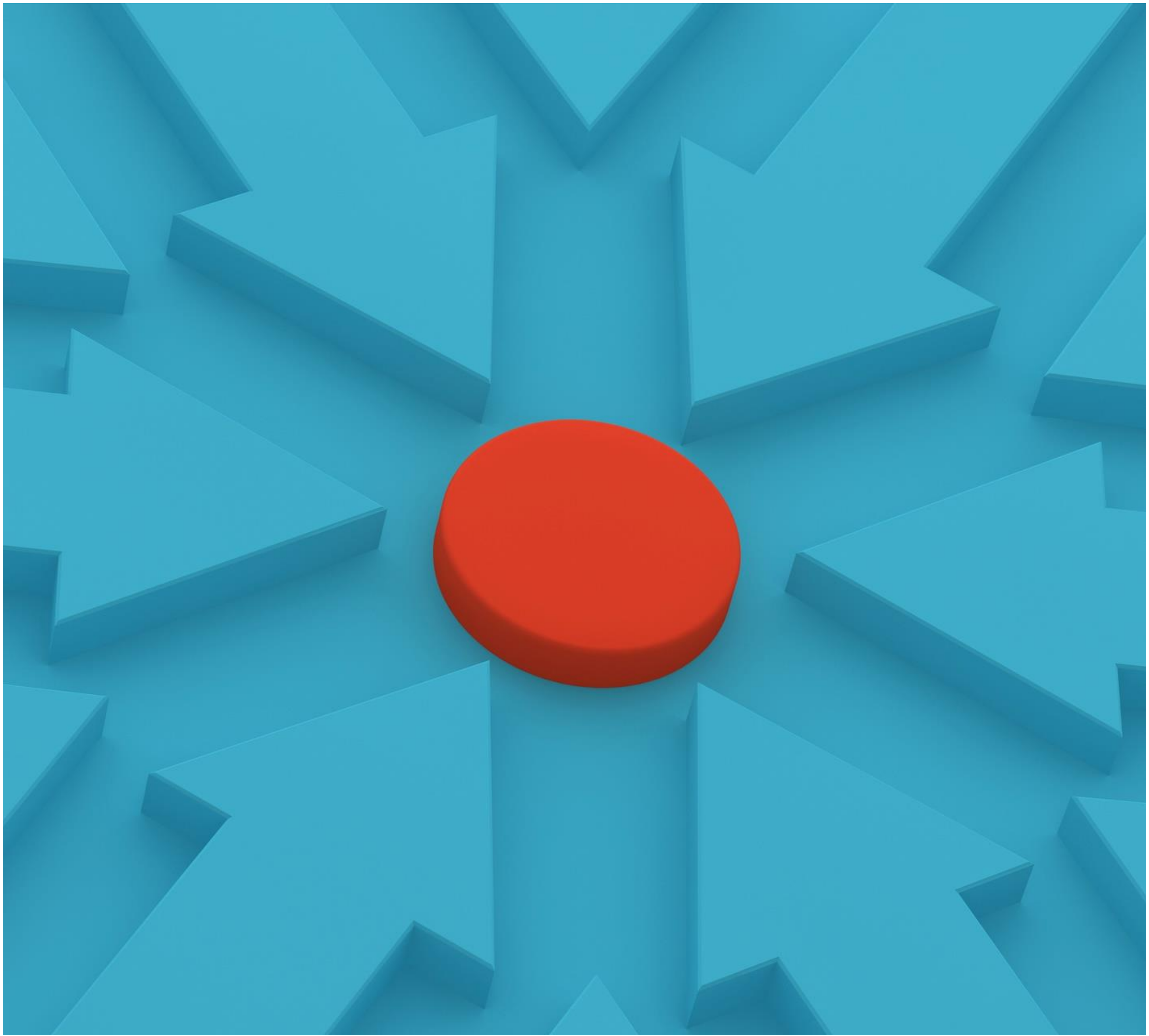
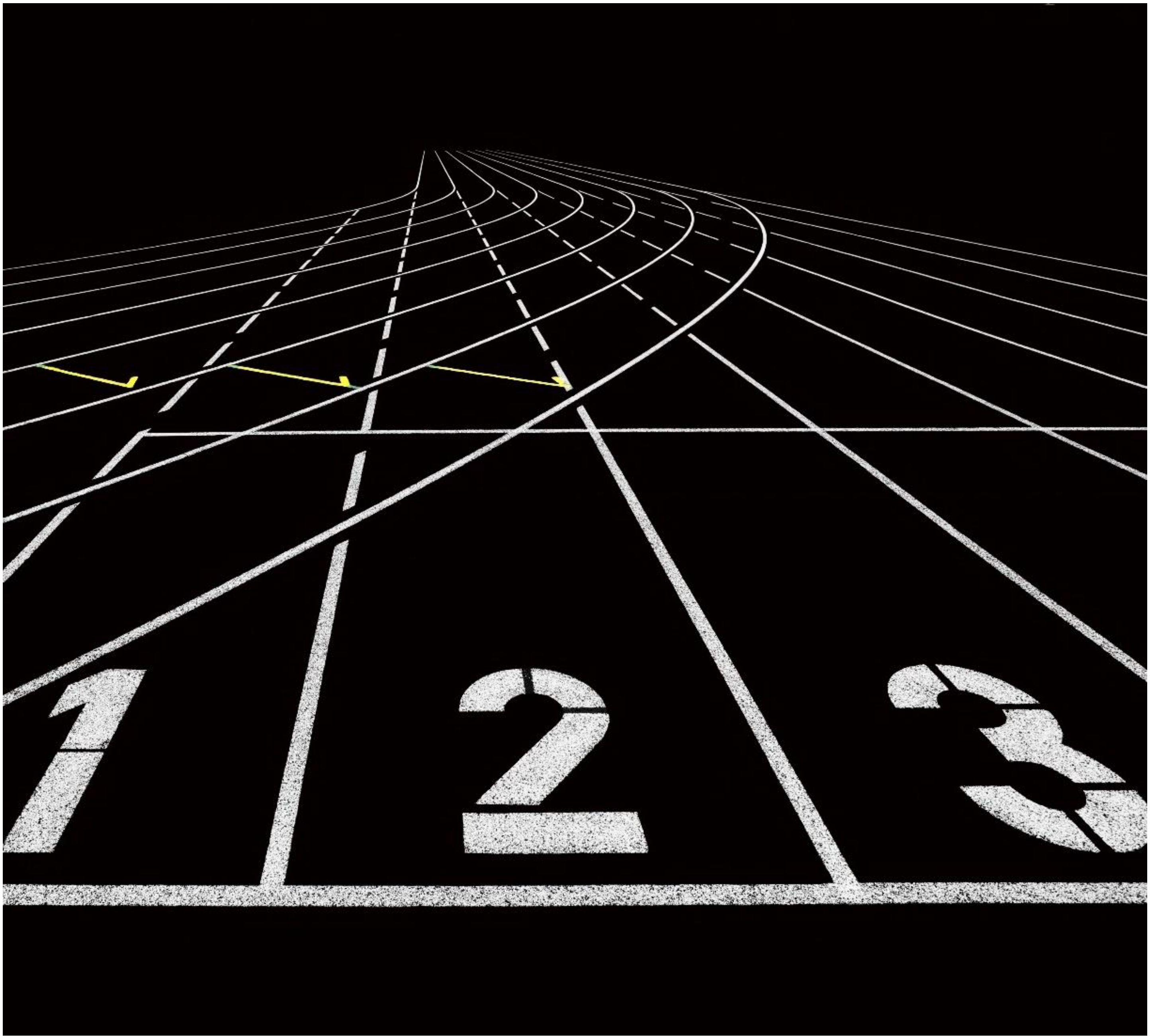


BIG MART SALES REPORT

ASSIGNMENT 1

PREPARED BY : BERNARD, LIONEL, JOEL, ETHAN



Contents

1.1 FRAMING THE PROBLEM

We must first understand the objective of working on this BigMart Sales Data. The first objective is to determine the products that has the potential to increase their sales based on their current sales value. The second objective is to determine the store that has the potential to increase their sales based on their current sales output on the products.

The problem of the missing data in for the products in some stores due to technical glitches has the impact of BigMart being unable to determine the store or products with the highest sales accurately. This will affect BigMart as they are not able to increase sales, so a good starting point would be to use the existing perfect data and work from there to determine the sales value of the product and the stores.

