Delta Sigma Technologies
Project Assignment

BIGMART 2013 SALES DATA FOR 1559 PRODUCTS ACROSS 10 STORES IN DIFFERENT CITIES.

Name: Mohammad Arshadulla Noor

Course ID: DST204A21

PROBLEM STATEMENT

The aim is to (visualize data) get insights from the dataset. Try to understand the how various features play a role in increasing the sales

SOLUTION

1

Data Preparation

Step 1: Explore
Step 2: Pre-Process

2

Data Visualization

Present Data in Graphical Format

3

Data Analysis and Model Building

Decision Tree
Step 1: Regression
Step 2: Build Model
Step 3: Evaluate

WHY GOOGLE COLABORATORY

Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

- You can share your Google Colab notebooks very easily
- Versioning You can save your notebook to Github with just one simple click
- Performance Use the computing power of the Google servers instead of your own machine. Running python scripts requires often a lot of computing power and can take time. By running scripts in the cloud, you don’t need to worry.
- can access or edit your notebooks any where even in your phone

First, we should import the libraries that we are going to use. In this case pandas, numpy, matplotlib, sklearn.tree, sklearn.model_selection Then store the file in a variable (url) and read the file in CSV [df = pd.read_csv(url)]

DATA PREPARATION

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis.

Cleaning the given data. Delete the columns that will not affect the sales like Item_Identifier, Outet_Identifier, Item_Weight, Item_type and replace all the null values with appropriate data then Label Encode the columns which have string datatype.

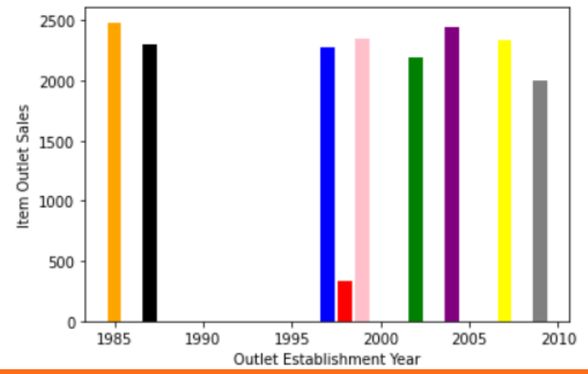
DATA VISUALIZATION

Data visualization is a graphical representation of information and data. By using visual elements like charts and graphs

By grouping item outlet sales and outlet establishment year and plotting the bar chart. Outlets opened in the year 1985 sales are higher and the outlets opened in the year 1998 are low by this we can say people are trusting the old-established outlets.

```
Item_Outlet_Sales_by_Outlet_Establishment_Year = df.groupby('Outlet_Establishment_Year').Item_Outlet_Sales.mean()
Item_Outlet_Sales_by_Outlet_Establishment_Year.sort_values(inplace=True)
x = Item_Outlet_Sales_by_Outlet_Establishment_Year.index.tolist()
y = Item_Outlet_Sales_by_Outlet_Establishment_Year.values.tolist()
plt.xlabel("Outlet Establishment Year")
plt.ylabel("Item Outlet Sales")
plt.bar(x,y,color=['red','grey','green','blue','black','yellow','pink','purple','orange'])
```

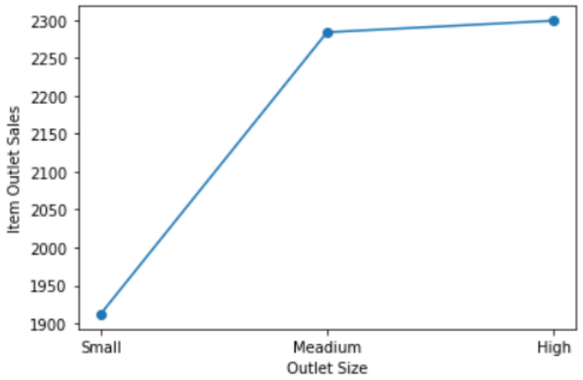
```
<BarContainer object of 9 artists>
```



By grouping item outlet sales and outlet size and plotting the line graph. Sales of small outlets are low compared to the medium and high.

```
Item_Outlet_Sales_by_Outlet_Size = df.groupby('Outlet_Size').Item_Outlet_Sales.mean()
Item_Outlet_Sales_by_Outlet_Size.sort_values(inplace=True)
x = Item_Outlet_Sales_by_Outlet_Size.index.tolist()
y = Item_Outlet_Sales_by_Outlet_Size.values.tolist()
plt.xlabel("Outlet Size")
plt.ylabel("Item Outlet Sales")
plt.xticks(ticks = (0,1,2), labels=["Small","Meadium","High"])
plt.plot(x,y,marker='o')
```

