STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Bigmart Sales Prediction

CS 593: Data Mining II -Advanced Algorithms for Mining Big Data

Instructor: Prof Khasha Dehnad

Team:

Arya Mane (10411369)

Debapriya Pal (10408893)

Akash Chawla (10406187)

Soham Mukherjee (10409945)

Introduction

Problem Statement : Predict sales of product across multiple stores

- The Data was collected from Analytics Vidhya ,an India based Data Science and Analytics Competition and Knowledge sharing website .
- The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities.
- The aim is to build a predictive model and find out the sales of each product at a particular store.
- Data Source: http://datahack.analyticsvidhya.com/contest/practice-problem-bigmart-sales-prediction

Objective

- Exploratory analysis
- Data Cleaning
- Machine Learning Algorithms:
  - Linear regression
  - Decision tree
  - Random Forest
- Limitation & Future scope
- Conclusion

DataSet

Bigmart Dataset

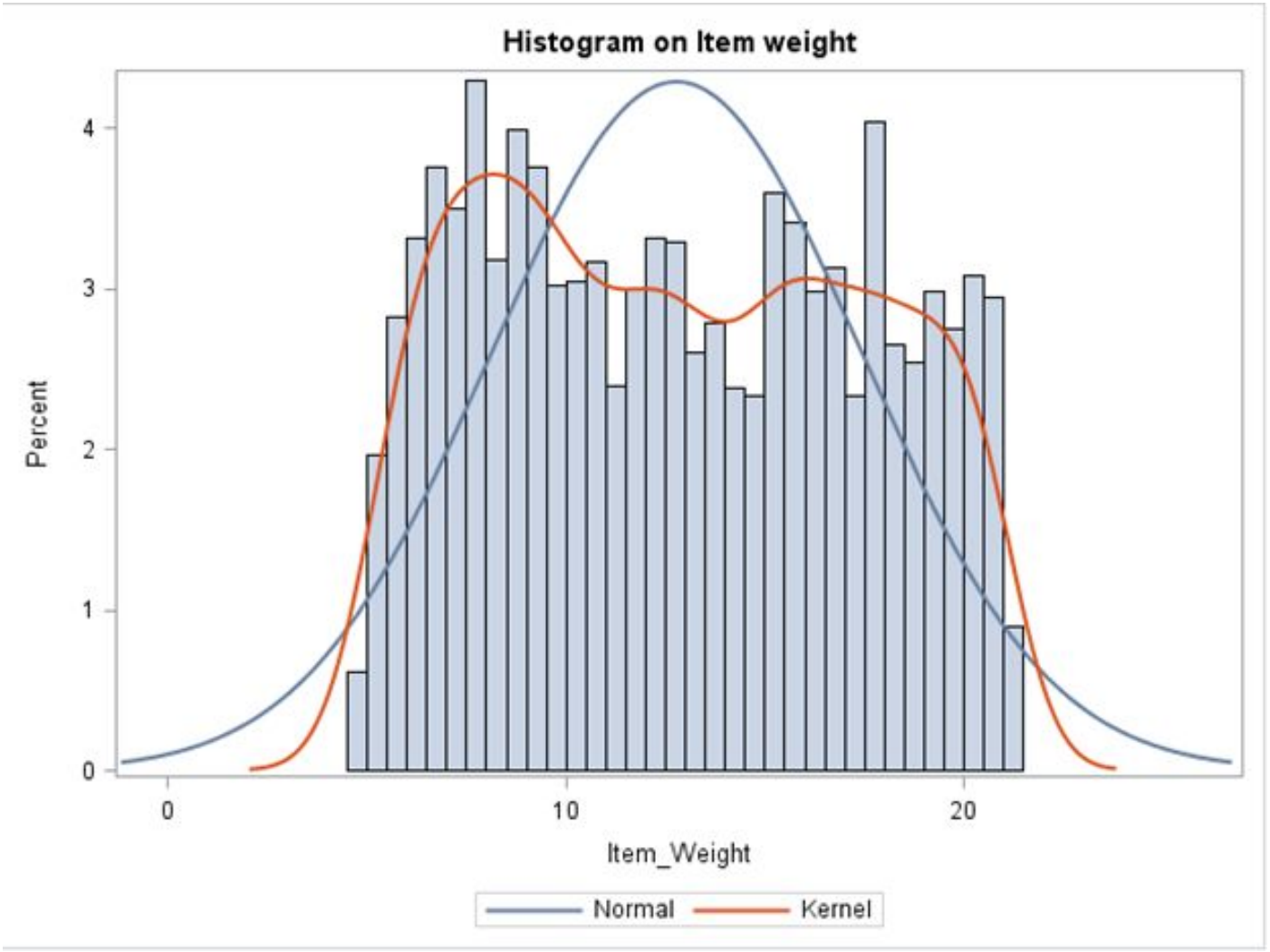
Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particulat store. This is the outcome variable to be predicted.