Based on the case study, because of the impacts are directed towards the revenue, we can assume that the stakeholder could be a leadership-level person on the marketing team, or it could also be someone from the executive level, which prioritizes revenue over the maintainability of the product.

In order to achieve the objectives, we must first determine the factors that affect the sales of the product. This can be ranged from the size and location of the store down to the product itself. Next, we can prepare the data by identifying the trends and patterns based on the cleaned data. The model will be built and trained to identify how much of an error can the business tolerate. Finally, we can make the initial predictions and to further refine them later. The expected outcome would be to successfully predict the products and stores that have potential to increase their sales.

a. Clean the dataset and record all the changes.

i. Read BigMart_Data.csv –
File Reader → Insert BigMart_Data.csv

ii. Remove the missing values from Outlet_size –
Row Filter → Exclude rows by attribute values, Column = Outlet_size, select “only missing values match”

iii. Remove the missing values from item_weight –
Row Filter → Exclude rows by attribute values, Column = item_size, select “only missing values match”

iv. Groping inconsistent values “LF, low fat and reg” to the new values “Low Fat and Regular” –
Rule Engine → $Item_Fat_Content$ LIKE "LF" => "Low Fat"
$Item_Fat_Content$ LIKE "low fat" => "Low Fat"
$Item_Fat_Content$ LIKE "reg" => "Regular"
$Item_Fat_Content$ LIKE "Low Fat" => "Low Fat"
$Item_Fat_Content$ LIKE "Regular" => "Regular"
Replace column “Item_Fat_Content”

v. Group by item fat content with their average and highest sales –
GroupBy → Include Item_Fat_Content, Aggregation → Item_Outlet_Sales (Mean), Item_Outlet_Sales (Max)

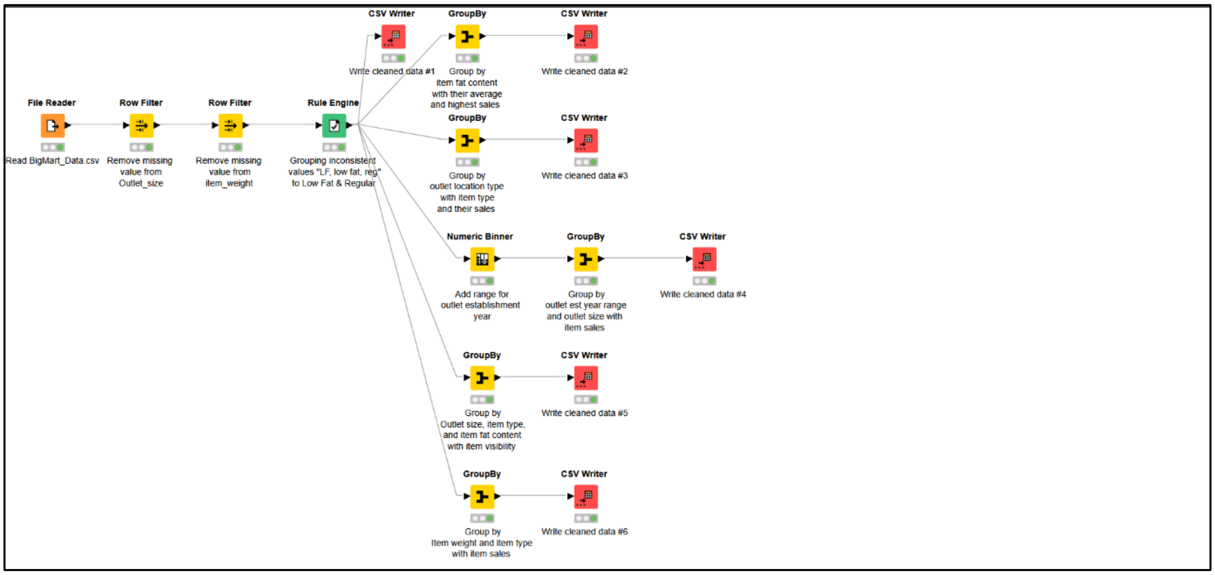
vi. Group by outlet location type with their item types and sales –
GroupBy → Include Item_Type + Outlet_Location_Type, Aggregation → Item_Outlet_Sales (Sum)

vii. Group by outlet establishment year and outlet size with item sales –
Numeric Binner → Select Outlet_Establishment_Year, Add bin (∞ … 1990), (1990 … 2000), (2000 … ∞)
GroupBy → Include Outlet_Establishment_Year + Outlet_Size, Aggregation → Item_Outlet_Sales (Mean), Item_Outlet_Sales (Sum)

viii. Group by outlet size, item type and item fat content with item visibility –
GroupBy → Item_Fat_Content + Item_Type + Outlet_Size, Aggregation → Item_Visibility (Max)

ix. Group by item weight and item type with item sales –
GroupBy → Item_Weight + Item_Type, Aggregation → Item_Outlet_Sales (Max)

<u>Screenshot of Workflow</u>


File Reader
Read BigMart_Data.csv
Row Filter
Remove missing value from Outlet_size
Row Filter
Remove missing value from item_weight
Rule Engine
Grouping inconsistent values "LF, low fat, reg" to Low Fat & Regular
CSV Writer
Write cleaned data #1
GroupBy
Group by item fat content with their average and highest sales
CSV Writer
Write cleaned data #2
GroupBy
Group by outlet location type with item type and their sales
CSV Writer
Write cleaned data #3
Numeric Binner
Add range for outlet establishment year
GroupBy
Group by outlet est year range and outlet size with item sales
CSV Writer
Write cleaned data #4
GroupBy
Group by Outlet size, item type, and item fat content with item visibility
CSV Writer
Write cleaned data #5
GroupBy
Group by Item weight and item type with item sales
CSV Writer
Write cleaned data #6

b. Identify dependent and independent variables + determine data type.

Dependent variable	Independent variable
Item_Outlet_Sales (Continuous)	Item_Identifier (Categorical) Item_Weight (Continuous) Item_Fat_Content (Categorical) Item_Type(Categorical) Item_MRP (Continuous) Outlet_Identifier (Categorical) Outlet_Establishment_Year (Continous) Outlet_Size (Categorical) Outlet_Location_Type (Categorical) Outlet_Type (Categorical)

	Item_Visibility (Continuous)

c. Find relevance of each independent variable for prediction of dependent variable.

i. Outlet_Location_Type + Item_Type → Item_Outlet_Sales
The sales can be affected by the outlet location type and the item type because we can find out the item types that sell best at every outlet location.

ii. Outlet_Establishment_Year + Outlet_Size → Item_Outlet_Sales
The sales can be affected by the year the outlet is established as well as the outlet size. This is because customers might prefer new and larger stores rather than older and smaller variants.

iii. Item_Fat_Content → Item_Outlet_Sales
The sales of a product may depend on its fat content. However, this does not mean that more is better or vice versa.

iv. Outlet_Size + Item_Type + Item_Fat_Content → Item_Visibility
The visibility of a product can be affected by the size of the outlet as well as the product type and its fat content. This is because larger outlets store more products, making them more available to the customers. However, the visibility alone will not determine the sales of a product as higher product visibility might be a sign that that product did not leave the shelf often.

v. Item_Weight + Item_Type → Item_Outlet_Sales
The sales of a product can be affected by its weight and type. This is because we can determine the weight of a given product that sells best.

d. Develop (5) store/product-level hypothesis.

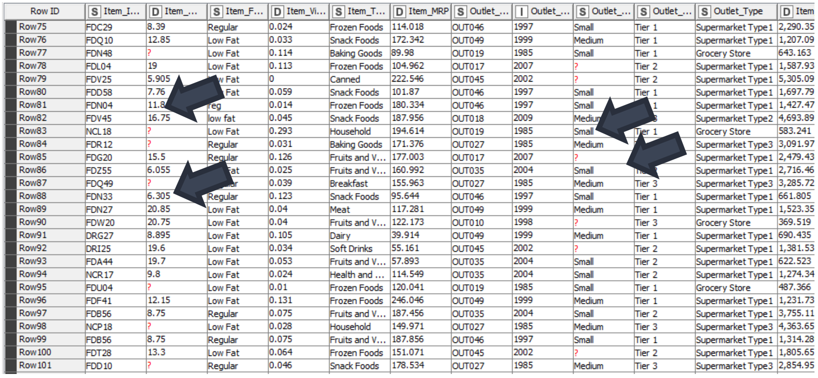
- Stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place.
- Low fat products have higher sales compared to product with regular fat content.
- Stores located in urban, or Tier 1 cities should have higher sales because of the higher income levels of people there.
- Medium sized stores have higher visibility of household items because the low demand for them.
- Products that weigh heavier have higher sales because they pack more content.

