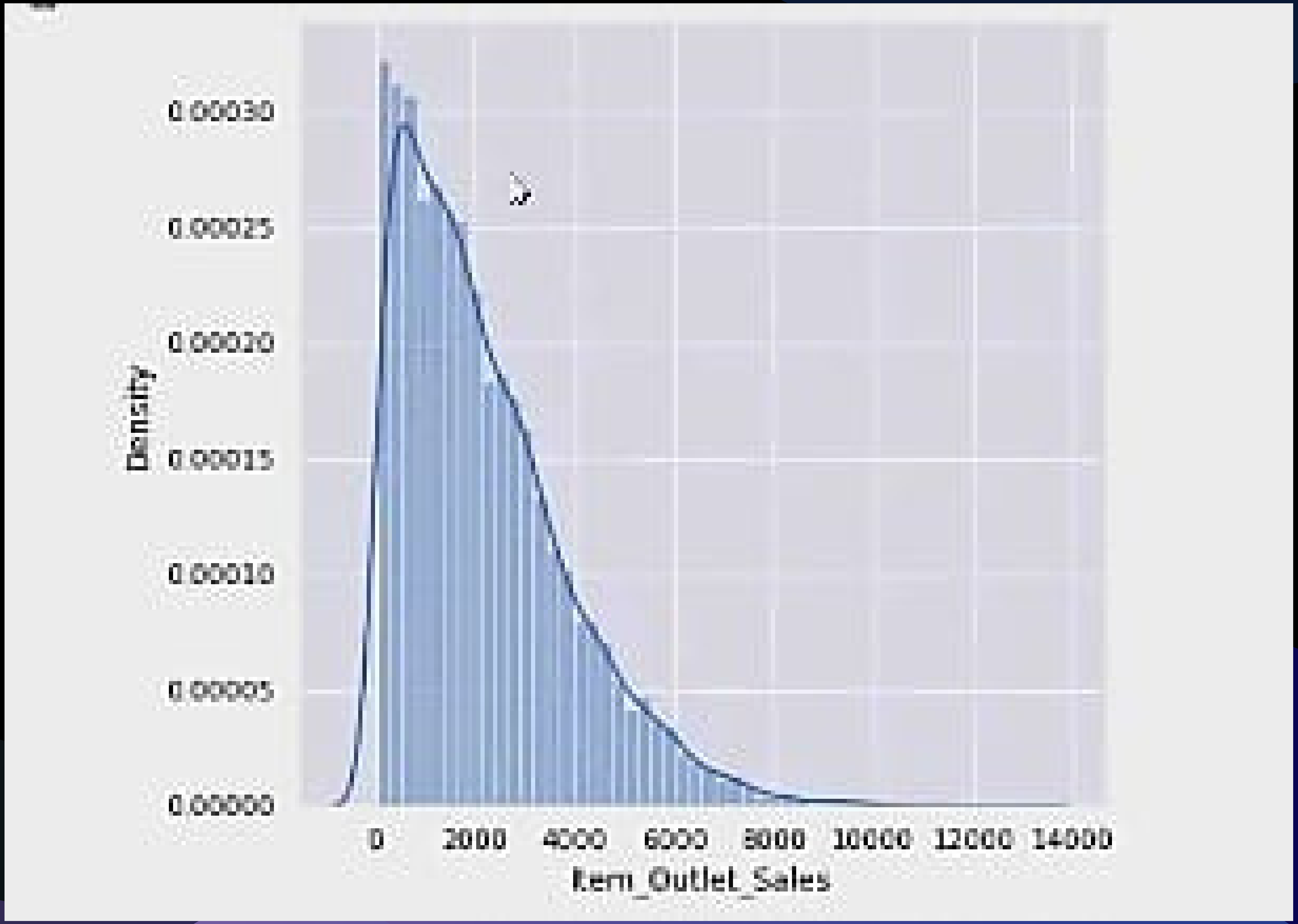
Distribution Based on Item MRP (Maximum Retail Price : -

Accompanying this analysis, we present a histogram overlaid with a Kernel Density Estimate (KDE) to illustrate the distribution of items based on their maximum retail price (MRP). The histogram provides a clear view of the frequency of items within specific price ranges, while the KDE offers a smooth representation of the data's overall distribution, highlighting trends and patterns that may not be immediately evident.