Bigmart Dataset

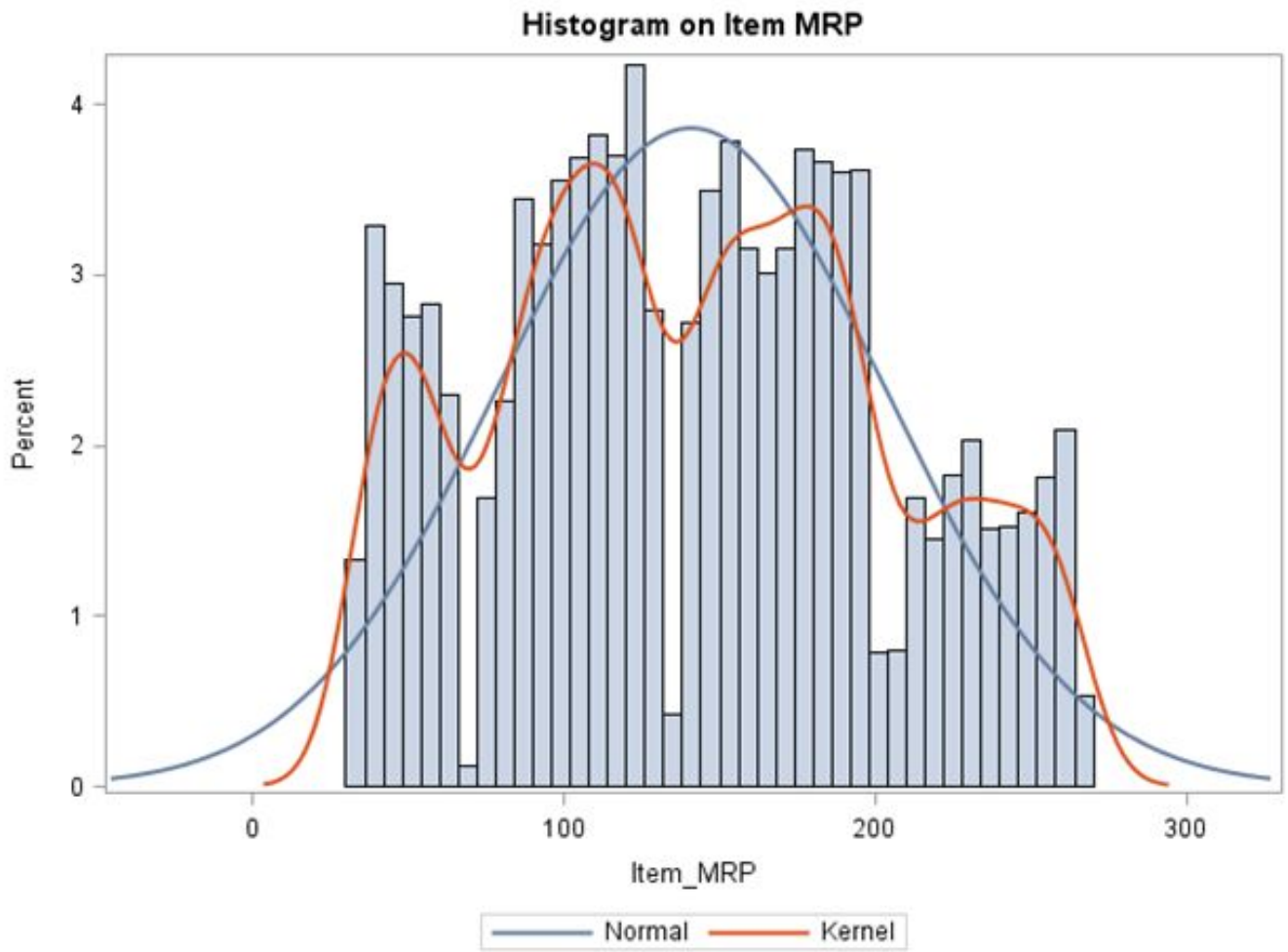
Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
FDX07	19.2	Regular	0	Fruits and Vegetables	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
FDP10		Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636
FDH17	16.2	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket Type1	1076.5986
FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007		Tier 2	Supermarket Type1	4710.535
FDY07	11.8	Low Fat	0	Fruits and Vegetables	45.5402	OUT049	1999	Medium	Tier 1	Supermarket Type1	1516.0266
FDA03	18.5	Regular	0.045463773	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket Type1	2187.153
FDX32	15.1	Regular	0.1000135	Fruits and Vegetables	145.4786	OUT049	1999	Medium	Tier 1	Supermarket Type1	1589.2646
FDS46	17.6	Regular	0.047257328	Snack Foods	119.6782	OUT046	1997	Small	Tier 1	Supermarket Type1	2145.2076
FDF32	16.35	Low Fat	0.0680243	Fruits and Vegetables	196.4426	OUT013	1987	High	Tier 3	Supermarket Type1	1977.426
FDP49	9	Regular	0.069088961	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket Type1	1547.3192
NCB42	11.8	Low Fat	0.008596051	Health and Hygiene	115.3492	OUT018	2009	Medium	Tier 3	Supermarket Type2	1621.8888
FDP49	9	Regular	0.069196376	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket Type1	718.3982
DRI11		Low Fat	0.034237682	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket Type3	2303.668

