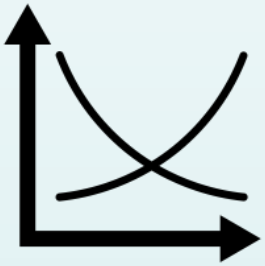
KOGOD SCHOOL of BUSINESS WASHINGTON, DC

Why Analytics in Retail?

Rich integrated sources of data

Clear, concise insights which can be used for marketing



Demand-Supply procurement optimization

Identify best-selling products efficiently

Fine tuning store performance

Accurately predict store Revenues

Dataset Snapshot

	ProductID	Weight	FatContent	ProductVisibility	ProductType	MRP	OutletID	EstablishmentYear	OutletSize	LocationType	OutletType	OutletSales
1	FDA15	9.300	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
2	DRC01	5.920	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
3	FDN15	17.500	Low Fat	0.016760075	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
4	FDX07	19.200	Regular	0.000000000	Fruits and Vegetables	182.0950	OUT010	1998	*NA*	Tier 3	Grocery Store	732.3800
5	NCD19	8.930	Low Fat	0.000000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
6	FDP36	10.395	Regular	0.000000000	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
7	FDO10	13.650	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
8	FDP10	*NA*	Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636
9	FDH17	16.200	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002	*NA*	Tier 2	Supermarket Type1	1076.5986
10	FDU28	19.200	Regular	0.094449590	Frozen Foods	187.8214	OUT017	2007	*NA*	Tier 2	Supermarket Type1	4710.5350
11	FDY07	11.800	Low Fat	0.000000000	Fruits and Vegetables	45.5402	OUT049	1999	Medium	Tier 1	Supermarket Type1	1516.0266
12	FDA03	18.500	Regular	0.045463773	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket Type1	2187.1530
13	FDX32	15.100	Regular	0.100013500	Fruits and Vegetables	145.4786	OUT049	1999	Medium	Tier 1	Supermarket Type1	1589.2646
14	FDS46	17.600	Regular	0.047257328	Snack Foods	119.6782	OUT046	1997	Small	Tier 1	Supermarket Type1	2145.2076
15	FDF32	16.350	Low Fat	0.068024300	Fruits and Vegetables	196.4426	OUT013	1987	High	Tier 3	Supermarket Type1	1977.4260
16	FDP49	9.000	Regular	0.069088961	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket Type1	1547.3192
17	NCB42	11.800	Low Fat	0.008596051	Health and Hygiene	115.3492	OUT018	2009	Medium	Tier 3	Supermarket Type2	1621.8888
18	FDP49	9.000	Regular	0.069196376	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket Type1	718.3982
19	DRI11	*NA*	Low Fat	0.034237682	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket Type3	2303.6680
20	FDU02	13.350	Low Fat	0.102492120	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermarket Type1	2748.4224

Training Set ~ 8620 rows
Test Set ~ 5682 rows

KOGOD SCHOOL of BUSINESS WASHINGTON, DC

Dataset Parameters in Focus

QUANTITATIVE DATA:

- **Weight:** Product weight in grams
- **Product Visibility:** An index to determine positioning of product in store
- **MRP:** Maximum Retail Price of product in nominal currency units

QUALITATIVE DATA:

- **Product ID:** Product weight in grams
- **Fat Content:** An index to determine positioning of product in store
- **Product Type:** Maximum Retail Price of product in nominal currency units
- **Location Type:** Whether the outlet is in a Tier 1,2 or 3 city
- **Outlet Type:** Whether it is a large hypermarket (type 1), supermarket (type 2), convenience store (type 3) or grocery store.

