Cover sheet for submission of work for assessment

SWINBURNE UNIVERSITY OF TECHNOLOGY

UNIT DETAILS

Unit name	Introduction to Data Science			Class day/time	Tuesday, 8:30 am	Office use only
Unit code	COS10022	Assignment no.	1	Due date	6/4/2021	
Name of lecturer/teacher	Pei-Wei Tsai					
Tutor/marker's name	Pei-Wei Tsai					Faculty or school date stamp

STUDENT(S)

	Family Name(s)	Given Name(s)	Student ID Number(s)
(1)	Bui	Minh Quynh Nhu	102443018
(2)	Syed	Omair Maqdoom Mohiuddin	102863768
(3)	Kakolyris	Anthony	103059636
(4)	de Groot	Henry	103058565
(5)			
(6)			

DECLARATION AND STATEMENT OF AUTHORSHIP

1. I/we have not impersonated or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment.
2. This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made.
3. No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/teacher concerned.
4. I/we have not previously submitted this work for this or any other course/unit.
5. I/we give permission for my/our assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.

I/we understand that:

6. Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited.

Student signature/s

I/we declare that I/we have read and understood the declaration and statement of authorship.

(1)	Syed Omair	(4)	Henry de Groot
(2)	Nhu Bui	(5)	
(3)	Anthony Kakolyris	(6)	

Further information relating to the penalties for plagiarism, which range from a formal caution to expulsion from the University is contained on the Current Students website at **www.swin.edu.au/student/** Copies of this form can be downloaded from the Student Forms web page at **www.swinburne.edu.au/studentforms/**

Optimizing Product Placement in Retail

INTROUDCTION TO DATA SCIENCE (COS10022)

Anthony Kakolyris (103059636), Henry de Groot (103058565), Syed Omair Maqdoom Mohiuddin (102863768), Minh Quynh Nhu Bui (102443018)

Introduction

BigMart Sales Problem is a common data related problem that is found on the Kaggle website. The goal of the training dataset that was given is to build a predictive model and evaluate the sales of the different items in the store. The dataset contains 8,523 sales records with 12 different attributed ranging from the items themselves and the outlet they are purchased from. The purpose for this task is to help BigMart understand what purchasable items and/or stores are helping them gain the largest profit.

Data Observation and Cleaning

Data Observation

	Nominal data	Ordinal data	Continuous data	Discrete data
Item_Identifier	X			
Item_Weight			X	
Item_Fat_Content	X			
Item_Visibility			X	
Item_Type	X			
Item_MRP			X	
Outlet_Identifier	X			
Outlet_Establishment_Year	X			
Outlet_Size		X		
Outlet_Location_Type		X		
Outlet_Type		X		
Item_Outlet_Sales			X	

Data Cleaning

In the dataset, there are 2 attributes which are Item_Weight and Outlet_Size contained missing values. Data must address specific business requirement in order to achieve expected outcome which is to create a predictive model and identify the sales of each product at a particular store. Therefore, compare to Item_Weight, Outlet_Size is more about an essential element to analyze strategic goal.

The changed dataset excluded all the missing values in attributes Outlet_Size.

Table "Assignment1_BigMart_Data.csv" - Rows: 8523 | 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
Row3	FDX07	19.2	Regular	0	Fruits and V...	182.095	OUT010	1998	?	Tier 3	Grocery Store	732.38
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
Row7	FDP10	?	Low Fat	0.127	Snack Foods	107.762	OUT027	1985	Medium	Tier 3	Supermarket Type3	4,022.764
Row8	FDH17	16.2	Regular	0.017	Frozen Foods	96.973	OUT045	2002	?	Tier 2	Supermarket Type1	1,076.599
Row9	FDU28	19.2	Regular	0.094	Frozen Foods	187.821	OUT017	2007	?	Tier 2	Supermarket Type1	4,710.535
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

Before – Rows: 8523/ Columns: 12

Table "Assignment1_BigMart_Data.csv" - Rows: 6113 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Item_I...	Item_...	Item_F...	Item_Vi...	Item_T...	Item_MRP	Outlet_...	Outlet_...	Outlet_...	Outlet_...	Outlet_Type	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
Row7	FDP10	?	Low Fat	0.127	Snack Foods	107.762	OUT027	1985	Medium	Tier 3	Supermarket Type3	4,022.764
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

Removing missing data in Outlet_Size After – Rows: 6113/ Columns: 12

File Edit Hilite Navigation View

Table "default" - Rows: 7060 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Item_I...	Item_...	Item_F...	Item_Vi...	Item_T...	Item_MRP	Outlet_...	Outlet_...	Outlet_...	Outlet_...	Outlet_Type	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
Row3	FDX07	19.2	Regular	0	Fruits and V...	182.095	OUT010	1998	?	Tier 3	Grocery Store	732.38
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
Row8	FDH17	16.2	Regular	0.017	Frozen Foods	96.973	OUT045	2002	?	Tier 2	Supermarket Type1	1,076.599
Row9	FDU28	19.2	Regular	0.094	Frozen Foods	187.821	OUT017	2007	?	Tier 2	Supermarket Type1	4,710.535
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
Row25	NCD06	13	Low Fat	0.1	Household	45.906	OUT017	2007	?	Tier 2	Supermarket Type1	838.908
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
Row28	FDE51	5.925	Regular	0.161	Dairy	45.509	OUT010	1998	?	Tier 3	Grocery Store	178.434
Row30	FDV38	19.25	Low Fat	0.17	Dairy	55.796	OUT010	1998	?	Tier 3	Grocery Store	163.787
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
Row33	FDO23	17.85	Low Fat	0	Breads	93.144	OUT045	2002	?	Tier 2	Supermarket Type1	2,174.503
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
Row45	FDM39	6.42	LF	0.089	Dairy	178.1	OUT010	1998	?	Tier 3	Grocery Store	358.2

Removing missing data in Item_Weight After – Rows: 7060/Colums: 12

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

Row ID	Item_I...	Item_...	Item_F...	Item_Vi...	Item_T...	Item_MRP	Outlet_...	Outlet_...	Outlet_...	Outlet_...	Outlet_Type	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

Removing all the missing data from the sample After – Rows: 4650/Columns: 12

In conclusion, the number of tuples decreased from 8523 to 4650.