2.1 CONSTRUCTING A PREDICTION DATA SET

Step 1 – File reading

1. KNIME was used to execute this task. The "File Reader" node was used to read the pre-processed.csv file.

Result of read file



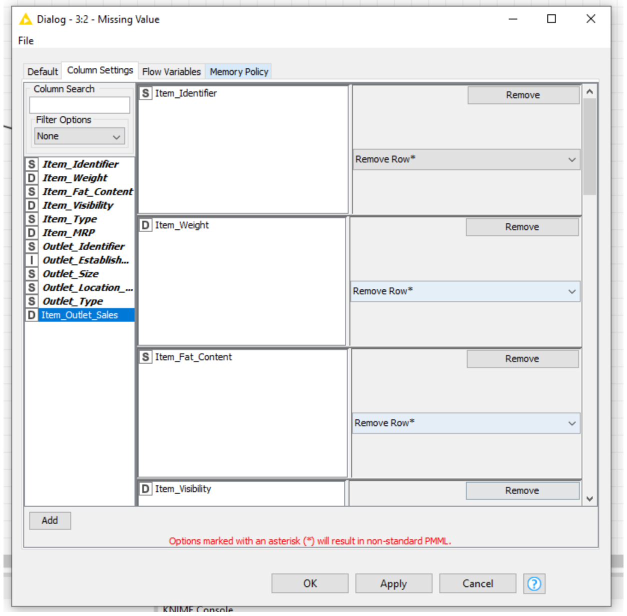
Row ID	S Item_I...	D Item_...	S Item_F...	D Item_Vi...	S Item_T...	D Item_MRP	S Outlet_...	I Outlet_...	S Outlet_...	S Outlet_...	S Outlet_Type	D Item
Row75	FDC29	8.39	Regular	0.024	Frozen Foods	114.018	OUT046	1997	Small	Tier 1	Supermarket Type1	2,290.35
Row76	FDQ10	12.85	Low Fat	0.033	Snack Foods	172.342	OUT049	1999	Medium	Tier 1	Supermarket Type1	1,207.09
Row77	FDN48	?	Low Fat	0.114	Baking Goods	89.98	OUT019	1985	Small	Tier 1	Grocery Store	643.163
Row78	FDL04	19	Low Fat	0.113	Frozen Foods	104.962	OUT017	2007	?	Tier 2	Supermarket Type1	1,587.93
Row79	FDV25	5.905	Low Fat	0	Canned	222.546	OUT045	2002	?	Tier 2	Supermarket Type1	5,305.09
Row80	FDD58	7.76	Low Fat	0.059	Snack Foods	101.87	OUT046	1997	Small	Tier 1	Supermarket Type1	1,697.79
Row81	FDN04	11.8	reg	0.014	Frozen Foods	180.334	OUT046	1997	Small	Tier 1	Supermarket Type1	1,427.47
Row82	FDV45	16.75	low fat	0.045	Snack Foods	187.956	OUT018	2009	Medium	Tier 3	Supermarket Type2	4,693.89
Row83	NCL18	?	Low Fat	0.293	Household	194.614	OUT019	1985	Small	Tier 1	Grocery Store	583.241
Row84	FDR12	?	Regular	0.031	Baking Goods	171.376	OUT027	1985	Medium	[illegible]	Supermarket Type3	3,091.97
Row85	FDG20	15.5	Regular	0.126	Fruits and V...	177.003	OUT017	2007	?	[illegible]	Supermarket Type1	2,479.43
Row86	FDZ55	6.055	Low Fat	0.025	Fruits and V...	160.992	OUT035	2004	Small	Tier 2	Supermarket Type1	2,716.46
Row87	FDQ49	?	Regular	0.039	Breakfast	155.963	OUT027	1985	Medium	Tier 3	Supermarket Type3	3,285.72
Row88	FDN33	6.305	Regular	0.123	Snack Foods	95.644	OUT046	1997	Small	Tier 1	Supermarket Type1	661.805
Row89	FDN27	20.85	Low Fat	0.04	Meat	117.281	OUT049	1999	Medium	Tier 1	Supermarket Type1	1,523.35
Row90	FDW20	20.75	Low Fat	0.04	Fruits and V...	122.173	OUT010	1998	?	Tier 3	Grocery Store	369.519
Row91	DRG27	8.895	Low Fat	0.105	Dairy	39.914	OUT049	1999	Medium	Tier 1	Supermarket Type1	690.435
Row92	DRI25	19.6	Low Fat	0.034	Soft Drinks	55.161	OUT045	2002	?	Tier 2	Supermarket Type1	1,381.53
Row93	FDA44	19.7	Low Fat	0.053	Fruits and V...	57.893	OUT035	2004	Small	Tier 2	Supermarket Type1	622.523
Row94	NCR17	9.8	Low Fat	0.024	Health and ...	114.549	OUT035	2004	Small	Tier 2	Supermarket Type1	1,274.34
Row95	FDU04	?	Low Fat	0.01	Frozen Foods	120.041	OUT019	1985	Small	Tier 1	Grocery Store	487.366
Row96	FDF41	12.15	Low Fat	0.131	Frozen Foods	246.046	OUT049	1999	Medium	Tier 1	Supermarket Type1	1,231.73
Row97	FDB56	8.75	Regular	0.075	Fruits and V...	187.456	OUT035	2004	Small	Tier 2	Supermarket Type1	3,755.11
Row98	NCP18	?	Low Fat	0.028	Household	149.971	OUT027	1985	Medium	Tier 3	Supermarket Type3	4,363.65
Row99	FDB56	8.75	Regular	0.075	Fruits and V...	187.856	OUT046	1997	Small	Tier 1	Supermarket Type1	1,314.28
Row100	FDT28	13.3	Low Fat	0.064	Frozen Foods	151.071	OUT045	2002	?	Tier 2	Supermarket Type1	1,805.65
Row101	FDD10	?	Regular	0.046	Snack Foods	178.534	OUT027	1985	Medium	Tier 3	Supermarket Type3	2,854.95

Input strategies:

- KNIME new workflow - pre-processed.csv was loaded into KNIME by importing it
- File Reader – This node needed some configuration, such as setting the file location so that the reader can locate it
- After the file is read, certain columns had missing values indicated with a ?
- To remove the missing values from the file, step 2 is used and explained below

Step 2- Removing missing values

2. Using the "Missing Value" node, missing values in all the rows were filtered out. This filter needed some parameters to be inputted such as (Remove Row) for all column. After setting the parameters, the node would automatically remove the missing values row



Filter Options: None

S Item_Identifier
D Item_Weight
S Item_Fat_Content
D Item_Visibility
S Item_Type
D Item_MRP
S Outlet_Identifier
I Outlet_Establish...
S Outlet_Size
S Outlet_Location_...
S Outlet_Type
D Item_Outlet_Sales

S Item_Identifier: Remove Row*

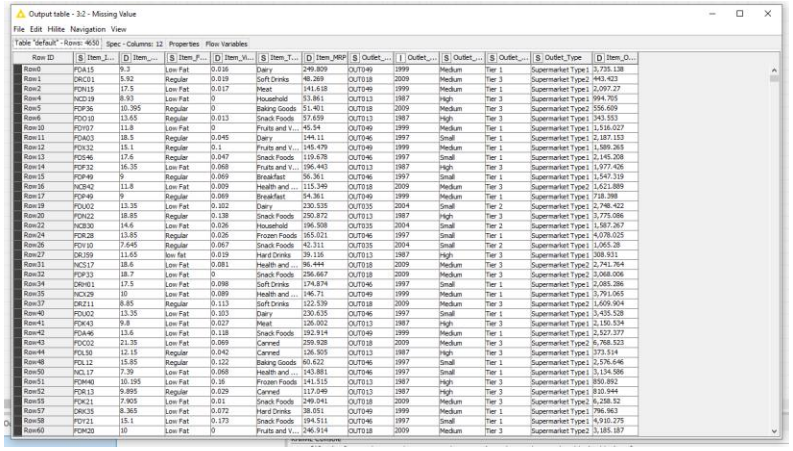
D Item_Weight: Remove Row*

S Item_Fat_Content: Remove Row*

D Item_Visibility

Options marked with an asterisk (*) will result in non-standard PMML.