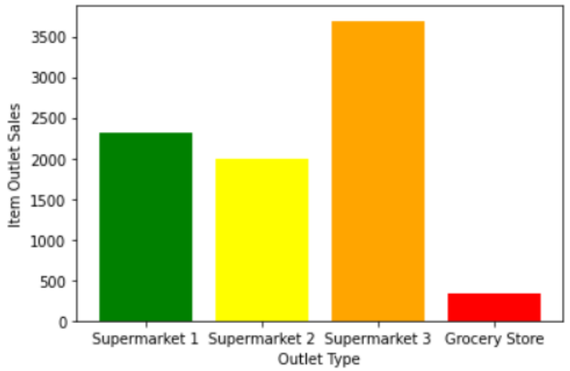
```
[<matplotlib.lines.Line2D at 0x7f7dcad3a710>]
```



By grouping item outlet sales and Outlet type and plotting the bar chart we can see that Supermarket Type 3 has high sales and grocery stores have low sales

```
Item_Outlet_Sales_by_Outlet_Type = df.groupby('Outlet_Type').Item_Outlet_Sales.mean()
Item_Outlet_Sales_by_Outlet_Type.sort_values(inplace=True)
x = Item_Outlet_Sales_by_Outlet_Type.index.tolist()
y = Item_Outlet_Sales_by_Outlet_Type.values.tolist()
plt.xlabel("Outlet Type")
plt.ylabel("Item Outlet Sales")
plt.xticks(ticks = (1,2,3,4), labels=["Supermarket 1","Supermarket 2","Supermarket 3","Grocery Store"])
plt.bar(x,y,color=['red','yellow',"green","orange"])
```

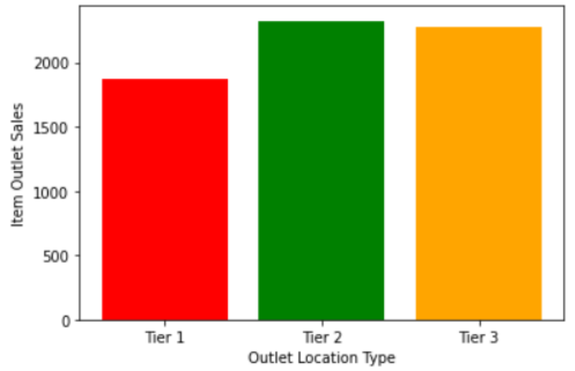
```
<BarContainer object of 4 artists>
```



By grouping item outlet sales and outlet location type and plotting the bar graph we can see that the sales are high in location Tier 2 and Tier 3 and low in Tier 1

```
Item_Outlet_Sales_by_Outlet_Location_Type = df.groupby('Outlet_Location_Type').Item_Outlet_Sales.mean()
Item_Outlet_Sales_by_Outlet_Location_Type.sort_values(inplace=True)
x = Item_Outlet_Sales_by_Outlet_Location_Type.index.tolist()
y = Item_Outlet_Sales_by_Outlet_Location_Type.values.tolist()
plt.xlabel("Outlet Location Type")
plt.ylabel("Item Outlet Sales")
plt.xticks(ticks = (1,2,3), labels=["Tier 1","Tier 2","Tier 3"])
plt.bar(x,y, color=['red','orange',"green"])
```

```
<BarContainer object of 3 artists>
```



DATA ANALYSIS AND MODEL BUILDING

```
x = df.drop(["Item_Outlet_Sales"],axis=1)
y = df["Item_Outlet_Sales"]
```

```
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.15, random_state = 100)
```

```
model = DecisionTreeRegressor()
model.fit(x_train, y_train)

y_pred = model.predict(x_test)

print("Training Accuracy: ",model.score(x_train, y_train))
print("Testing Accuracy: ",model.score(x_test,y_test))
```

```
Training Accuracy:  1.0
Testing Accuracy:  0.19393768656208768
```

x = All the columns leaving Item outlet sales

y = Item outlet sales

Testing and training x & y by Regressor technique. specifying test size as 0.15 and the random state as 100 getting training accuracy as 1.0 and testing accuracy as 0.193

By x test of 10 values and printing y pred of it, we can have an idea about how sales are affected by various factors

```
x_test[:10]
```

	Item_Fat_Content	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type
3454	0	16.966714	183.5266	1987	2	3	1
3386	0	3.546967	256.3014	2007	1	2	1
235	0	4.357366	192.8846	1985	1	3	3
7201	0	5.233528	108.1280	2007	1	2	1
7782	1	4.419237	180.3318	1997	0	1	1
3960	0	19.254067	160.2236	1998	1	3	4
2688	0	5.762690	83.3566	1999	1	1	1
1867	0	16.507364	207.7270	1998	1	3	4
6309	0	4.460638	174.2054	1985	1	3	3
4202	0	3.786485	63.0826	1987	2	3	1

```
[ ] y_pred[:10]
```

```
array([5163.9448, 6358.39  , 3163.8816, 1965.4416, 2189.1504,  633.8416,
       1606.5754,  621.1914, 4012.1108, 1661.8368])
```

By the analysis of the given data. Outlets of medium and high size, supermarket type 3 in tier 2 location can expect higher sales. visibility can be considered as stock. less visibility more sell-outs means higher sales old-established outlets have a slight advantage.

Thank you

Mohammad Arshadulla Noor