Final Project Report(Technocolab)

Presented by – Shayantan Dutta Choudhury

This report consists of the Data Analysis of Train.csv data csv file, assigned as a task by Technocolab, which contains a number of Items along with its outlet information. Following is the preview of the data file—

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
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

The following python libraries were used to perform the analysis task—

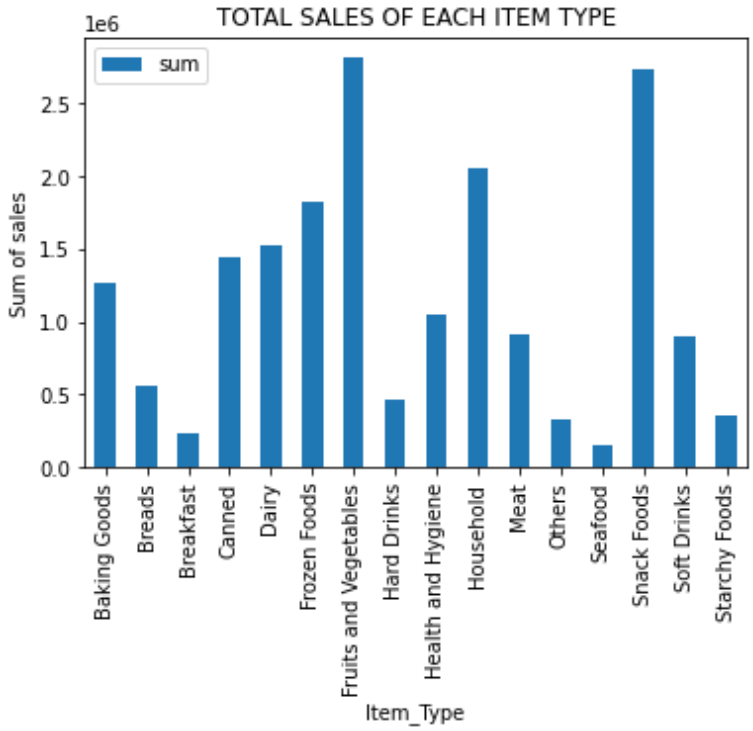
1. Pandas
2. matplotlib.pyplot
3. seaborn
4. numpy
5. matplotlib.gridspec
6. sklearn.linear_model
7. sklearn.model_selection
8. metrics(sklearn)

Thus we perform Exploratory Data Analysis on the data file after successfully cleaning the data file and then train and test the data for making predictions on future values.

Exploratory Data Analysis

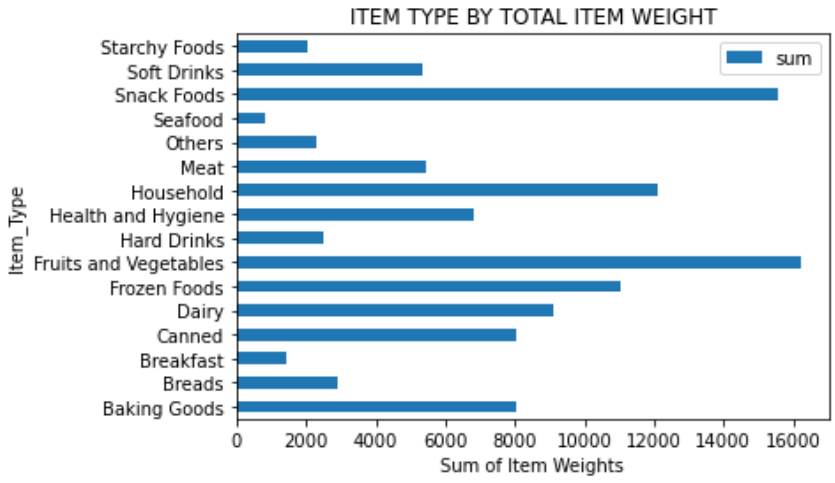
We perform exploratory data analysis on a number of findings from the data table which were important from business point of view.