Data Preparation

Shuffling

KNIME Analytics Platform provides us a manipulation to shuffle data, Row Shuffle.

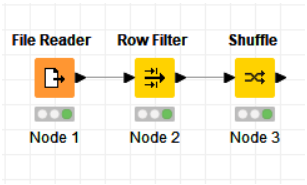


Table "default" - Rows: 6113 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Item_I...	Item_...	Item_F...	Item_Vi...	Item_T...	Item_MRP	Outlet_...	Outlet_...	Outlet_...	Outlet_...	Outlet_Type	Item_O...
Row7717	NCC30	16.6	Low Fat	0.028	Household	176.634	OUT035	2004	Small	Tier 2	Supermarket Type1	2,676.516
Row6543	FDS43	11.65	Low Fat	0.041	Fruits and V...	186.924	OUT035	2004	Small	Tier 2	Supermarket Type1	2,237.088
Row2668	FDV16	7.75	Regular	0.083	Frozen Foods	34.956	OUT035	2004	Small	Tier 2	Supermarket Type1	645.16
Row7826	FDM39	?	Low Fat	0.053	Dairy	177.6	OUT027	1985	Medium	Tier 3	Supermarket Type3	8,417.709
Row3547	FDO01	21.1	Regular	0.021	Breakfast	129.799	OUT018	2009	Medium	Tier 3	Supermarket Type2	1,284.994
Row2930	FDW15	15.35	Regular	0.055	Meat	149.773	OUT035	2004	Small	Tier 2	Supermarket Type1	2,820.995
Row4516	FDV23	?	Low Fat	0.105	Breads	125.705	OUT027	1985	Medium	Tier 3	Supermarket Type3	3,237.12
Row4348	FDU11	4.785	Low Fat	0.093	Breads	122.01	OUT018	2009	Medium	Tier 3	Supermarket Type2	2,048.667
Row8512	FDR26	20.7	Low Fat	0.043	Dairy	178.303	OUT013	1987	High	Tier 3	Supermarket Type1	2,479.439
Row1945	NCF18	18.35	LF	0.089	Household	191.95	OUT018	2009	Medium	Tier 3	Supermarket Type2	5,369.011
Row3772	FDU39	18.85	Low Fat	0.036	Meat	58.556	OUT018	2009	Medium	Tier 3	Supermarket Type2	770.331
Row1410	FDL21	15.85	Regular	0.007	Snack Foods	40.848	OUT013	1987	High	Tier 3	Supermarket Type1	679.116
Row7046	DRQ35	9.3	Low Fat	0.042	Hard Drinks	125.439	OUT035	2004	Small	Tier 2	Supermarket Type1	3,715.164
Row1047	DRN11	7.85	Low Fat	0.163	Hard Drinks	145.244	OUT046	1997	Small	Tier 1	Supermarket Type1	1,451.444
Row2688	FDT38	18.7	Low Fat	0.058	Dairy	83.357	OUT049	1999	Medium	Tier 1	Supermarket Type1	1,860.245
Row1907	FDN25	7.895	Regular	0.061	Breakfast	59.259	OUT035	2004	Small	Tier 2	Supermarket Type1	801.623
Row8012	FDA21	13.65	Low Fat	0.036	Snack Foods	185.292	OUT035	2004	Small	Tier 2	Supermarket Type1	3,146.571
Row1062	NCO18	13.15	Low Fat	0.025	Household	177.469	OUT035	2004	Small	Tier 2	Supermarket Type1	5,510.827
Row6487	FDA49	19.7	Low Fat	0.065	Canned	88.52	OUT035	2004	Small	Tier 2	Supermarket Type1	1,308.297
Row4968	FDD39	?	Low Fat	0.123	Dairy	217.685	OUT019	1985	Small	Tier 1	Grocery Store	432.77
Row1166	FDT22	10.395	Low Fat	0.112	Snack Foods	58.022	OUT035	2004	Small	Tier 2	Supermarket Type1	659.142

The shuffled dataset.

