

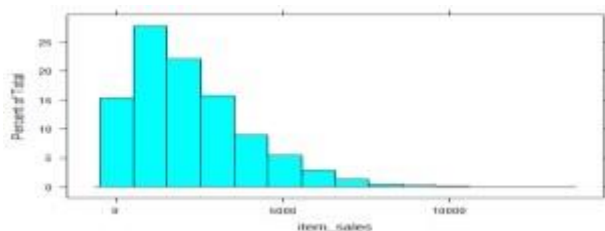
# Big Mart Sales – Multi Level Regression Analysis

## Relevant Independent variables

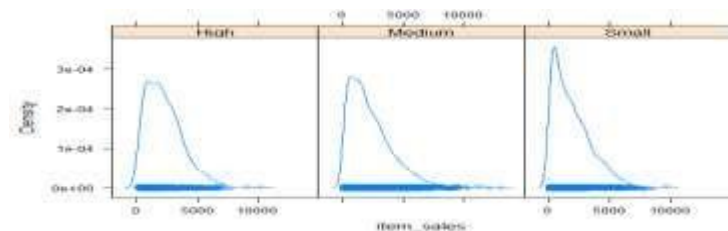
Predictors	Effect	Rationale
Item_Visibility	+	Item which has high visibility tend to sale higher as people prefer to buy that item.
Item_Fat_Content	+/-	Items with less fat content may sale more as people may prefer to buy less fat content products but it depends on item type and item mrp resulting into both positive or negative effect.
Item_Type	+/-	Different types of items such as fruits, vegetables and dairy products are necessities and hence may get sold more than other item types.
Item_Weight	+/-	People may prefer light weight items or heavy items based on item types, item visibility and item mrp.
Itemprice_perweight	+/-	I have generated one feature which is the ratio of item_price and item_weight. Generally people prefer getting more quantity of items in less price. Also in some cases people like to stick to one brand irrespective of price and weight.
Outlet_size	+/-	People may prefer different outlet sizes depending on city type, outlet types.
Outlet_type	+/-	Item sales depends on the type of the outlets along with the outlet sizes and item mrp of the respective item types.
Outlet_id	+/-	I don't think this variable affect directly to item sales bu I have included this as a random effect to answer question number 3.
City_Type	+/-	I don't think this variable affect directly to item sales but I have included this as a random effect to answer question number 2.
Store_years_of_operation	+/-	I have generated a new column which represents number of years the store is operating. Considering our dataset is from 2013 I have subtracted each Outlet_Year from 2013. Generally people prefer older stores which they visit frequently. In some cases people like to explore new stores as they tend to give some offers for promotion.

I have not used itemID as we are mostly focusing on outlet and city. Item\_price and item\_mrp become irrelevant as they have been already used to form new features.

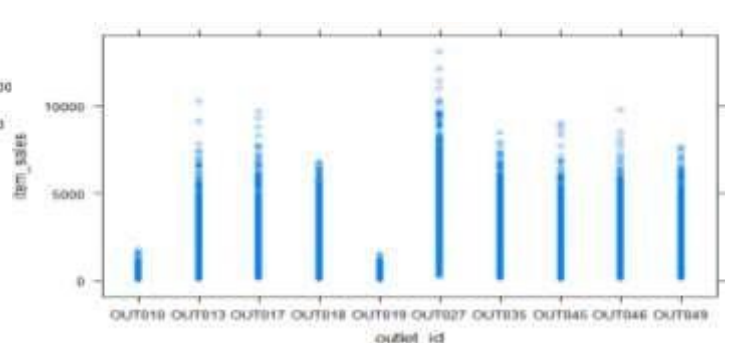
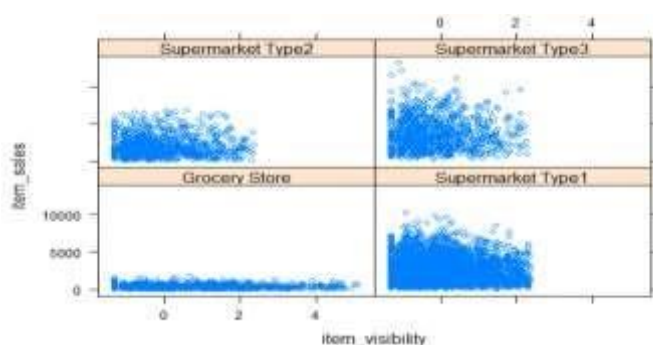
## Exploratory Data Analysis.