After filtering the missing values



Output table - 3:2 - Missing Value

File Edit Hilite Navigation View

Table "default" - Rows: 4650 | Spec - Columns: 12 | Properties | Flow Variables

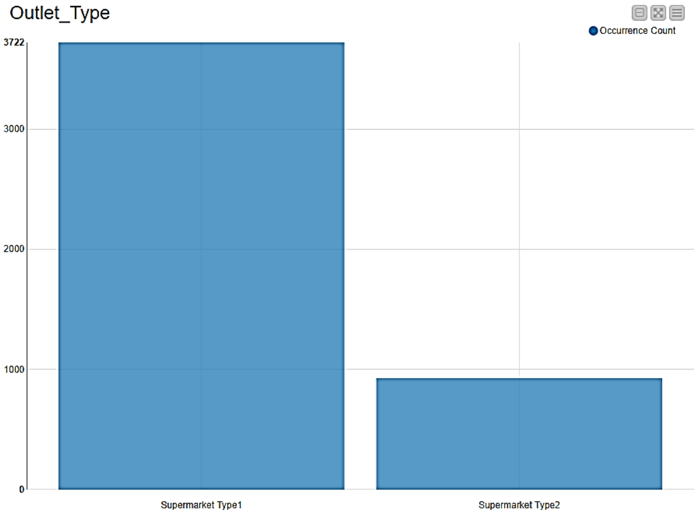
Row ID	S Item_I...	D Item_...	S Item_F...	D Item_Vi...	S Item_T...	D Item_MRP	S Outlet_...	I Outlet_...	S Outlet_...	S Outlet_...	S Outlet_Type	D Item_O...
Row0	FDA15	9.3	Low Fat	0.016	Dairy	249.809	OUT049	1999	Medium	Tier 1	Supermarket Type1	3,735.138
Row1	DRC01	5.92	Regular	0.019	Soft Drinks	48.269	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.423
Row2	FDN15	17.5	Low Fat	0.017	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2,097.27
Row4	NCD19	8.93	Low Fat	0	Household	53.861	OUT013	1987	High	Tier 3	Supermarket Type1	994.705
Row5	FDP36	10.395	Regular	0	Baking Goods	51.401	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.609
Row6	FDO10	13.65	Regular	0.013	Snack Foods	57.659	OUT013	1987	High	Tier 3	Supermarket Type1	343.553
Row10	FDY07	11.8	Low Fat	0	Fruits and V...	45.54	OUT049	1999	Medium	Tier 1	Supermarket Type1	1,516.027
Row11	FDA03	18.5	Regular	0.045	Dairy	144.11	OUT046	1997	Small	Tier 1	Supermarket Type1	2,187.153
Row12	FDX32	15.1	Regular	0.1	Fruits and V...	145.479	OUT049	1999	Medium	Tier 1	Supermarket Type1	1,589.265
Row13	FDS46	17.6	Regular	0.047	Snack Foods	119.678	OUT046	1997	Small	Tier 1	Supermarket Type1	2,145.208
Row14	FDF32	16.35	Low Fat	0.068	Fruits and V...	196.443	OUT013	1987	High	Tier 3	Supermarket Type1	1,977.426
Row15	FDP49	9	Regular	0.069	Breakfast	56.361	OUT046	1997	Small	Tier 1	Supermarket Type1	1,547.319
Row16	NCB42	11.8	Low Fat	0.009	Health and ...	115.349	OUT018	2009	Medium	Tier 3	Supermarket Type2	1,621.889
Row17	FDP49	9	Regular	0.069	Breakfast	54.361	OUT049	1999	Medium	Tier 1	Supermarket Type1	718.398
Row19	FDU02	13.35	Low Fat	0.102	Dairy	230.535	OUT035	2004	Small	Tier 2	Supermarket Type1	2,748.422
Row20	FDN22	18.85	Regular	0.138	Snack Foods	250.872	OUT013	1987	High	Tier 3	Supermarket Type1	3,775.086
Row22	NCB30	14.6	Low Fat	0.026	Household	196.508	OUT035	2004	Small	Tier 2	Supermarket Type1	1,587.267
Row24	FDR28	13.85	Regular	0.026	Frozen Foods	165.021	OUT046	1997	Small	Tier 1	Supermarket Type1	4,078.025
Row26	FDV10	7.645	Regular	0.067	Snack Foods	42.311	OUT035	2004	Small	Tier 2	Supermarket Type1	1,065.28
Row27	DRJ59	11.65	low fat	0.019	Hard Drinks	39.116	OUT013	1987	High	Tier 3	Supermarket Type1	308.931
Row31	NCS17	18.6	Low Fat	0.081	Health and ...	96.444	OUT018	2009	Medium	Tier 3	Supermarket Type2	2,741.764
Row32	FDP33	18.7	Low Fat	0	Snack Foods	256.667	OUT018	2009	Medium	Tier 3	Supermarket Type2	3,068.006
Row34	DRH01	17.5	Low Fat	0.098	Soft Drinks	174.874	OUT046	1997	Small	Tier 1	Supermarket Type1	2,085.286
Row35	NCX29	10	Low Fat	0.089	Health and ...	146.71	OUT049	1999	Medium	Tier 1	Supermarket Type1	3,791.065
Row37	DRZ11	8.85	Regular	0.113	Soft Drinks	122.539	OUT018	2009	Medium	Tier 3	Supermarket Type2	1,609.904
Row40	FDU02	13.35	Low Fat	0.103	Dairy	230.635	OUT046	1997	Small	Tier 1	Supermarket Type1	3,435.528
Row41	FDK43	9.8	Low Fat	0.027	Meat	126.002	OUT013	1987	High	Tier 3	Supermarket Type1	2,150.534
Row42	FDA46	13.6	Low Fat	0.118	Snack Foods	192.914	OUT049	1999	Medium	Tier 1	Supermarket Type1	2,527.377
Row43	FDC02	21.35	Low Fat	0.069	Canned	259.928	OUT018	2009	Medium	Tier 3	Supermarket Type2	6,768.523
Row44	FDL50	12.15	Regular	0.042	Canned	126.505	OUT013	1987	High	Tier 3	Supermarket Type1	373.514
Row48	FDL12	15.85	Regular	0.122	Baking Goods	60.622	OUT046	1997	Small	Tier 1	Supermarket Type1	2,576.646
Row50	NCL17	7.39	Low Fat	0.068	Health and ...	143.881	OUT046	1997	Small	Tier 1	Supermarket Type1	3,134.586
Row51	FDM40	10.195	Low Fat	0.16	Frozen Foods	141.515	OUT013	1987	High	Tier 3	Supermarket Type1	850.892
Row52	FDR13	9.895	Regular	0.029	Canned	117.049	OUT013	1987	High	Tier 3	Supermarket Type1	810.944
Row55	FDK21	7.905	Low Fat	0.01	Snack Foods	249.041	OUT018	2009	Medium	Tier 3	Supermarket Type2	6,258.52
Row57	DRK35	8.365	Low Fat	0.072	Hard Drinks	38.051	OUT049	1999	Medium	Tier 1	Supermarket Type1	796.963
Row58	FDY21	15.1	Low Fat	0.173	Snack Foods	194.511	OUT046	1997	Small	Tier 1	Supermarket Type1	4,910.275
Row60	FDM20	10	Low Fat	0	Fruits and V...	246.914	OUT018	2009	Medium	Tier 3	Supermarket Type2	3,185.187

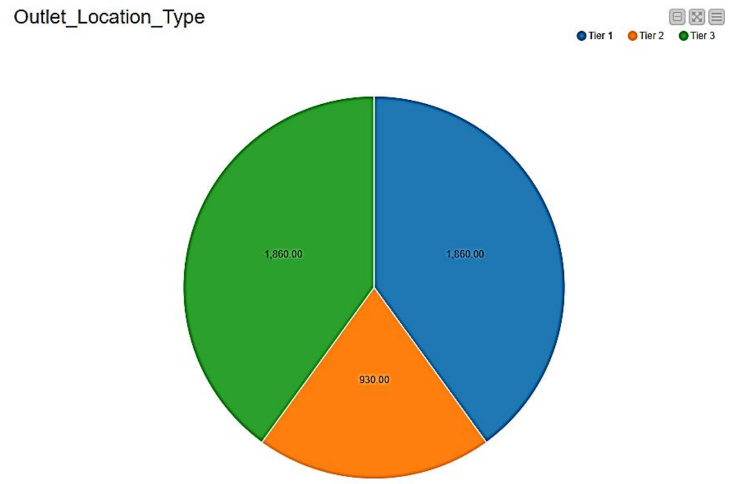
Input strategies:

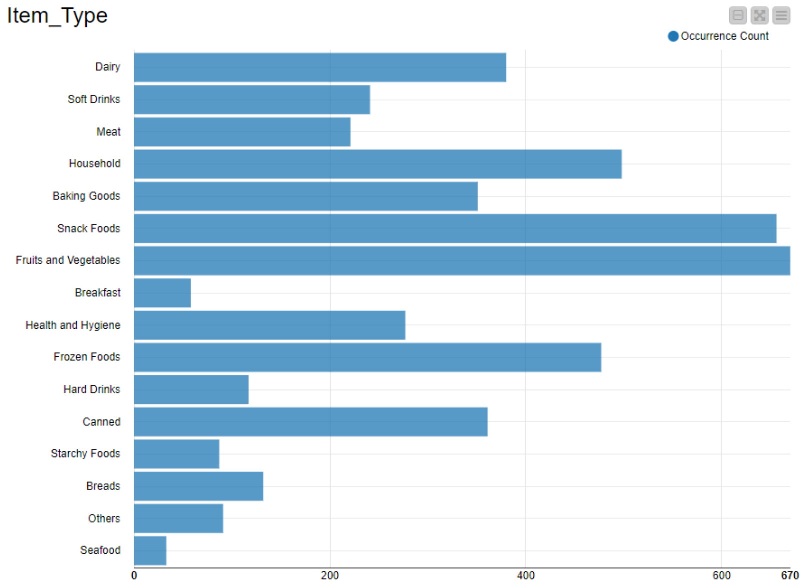
- Missing value – The parameters to remove the missing values in the file is explained above.
- Results – In the 2nd picture, we can see that there is no "?" anymore, which confirms that missing values have been removed
- To help with the visualization process, visual representations are used to assist in interpreting the data because reading from a csv file may be confusing for certain people
- The visualization representations are shown in step 3 such as using charts

<u>Step 3 – Visual representations</u>

3. Both the bar & pie chart was used to represent the data visually. There simply is too many data to show so 3 category types were chosen. The 3 categories are as below
   Outlet type, Location type, and Item type followed by the graphs below.







Explanations

These explanations help to describe the transformation of data from numerical to categorical data better. For the **outlet type,** bar charts are used to represent this data type

<u>Outlet Type</u>

Based on this data type, there are only 2 types of supermarkets which consists of (Type 1 & Type 2). From the graph, we can see that **Supermarket Type 1** has higher occurrences compared to Type 2 from various aspects such as the item type, outlet establishment year, outlet size, and many other attributed that can be found from the file. However, this graph gives an overview of the data.

Outline Location Type

Based on this data type, there are 3 main outlet location types (Tier 1 , Tier 2, Tier 3). Both Tier 1 and Tier 3 has almost an equal amount of 1860 occurrences. Tier 2 has the least occurrences which is 930. Based on this data, we can conclude that this company has more Tier 1 and Tier 3 supermarkets in the particular region probably due to the amount of customers around the area. In addition, Tier 2 supermarkets consist of a smaller demographic which can explain why Tier 2 has less supermarkets

Item Type

Based on this data type, there are 3 mostly bought items from the store consisting of (Fruits & Vegetables , Snack foods, Household).The highest bought is fruits and vegetables followed by snack foods, and household items. Based on this data, we can conclude that there's a high number of healthy eaters and unhealthy eaters at the same time. Healthy eaters are slightly more than unhealthy eaters.

From the household items, we can conclude there the area is a matured area, with more people living in the area which is why household items is the 3rd most bought items.

Step 4. Partitioning the data

The following step was to begin partitioning the data set. Why partition the data set ? According to *(Data Partition - Statistics.com: Data Science, Analytics & Statistics Courses, 2021)*, partitioning a data is used when the data scientist is trying to choose a type of model from a broad list of models. The main concept of partitioning the dataset, is to keep a small percentage of the data to be used for verification purposes after the model has been developed.

Moreover, data partitioning is normally used in supervised learning in data mining which is an overarching concept under the **Data Preparation** phase. Normally, the performance of the model will be evaluated using the training data set.

To put this into context of the project, the partitioning step for the data is explained below.

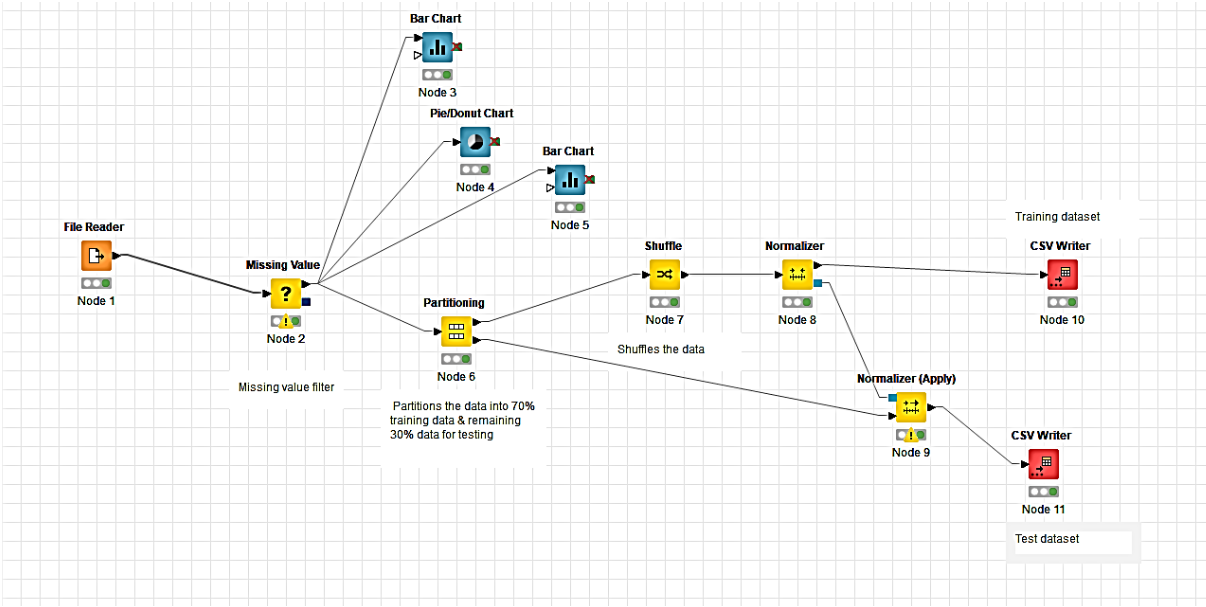
4. The data is partitioned into 70% training data and 30% test data. The “Partition" node then links to the "Shuffle" node to randomize the data

   Normalization node was used to normalize data.

   - The upper gate puts standardized data into the training set.
   - The light blue square gate below is to repeat the same normalized data for another node "Normalization (Apply)" node applies normalization from "Normalization" and get 30% of original file data from the lower node of "Partitioning".

   CSV Writer: 2 nodes

   - Top node is connected to the arrow gate of Normalization to get 70% of data has been shuffled and normalized -> train.csv.
   - Bottom node is connected to the arrow gate of “Normalization (Apply)” to get 30% of normalized data ->test.csv file



Missing value filter

Partitions the data into 70% training data & remaining 30% data for testing

Shuffles the data

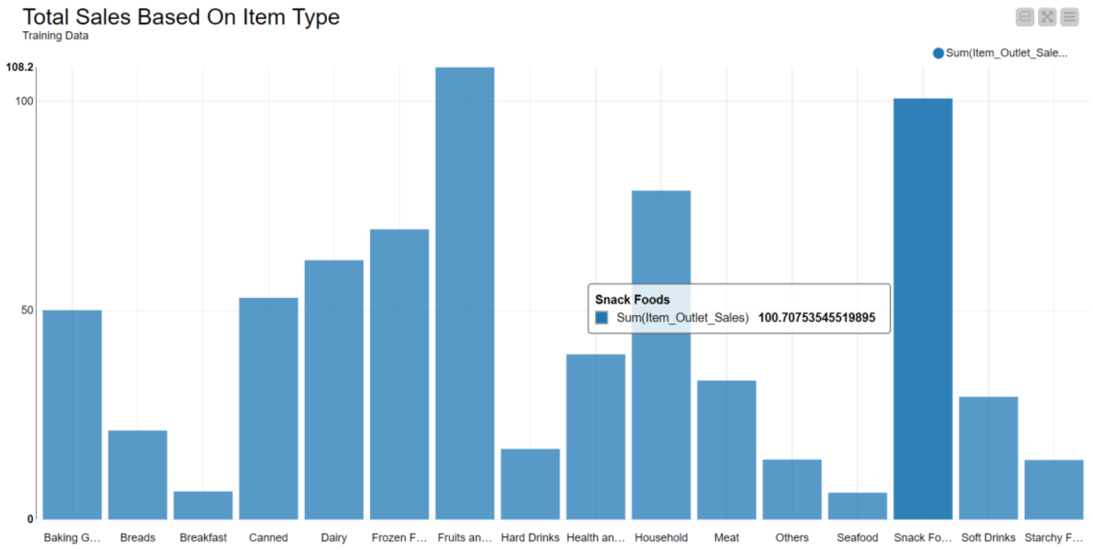
This image shows the entire workflow from reading the file to partitioning the dataset.

