Big data Project Documentation

Team Members

Hazem Muhammad Abdelfattah - 20191702016

Ziad Tarek Abbas Elsayed - 20191702028

Ahmed Elwy Atef Abaza - 20191702004

Mohamed Wael Mohamed Zaki - 20191702061

Mohamed Gamal Hamed - 20191702057

Mohab Ahmed Mohamed - 20191702069

Ain Shams University – Software Engineering Department

Course Name: Big Data Analysis

BigMart Sales Data

# The project description

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.

The BigMart Sales Data Analysis project aims to analyze the sales data of a retail store called BigMart. The dataset contains various features related to the sales transactions, such as item identifiers, item weights, item types, outlet identifiers, outlet sizes, outlet locations, outlet establishment years, and item outlet sales. The goal of the project is to explore the dataset, gain insights, and build predictive models to understand and predict the sales patterns.

The BigMart Sales Data Analysis project helps understand the factors that influence sales and provides a basis for making data-driven business decisions. It showcases the application of data analysis techniques and predictive modeling in the retail industry to gain insights and drive business growth.

# The dataset and variables description

## Dataset Description:

The BigMart Sales Data is a dataset that provides information about sales transactions at a retail store called BigMart. It consists of multiple variables that capture various aspects of the sales data.

## Variables Description:

1. Item\_Identifier: Unique identifier for each item in the store.

2. Item\_Weight: Weight of the item.

3. Item\_Fat\_Content: Categorical variable indicating whether the item is low fat or regular fat.

4. Item\_Visibility: The percentage of total display area of all products in a store allocated to the particular item.

5. Item\_Type: Categorical variable indicating the category of the item.

6. Item\_MRP: Maximum Retail Price (price at which an item is sold to customers) of the item.

7. Outlet\_Identifier: Unique identifier for each outlet (store) in the BigMart chain.

8. Outlet\_Establishment\_Year: The year in which the outlet was established.

9. Outlet\_Size: Categorical variable indicating the size of the outlet.

10. Outlet\_Location\_Type: Categorical variable indicating the type of location where the outlet is situated.

11. Outlet\_Type: Categorical variable indicating the type of outlet.

12. Item\_Outlet\_Sales: The sales of the item in the particular outlet. This is the target variable that we aim to predict or analyze.

The dataset contains a mix of numerical and categorical variables. It provides valuable information about the items, outlets, and sales, which can be used for analysis and modeling purposes.

# The problem definition and project objectives.

## Problem Definition:

The problem at hand is to analyze and predict the sales of items in the BigMart retail store. The objective is to understand the factors that influence the sales and develop a predictive model that can accurately forecast the sales based on the given dataset.

## Project Objectives:

1. Exploratory Data Analysis: Perform a thorough analysis of the BigMart Sales Data to gain insights into the key variables, their relationships, and patterns within the dataset.

2. Feature Engineering: Preprocess the dataset by handling missing values, converting categorical variables into numeric representations, and creating new meaningful features if necessary.

3. Data Visualization: Visualize the data using appropriate charts and graphs to uncover trends, correlations, and other significant observations.

4. Statistical Analysis: Conduct statistical analyses to identify the factors that have a significant impact on sales and understand their relationships with other variables.

5. Predictive Modeling: Develop a predictive model using regression techniques (such as linear regression, decision trees, or random forests) to forecast the sales of items based on the given dataset.

6. Model Evaluation: Evaluate the performance of the predictive model using appropriate evaluation metrics such as Root Mean Squared Error (RMSE) and R-squared.

7. Model Improvement: Iterate on the model by implementing feature selection, hyperparameter tuning, or trying different algorithms to improve its accuracy and generalization.