Tidying the Mess

CHALLENGE

- Redundant factorization for 'FatContent' variable
- Missing values for 'Weight' for a given 'ProductID'
- Presence of qualitative information which cannot be directly processed by algorithms

APPROACH

- Use forcats in R to combine redundant categories
- Replace missing values with mean of 'Weight' for observable rows
- Employ dummy variables to ensure we can run classification/regression methods

Final Dataset after Consideration

	ProductID	Weight	ProductVisibility	MRP	OutletSales	FatContent_Low Fat	FatContent_Regular	ProductType_Baking Goods	ProductType_Breads	ProductType_Breakfast
1	DRA12	11.60	0.041177505	140.3154	2552.6772	1	0	0	0	0
2	DRA12	11.60	0.000000000	141.6154	3829.0158	1	0	0	0	0
3	DRA12	11.60	0.040911824	142.3154	2552.6772	1	0	0	0	0
4	DRA12	11.60	0.000000000	141.9154	992.7078	1	0	0	0	0
5	DRA12	11.60	0.041112694	142.0154	850.8924	1	0	0	0	0
6	DRA12	11.60	0.068535039	143.0154	283.6308	1	0	0	0	0
7	DRA24	19.35	0.040154087	164.6868	1146.5076	0	1	0	0	0
8	DRA24	19.35	0.069909188	163.2868	491.3604	0	1	0	0	0
9	DRA24	19.35	0.066831682	163.8868	327.5736	0	1	0	0	0
10	DRA24	19.35	0.039734882	165.7868	4913.6040	0	1	0	0	0
11	DRA24	19.35	0.039920687	163.3868	3439.5228	0	1	0	0	0
12	DRA24	19.35	0.039990314	165.0868	982.7208	0	1	0	0	0
13	DRA24	19.35	0.039895009	162.4868	4422.2436	0	1	0	0	0
14	DRA59	8.27	0.127927931	184.8924	4442.2176	0	1	0	0	0
15	DRA59	8.27	0.128126825	183.6924	1295.6468	0	1	0	0	0
16	DRA59	8.27	0.127821472	185.9924	555.2772	0	1	0	0	0
17	DRA59	8.27	0.000000000	183.2924	2406.2012	0	1	0	0	0
18	DRA59	8.27	0.127308434	186.6924	7033.5112	0	1	0	0	0
19	DRA59	8.27	0.223985293	186.2924	555.2772	0	1	0	0	0
20	DRA59	8.27	0.128449055	186.5924	4442.2176	0	1	0	0	0