Exploratory Analysis

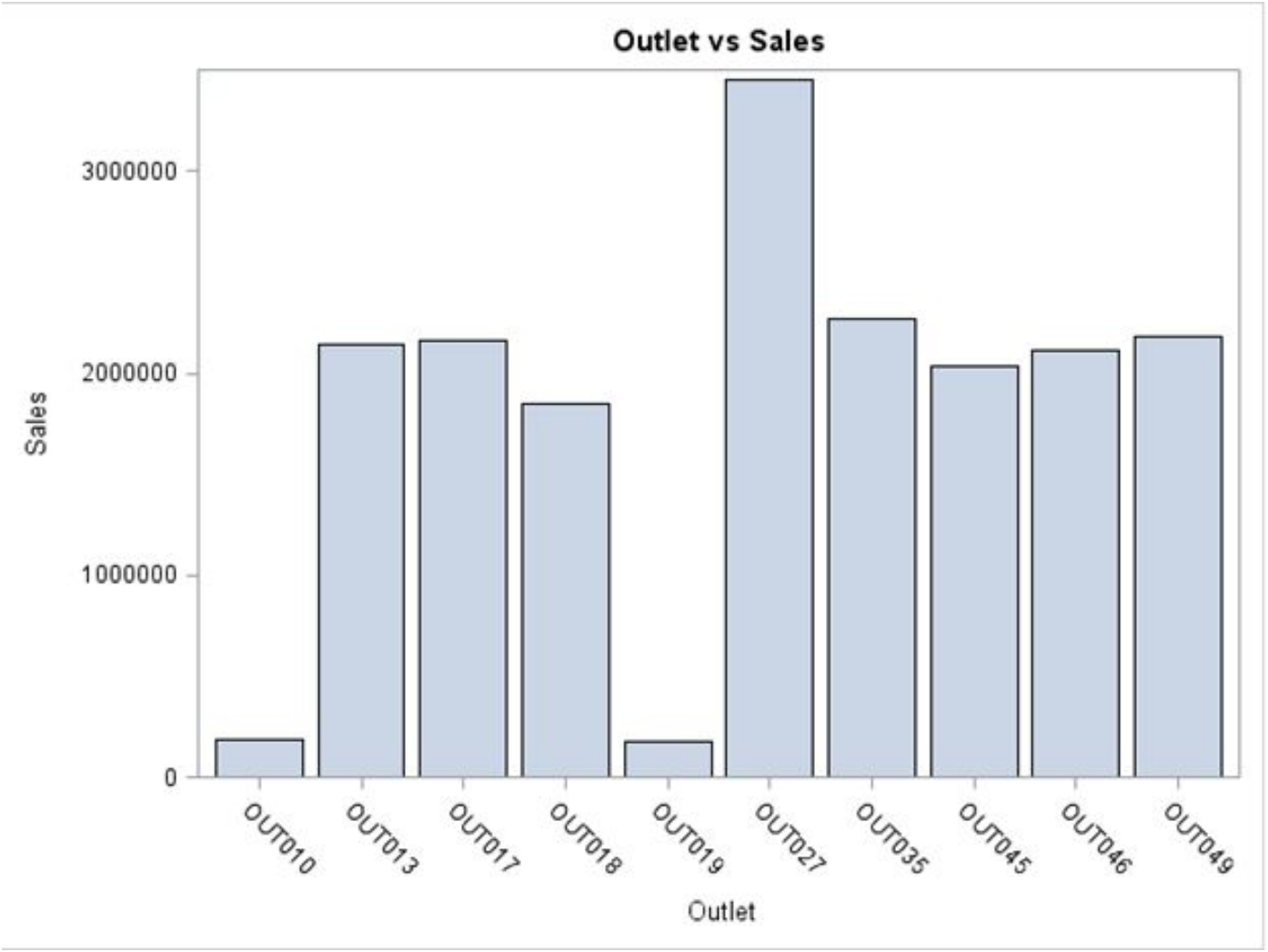
Exploratory Analysis



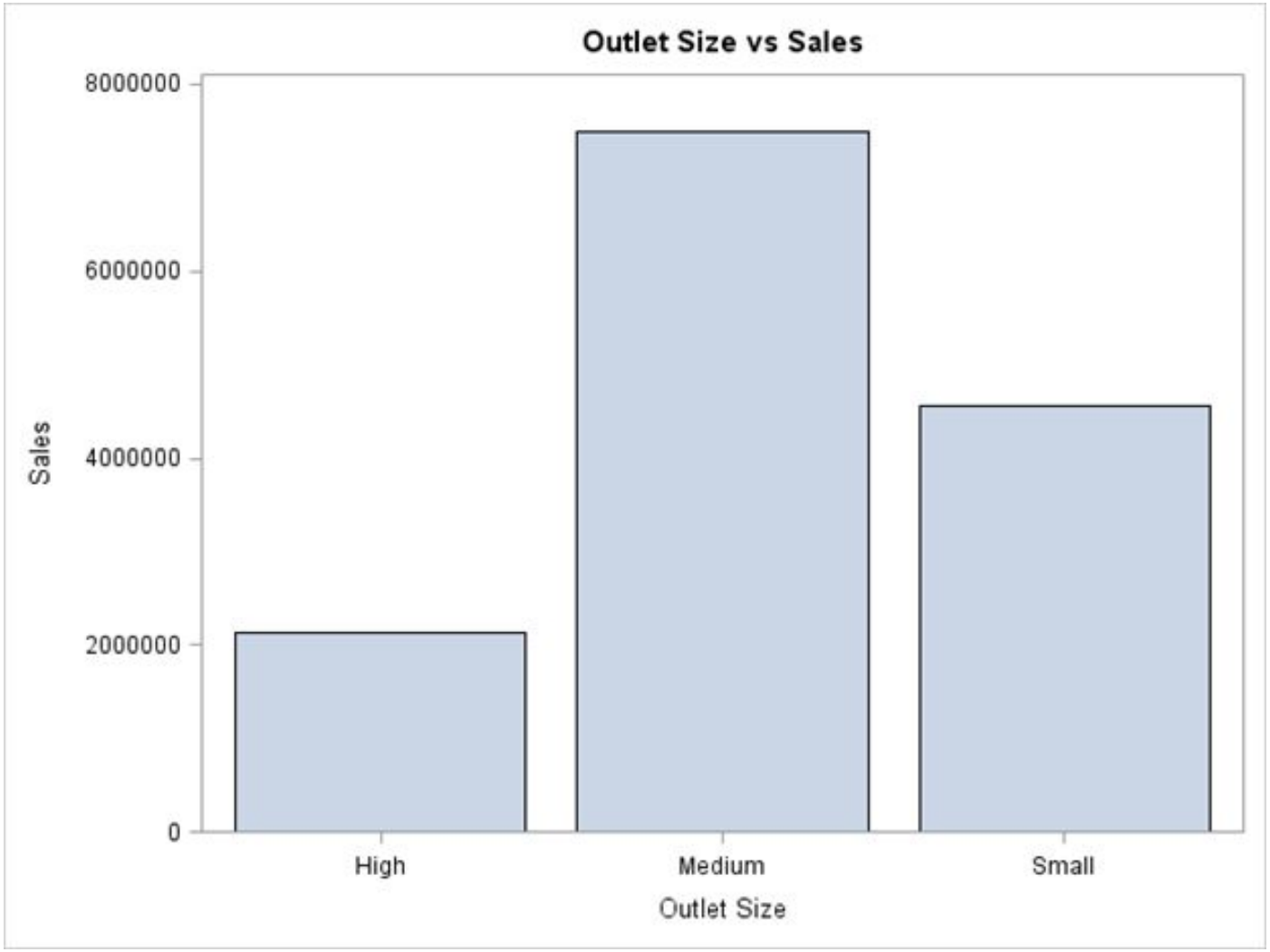
Exploratory Analysis



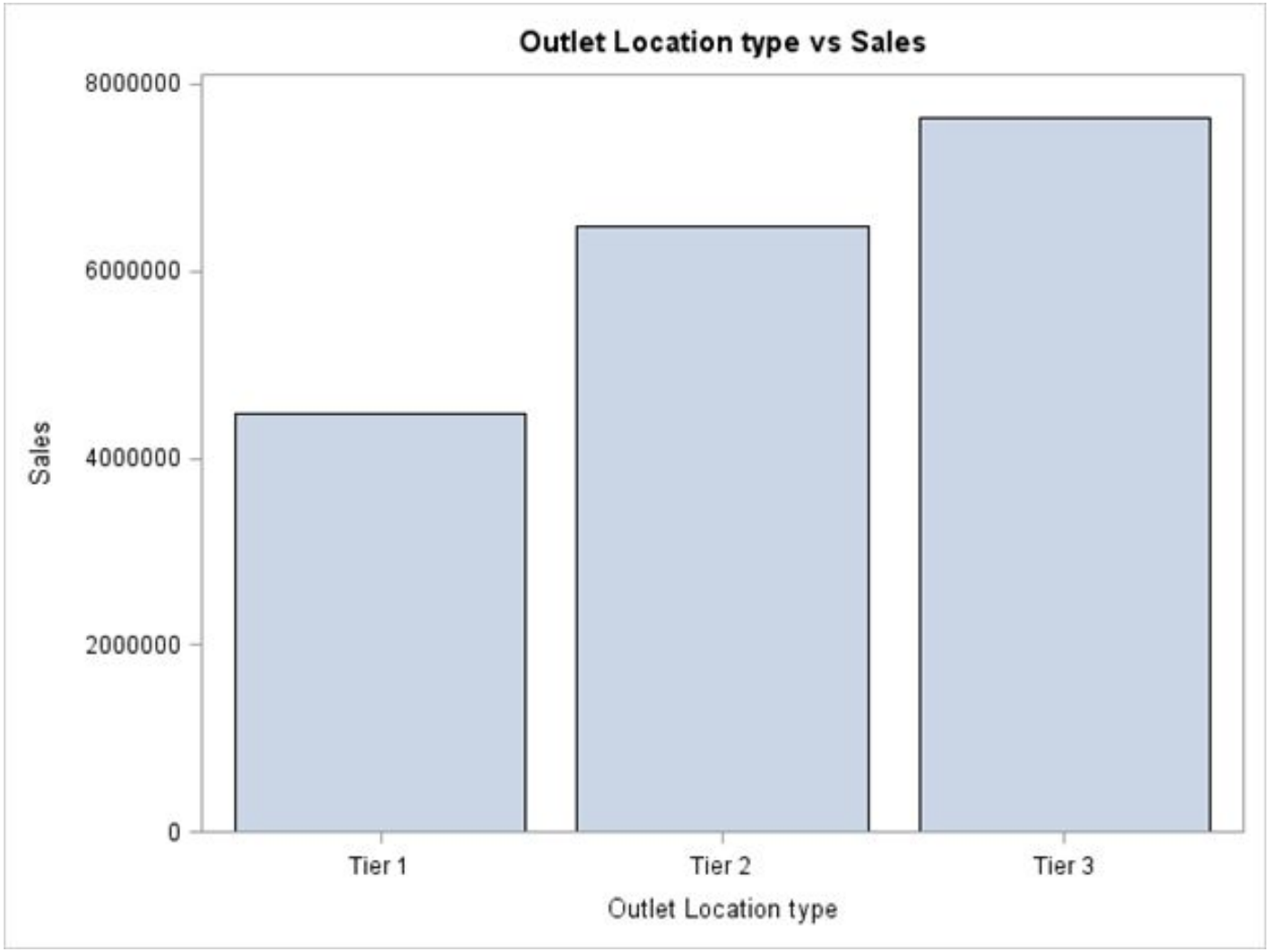
Exploratory Analysis



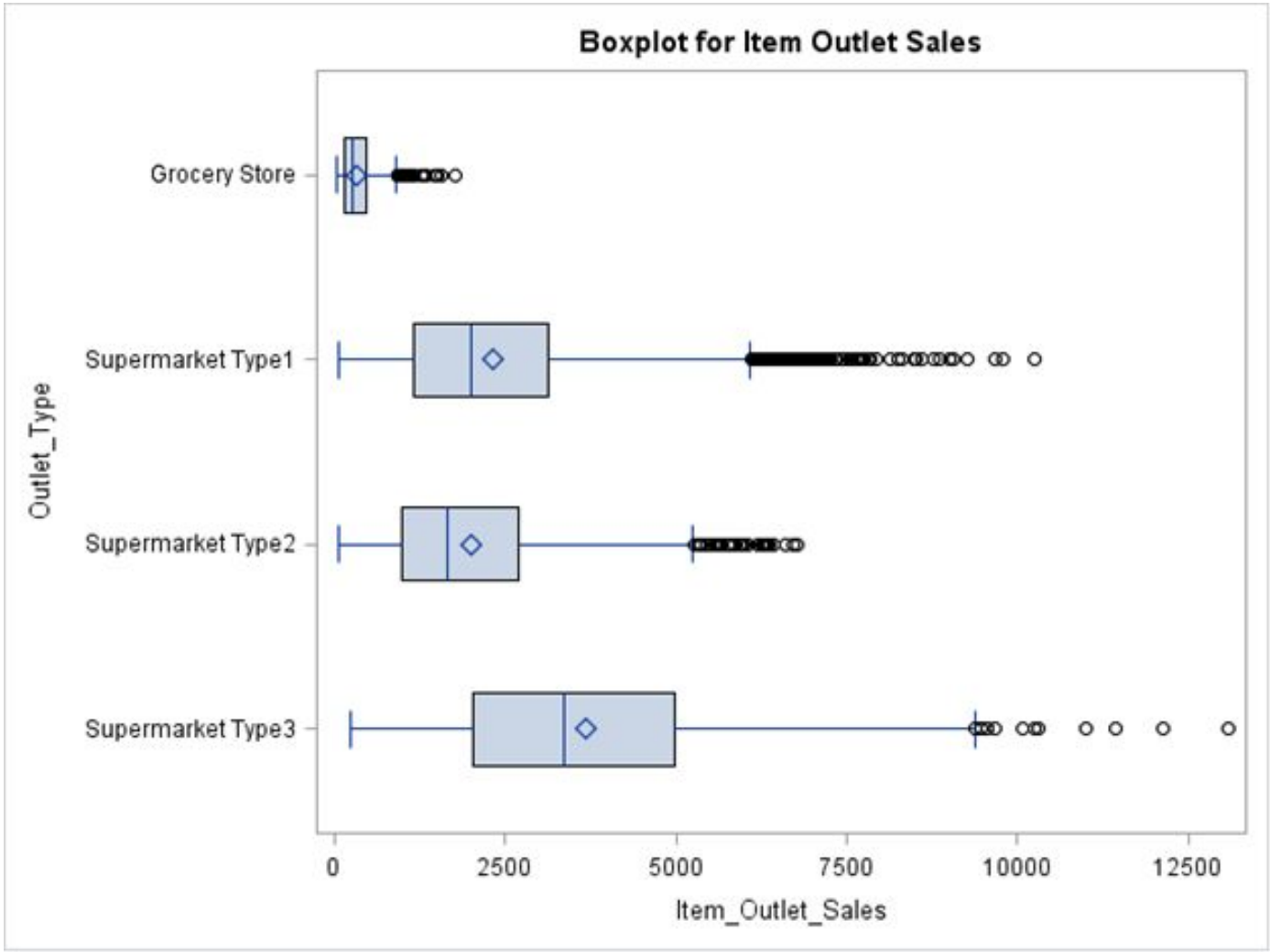
Exploratory Analysis



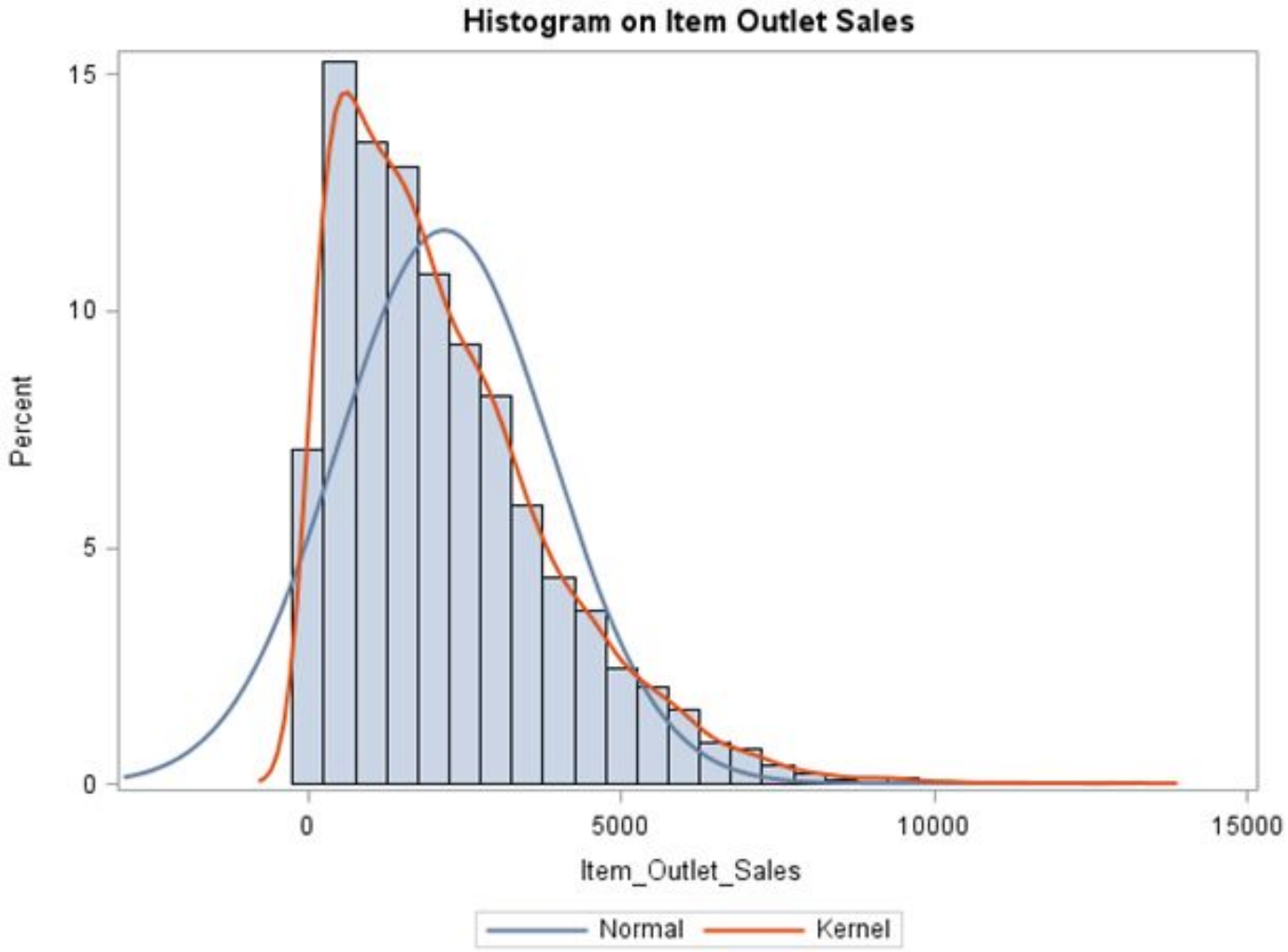
Exploratory Analysis



Exploratory Analysis



Exploratory Analysis



Data Cleaning

Add new column: item_type_cat

- Insert item_type_cat column as per item_identifier. Eg : item_type_cat = Food for item_identifier like 'FD%'

Item_Identifier	item_type_cat
FDA15	Food
DRC01	Drinks
FDN15	Food
FDX07	Food
NCD19	NonConsumable