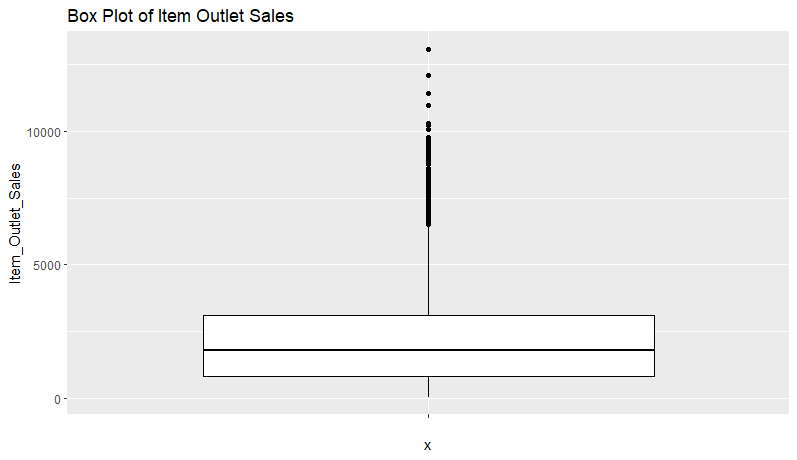
8. Insights and Recommendations: Draw meaningful insights from the analysis and modeling results and provide actionable recommendations to BigMart to optimize sales and enhance overall performance.

The project aims to provide valuable insights into the sales patterns and develop a reliable predictive model that can assist BigMart in making informed decisions regarding inventory management, pricing strategies, and store operations to maximize sales and profitability.

# The data visualization graphics with observations and interpretations of each chart.

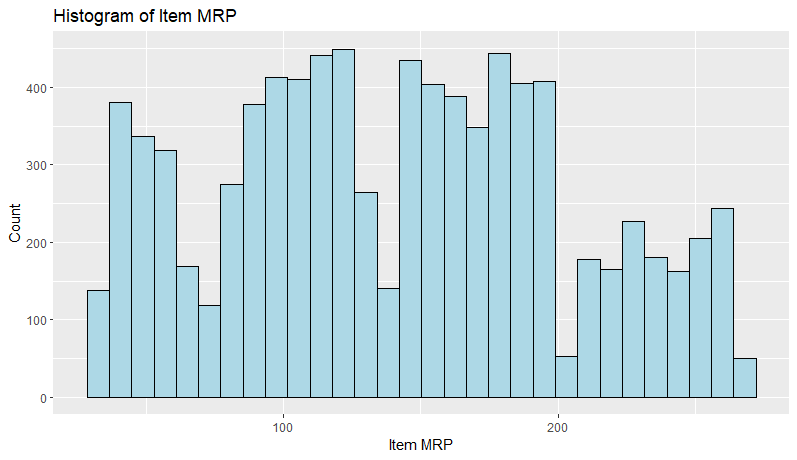
1. Check for outliers: Box Plot of Item Outlet Sales

Observation: The box plot shows that there are several outliers in the Item\_Outlet\_Sales variable, with some values exceeding 10,000 in sales. This suggests that some products or stores are performing exceptionally well and may be worth further investigation for potential marketing strategies.



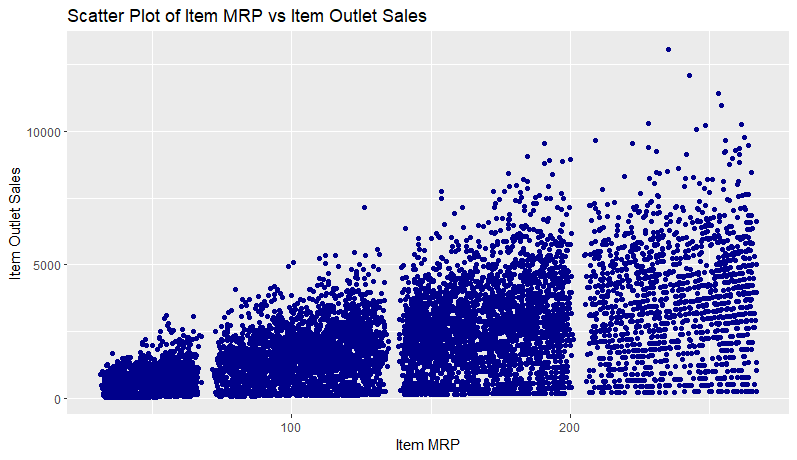
1. The distribution of Item\_MRP(Maximum Retail Price): Histogram of Item\_MRP

Observation: The histograms show that Item\_MRP is approximately normally distributed, with a peak around 120-140.