Training Set ~ 8529 rows
Test Set ~ 5682 rows

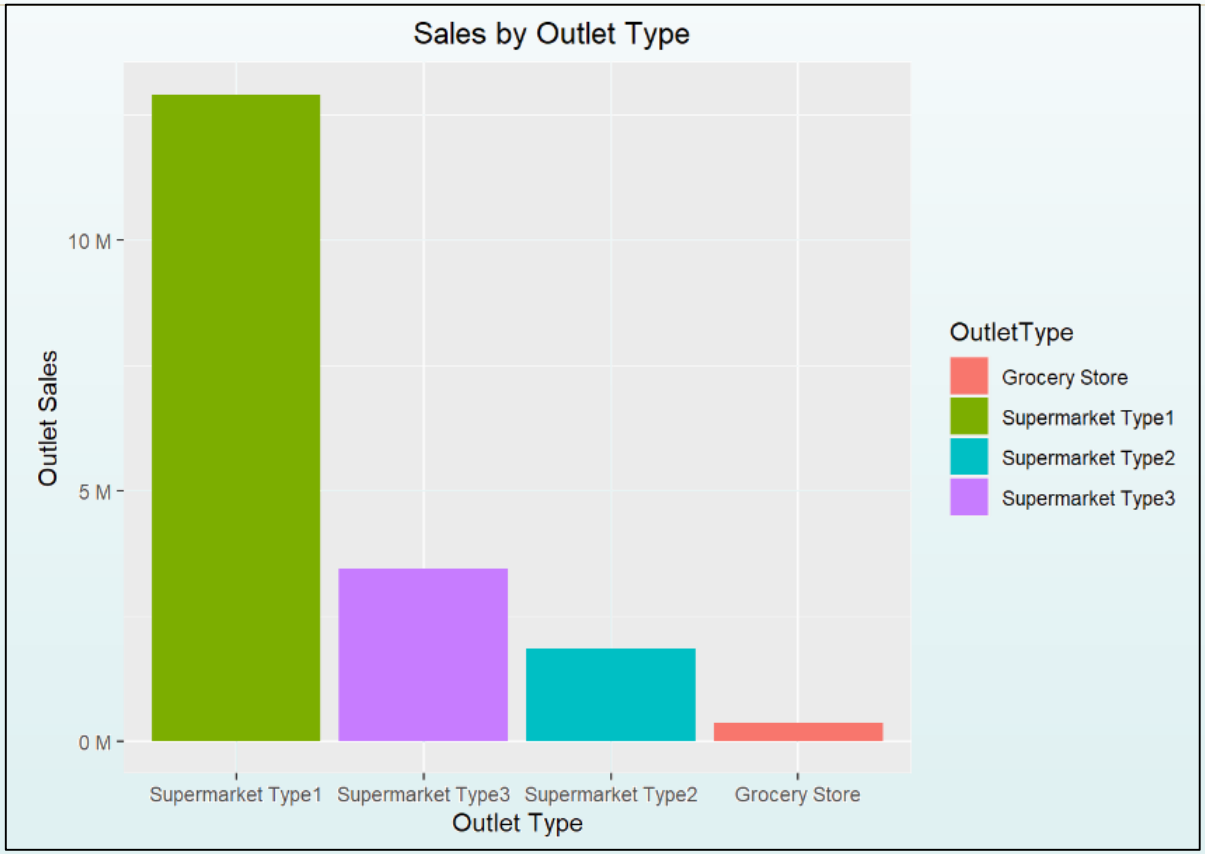
Preliminary Analysis

Identifying basic trends in the data by:

- Generating **scatter plots** and **correlation** coefficient tables to highlight trends.

- Performing **k-means clustering** and **PCA** to observe related product clusters and the dimensions contributing to the clusters.

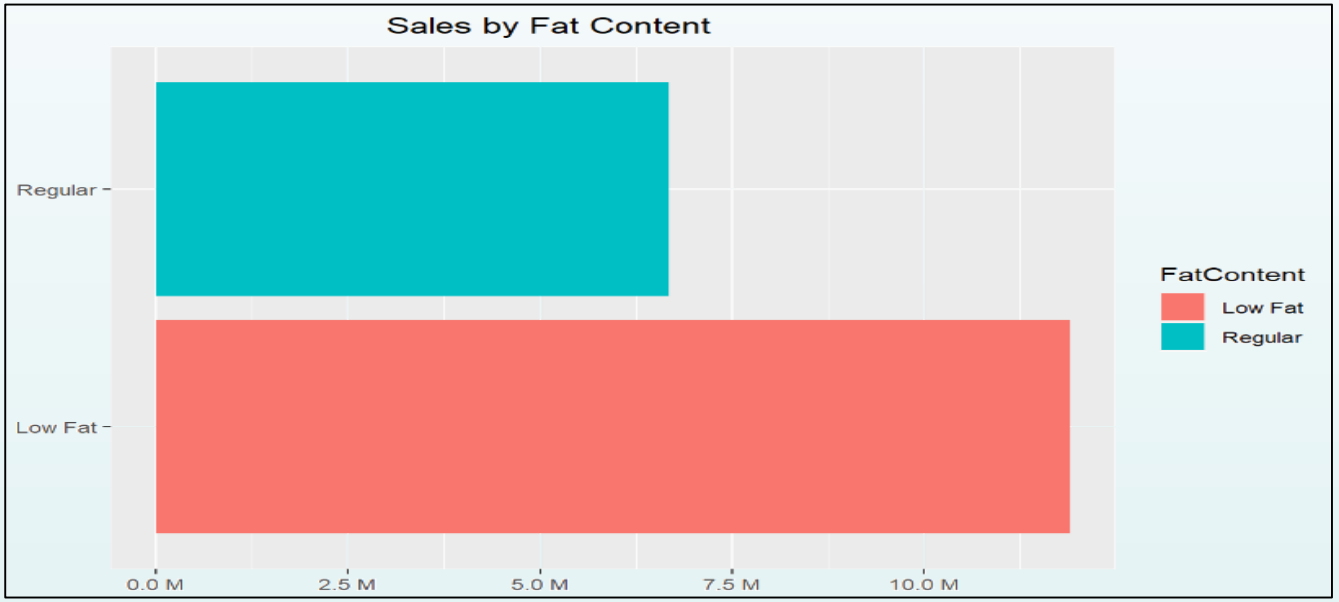
Type 1 Supermarkets outperform other three categories combined