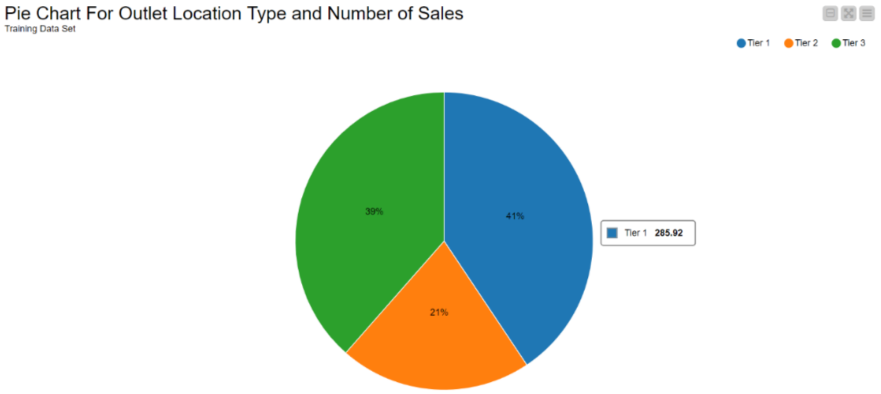
3.1 PROS AND CONS OF TREATING THE OUTPUT VARIABLE AS EITHER CATEGORICAL OR CONTINUOUS

Categorical Data.

Pros

Among the pros of treating the output variable as a Categorical Variable is that it will be easier to present it to the stakeholders. If the output variable is treated as a categorical variable the most probable way to analyse it will be through grouping. By grouping these data we get to compare them to the predictor variables(independent variables) which will allow for an easier understanding or visualization of the data. As data visualization with grouped data can be done in the form of pie charts and bar charts. Hence, it will be easier for the stakeholders to understand the data that is being presented to them. Below are examples.



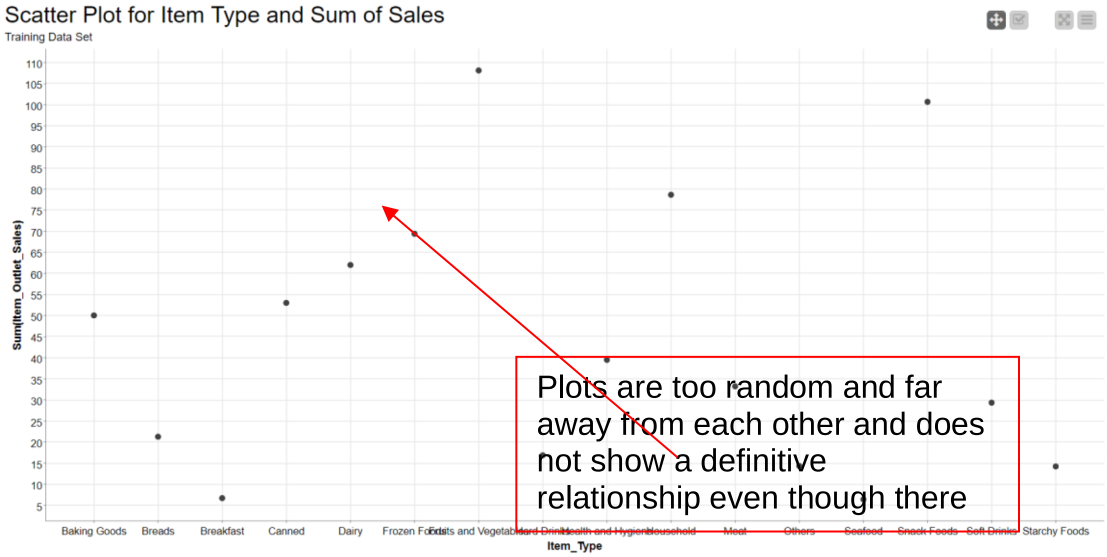


Furthermore, it is easier to identify relationships between the independent variable and output variable when we make the output variable as a categorical variable. This is also because of how the data has been reduced from a large amount of data into small group which makes analysis of the relationship easier.

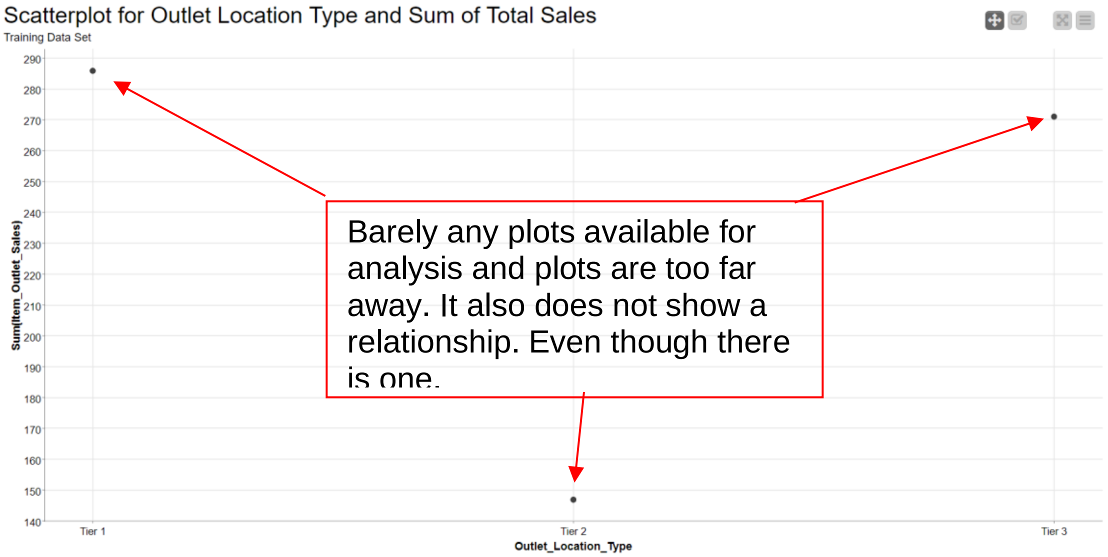
Cons

Large amounts of data goes missing. This is because categorical variables usually show us the data/value in categories. In this case it only shows us the total sum of each item type based on the total number of occurrences of the said item type. For example, let us say that in the year 2005 due to economic problems in the country the value of bread was lower than its usual. This value will not be shown as all the values would be added up together when the data is presented. This could lead to predicting future sales of these kind of item to be affected.