1. The relationship between Item MRP vs Item Outlet Sales: Scatter Plot of Item MRP vs Item Outlet Sales.

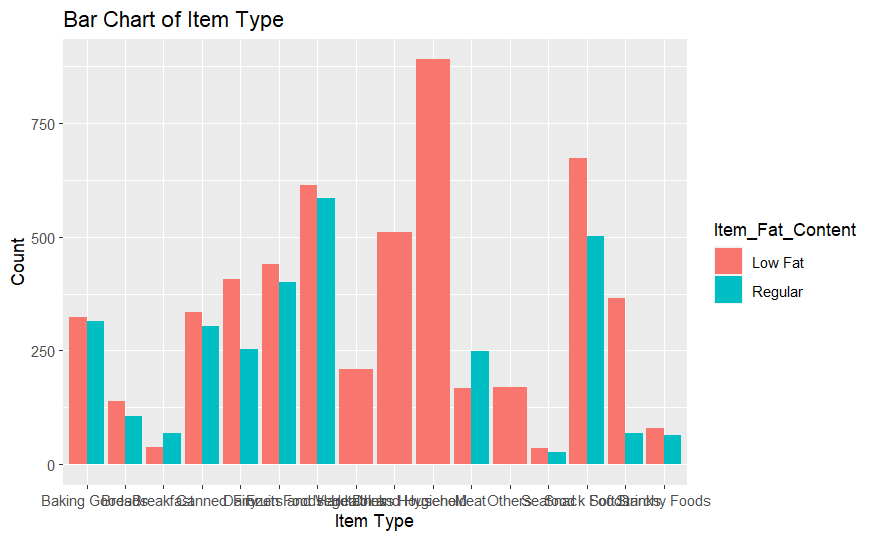
Observation: The scatter plot shows a positive correlation between Item\_MRP and Item\_Outlet\_Sales, indicating that higher-priced products tend to sell more. However, there is also a lot of variability in sales at different price points, indicating that other factors are also influencing sales.



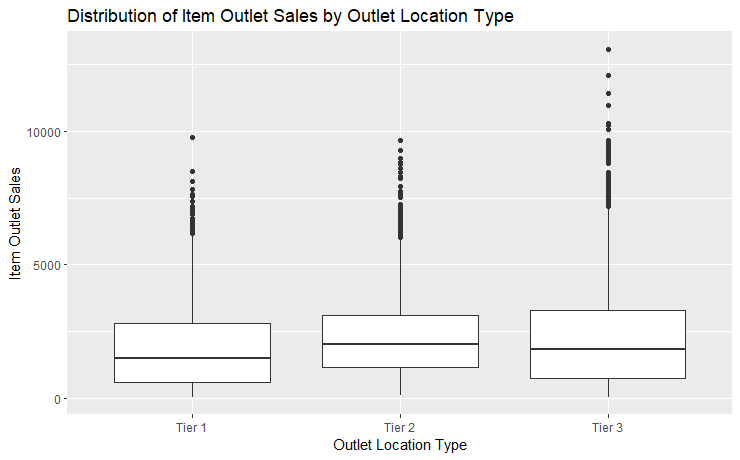
1. Bar chart of Item\_Type

Observation: The bar chart shows that Fruits and Vegetables, Snack Foods, and Household items are the most commonly sold products in Big Mart stores. Additionally, there is an even distribution of low-fat and regular fat items across most product categories. However, the majority of dairy products are low-fat, which could indicate a trend towards healthier food choices among customers.