Outlet Type	Count of ProductID
Supermarket Type1	5,577
Supermarket Type3	932
Supermarket Type2	928
Grocery Store	1,082
Grand Total	**8,519**

This indicates that Products that are sufficiently placed at these supermarkets are **more likely to grab higher market share.**

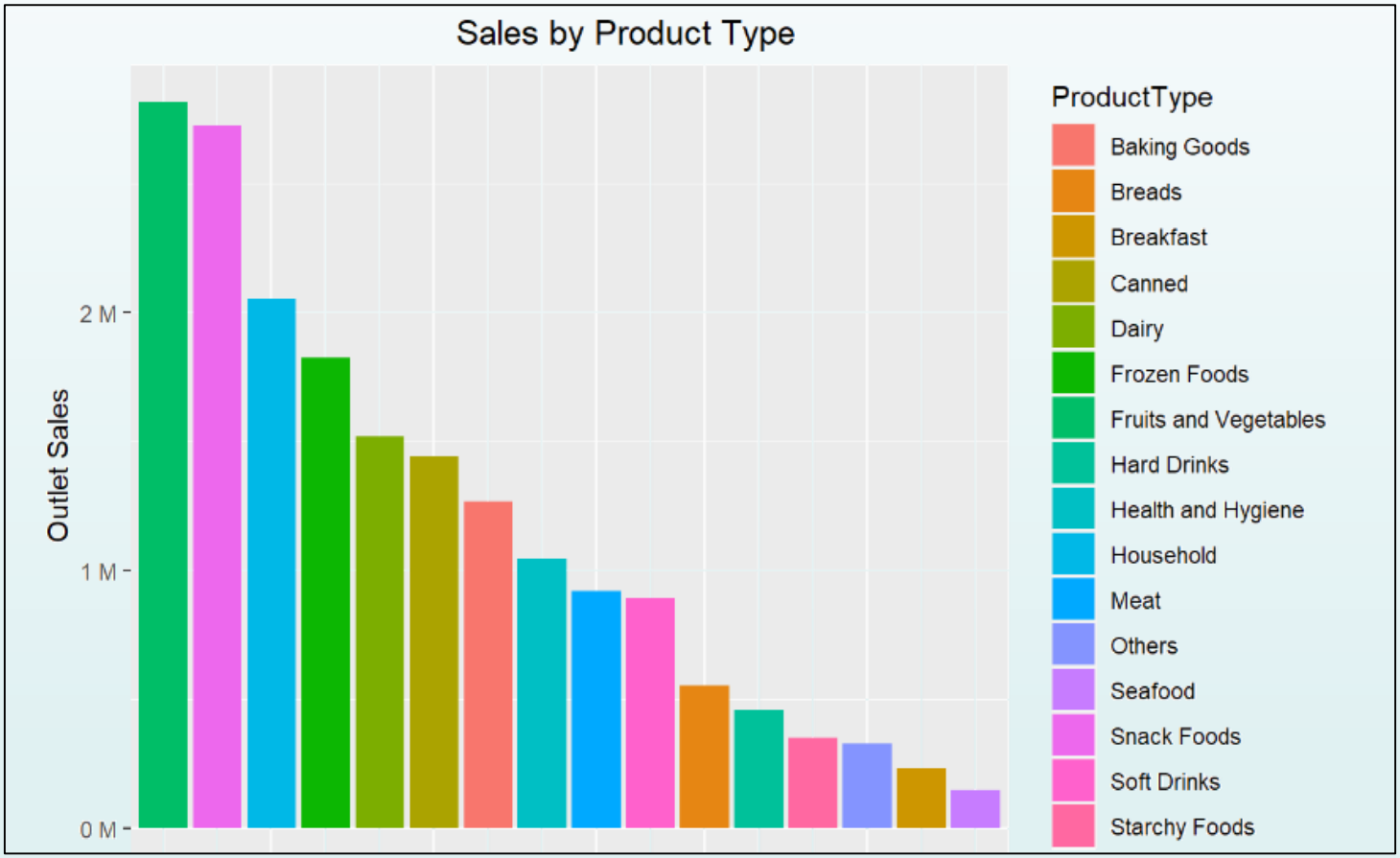
Low Fat foods sell at twice the volume than Regular Fat foods