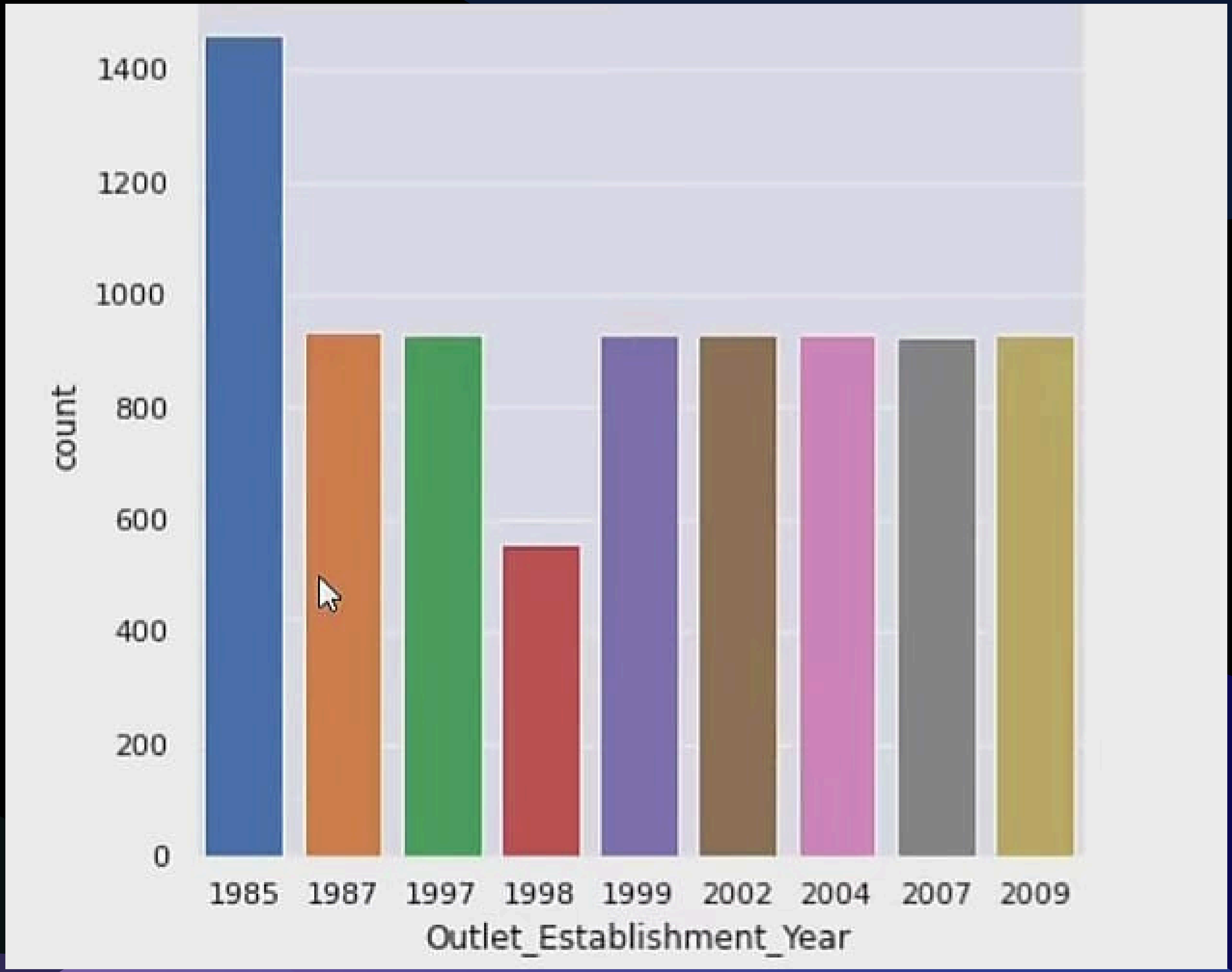
Distribution Based on Item Outlet Sales : -

Given alongside we have a histogram with a KDE (Kernel Density Estimate) that displays the distribution of Item's of the basis of their sales from the outlet store.