Partition

Partitioning the data into 2 sets. A Training Set and a Test Set at a 7:3 ratio

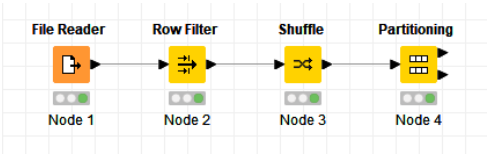


Table "default" - Rows: 4279 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Item_I...	Item_...	Item_F...	Item_Vi...	Item_T...	Item_MRP	Outlet_...	Outlet_...	Outlet_...	Outlet_...	Outlet_Type	Item_O...
Row5674	NCP50	17.35	Low Fat	0.021	Others	78.762	OUT046	1997	Small	Tier 1	Supermarket Type1	1,288.989
Row4884	NCL42	18.85	Low Fat	0.04	Household	244.314	OUT049	1999	Medium	Tier 1	Supermarket Type1	5,635.331
Row5277	FDD57	18.1	Low Fat	0.022	Fruits and V...	96.509	OUT035	2004	Small	Tier 2	Supermarket Type1	1,332.932
Row5266	FDQ26	?	Regular	0.068	Dairy	57.256	OUT027	1985	Medium	Tier 3	Supermarket Type3	2,073.967
Row1016	FDZ25	?	Regular	0.027	Canned	169.379	OUT027	1985	Medium	Tier 3	Supermarket Type3	5,602.707
Row6432	DRE12	4.59	Low Fat	0.071	Soft Drinks	113.286	OUT013	1987	High	Tier 3	Supermarket Type1	1,471.418
Row715	FDA14	?	Low Fat	0.114	Dairy	147.176	OUT019	1985	Small	Tier 1	Grocery Store	585.904
Row1183	FDW09	13.65	reg	0.026	Snack Foods	81.13	OUT013	1987	High	Tier 3	Supermarket Type1	792.302
Row4614	FDH45	15.1	Regular	0.106	Fruits and V...	43.28	OUT018	2009	Medium	Tier 3	Supermarket Type2	123.839
Row2420	FDU14	17.75	Low Fat	0.035	Dairy	248.275	OUT035	2004	Small	Tier 2	Supermarket Type1	998.7
Row6520	FDQ25	8.63	Regular	0.028	Canned	170.542	OUT035	2004	Small	Tier 2	Supermarket Type1	2,069.306
Row6222	DRL49	13.15	Low Fat	0.056	Soft Drinks	144.281	OUT046	1997	Small	Tier 1	Supermarket Type1	1,282.331
Row8046	FDP07	18.2	Low Fat	0.09	Fruits and V...	197.511	OUT046	1997	Small	Tier 1	Supermarket Type1	392.822
Row2744	NCB30	14.6	Low Fat	0.026	Household	197.108	OUT049	1999	Medium	Tier 1	Supermarket Type1	4,761.802
Row6807	FDZ03	13.65	Regular	0.079	Dairy	186.724	OUT049	1999	Medium	Tier 1	Supermarket Type1	4,474.176
Row1987	NCV06	?	Low Fat	0.066	Household	195.248	OUT027	1985	Medium	Tier 3	Supermarket Type3	7,168.669
Row2227	FDA38	5.44	Low Fat	0.026	Dairy	239.154	OUT018	2009	Medium	Tier 3	Supermarket Type2	480.708
Row4681	FDW34	?	Low Fat	0.035	Snack Foods	244.317	OUT027	1985	Medium	Tier 3	Supermarket Type3	8,262.578
Row5670	FDA13	?	Low Fat	0.138	Canned	38.851	OUT019	1985	Small	Tier 1	Grocery Store	37.951
Row2231	FDY25	12	Low Fat	0.034	Canned	181.898	OUT046	1997	Small	Tier 1	Supermarket Type1	4,527.44
Row6286	FDT15	12.15	Regular	0.043	Meat	183.695	OUT049	1999	Medium	Tier 1	Supermarket Type1	2,929.52

The first partition (Training Set)

Table "default" - Rows: 1834 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Item_I...	Item_...	Item_F...	Item_Vi...	Item_T...	Item_MRP	Outlet_...	Outlet_...	Outlet_...	Outlet_...	Outlet_Type	Item_O...
Row6777	FDV60	20.2	Regular	0.118	Baking Goods	195.211	OUT018	2009	Medium	Tier 3	Supermarket Type2	2,356.932
Row2096	FDO04	16.6	Low Fat	0.027	Frozen Foods	53.561	OUT018	2009	Medium	Tier 3	Supermarket Type2	939.444
Row4282	NCU41	18.85	Low Fat	0.052	Health and ...	190.385	OUT013	1987	High	Tier 3	Supermarket Type1	3,630.607
Row3475	FDN58	13.8	Regular	0.057	Snack Foods	230.998	OUT013	1987	High	Tier 3	Supermarket Type1	4,633.968
Row6659	FDK21	7.905	Low Fat	0.01	Snack Foods	250.441	OUT046	1997	Small	Tier 1	Supermarket Type1	3,004.09
Row8327	FDT24	?	Regular	0	Baking Goods	75.933	OUT027	1985	Medium	Tier 3	Supermarket Type3	3,012.079
Row2817	DRF03	19.1	Low Fat	0.045	Dairy	42.414	OUT049	1999	Medium	Tier 1	Supermarket Type1	243.683
Row6151	FDO22	13.5	Regular	0.018	Snack Foods	78.796	OUT018	2009	Medium	Tier 3	Supermarket Type2	239.688
Row4938	NCD43	?	Low Fat	0.028	Household	106.196	OUT019	1985	Small	Tier 1	Grocery Store	210.393
Row7509	FDT40	?	Low Fat	0.095	Frozen Foods	125.568	OUT027	1985	Medium	Tier 3	Supermarket Type3	2,543.356
Row4720	FDT10	16.7	Regular	0.062	Snack Foods	60.656	OUT018	2009	Medium	Tier 3	Supermarket Type2	592.562
Row4618	NCB31	6.235	Low Fat	0.119	Household	263.791	OUT018	2009	Medium	Tier 3	Supermarket Type2	2,103.928
Row6749	FDJ46	11.1	Low Fat	0.045	Snack Foods	174.205	OUT046	1997	Small	Tier 1	Supermarket Type1	1,926.159
Row3385	FDQ32	17.85	Regular	0.047	Fruits and V...	121.939	OUT049	1999	Medium	Tier 1	Supermarket Type1	3,962.842
Row1280	FDI04	?	Regular	0.128	Frozen Foods	198.543	OUT019	1985	Small	Tier 1	Grocery Store	790.97
Row3649	FDJ48	?	Low Fat	0.056	Baking Goods	246.912	OUT027	1985	Medium	Tier 3	Supermarket Type3	5,681.271
Row3752	FDW21	5.34	Regular	0.006	Snack Foods	102.436	OUT035	2004	Small	Tier 2	Supermarket Type1	2,010.716
Row2395	FDV38	?	Low Fat	0.101	Dairy	55.096	OUT027	1985	Medium	Tier 3	Supermarket Type3	2,020.037
Row1011	FDC29	8.39	Regular	0.024	Frozen Foods	114.018	OUT018	2009	Medium	Tier 3	Supermarket Type2	1,488.729
Row2708	NCE30	16	Low Fat	0.099	Household	210.79	OUT035	2004	Small	Tier 2	Supermarket Type1	4,460.194
Row1547	NCB55	15.7	Low Fat	0.161	Household	57.556	OUT018	2009	Medium	Tier 3	Supermarket Type2	829.587