1. Total Sales of Each Item Type



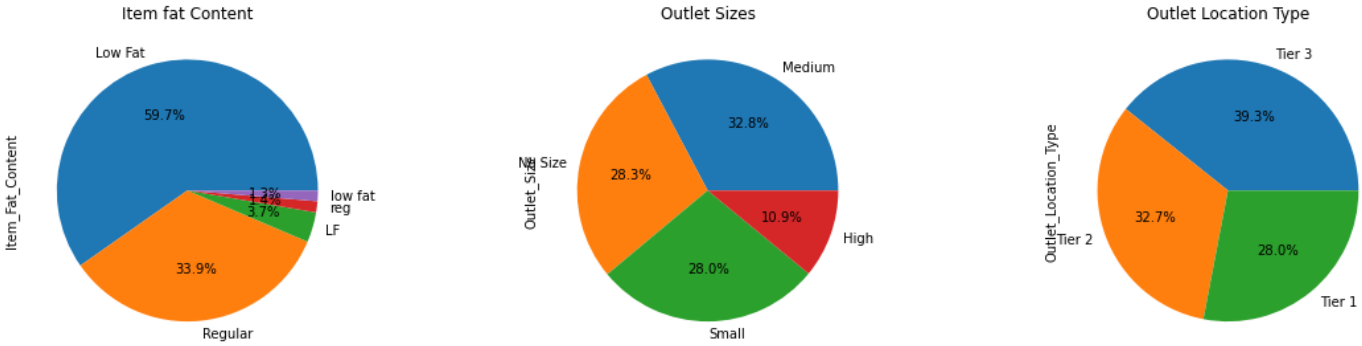
This bar graph gives us a clear information about the total sales of each Item type. As it can be seen Fruits and Vegetables tends to have the highest total count of sales as compared to other item types.

2. Item type by total item weight



This horizontal bar graph gives us the information about the total count of weights for each item type. Thus, it can be observed that Fruits and Vegetables bought into the outlets seems to have the highest total weight among other item types.

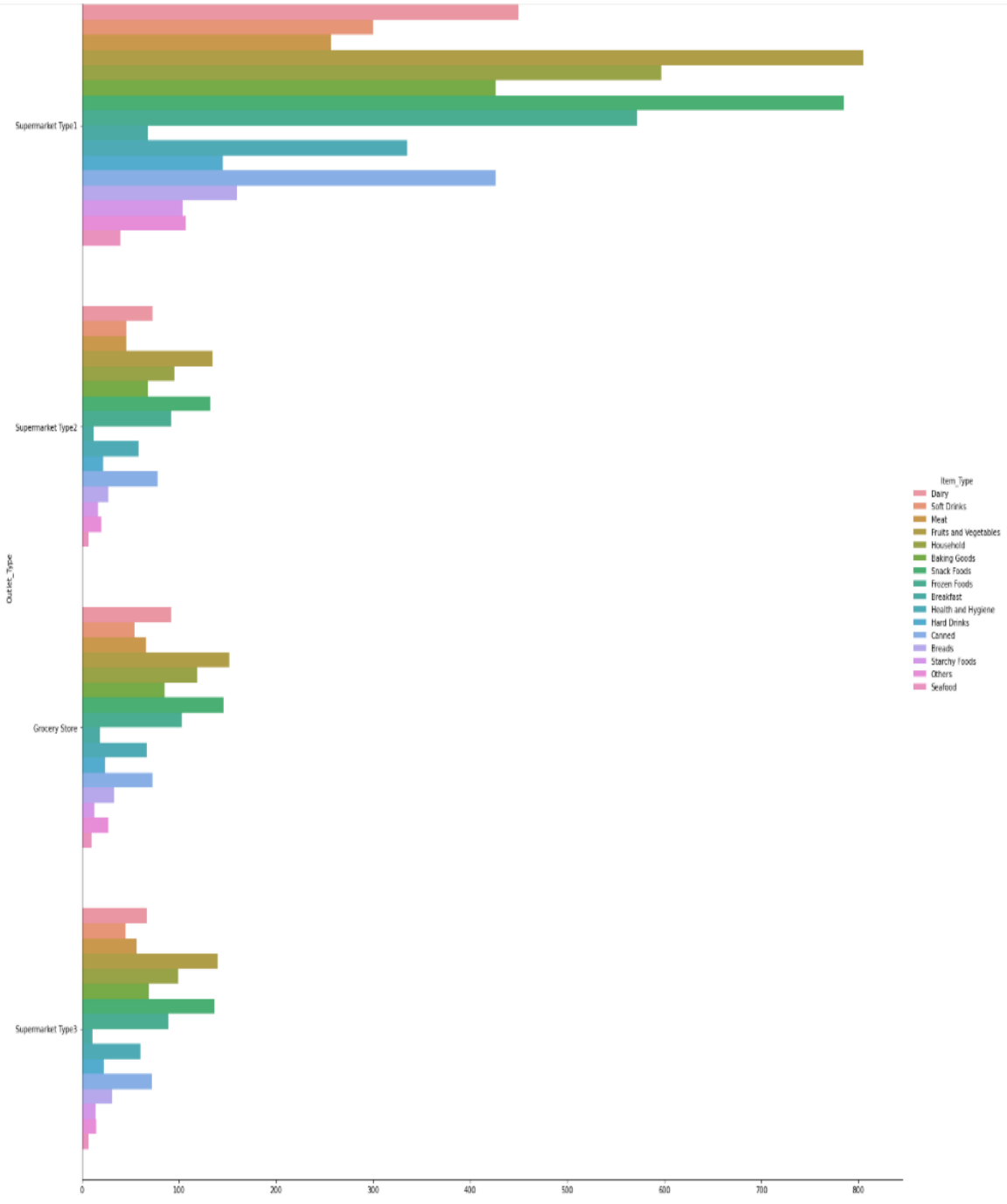
3. Pie Charts of Item fat content, Outlet Sizes and Outlet location type



