**Dataset description**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Description** | **Data Type** | **Example** |
| Store ID | An identifier for each store | Integer | 1, 2, 3, 4, 5 |
| Store\_Area | The area size of each store in square units | Integer | 1659, 1461, 1340 |
| Items\_Available | The number of items available in each store | Integer | 1961, 1752, 1609 |
| Daily\_Customer\_Count | The count of customers visiting each store per day | Integer | 530, 210, 720 |
| Store\_Sales | The total sales generated by each store | Integer | 66490, 39820 |

* 896 records in our dataset.

**Data Preprocessing**

* As part of preprocessing, the Store ID column has been removed as it does not directly impact any other column values.
* No null values were detected in any of the columns.
* A few outliers were identified and removed from each column to improve analysis and visualization.
* Upon analyzing the correlation among the features, most of the features are not correlated with each other, except for the Store\_Area feature and the Items\_Available feature.

Detecting outliers:

|  |
| --- |
|  |
|  |
|  |
|  |

With the above visualisation of each feature using violin plot, we could say that now of the distribution plot in the violin is skewed towards either of the direction. Therefore, there are no outliers in this dataset. In the last violin plot of Store\_Sales feature, it is slightly skewed but, there not much difference between maximum value and upper fence value.

**Chart 1. Distribution of Store Area**

A graph of colorful squares

Description automatically generated

We can visualise the frequency of stores with same Store area.

A screenshot of a graph

Description automatically generated

**Chart 2. Relationship between Store Area and Daily Customer Count (Scatter Plot)**

**A screen shot of a computer screen

Description automatically generated**

On plotting the Store\_Area and Daily\_Customer\_Count feature over the scatter plot, we can observe that the relationship between those features is Non-Linear and correlation value of these feature is (-0.041), therefore we can’t use Linear regression technique. To understand pattern between these features, we have K-means clustering technique.

**A diagram of clustering dots

Description automatically generated with medium confidence**

With this visualization, we could observe that when the size of the store in terms of Store Area is less or more, Daily customer count remains the same it is not more or less.

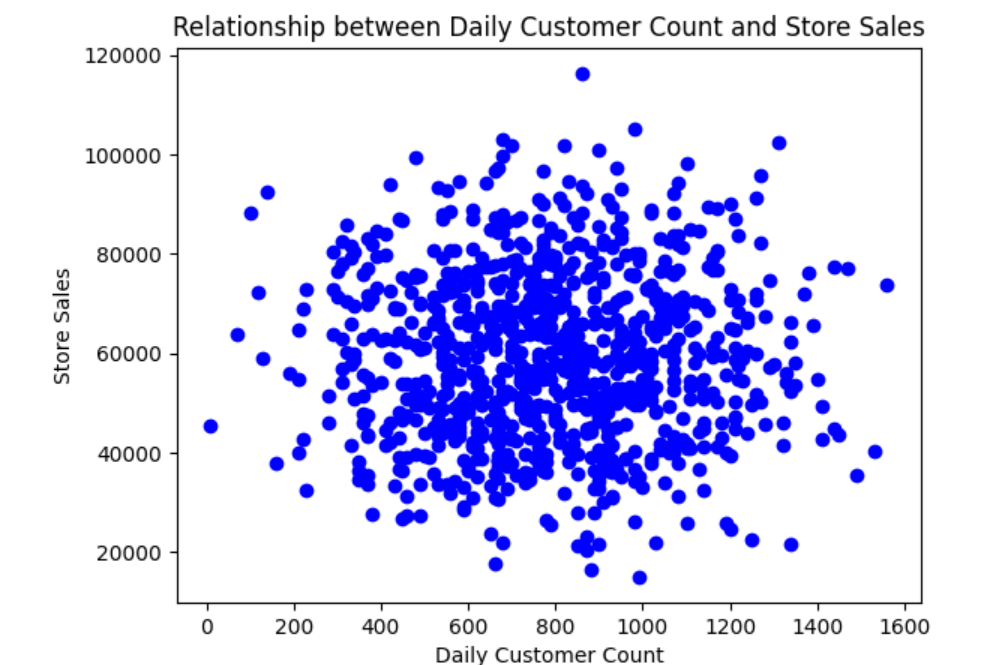
Majority of the stores with store area size between 1100 to 1900 has good number of daily customers between 400 to 1000 respectively.