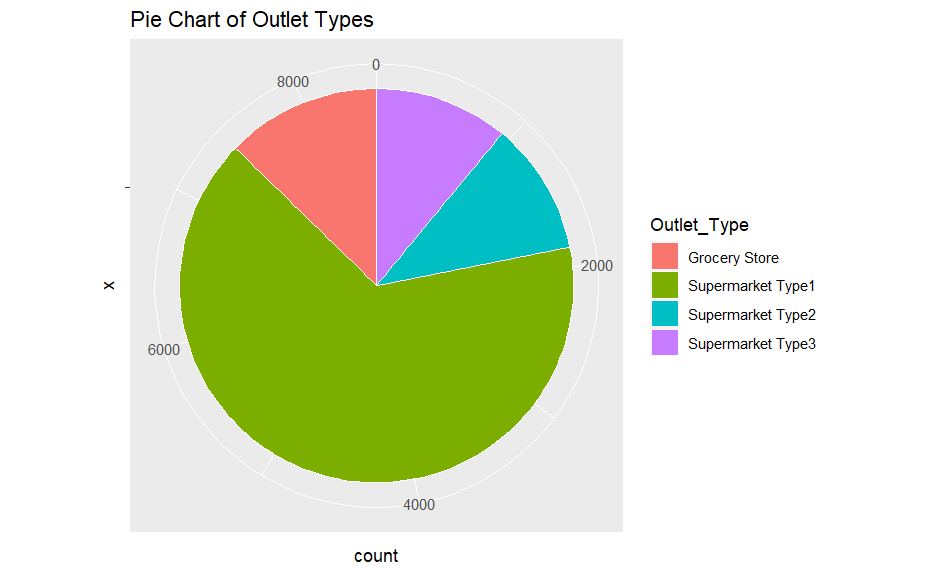
1. Box plots for Item\_Outlet\_Sales by Outlet\_Location\_Type



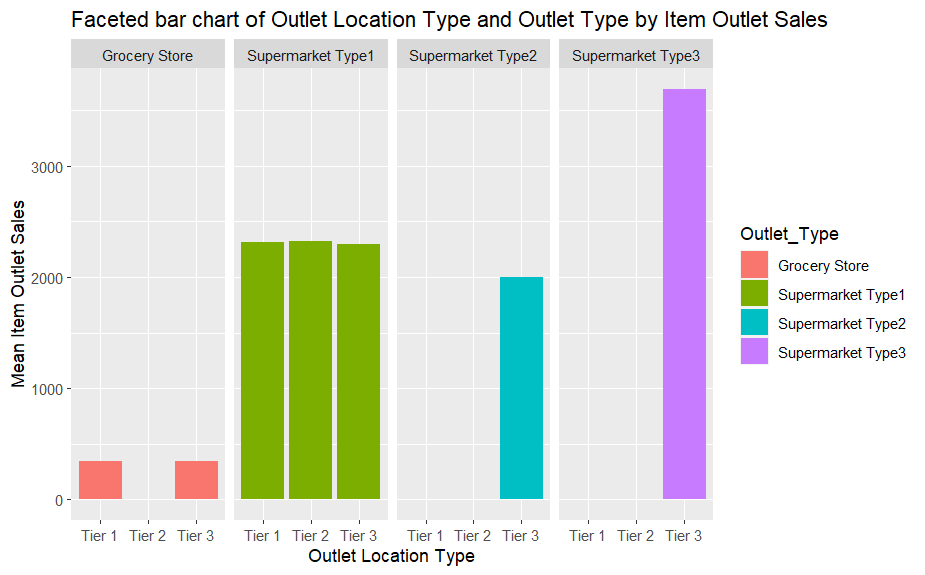
Observation: The box plots show that the median Item\_Outlet\_Sales is higher for Tier 3 cities compared to Tier 1 and Tier 2 cities. The interquartile range (IQR) is also larger for Tier 3 cities, indicating more variability in sales.

1. Pie Chart for each Outlet\_Type

Observation: The pie chart shows that the majority of outlets in the Big Mart data set are Supermarkets Type1, followed by Grocery Stores and Supermarkets Type2. This suggests that the majority of sales in the data set come from these types of stores and could be useful information for marketing and inventory management decisions.