FatContent	Sum of OutletSales	Count of ProductID	Percent of Total Sales	Percent of Total Product
Low Fat	11,899,660.31	5,516.00	64%	65%
Regular	6,681,886.91	3,003.00	36%	35%
Grand Total	**18,581,547.21**	**8,519.00**	**100%**	**100%**

This indicates that additional nutritional dimensions could reveal insights about customer preferences.

Retail Sales are dominated by Frozen and Snack Foods



Noticeable dips in Sales are seen for Breakfast and Seafood.

Correlation reveals one trend

Generating a simple correlation table between the quantitative variables in the dataset shows:

- A general positive trend between a Product's MRP and its Outlet Sales for a given OutletID

	Weight	ProductVisibility	MRP	OutletSales
Weight	1.00000000	-0.014047726	0.027141154	0.01412274
ProductVisibility	-0.01404773	1.000000000	-0.006061148	-0.08533404
MRP	0.02714115	-0.006061148	1.000000000	0.62096132
OutletSales	0.01412274	-0.085334041	0.620961316	1.00000000

k-means Clustering

k=3 since we have three main variables of interest

PCA analysis reveals ProductType and FatContent contribute most to Dim1 and FatContent contributes most to Dim2

k-means Clustering

	ProductVisibility	MRP	OutletSales
1	0.06072282	141.2139	2316.1811
2	0.10478223	140.3123	340.0312
3	0.05977723	140.8046	2847.4684

Three clusters reveal:

- Lower visibility products with higher MRPs contributing to higher OutletSales

- Higher visibility products with higher MRPs strangely having lower OutletSales.

Intermediate Analysis

Estimating Dependent Variable 'OutletSales' by:

- Utilizing Linear Multivariate Regression to identify a basic equation to estimate OutletSales for a given ProductID and OutletID.

- Eliminate variables that are beyond a level of significance of 0.05 and re-running the regression model to obtain a more fine-tuned equation to predict OutletSales.

- Multivariate regression model **yielded an RMSE of 1142.497**.

Predictive Analysis

Predicting Sales Prices of Newer Product Categories by:

- Employing **kNN and Regression Trees** to predict Sales Estimates for the test data.
- Identifying **values of k and mindev to optimize RMSE** for the aforementioned methods.
- Utilizing optimal k and mindev models to estimate product sales for newer product categories.