FDP36	Food
FDO10	Food
FDP10	Food
FDH17	Food
FDU28	Food
FDY07	Food
FDA03	Food
FDX32	Food
FDS46	Food
FDF32	Food
FDP49	Food
NCB42	NonConsumable
FDP49	Food
DRI11	Drinks

Clean and update item_fat_content column

- Eg: Replace **LF** and **low fat** with **Low Fat**

Item_Fat_Content
LF
Low Fat
Regular
low fat
reg

Before cleaning the data

Item_Fat_Content
Low Fat
NonEdib
Regular

After cleaning the data

Add new column: years_of_operation

- Calculate the year of operation till 2013

Outlet_Establishment_Year	years_of_operation
1999	14
2009	4
1999	14
1998	15
1987	26
2009	4
1987	26
1985	28
2002	11
2007	6
1999	14
1997	16
1999	14
1997	16
1987	26
1997	16
2009	4
1999	14
1985	28

Update outlet_size column

If outlet_size = **blank,** then update outlet_size column to "small" for outlet_type = Grocery Store and Supermarket Type1

Outlet_Type	Outlet_Size	count_obs
Grocery Store		0
Grocery Store	Small	880
Supermarket Type1		0
Supermarket Type1	High	1553
Supermarket Type1	Medium	1550
Supermarket Type1	Small	3100
Supermarket Type2	Medium	1546
Supermarket Type3	Medium	1559

Outlet_Type	Outlet_Size	count_obs
Grocery Store	Small	1805
Supermarket Type1	High	1553
Supermarket Type1	Medium	1550
Supermarket Type1	Small	6191
Supermarket Type2	Medium	1546
Supermarket Type3	Medium	1559