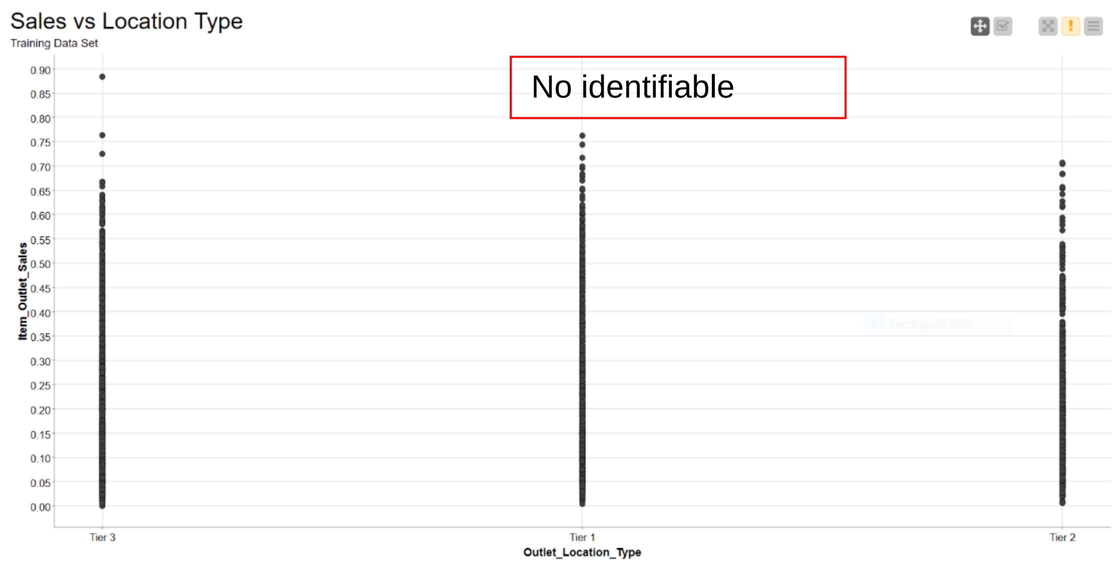
Besides that, most prediction models require a numeric form of data to conduct analysis. For example, most of the algorithms (or ML libraries) produce better result with numerical variable. In python, library “sklearn” requires features in numerical arrays. If a categorical variable is added there will be an error and the programme would fail. Since, it is harder to get categorical variables to fit in a regression equation and regression equations are essential to machine learning models(prediction models). In this case all the categorical variables have very large gaps between each other and have very little plots on the regression charts therefore it would be hard to plot a line of best fit or so on to later have the model use to analyse and predict an outcome from. Even if we did not categorized(group) the output variable there will still be no pattern in the regression equation to analysis to identify a relationship from. Below are the examples.



Plots are too random and far away from each other and does not show a definitive relationship even though there



Barely any plots available for analysis and plots are too far away. It also does not show a relationship. Even though there is one.



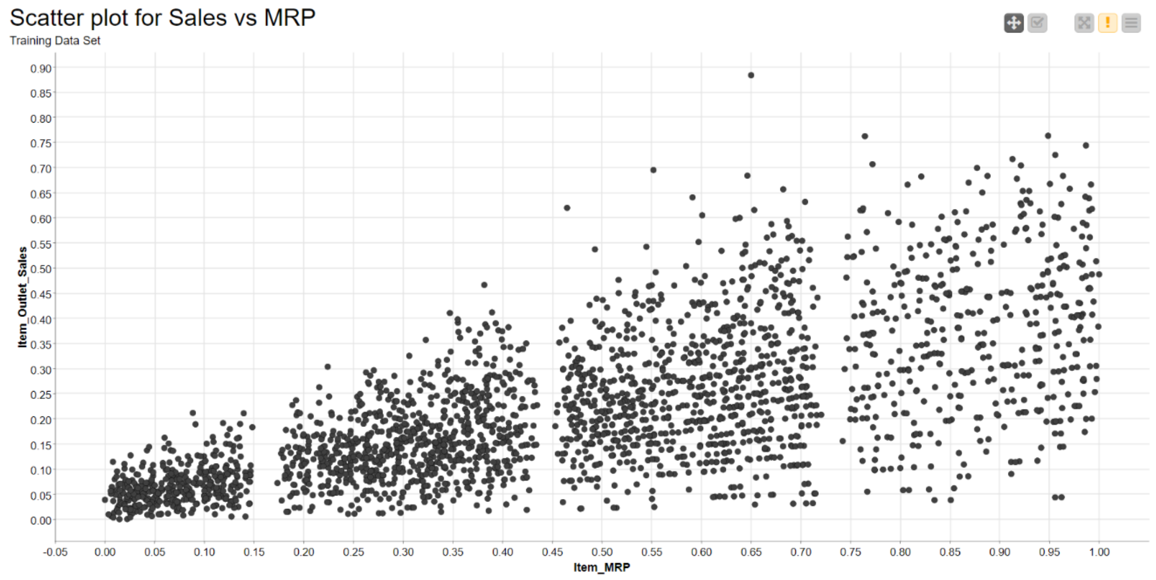
No identifiable

Furthermore, the task of converting categorical data to numeric data through encoding is too tedious and time consuming which in turn may increase operational cost for creating a prediction model as the data science team may need to put in more manpower and time to convert all the categorical variables to numeric variables(binary for example) via an encoding method such one hot encoding since there are some categorical variables that have ordinal attributes.

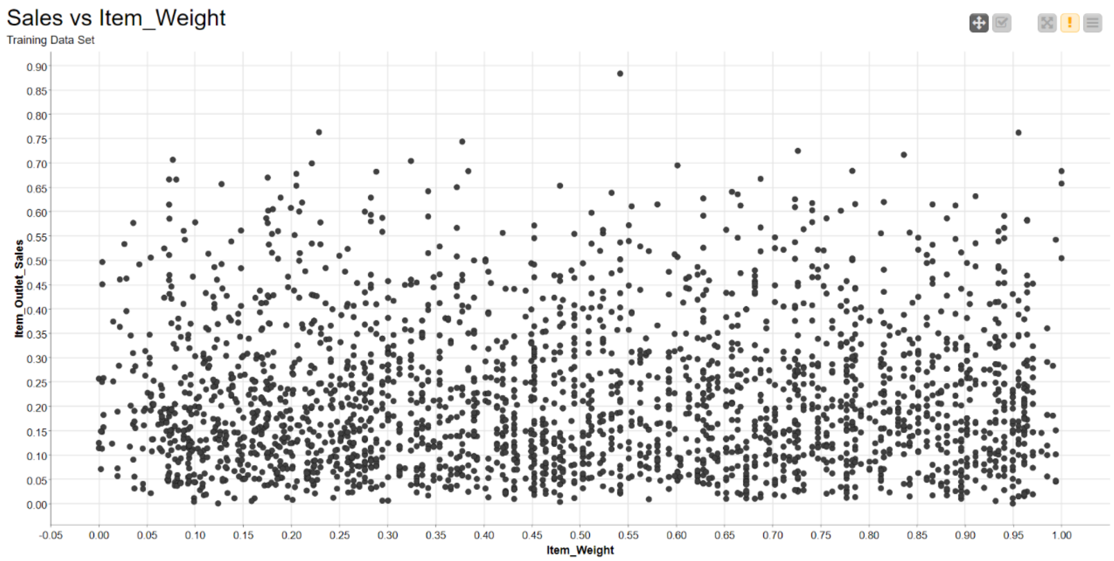
Continuous Variable

Pros

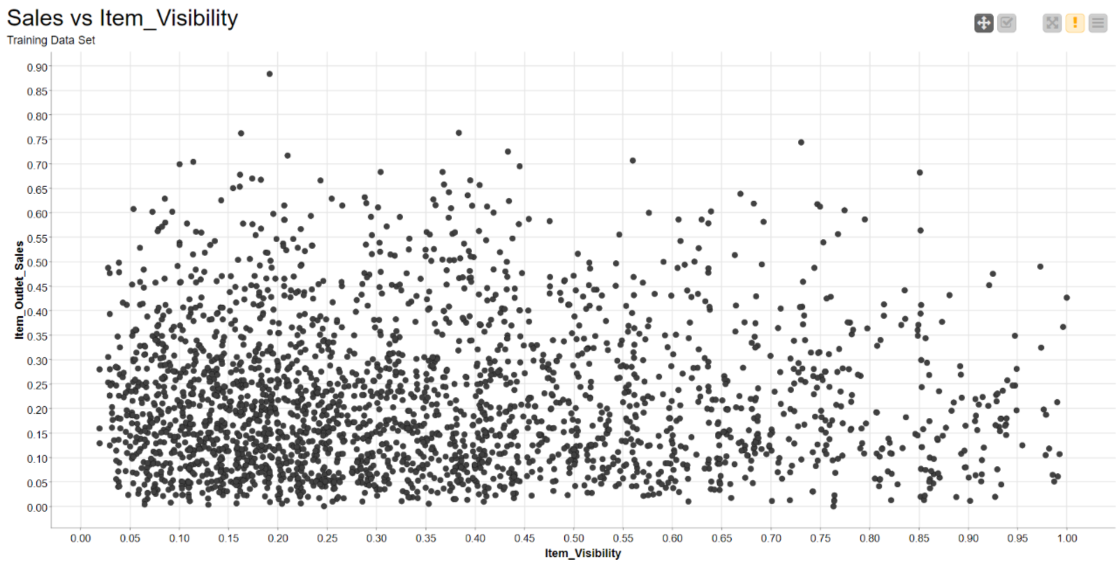
Among the pros of treating the output variable as a continuous variable is that it can be fitted directly into a regression equation which can later be used for prediction modelling. Since regression equations are best or more efficient when using continuous variables(numeric data).This is because the regression models are mainly used to compare two numeric variables and these two variables will produce a pattern which can later be analysed to identify the relationship unlike categorical variables. Below are the examples.