This pie charts contains details about the Item fat content, Outlet Sizes and Outlet location type.

- In the Item fat Content, it can be seen that most items are made as Low Fat.
- In Outlet Sizes chart, Medium sized outlets are more on the list as compared to other outlet types.
- In Outlet Location Type, Tier 3 dominates among other location types of the items.

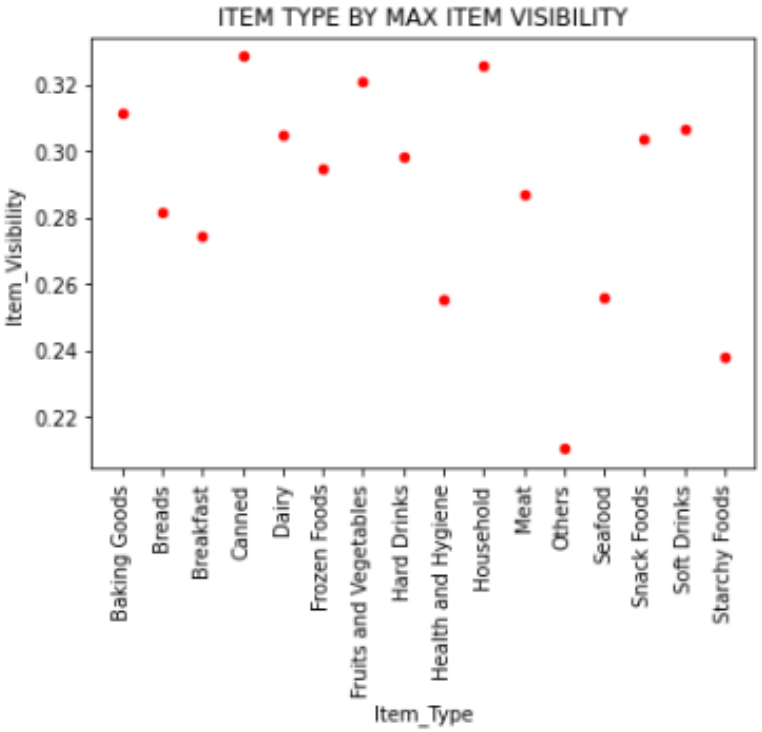
4. Different Outlet type by different item types.