Update item_weight

If item_weight is **missing** , **then** update with the **median** value

Before Cleaning the Data

Missing Values			
Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	2439	17.17	100.00

Basic Statistical Measures			
Location		Variability	
Mean	12.79285	Std Deviation	4.65250
Median	12.60000	Variance	21.64578
Mode	17.60000	Range	16.79500
		Interquartile Range	8.04000

After Cleaning the Data

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
4.555	12421	21.35	11847
4.555	9531	21.35	12185
4.555	9097	21.35	12592
4.555	7809	21.35	13484
4.555	4431	21.35	13531

Basic Statistical Measures			
Location		Variability	
Mean	12.75974	Std Deviation	4.23485
Median	12.60000	Variance	17.93396
Mode	12.60000	Range	16.79500
		Interquartile Range	6.70000

Update item_visibility

If item_visibility is 0 update with median value

Before Cleaning the data

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	14203	0.313935	12732
0	14185	0.321115	1806
0	14178	0.323637	11133
0	14172	0.325781	3751
0	14167	0.328391	855

Basic Statistical Measures			
Location		Variability	
Mean	0.065953	Std Deviation	0.05146
Median	0.054021	Variance	0.00265
Mode	0.000000	Range	0.32839
		Interquartile Range	0.06700

After Cleaning the data

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.00357470	7552	0.313935	12732
0.00358910	3863	0.321115	1806
0.00359141	11423	0.323637	11133
0.00359209	12573	0.325781	3751
0.00359768	7465	0.328391	855

Basic Statistical Measures			
Location		Variability	
Mean	0.069296	Std Deviation	0.04875
Median	0.054023	Variance	0.00238
Mode	0.054021	Range	0.32482
		Interquartile Range	0.06090

Machine Learning Algorithms

Linear Regression

In statistics, **linear regression** is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables)

```
Title1 "Proc Reg for Item_outlet_sales";
proc reg data=Bigmart_cat_data_train;
    model log_Item_outlet_sales =   Item_MRP
                                Outlet_size_Medium
                                Outlet_Type_Grocery Outlet_Type_Super_3  / VIF dwProb ;
    OUTPUT OUT = reg_bigmart_OUT RESIDUAL=c_Res h=lev cookd=Cookd dffits=dffit;
    run;
quit;
```

Linear Regression

First Iteration

Proc Reg for Item_outlet_sales

The REG Procedure
Model: MODEL1
Dependent Variable: Item_Outlet_Sales

Number of Observations Read	8523
Number of Observations Used	8523

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	13984243784	932282919	732.11	<.0001
Error	8507	10833021294	1273424		
Corrected Total	8522	24817265078			

Root MSE	1128.46107	R-Square	0.5635
Dependent Mean	2181.28891	Adj R-Sq	0.5627
Coeff Var	51.73368		

Linear Regression

First Iteration

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Outlet_09 =	- 7 * Intercept + 0.5 * years_of_operation - 0.5 * Outlet_01 - 6 * Outlet_02 + 4 * Outlet_03 + 5 * Outlet_04 - 7 * Outlet_05 - 7 * Outlet_06 + 2.5 * Outlet_07 + 1.5 * Outlet_08
Outlet_10 =	8 * Intercept - 0.5 * years_of_operation - 0.5 * Outlet_01 + 5 * Outlet_02 - 5 * Outlet_03 - 6 * Outlet_04 + 6 * Outlet_05 + 6 * Outlet_06 - 3.5 * Outlet_07 - 2.5 * Outlet_08
Outlet_Size_High =	Outlet_02
Outlet_Size_Medium =	8 * Intercept - 0.5 * years_of_operation - 0.5 * Outlet_01 + 5 * Outlet_02 - 5 * Outlet_03 - 5 * Outlet_04 + 6 * Outlet_05 + 7 * Outlet_06 - 3.5 * Outlet_07 - 2.5 * Outlet_08
Outlet_Size_Small =	- 7 * Intercept + 0.5 * years_of_operation + 0.5 * Outlet_01 - 6 * Outlet_02 + 5 * Outlet_03 + 5 * Outlet_04 - 6 * Outlet_05 - 7 * Outlet_06 + 3.5 * Outlet_07 + 2.5 * Outlet_08
Outlet_Location_Tier_1 =	Intercept - Outlet_01 - Outlet_02 - Outlet_03 - Outlet_04 - Outlet_06 - Outlet_07 - Outlet_08
Outlet_Location_Tier_2 =	Outlet_03 + Outlet_07 + Outlet_08
Outlet_Location_Tier_3 =	Outlet_01 + Outlet_02 + Outlet_04 + Outlet_06
Outlet_Type_Grocery =	Outlet_01 + Outlet_05
Outlet_Type_Super_1 =	Intercept - Outlet_01 - Outlet_04 - Outlet_05 - Outlet_06
Outlet_Type_Super_2 =	Outlet_04
Outlet_Type_Super_3 =	Outlet_06

Linear Regression

Second Iteration

Proc Reg for Item_outlet_sales

The REG Procedure
Model: MODEL1
Dependent Variable: Item_Outlet_Sales

Number of Observations Read	8523
Number of Observations Used	8523

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	13970729390	1074671492	843.07	<.0001
Error	8509	10846535688	1274713		
Corrected Total	8522	24817265078			

Root MSE	1129.03203	R-Square	0.5629
Dependent Mean	2181.28891	Adj R-Sq	0.5623
Coeff Var	51.75986		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2342.59530	323.16104	7.25	<.0001	0
Item_Weight	1	-0.59928	2.90073	-0.21	0.8363	1.00521
Item_Visibility	1	-239.97800	262.71091	-0.91	0.3610	1.10292
Item_MRP	1	15.56437	0.19672	79.12	<.0001	1.00339
years_of_operation	1	-30.56543	10.41848	-2.93	0.0034	50.85942
Fat_Regular	1	50.48447	26.05691	1.94	0.0527	1.03641
Outlet_Size_High	1	752.54104	254.70518	2.95	0.0031	42.24598
Outlet_Size_Medium	1	36.71639	56.35214	0.65	0.5147	4.67780
Outlet_Location_Tier_1	1	415.61434	151.85473	2.74	0.0062	31.09571
Outlet_Location_Tier_2	1	251.93587	100.17044	2.52	0.0119	14.75910
Outlet_Type_Grocery	1	-3720.09335	175.87425	-21.15	<.0001	22.94039
Outlet_Type_Super_1	1	-2196.48747	294.23768	-7.47	<.0001	130.92572
Outlet_Type_Super_2	1	-2460.58796	255.45321	-9.63	<.0001	42.33441
Item_Type_Drinks	1	-8.72596	42.76085	-0.20	0.8383	1.03867