The second partition (Test Set)

Visualizations

Data visualization is essential in data analysis which help in building a prediction model.

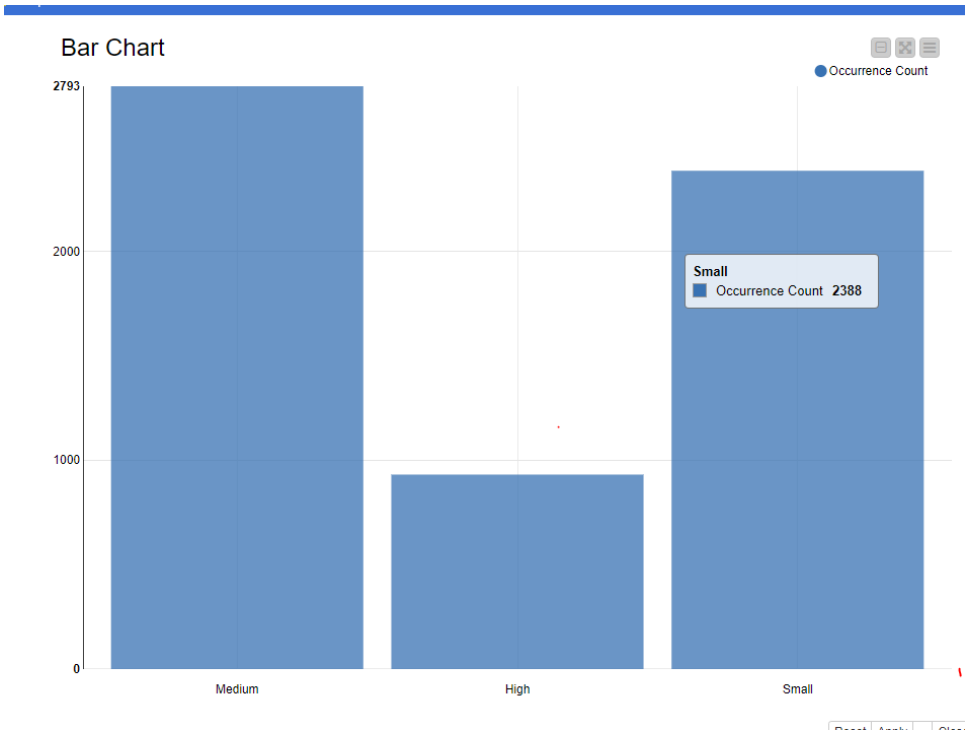


Figure 1: Bar Chart for different size of outlets

Figure 1 shows the sale (8523) in different outlets based on their size (high, medium, and low) in terms of ground area covered. As per the data collected (6113), major percent of sale was recorded from medium sized (46%, 2793), followed by small sized (39%, 2388) and least contribution from high sized (15%, 932) outlets. Whilst there was some missing data recorded (2410).