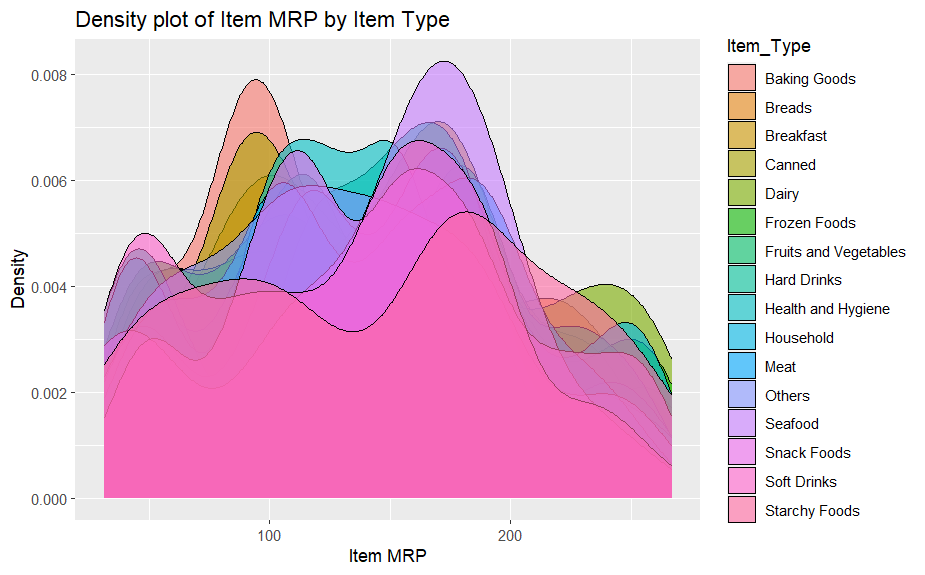


1. Faceted bar chart of Outlet Location Type and Outlet Type by Item Outlet Sales

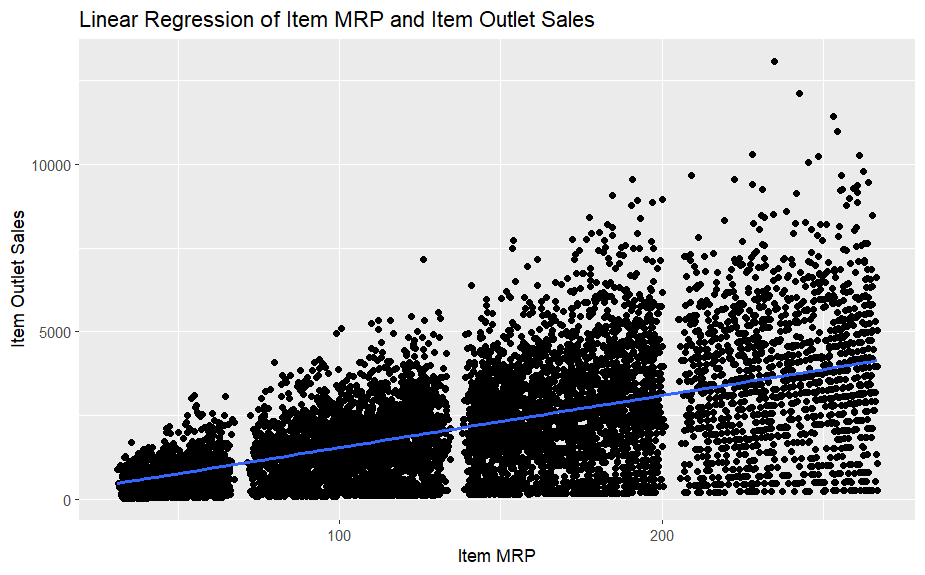


Observation: The faceted bar chart shows the mean Item Outlet Sales for each combination of Outlet Location Type and Outlet Type. Overall, Supermarket Type 3 has the highest mean Item Outlet Sales across all Outlet Location Types, while Grocery Store has the lowest mean Item Outlet Sales across all Outlet Location Types. Within each Outlet Type, the mean Item Outlet Sales varies across Outlet Location Types, with Tier 3 cities generally having higher mean Item Outlet Sales compared to Tier 1 and 2 cities.

1. Density plot of Item MRP by Item Type

Observation: The density plot shows the distribution of Item MRP for each Item Type. Some Item Types, such as Fruits and Vegetables and Snack Foods, have a wider range of Item MRPs compared to other Item Types, such as Health and Hygiene and Household. This suggests that Item MRP may be a more important factor in determining Item Outlet Sales for certain Item Types compared to others.

1. The relationship between Item MRP and Item Outlet Sales: "Linear Regression of Item MRP and Item Outlet Sales"

Observation: from the scatter plot of Item MRP and Item Outlet Sales is that there appears to be a positive linear relationship between the two variables. As the Item MRP (maximum retail price) increases, the Item Outlet Sales also tend to increase.