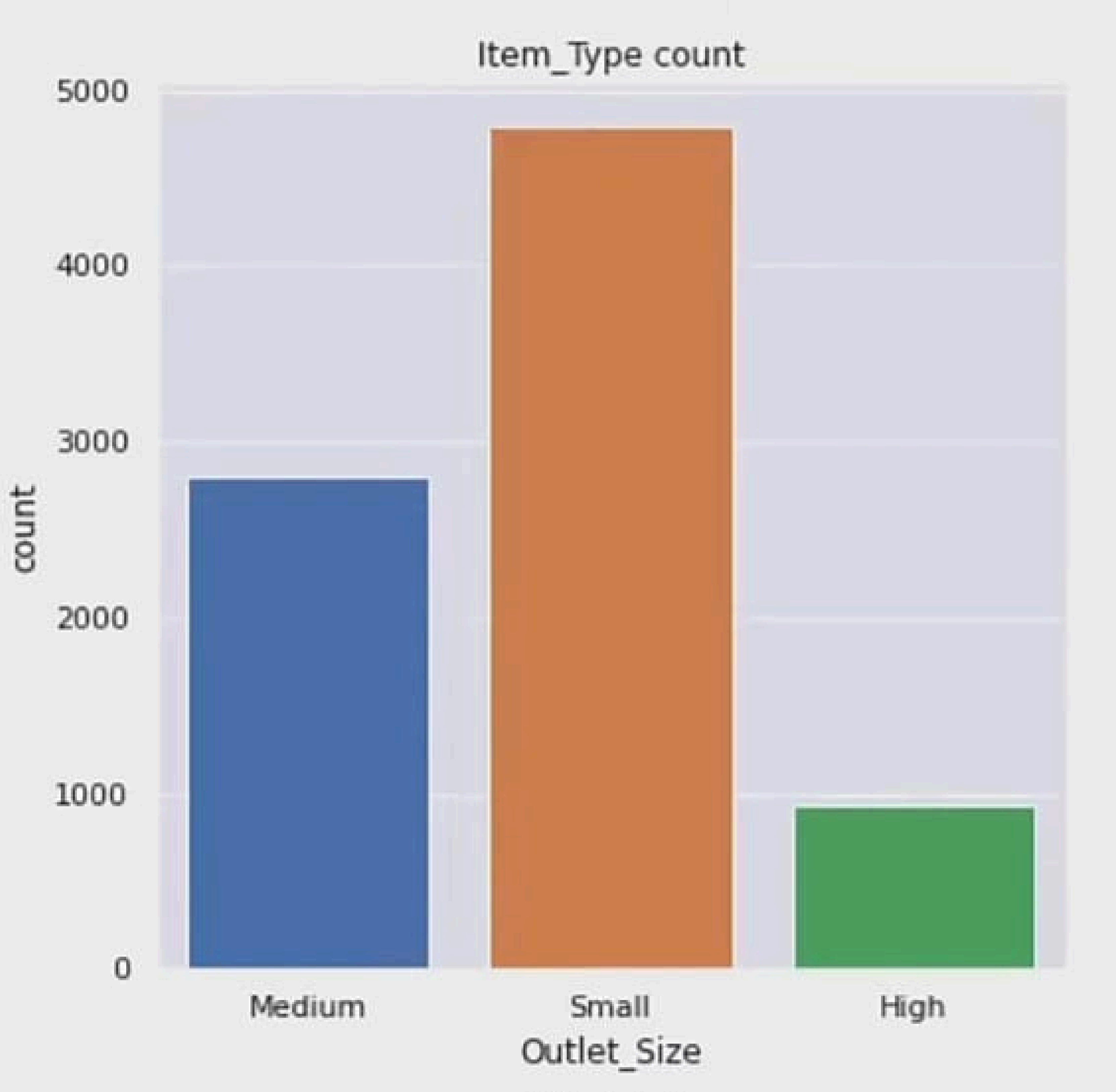
Distribution Graph Showing the Year of Establishment : -

Given alongside we have a counting bar graph that shows the year of establishment of different outlets.