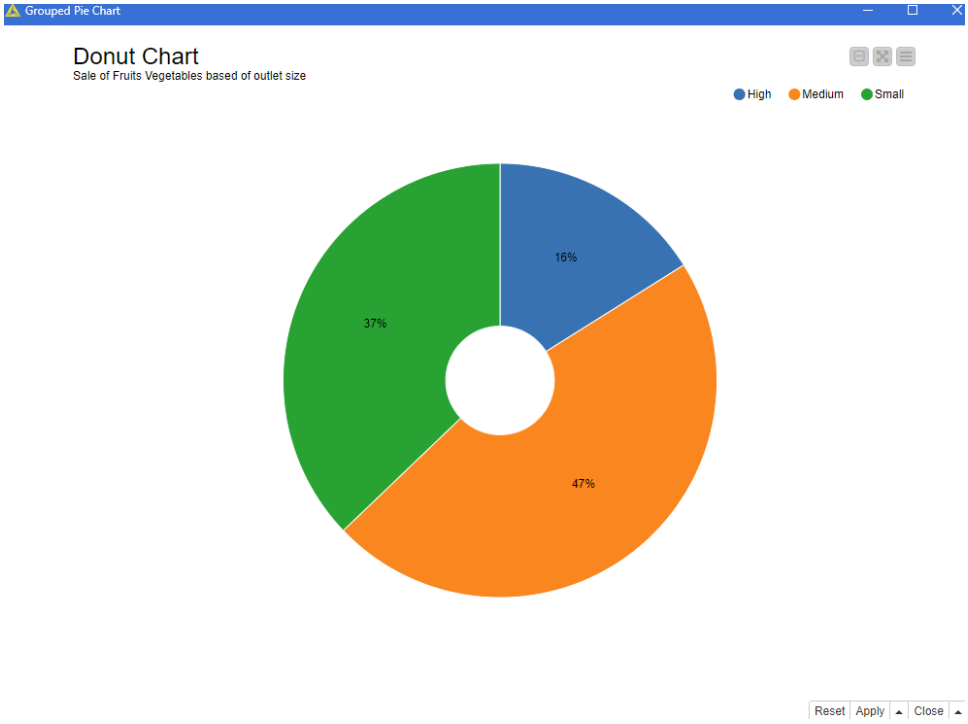


Figure 2: Donut Chart based on sale of Fruits and Vegetables in different sized outlets.

The Donut Chart shows the sale (1232) in fruits and vegetables in the outlets based on their size (High, Medium, and Low). It was recorded that major sale of fruits and vegetables were from medium sized outlets (47%, 413), followed by small sized outlets (37%, 328) and minor from high sized (16%, 142). Although there was some missing data (349) about the size of the data.

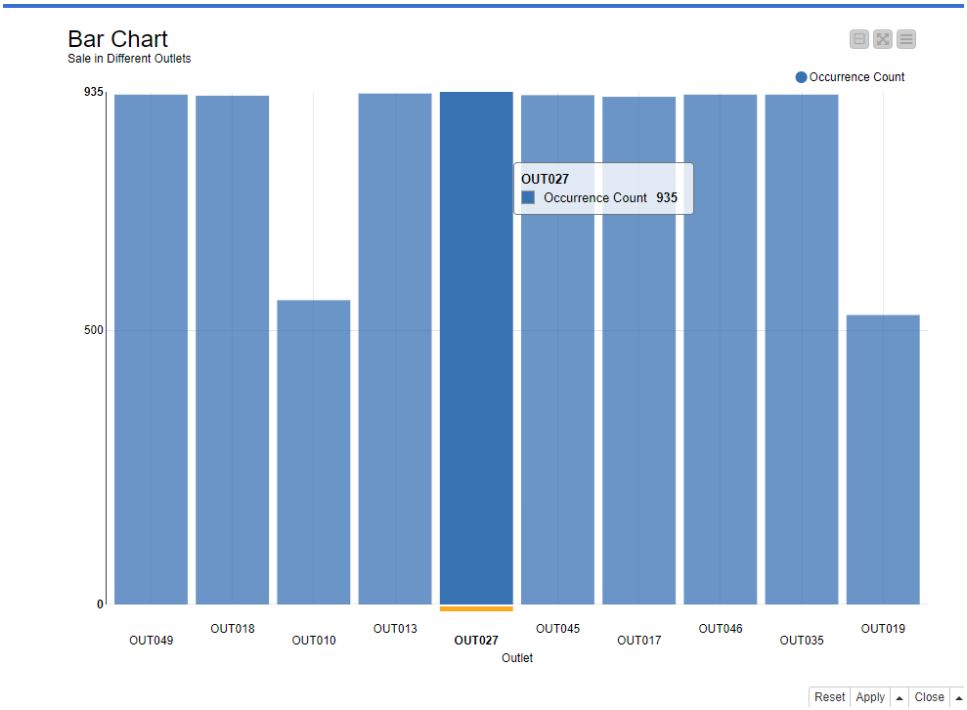


Figure 3: Bar Chart based on sale of the outlets.

The Bar Chart illustrates sale (8523) in different outlets. Most of the outlets have good figure in sales, although OUT010 and OUT019 (lowest) recorded least sale when compared with the other outlets. OUT027 recorded the highest sale (935), while many outlets were close to OUT027.

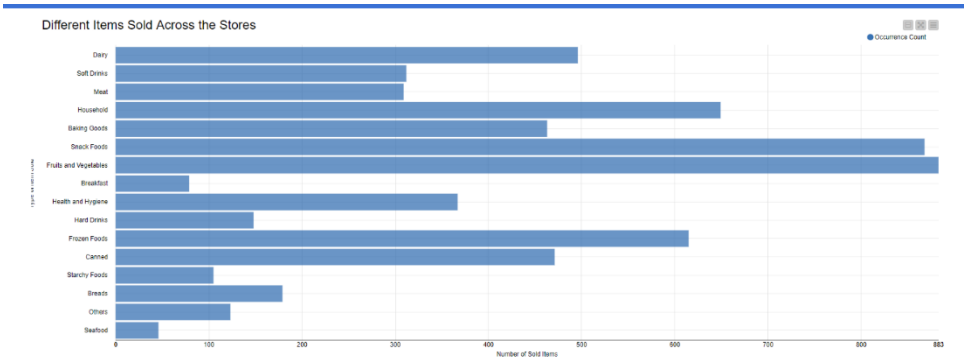


Figure 4: Bar Chart based on the sale of items in the outlets.

The Bar Chart shows the sale (6113) of items in particular sector (example: Dairy, Meat etc.) As per the data provided fruits and vegetables (883), and Snack Foods (868) both contributed 14% each adding up to 18%. Followed by Household (649) 11%, Frozen Foods (615) 10% also, Dairy (496), Canned (471) and Baking Goods (463) contribute 8% each. Health and Hygiene (367) 6%, Soft drinks (312) 5% and least contribution from Breads (179), Hard Drinks (148), Others (123), Breakfast (79), Seafood (46).