Linear Regression

Third Iteration

Proc Reg for Item_outlet_sales

The REG Procedure
Model: MODEL1
Dependent Variable: Item_Outlet_Sales

Number of Observations Read	8523
Number of Observations Used	8523

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	13885791348	3471447837	2705.01	<.0001
Error	8518	10931473730	1283338		
Corrected Total	8522	24817265078			

Root MSE	1132.84512	R-Square	0.5595
Dependent Mean	2181.28891	Adj R-Sq	0.5593
Coeff Var	51.93467		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	110.19609	32.44331	3.40	0.0007	0
Item_MRP	1	15.55579	0.19706	78.94	<.0001	1.00008
Outlet_Size_Medium	1	-131.25050	31.09471	-4.22	<.0001	1.41470
Outlet_Type_Grocery	1	-1952.76255	38.22560	-51.09	<.0001	1.07641
Outlet_Type_Super_3	1	1540.36541	45.42376	33.91	<.0001	1.33836

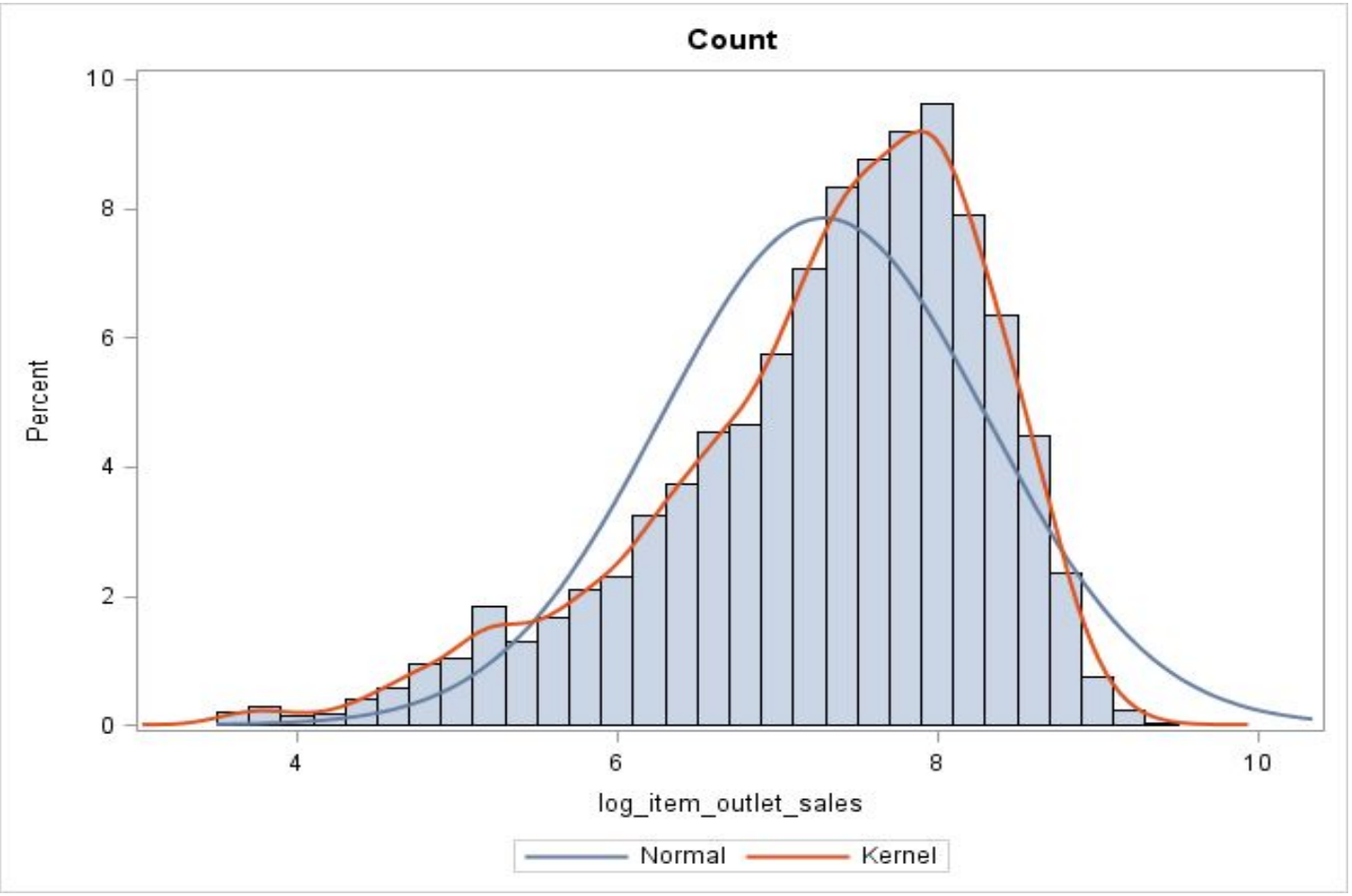
Proc Reg for Item_outlet_sales

The REG Procedure
Model: MODEL1
Dependent Variable: Item_Outlet_Sales

Durbin-Watson D	2.005
Pr < DW	0.5926
Pr > DW	0.4074
Number of Observations	8523
1st Order Autocorrelation	-0.003

Linear Regression

Log Transformed Item_Outlet_Sales



Linear Regression

Fourth Iteration : Proc Reg on Log transformed Item_outlet_sales

Proc Reg for Item_outlet_sales

The REG Procedure
Model: MODEL1
Dependent Variable: log_item_outlet_sales

Number of Observations Read	8523
Number of Observations Used	8523

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6325.31669	1581.32917	5407.87	<.0001
Error	8518	2490.77063	0.29241		
Corrected Total	8522	8816.08731			