There is a pattern that corelates to a positive linear regression.



Pattern is present however to deduce relationship a line of best fit should be drawn.



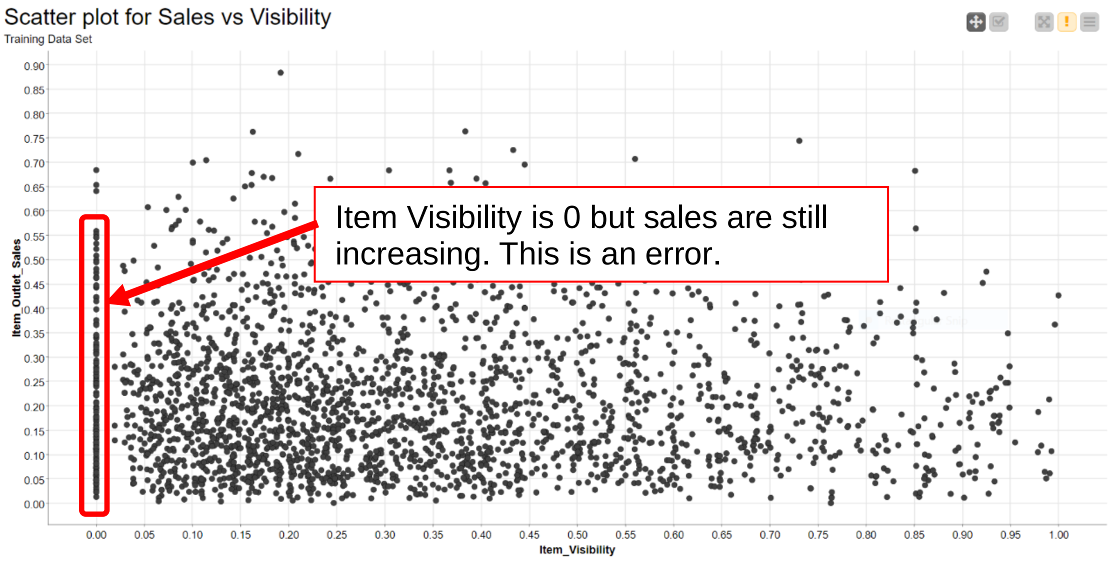
There is also an identifiable pattern in this regression equation.

In conjunction with that, we can say that using a continuous variable will save time as we would not have to encode the variables manually into a format that the machine would understand. This would direct improve the time taken to make a prediction model. Which in turn would lead to increased productivity reduced cost of production as the time taken to create the prediction model would be significantly increased.

Not only that but continuous variable data can also be visualized in a simpler form to allow stakeholders with no technical knowledge in data science to understand the data. Such forms of visualizations are histograms and box-plots.

Cons

Not many cons when compared to categorical data. However, the only disadvantage that cam up during the analysis is that it requires a lot of preparation and have to be really thorough as continuous variables/data are highly sensitive. For example, in the data there might be values that do not carry any significance for example Item_Visibility = 0 (This cannot be the case as if the item is not visible then nobody will know that it can be bought). In categorical variables this is not a problem as when the values are grouped(summed up/averaged/mean) the 0 would gradually disappear as it would be added to another value but this is not the case in continuous variables as it is really sensitive to the input that is given to it therefore even the smallest change it can detect it. Therefore, if not careful then it can effect the prediction model and give the incorrect output. Below is an example.



Item Visibility is 0 but sales are still increasing. This is an error.

4.Justify several business values to be gained from the ability to automatically predict the expected sales of a product [15 marks]

Thanks to the rising popularity of Data Science these days, businesses are starting to get more and more interested in being able to use data to predict expected sales on products they sell and with good reason. Being able to predict the expected sales of a product brings numerous benefits for a business. Below are just a few business values that businesses can gained from the ability to automatically predict the expected sales of a product:

Better Allocation of Resources

With the ability to be able to predict the expected sales of a product, a business can work around this knowledge and maximize the resources available to them. All businesses have a limited amount of resources and capital that is available to them. By having the ability to predict the expected sales of a product, the company can better utilize what finite resources they have to improve their profits.

Example if a business is putting the same amount of resources into two products, but one product is expected to do have better sales, the business can now choose to put more resources into the more profitable product as that will be a better use of its resources or choose to reduce resources spent on hiring people. Example of these resources could be Man power, as in the number of workers working to produce that product, and machinery that is used to create that product.

Improve Manufacturing and Production

If a businesses is able to predict the amount of sales a product will have, they can plan their manufacturing to suit accordingly. Appropriate planning in resources in both Man Power and Raw materials can be forecasted in line with the predicted expected sales. Once they know how much a product is expected to sell, they have no use in overbuying raw materials for manufacturing of the product, because that would just be a waste in capital. The company then can able to capitalize the "just in time" inventory to improve their cash flow.

By being able to predict the expected sales businesses can also calculate how much man power and machinery that they need to meet said sales. That means businesses can allocate the excess manpower and machinery to other parts of the business to improve efficiency.

An example we can use is with a furniture company. The furniture company makes assembled furniture and sells it to customers. To make the furniture we need workers and machinery to cut and design the raw material into furniture to sell. If the furniture company has the ability to predict expected sales of each furniture, They can allocate workers/machinery to help with assembling the expected more popular furniture.

Improve Negotiations with Vendors

Once a business knows the expected sales of a product, they are able to have a more informed decision when negotiating with vendors for supplies. With the ability to better forecast their Sales from , there can better negotiate with their respective Vendors for both The Cost (Volume Discount) and also the delivery of their raw materials.

This will enable the Business to have upper hand when negotiating with vendors. By being able to get a grasp on the expected sales of a product, a business can schedule with vendors and work together with them to get a lower cost and supplies they need on time.

Better Utilization of Capital

Capital in business terms is the amount of money they have available to use for daily operations and funding for future growths. With being able to know expected sales of a product, businesses can better utilize capital available to them to maximize their profits and planned for future growth.

Example with knowing the expected sales of a product, a business can estimate the amount of capital needed to be allocated to manufacture that product and choose to allocate excess capital elsewhere. In addition, the company can also use the forecast Sales to plan for future expanding accordingly.

With McDonalds for example, if they are able to predict the expected sales from each branch, they allocate any capital that some branches don't need and use that capital in future projects example creating more branches.

Improve on Cashflow and Forecast Revenue

Cash flow in layman terms is the amount of money going in and coming out from a business. By knowing expected sales, a business can get a better understanding on its cashflow specifically the amount of profit that is going to come back to the business. With this knowledge businesses can more confidently make decisions on putting more money in to the business knowing that it will be able to get its money back. Cash flow is harder to give example because it is more of a business concept.

Marketing

With the knowledge on expected sales, a business can make a better decision regarding marketing of a product, both in advertising and promotion. Once a business knows the expected sales, the business will now know which product to focus its marketing on. If a product is expected to have higher sales compared to other products, a business will naturally focus more on advertising and promoting the product as it will be overall more profitable for the business.

An example of this could be with the car brand Toyota. Toyota like every other company has a budget on the marketing they can do on their cars. With the ability to predict expected sales of each car, they can focus their budget on marketing the less popular models to enhance sales.

4.1 REFERENCES:

Statistics.com: Data Science, Analytics & Statistics Courses. 2021. *Data Partition - Statistics.com: Data Science, Analytics & Statistics Courses*. [online] Available at: <https://www.statistics.com/glossary/data-partition/#:~:text=Partitioning%20is%20normally%20used%20when,for%20verification%20of%20the%20model.> [Accessed 4 May 2021].

Contribution	Members
Part I	Ethan
Part II	Joel
Part III	Bernard
Part IV	Lionel
Part V	Joel, Bernard, Lionel