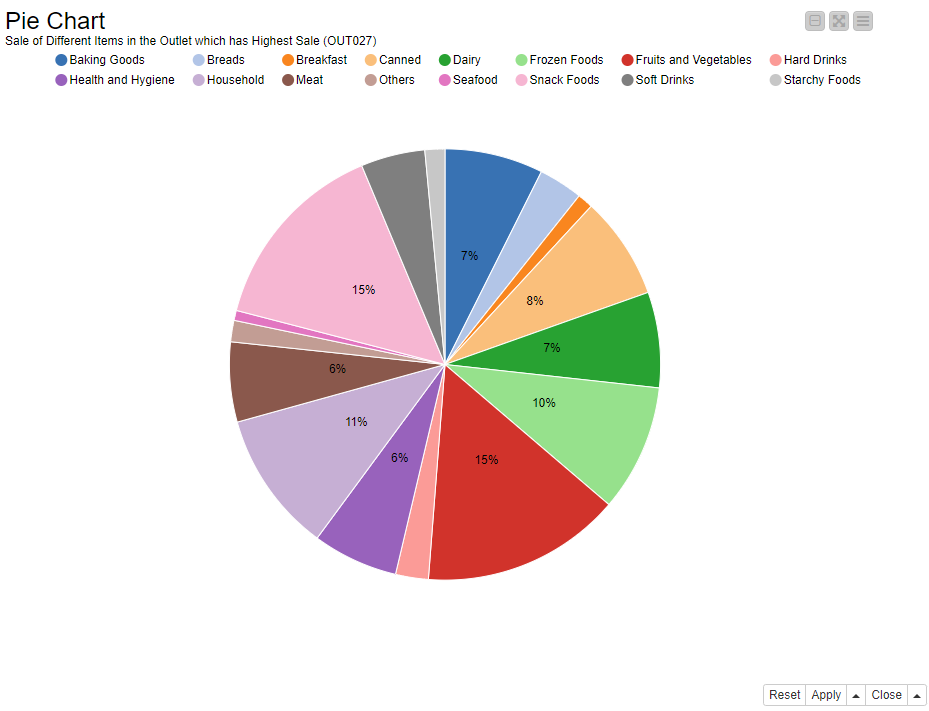


Figure 5: Pie Chart based on sale of the outlet with highest sale.

The Pie Chart illustrates the sale (935) of the OUT027 which has the highest sale. According to the received data Fruits and vegetable (140) and Snack Foods (137) sectors have the highest sale 14% when compared to other sectors. Household (99) 11%, Frozen Foods (89) 10%, Canned (72) and Dairy (67) and Baking Goods (69) contributed 7% each, Health and Hygiene (60) and Meat (56) contributed 6% each Soft Drinks (45). Hard drinks (23), Breads (31), Break Fast (11), Others (15) and Seafood (7) have little contribution to sale.

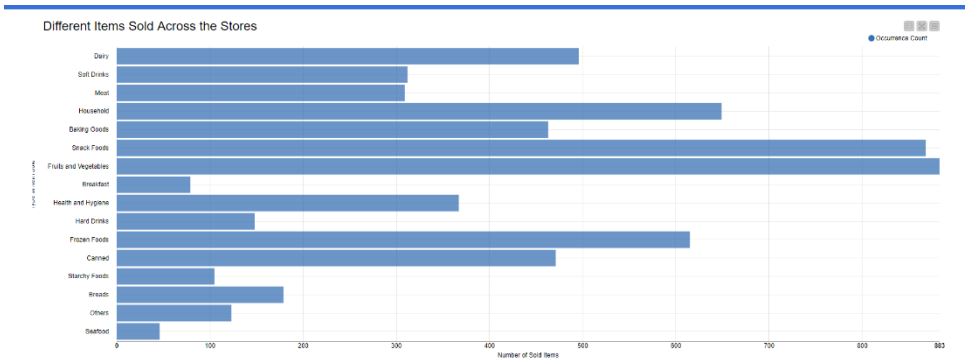


Figure 6: Pie Chart based for sale of outlets based on their type (grocery, supermarket).

The Pie Chart shows the sale (8523) based on type of outlet (Grocery store, Supermarket Type1, Supermarket Type2 and Supermarket Type3). Major contribution was from Supermarket Type 1(5577) 69%, followed by Grocery store (1083) 13% and least from Supermarket Type2 (928) and Supermarket Type3 (935) 11% each.

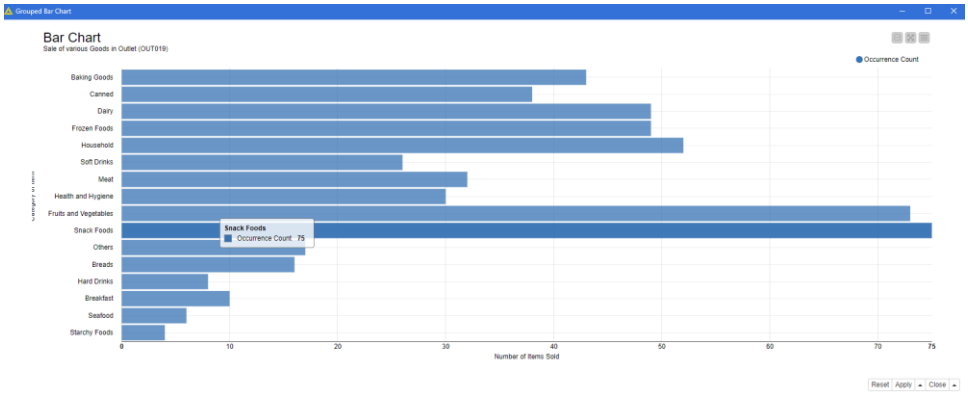


Figure 7: Bar Chart based on sale of the outlet with lowest sale (OUT019)

The Bar Graph illustrates the sale (528) of OUT019 based on the items. Snack Foods (75) and Fruits and Vegetables (73) contributed much of the sale. Followed by other items (Household (52), Dairy (49), Frozen Foods (49), Baking Goods (43), Canned (38), Meat (32) and others). Least contribution was from Breakfast (10), Hard drinks (8), Sea Food (6), and Starchy Foods (4).