Root MSE	0.54075	R-Square	0.7175
Dependent Mean	7.29654	Adj R-Sq	0.7173
Coeff Var	7.41107		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6.33099	0.01549	408.81	<.0001	0
Item_MRP	1	0.00830	0.00009407	88.21	<.0001	1.00008
Outlet_Size_Medium	1	-0.07024	0.01484	-4.73	<.0001	1.41470
Outlet_Type_Grocery	1	-1.94923	0.01825	-106.83	<.0001	1.07641
Outlet_Type_Super_3	1	0.60477	0.02168	27.89	<.0001	1.33836

Proc Reg for Item_outlet_sales

The REG Procedure
Model: MODEL1
Dependent Variable: log_item_outlet_sales

Durbin-Watson D	2.010
Pr < DW	0.6770
Pr > DW	0.3230
Number of Observations	8523
1st Order Autocorrelation	-0.005

RMSE
1146.17

Decision Tree (HPSPLIT)

Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree.

```
proc hpsplit data=Bigmart_cat_data_train ;
target Item_outlet_sales / level= INT;
input outlet_size outlet_type /level=NOM;
input item_MRP /level=INT;
criterion variance;
PRUne costcomplexity;
rules file='bigmart_tree_rules.txt';
score out=Dtree_out;

run;
quit;
```

	Item_Outlet_Sales	Node number	Leaf number	Predicted: Item_Outlet_Sales
1	3735.138	69	31	4050.9621882
2	443.4228	19	4	707.92498563
3	2097.27	44	15	2287.2814871
4	732.38	6	1	473.53079481
5	994.7052	19	4	707.92498563
6	556.6088	19	4	707.92498563
7	343.5528	19	4	707.92498563
8	4022.7636	21	6	2540.52635
9	1076.5986	55	20	1589.9551612
10	4710.535	78	38	3165.2586781
11	1516.0266	19	4	707.92498563
12	2187.153	44	15	2287.2814871
13	1589.2646	59	24	2643.4061576
14	2145.2076	58	23	2065.7554667
15	1977.426	64	28	2903.2977231
16	1547.3192	19	4	707.92498563
17	1621.8888	33	12	1736.4983855
18	718.3982	19	4	707.92498563
19	2303.668	22	7	3260.702236

Decision Tree

	Item_Outlet_Sales	Node number	Leaf number	Predicted: Item_Outlet_Sales	Squared_error
1	3735.138	69	31	4050.9621882	99744.917874
2	443.4228	19	4	707.92498563	69961.406203
3	2097.27	44	15	2287.2814871	36104.365214
4	732.38	6	1	473.53079481	67002.91103
5	994.7052	19	4	707.92498563	82242.891354
6	556.6088	19	4	707.92498563	22896.588034
7	343.5528	19	4	707.92498563	132767.08966
8	4022.7636	21	6	2540.52635	2197027.2653
9	1076.5986	55	20	1589.9551612	263534.95888
10	4710.535	78	38	3165.2586781	2387878.9109
11	1516.0266	19	4	707.92498563	653028.21915
12	2187.153	44	15	2287.2814871	10025.713921
13	1589.2646	59	24	2643.4061576	1111214.4236
14	2145.2076	58	23	2065.7554667	6312.6414912
15	1977.426	64	28	2903.2977231	857238.44759
16	1547.3192	19	4	707.92498563	704582.64712
17	1621.8888	33	12	1736.4983855	13135.357095
18	718.3982	19	4	707.92498563	109.68821924
19	2303.668	22	7	3260.702236	915914.52888

RMSE
1067.33

Random Forest (HPFOREST)

Random forests is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

```
proc hpforest data = Bigmart_cat_data_train;
    target log_Item_outlet_sales/level = interval;

    input outlet_size outlet_type / level=nominal;
    input item_MRP / level = interval;

    ods output fitstatistics = fitstats;

    save file = "D:\model.bin";

run;
```

```
proc hp4score data = Sample_Bigmart_Test;
    id log_Item_outlet_sales;
    score file = "D:\model.bin" out = Big_scored;

run;
quit;
```

	Item_Outlet_Sales	Predicted: Item_Outlet_Sales
1	994.7052	820.24657559
2	343.5528	802.1418576
3	4022.7636	2609.6565955
4	1589.2646	2509.4107632
5	1547.3192	888.10078848
6	718.3982	867.69618564
7	2748.4224	3832.2074736
8	3775.086	4116.5123767
9	214.3876	254.26518947
10	1065.28	679.53475784
11	125.8362	129.01524791
12	2797.6916	2966.4535237
13	388.1614	1357.286461
14	2527.3768	3122.0246658
15	810.9444	1930.0978372

Random Forest

	Item_Outlet_Sales	Predicted: Item_Outlet_Sales	Warnings	squared_difference
1	994.7052	820.24657559		30435.811632
2	343.5528	802.1418576		210303.92375
3	4022.7636	2609.6565955		1996871.4062
4	1589.2646	2509.4107632		846668.96172
5	1547.3192	888.10078848		434568.91409
6	718.3982	867.69618564		22289.888516
7	2748.4224	3832.2074736		1174590.0857
8	3775.086	4116.5123767		116571.9707
9	214.3876	254.26518947		1590.2221418
10	1065.28	679.53475784		148799.39185
11	125.8362	129.01524791		10.106345612
12	2797.6916	2966.4535237		28480.586889
13	388.1614	1357.286461		939203.38378
14	2527.3768	3122.0246658		353606.08435
15	810.9444	1930.0978372		1252504.4159
16	6258.52	3605.7680512		7037092.9019
17	2117.244	1669.6075007		200378.4355
18	4910.275	3075.2392567		3367356.1794
19	1062.6168	1415.5484872		124560.77585
20	484.7024	1357.286461		761402.94344

RMSE
1062.467

Limitations and Future Scope

- Most of the independent variables are categorical whereas dependent variable is continuous - Linear Regression is not a good fit.
- Amount of data is less which results in partial training of model and overfitting.
- Use of KNN regression to predict the sales.
- Carefully pruning and adjusting max and min depth of the decision tree.
- Using different splitting criterion for HPSPLIT.
- Deeper analysis of data will give better insights.

Conclusion

HPFOREST or High Performance Forest which is used to create Random Forest gives the best RMSE (Root mean square error) score based on our analysis.

Linear Regression on
Log transformed Data

RMSE
1146.17

Decision Tree (HPSPLIT)

RMSE
1067.33

Random Forest Regression (HPFOREST)

RMSE
1062.467

Thank you!
Questions?