**Chart 3. Distribution of Items Available in stores**

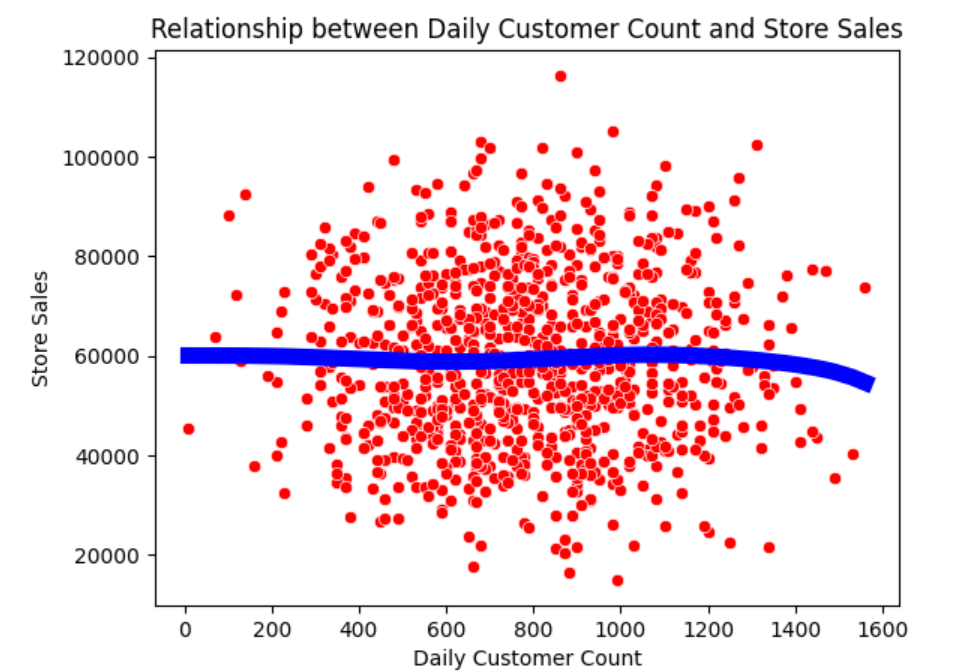
****

The bar chart shows that most stores have between 1432 and 1931 items available, indicating a common trend. However, there are fewer stores with items in the range of 2432 to 2931, suggesting a disparity in inventory levels.

**Chart 4. Relationship between Daily Customer Count and Store Sales**

****

The scatter plot illustrates a lack of correlation between store sales and daily customer count, revealing a non-linear pattern.

****

Reason to choose polynomial regression:  
The dataset contained continuous and labeled features, necessitating the use of a supervised regression model. Examined the association between daily customer count and store sales, revealing a non-linear relationship. Hence, polynomial regression was employed to establish this connection.

This model was constructed to determine the optimal fitted line but encountered difficulty in generating the regression line.

**A diagram of a cluster of blue and green dots

Description automatically generated with medium confidence**

Reason to choose K-means:

As the polynomial regression failed to capture the relationship between the features and the target variable due to non-linearity. K-means was implemented which works by partitioning the data into clusters based on similarity, which can provide insights into the underlying structure of the data, potentially revealing patterns or groupings.

K-means algorithm was implemented with two distinct centroids. The achieved accuracy of 58.3% indicates superior performance compared to alternative models.

**Chart 5. Distribution of Store Sales**

A graph showing a line of red dots

Description automatically generated with medium confidence

The reason to choose multiple linear regression:

The data values are continuous and multiple linear regression aims to find a linear relationship between independent and dependent features. But it did not perform well.

The substantial disparity between the actual and predicted values is apparent due to the scattered distribution of data points for each independent variable concerning the dependent variable. This creates challenges in accurately capturing the relationship between the independent variables and the dependent variable.

A screenshot of a graph

Description automatically generated

Reason to choose K-means:

In here, the multiple-linear regression model failed to capture linear relationship. So, K-means clustering approach was used here. The K-means clustering algorithm has three clusters based on four features. Achieved an accuracy of 52.8%, outperforming traditional machine learning algorithms.

**A graph with a blue rectangle

Description automatically generated**

The box plot serves as a tool to identify outliers, representing data points that deviate from the typical distribution of the sample. In the case of the "store sales" feature, the box plot reveals the presence of a solitary outlier, with a median value recorded at 58605.0.

**Chart 6. Average Daily Customer Count by Store Area**

The plot illustrates the distribution of customer counts across various ranges of store areas, namely less than 1500, between 1500 and 1800, between 1800 and 2000, and greater than 2000. It demonstrates that the relationship between customer count and store areas is non-linear. Thus, an expansion in store area does not invariably lead to a proportional increase in the number of customers.

A graph of a customer

Description automatically generated with medium confidence

**Chart 8. Correlation Matrix Heatmap**

Correlation matrices explore the relationships between different variables (store area, items available, daily customer count, store sales).

Store Area and Items Available have a very strong positive correlation of approximately 0.99. This suggests that stores with larger areas tend to have more items available.

Daily Customer Count shows a very weak negative correlation with both Store Area and Items Available, suggesting that there's no significant relationship between these variables.

Store Sales have a very weak positive correlation with Store Area and Items Available, indicating that larger stores with more items available may have slightly higher sales, but the correlation is not strong.

A diagram of a store area

Description automatically generated with medium confidence