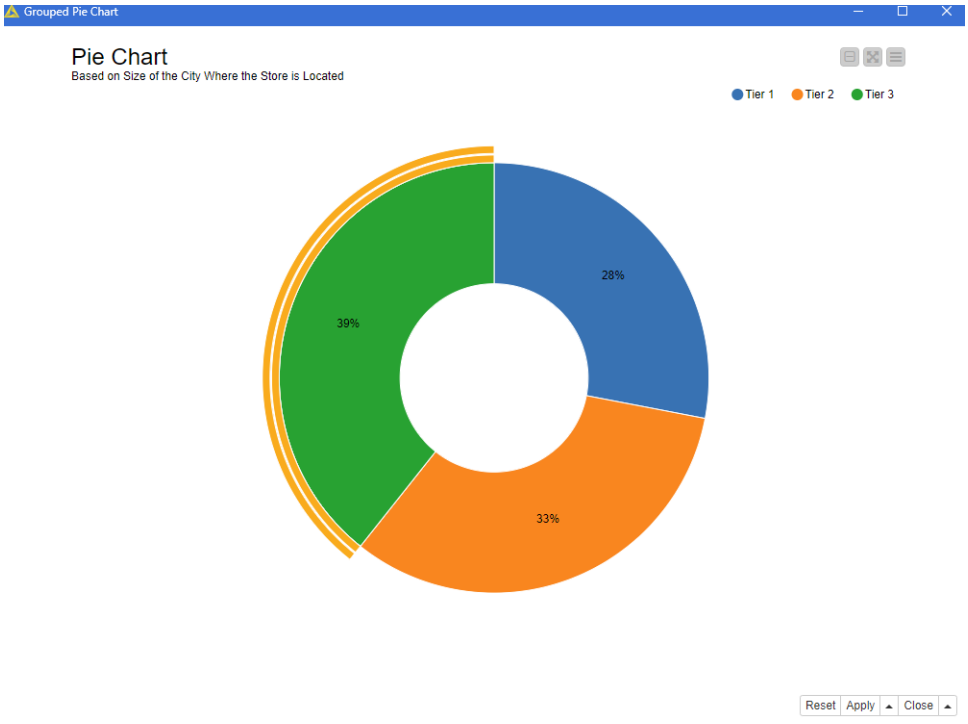
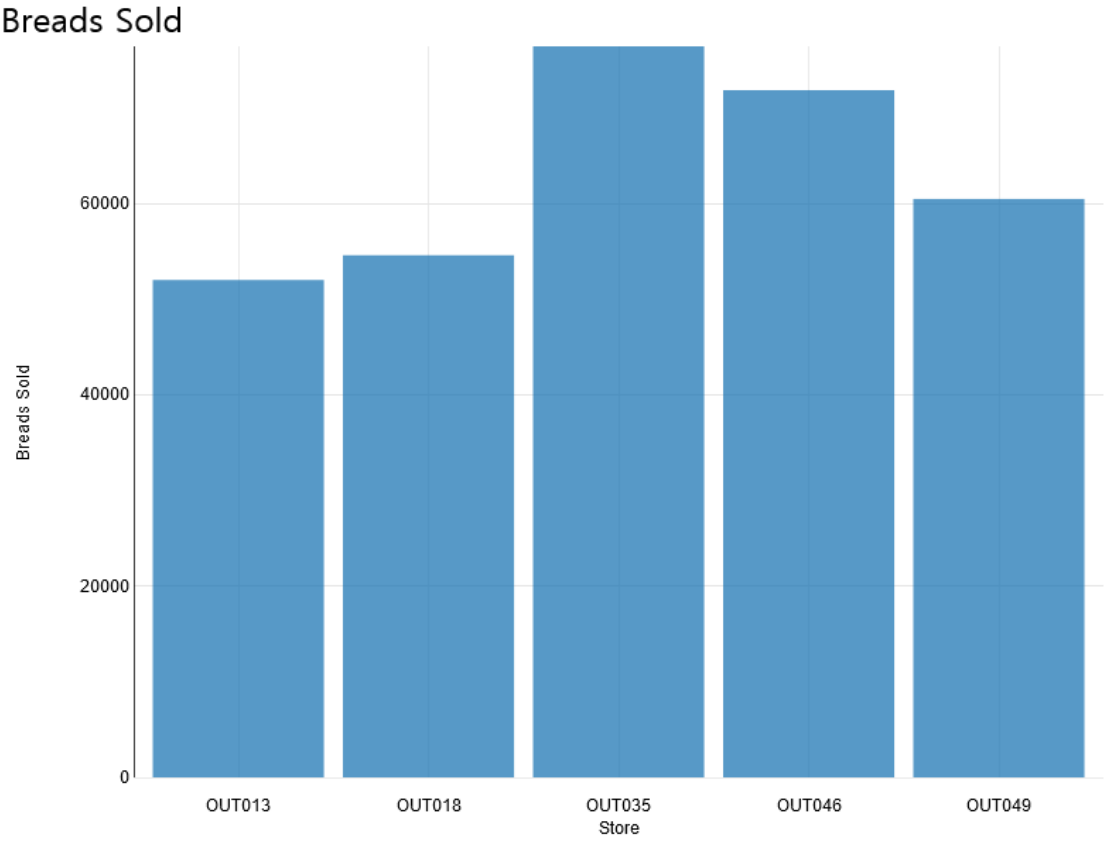