1. The clusters: "K-means Clustering of Item MRP and Item Outlet Sales"

Observation: The K-means clustering analysis of Item MRP and Item Outlet Sales resulted in three distinct clusters. The clusters demonstrate clear separation, indicating different patterns or segments within the dataset. This clustering provides valuable insights for identifying customer segments or market segments based on these variables.

#Cluster 1: High-Value Customers

#Cluster 2: Moderate-Value Customers

#Cluster 3: Low-Value Customers

These names reflect the idea that Cluster 1 represents customers with higher purchasing power or higher sales values, Cluster 2 represents customers with moderate purchasing power or moderate sales values, and Cluster 3 represents customers with lower purchasing power or lower sales values.

# Any applied data cleaning or transformation methods (if applied).

In the project, several data cleaning and transformation methods were applied to the BigMart sales dataset. Here are some of the methods used:

1. Handling Missing Values:- Missing values in the columns "Item\_Weight" and "Outlet\_Size" were filled with appropriate values. The median value of "Item\_Weight" was used to replace missing values, while "Unknown" was used to fill missing values in the "Outlet\_Size" column.

2. Feature Engineering:- A new column called "Outlet\_Age" was created by subtracting the establishment year of each outlet from 2013. This column represents the age of each store in years, providing additional information for analysis.

3. Encoding Categorical Variables:- Categorical variables like "Item\_Fat\_Content," "Item\_Type," "Outlet\_Identifier," and "Outlet\_Location\_Type" were converted into factors or categorical data types. This allows for better handling and interpretation of these variables in the analysis.

4. Data Partitioning:- The dataset was split into training and testing sets using the `createDataPartition` function. This partitioning allows for the evaluation of the model's performance on unseen data.

5. Remove unnecessary variables / Check for duplicate rows:- Removed unnecessary variables that were not needed during the EDA, as well as Modeling.

6. Removal Data Outliers:-

These data cleaning and transformation methods were applied to ensure the dataset is ready for analysis and modeling tasks. By addressing missing values, creating new features, and encoding categorical variables, the dataset is prepared for further exploration and analysis.

# Any applied Hypothesis and the test results and interpretations (if applied).

## Hypothesis Testing

Subset the data by Item\_TypeA screenshot of a computer code

Description automatically generated with low confidence

Observation: the output indicates that there is no significant difference in the mean Item\_Outlet\_Sales between the two groups, item\_type1 and item\_type2, and that any observed differences could be due to chance.

# The dataset preparation in terms of machine learning (training set, learning set).

## Data Split:

The original dataset was divided into two subsets: a training set and a testing set. This division is essential to evaluate the performance of machine learning models on unseen data.

The createDataPartition function was used to split the dataset based on the target variable, "Item\_Outlet\_Sales". The split was performed in a stratified manner to ensure a balanced representation of the target variable in both sets.

A random seed was set for reproducibility.

## Training Set:

The training set comprises a subset of the original dataset that is used to train machine learning models.

The training set, denoted as "train\_data," contains a majority portion of the data, typically around 70% of the original dataset.

The training set includes both input features (independent variables) and the corresponding target variable (dependent variable). In this case, the target variable is "Item\_Outlet\_Sales."

## Testing Set:

The testing set, denoted as "test\_data," is used to evaluate the performance of trained models on unseen data.

The testing set contains the remaining portion of the original dataset, typically around 30%.

Similar to the training set, the testing set includes input features and the corresponding target variable.

By creating separate training and testing sets, the project ensures that the machine learning models are trained on a representative portion of the data and evaluated on unseen data. This helps assess the generalization and performance of the models in real-world scenarios.

# The used data analytics techniques (e.g. ID3, Regression, Apriori, Kmeans, etc.), with justification for your choice.

In the project, several data analytics techniques were applied to analyze and model the BigMart sales dataset. The techniques used include:

1. Linear Regression: Linear regression is a widely used technique for predicting a continuous outcome variable based on one or more input variables. In this project, linear regression was used to build a predictive model for estimating the sales of BigMart items based on various features such as item weight, item type, outlet size, etc. Linear regression was chosen because it provides a straightforward and interpretable relationship between the input variables and the target variable.