kNN and Regression Tree Analysis

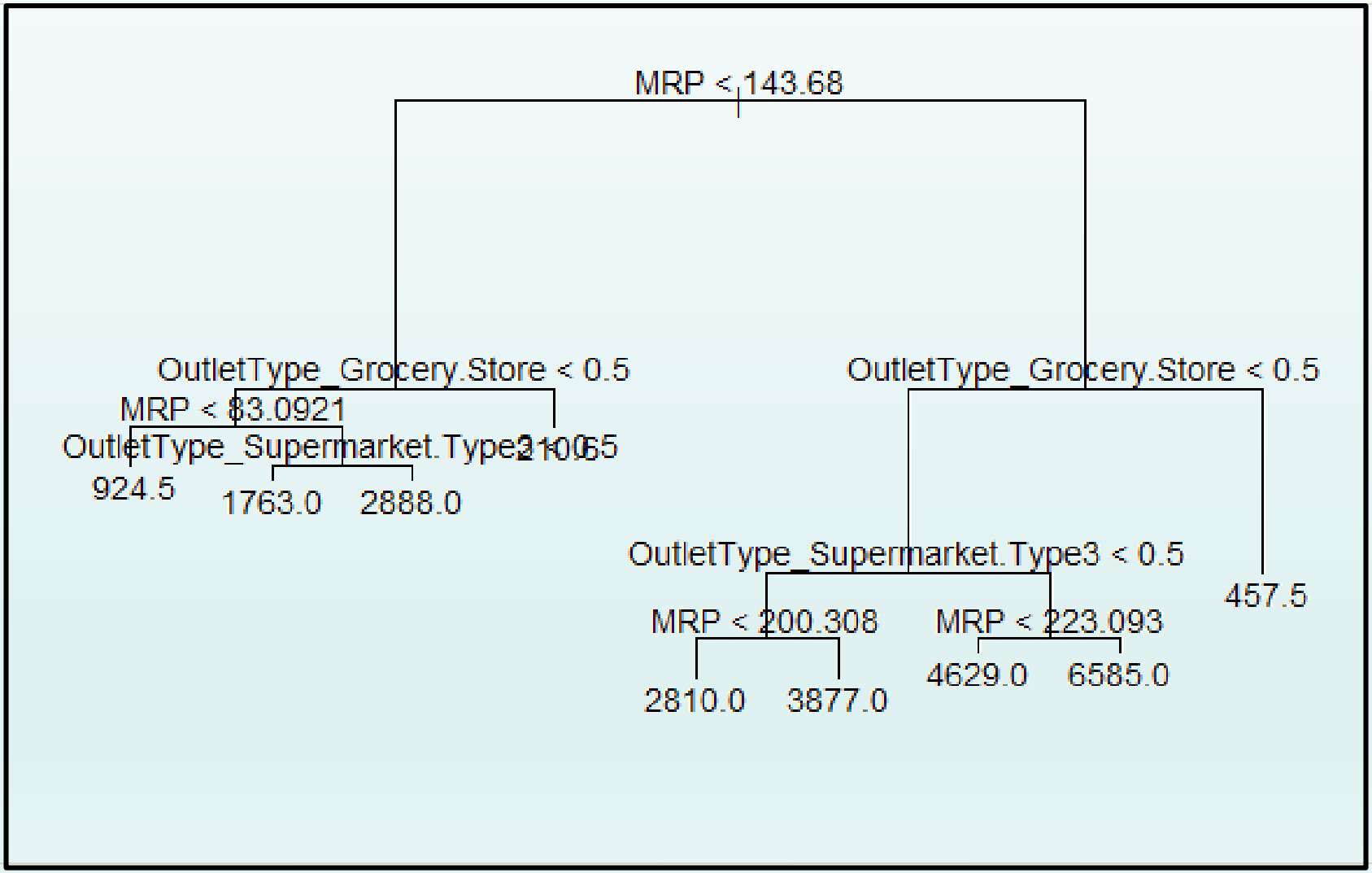
Performing analysis using 60% of the Train Set data to train the model and the remaining 40% for validation:

- Results in an **optimal k-value of 134** with an **RMSE of 1419.43** during kNN analysis.

- Results in an **optimal mindev value of 0.0015** with a **RMSE of 1099.4** .

- Thus, we implemented Regression trees to estimate sales values for newer products in the test set.

Regression Tree Output



KOGOD SCHOOL of BUSINESS WASHINGTON, DC

Conclusion

Bigmart can use these insights to improve Outlet Sales:

- Stocking inventory with product MRPs in the medium-to-high industry average category as these contribute to higher Outlet Sales.

- Focusing on scaling the Grocery and Hypermarkets segment since these tend to maximize higher MRP products being sold.

- Updating inventory to track products along additional nutritional dimensions (gluten-free, vegan, omega-3 content) for more accurate clustering and prediction results.