Figure 8: Donut Chart on sale based on size of city where the store is located (Tier 1, Tier 2 and Tier 3)

The Donut Chart shows the sale (8523) based on the size of the city in which the store is located. There was a tough competition between the categories. Tier 3 (3350) has 39% of sale which is 6% higher than the sale of Tier 2 (2785) 33%, these two make up a major part of the sale and Tier 1 (2388) 28% has least contribution to sale when compared with rest categories.

Conventional Data Analysis



This Bar Chart shows the revenue from bread sales per outlet. After the data cleaning in the earlier steps, we only have 5 stores to compare from. This Chart shows that OUT035 had the greatest revenue from bread sales, while OUT013 has the lowest revenue.

Row ID	Outlet_Identifier	Set(Outlet_Size)	Set(Outlet_Type)
Row0	OUT013	[High]	[Supermarket Type1]
Row1	OUT018	[Medium]	[Supermarket Type2]
Row2	OUT035	[Small]	[Supermarket Type1]
Row3	OUT046	[Small]	[Supermarket Type1]
Row4	OUT049	[Medium]	[Supermarket Type1]

Looking at the store sizes and types, it appears that the smaller the store, the more revenue it made from breads.

Data Verification

A senior data scientist highlighted potential concerns with data, hence, an investigation into the data was put into place to identify any potential issues. To begin, looking through the file and using logical reasoning to reassure that there are no issues with the outlet data collected. Looking at the attributes of the data, it can be assumed that for each unique outlet (Outlet_Identifer) they only have one size (Outlet_Size), type (Outlet_Type), location type (Outlet_Location_Type) and establishment year (Outlet_Establishment_Year). Using the KNIME platform to read and group the data, it can be seen below that this is true.

Row ID	S Outlet_Identifier	I Unique count(Outlet_Size)	I Unique count(Outlet_Type)	I Unique count(Outlet_Location_Type)	I Unique count*(Outlet_Establishment_Year)
Row0	OUT013	1	1	1	1
Row1	OUT018	1	1	1	1
Row2	OUT035	1	1	1	1
Row3	OUT046	1	1	1	1
Row4	OUT049	1	1	1	1

Because each unique outlet has only one unique count in these attributes, the data collected here is correct and would not cause any issues. However, it should also be assumed that item visibility (Item_Visibility) is greater than zero if there are outlet sales for that item (Item_Outlet_Sales). Using KNIME once again this can be investigated.

Row ID	S Item_Identifier	S Outlet_Identifier	[...] Set*(Item_Visibility)	[...] Set*(Item_Outlet_Sales)
Row999	FDE22	OUT018	[0.029693277]	[2716.464]
Row998	FDE21	OUT049	[0.022980361]	[1274.3412]
Row997	FDE20	OUT049	[0.005539114]	[4074.696]
Row996	FDE20	OUT046	[0.005530516]	[4923.591]
Row995	FDE20	OUT035	[0.00552947]	[2376.906]
Row994	FDE20	OUT013	[0.005525913]	[2376.906]
Row993	FDE17	OUT049	[0.054540159]	[755.683]
Row992	FDE17	OUT046	[0.054455495]	[2871.5954]
Row991	FDE17	OUT035	[0.054445198]	[4836.3712]
Row990	FDE17	OUT013	[0.054410179]	[2720.4588]
Row99	DRE27	OUT046	[0.13267058]	[978.726]
Row989	FDE16	OUT049	[0.026384672]	[4584.6988]
Row988	FDE16	OUT018	[0.026451028]	[5001.4896]
Row987	FDE14	OUT049	[0.031494041]	[2197.14]
Row986	FDE14	OUT035	[0.031439206]	[2596.62]
Row985	FDE14	OUT018	[0.031573246]	[299.61]
Row984	FDE14	OUT013	[0.031418984]	[2596.62]
Row983	FDE11	OUT049	[0.135306012]	[2221.1088]
Row982	FDE11	OUT046	[0.0]	[4257.1252]
Row981	FDE11	OUT035	[0.0]	[7033.5112]
Row980	FDE11	OUT018	[0.135646297]	[3516.7556]
Row98	DRE27	OUT035	[0.132645493]	[1174.4712]
Row979	FDE11	OUT013	[0.13498355]	[3146.5708]
Row978	FDE10	OUT049	[0.0]	[1573.9512]
Row977	FDE10	OUT018	[0.09031453]	[1180.4634]
Row976	FDE09	OUT049	[0.0]	[1436.7964]
Row975	FDE08	OUT049	[0.049396364]	[3563.3616]
Row974	FDE08	OUT018	[0.049520593]	[2375.5744]

In the photo above it can be seen that this is not true and that there are some values of an item visibility of 0, this would make sense if there were no sales made from the specific outlet, however, even with the values of 0 there are still outlet sales. This could cause potential issues as the information is obviously incorrect as there can not be any sales if there is not any of that item in the store.

Distribution of Work

	TASK 1	TASK 2	TASK 3	TASK 4	TASK 5
Anthony Kakolyris					
Henry de Groot					
Syed Omair Maqdoom Mohiuddin					
Minh Quynh Nhu Bui					