This multiple bar graph depicts count of each item being sold in each outlet type. As it can be seen Supermarket Type 1 seems to be most populous

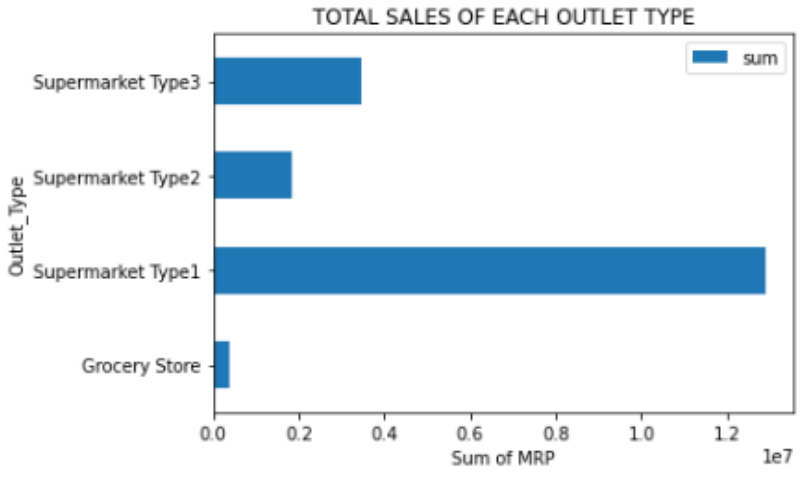
among other outlet type with Fruits and Vegetables being sold at maximum count.

5. Item type by maximum Item visibility



This scatter plot depicts the item type having maximum visibility. As it can be observed, Canned item type have the maximum Item visibility.

6. Total Sales at each Outlet type



This horizontal bar graph depicts the Outlet type having maximum count of sales. Thus, after observation, it can be seen that Supermarket Type1 has the highest count of sales among other Outlet types.

Training and Testing Data (using Linear Regression)

After performing the Exploratory Data Analysis, we started to train and test data for future predictions(in business point of view to state the KPI) of certain integer column of the data table. In this task I would be training MRP of items for Item Weights, Item visibility and Item outlet sales to calculate the Regression score and also to find out the Mean Squared error, Mean Absolute error and R^2 score of the predictions.

1. Regression score

   After fitting the x_train and y_train variables into Linear Regression Model, we calculated the Regression score of x_test and y_test, which comes out to be—

   0.29915893887725153

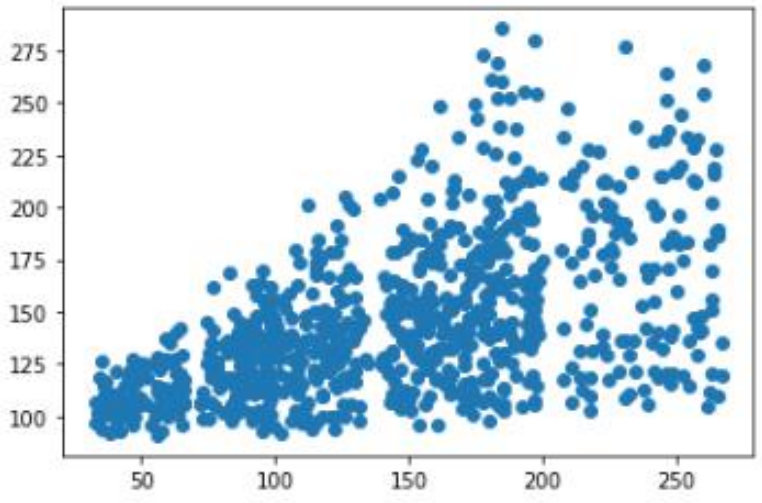
2. Actuals vs. Predicted Data frame and Scatter graph

   After predicting the values from the x_test values we form a data fram e containing the actual values and predicted values of MRP of Items.

	Actual	predicted
1112	75.2328	130.325549
1751	246.8460	236.196259
7648	87.8172	120.384990
7362	125.0730	105.703127
5332	102.5016	92.215062
...	...	...
8142	98.6068	142.872749
1190	178.3002	141.051784
7635	259.5962	148.240929
6156	256.1014	147.053340
4398	154.4630	151.329446

853 rows × 2 columns

The graph of Actual vs. Predicted that we get is—



3. Mean Squared error, Mean Absolute error and R^2 score of the predictions

   From sklearn we imported the metrics library to calculate the MSE, MAE and R^2 score of the regressions we perform—

   - MSE score- 51.31532246492415
   - MAE score- 41.82659848298155
   - R^2 score- 0.29915893887725153