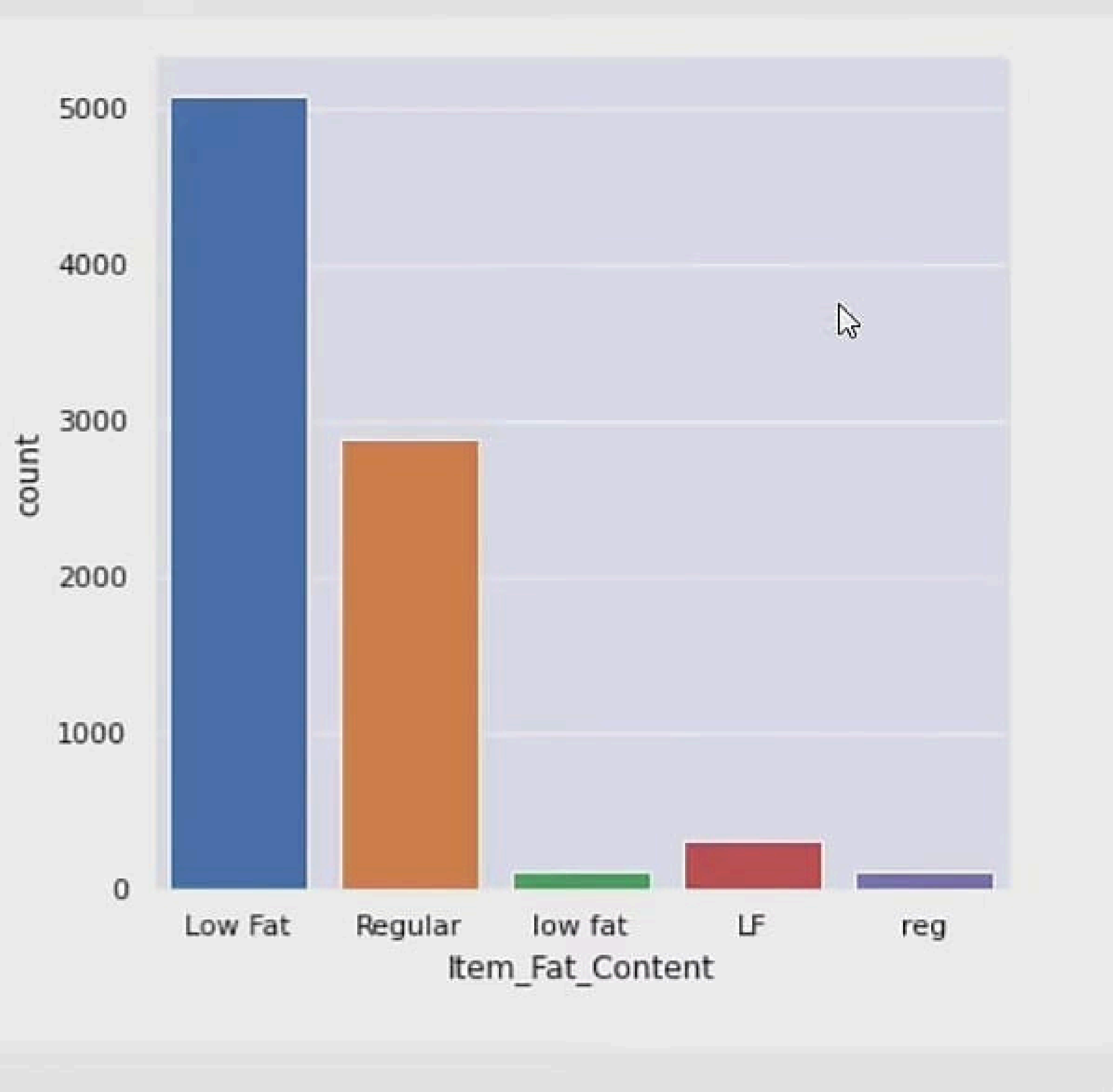
Distribution of Outlets Based on Their Size : -

The bar graph presented alongside illustrates the distribution of outlet stores categorized by their size as medium, small and high.