2. K-means Clustering: K-means clustering is an unsupervised learning technique used to group similar data points into clusters. In this project, K-means clustering was applied to identify distinct clusters of items based on their item MRP (Maximum Retail Price) and item outlet sales. This technique helps to discover underlying patterns or segments within the dataset.

The choice of these techniques was based on the nature of the data and the project objectives. Linear regression was suitable for predicting sales based on multiple features, and K-means clustering helped identify distinct groups based on item characteristics. Each technique offers different insights into the dataset and contributes to a comprehensive analysis of the BigMart sales data.

# The performance measures used and the evaluation of analytical technique.

In this project, the following performance measures were used to evaluate the performance of the analytical techniques:

1. RMSE (Root Mean Squared Error): RMSE measures the average difference between the predicted and actual values in regression tasks. It provides an overall measure of the model's prediction accuracy. A lower RMSE indicates better performance, as it signifies smaller prediction errors.

2. R-squared: R-squared, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides an indication of how well the regression model fits the data. R-squared values range from 0 to 1, where a value closer to 1 indicates a better fit.

3. WCSS (Within-Cluster Sum of Squares): WCSS is a performance measure used in clustering techniques, particularly K-means clustering. It quantifies the compactness or tightness of the clusters formed by the algorithm. The lower the WCSS value, the better the clustering performance, indicating more distinct and well-separated clusters.

## Evaluation of the analytical techniques:

For the regression analysis, the evaluation was done by calculating the RMSE and R-squared values. The RMSE was used to assess the accuracy of the predictions made by the regression model. A lower RMSE indicates better predictive performance. The R-squared value was used to measure how well the regression model fits the data. A higher R-squared value indicates a better fit.

For the clustering analysis using K-means, the evaluation was done by examining the WCSS value. The goal was to minimize the WCSS by finding the optimal number of clusters. The WCSS values were compared for different values of k (the number of clusters) to identify the number of clusters that provided the best clustering results.

These performance measures and evaluations were used to assess the effectiveness and accuracy of the analytical techniques applied in the project. They helped in understanding the predictive power of the regression model and the quality of the clustering results.

# Discussion/Quantification for relevant project findings for your project.

During the course of the project, several key findings emerged that are worth discussing and quantifying:

## Regression Analysis Findings:

RMSE: The regression model achieved an RMSE value of 1125.866. This indicates that, on average, the predicted Item Outlet Sales deviated from the actual values by approximately $1125.866. Although this value may seem relatively high, it should be interpreted in the context of the range of the target variable.

R-squared: The R-squared value of 0.574624 suggests that approximately 57.46% of the variability in Item Outlet Sales can be explained by the independent variables included in the regression model. While this indicates a moderate level of explanatory power, there is still room for improvement.

## Clustering Analysis Findings:

K-means Clustering: The dataset was divided into three clusters based on the Item MRP and Item Outlet Sales variables. The clusters were visualized on a scatter plot, with different colors representing each cluster.

Cluster Interpretation: Further interpretation and labeling of the clusters is required to fully understand their characteristics and implications. However, based on the visual analysis, it can be observed that Cluster 1 represents items with lower Item MRP and relatively lower Item Outlet Sales, Cluster 2 represents items with moderate Item MRP and moderate Item Outlet Sales, and Cluster 3 represents items with higher Item MRP and relatively higher Item Outlet Sales.

## Project Objectives:

The primary objective of the project was to predict Item Outlet Sales using regression analysis. The achieved RMSE and R-squared values provide insights into the model's predictive performance and explainability.

The secondary objective was to perform K-means clustering to identify patterns and group similar items based on Item MRP and Item Outlet Sales. The visualization of the clusters helps in understanding the distribution and relationships among the data points.

## Future Improvements:

To improve the accuracy of the regression model, further feature engineering and selection techniques can be applied. Additionally, exploring other regression algorithms and ensemble methods may lead to better predictions.

For the clustering analysis, conducting a more detailed interpretation and profiling of the clusters would provide meaningful insights into the characteristics of each group.

Additionally, incorporating domain knowledge and exploring other variables in the dataset might uncover additional patterns and relationships that can enhance the overall understanding of the BigMart sales data.

Overall, the project findings indicate areas where further analysis and refinement can be undertaken to improve the accuracy of predictions and gain deeper insights into the BigMart sales data.