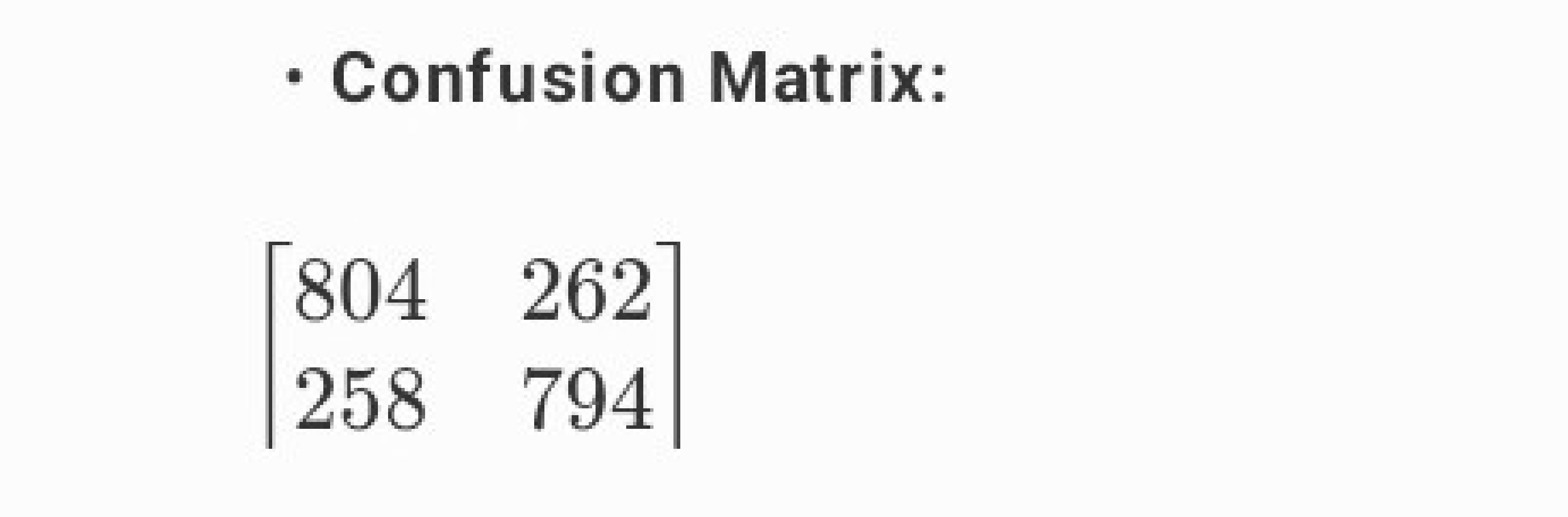


Distribution of Items Based on Their Fat Content : -

The bar graph displayed alongside illustrates the distribution of items in the store, categorized by their fat content into two groups: low fat and regular fat.

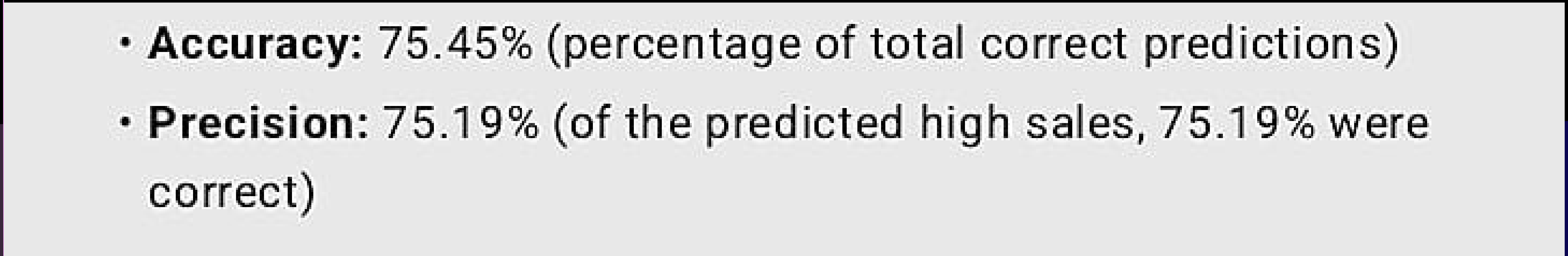


ACCURACY, PRECISION AND CONFUSION MATRIX :-



- **Confusion Matrix:**

$$\begin{bmatrix} 804 & 262 \\ 258 & 794 \end{bmatrix}$$



- **Accuracy:** 75.45% (percentage of total correct predictions)
- **Precision:** 75.19% (of the predicted high sales, 75.19% were correct)

CONCLUSION : -

Precision : 0.802

R-Square value : 0.802

The model has a precision of approximately 80.8%, meaning that among all items predicted as having high sales, around 80.8% were correctly identified. The R-squared value indicates that the model explains 80.2% of the variance in the binary sales classification.

THANK YOU