Histogram of Item\_Sales



Item\_Sales based on Outlet\_Size



## Item\_Sales based on Item visibility and Outlet\_types.

## Item\_Sales based on OutletID

## 1. What type of outlet will return him the best sales: Grocery store or Supermarket Type 1, 2, or 3.

## Model:

```
re3 <- lmer(item_sales ~ item_fat_content + item_visibility + itemprice_perweight + store_years_of_operation + outlet_size + (1 | outlet_type), data=master.dataset, REML=FALSE)
```

I tried various combinations of independent variables and built 3 different models. This model gave me best AIC, Residual Variance, Log Likelihood and Beta Coefficients values, hence I have selected this model as my best model.

```
Random effects:
Groups      Name      Variance Std.Dev.
outlet_type (Intercept) 1522986 1234
Residual    1745926 1321
Number of obs: 8523, groups: outlet_type, 4

Fixed effects:
              Estimate Std. Error t value
(Intercept)    2227.643    636.789    3.498
item_fat_contentRegular    35.703    29.999    1.190
item_visibility    -16.068    14.961   -1.074
itemprice_perweight    685.370    14.359   47.730
store_years_of_operation    -3.501     4.701   -0.745
outlet_sizeMedium    -84.848    92.037   -0.922
outlet_sizeSmall    -6.663    77.507   -0.086

Correlation of Fixed Effects:
              (Intr) itm_R itm_vs itmpr_ str__ otlt_M
itm_ft_cntR  -0.016
item_vsblty   0.001 -0.050
itmprc_prwg  -0.004 -0.020 -0.001
str_yrs_f_p  -0.234  0.000 -0.009  0.025
otlt_szMdm   -0.235 -0.002 -0.009  0.015  0.843
otlt_szSml   -0.214 -0.001 -0.013  0.014  0.742  0.898

> ranef(re3)
$outlet_type
(Grocery store) -1717.9362
(Supermarket Type1) 142.3013
(Supermarket Type2) -180.0143
(Supermarket Type3) 1755.6492

with conditional variances for "outlet_type"
```

## Interpretations and Recommendations.

- Looking at the random effect coefficients we can infer that **Supermarket Type 3** outlet type has **1755.6492 more sales** than the mean. Hence **Supermarket Type 3** is the best performing outlet type in the data.
- On the other hand **Grocery store** has **1717.9362 less sales** than the mean which is least among all outlet types hence **Grocery store is least performing** outlet type among all other outlet types.

## 2. What type of city will return him the best sales: Tier 1, 2 or 3.

## Model:

```
ct3 <- lmer(item_sales ~ itemprice_perweight + store_years_of_operation + item_type + outlet_size + item_visibility + (1 | city_type), data=master.dataset, REML=FALSE)
```

I tried various combinations of independent variables and built 3 different models. This model gave me best AIC, Residual Variance, Log Likelihood and Beta Coefficients values, hence I have selected this model as my best model.

```
Random effects:
Groups      Name      Variance Std.Dev.
city_type (Intercept) 63060 252.9
Residual    2306333 1518.7
Number of obs: 8523, groups: city_type, 3

Fixed effects:
              Estimate Std. Error t value
(Intercept)    1142.830    177.791    6.428
itemprice_perweight    683.721    18.592   36.801
store_years_of_operation    39.561     2.614   15.132
item_typeBreakfast    17.982    11.051    1.626
item_typeBreakFast    155.747    156.718    0.993
item_typeCanned    131.616    84.408    1.559
item_typeDairy    227.219     85.342    2.663
item_typeFrozen Foods    109.657     79.106    1.386
item_typeFruits and Vegetables    235.045     73.733    3.188
item_typeHard Drinks    -39.645    119.851   -0.331
item_typeHealth and Hygiene    14.079     89.532    0.157
item_typeHousehold    132.593     78.174    1.696
item_typeMeat    92.008     94.863    0.970
item_typeOthers    21.402     91.290    0.234
item_typeSeafood    179.098    130.073    1.376
item_typeSnack Foods    206.586     74.082    2.789
item_typeSoft Drinks    -59.178     91.540   -0.646
item_typeStarchy Foods    371.273    135.384    2.743
outlet_sizeMedium    434.169     64.518    6.729
outlet_sizeSmall     81.250     75.614    1.075
item_visibility    -221.443    16.656   -13.295

AIC      BIC      logLik deviance df.resid
149118.3 149280.4 -74536.1 149072.3    8500

Scaled residuals:
      Min       IQ   Median       3Q      Max
-3.1536 -0.6747 -0.1600  0.5261  6.2469

> ranef(ct3)
$city_type
(Tier 1) -289.75251
(Tier 2) 323.05228
(Tier 3) -33.29977

with conditional variances for "city_type"
```

## Interpretations and Recommendations.

- Looking at the random effect coefficients we can infer that **City Type Tier 2** has highest item sales of approximately **323.05 higher than the mean**.

2) On the other hand **City Type Tier 1** has lowest item sales of approximately 289.75 lower than the mean.

3. What are the top 3 highest performing and lowest performing stores in the sample.

#### Model

```
strel3 <- lmer(item_sales ~ itemprice_perweight + item_fat_content + item_visibility + item_type + outlet_size + city_type + (1 | outlet_id), data=master.dataset, REML=FALSE)
```

I tried various combinations of independent variables and built 3 different models. This model gave me best AIC, Residual Variance, Log Likelihood and Beta Coefficients values, hence I have selected this model as my best model.

```
Random effects:
Groups      Name      Variance Std.Dev.
outlet_id (Intercept) 805125  897.3
Residual      1735445 1317.4
Number of obs: 8523, groups: outlet_id, 10
```

```
Fixed effects:
              Estimate Std. Error t value
(Intercept)    2077.90    1271.57   1.634
itemprice_perweight 686.55     14.36  47.744
item_fat_contentRegular 20.09     32.94   0.610
item_visibility  -19.21     14.97  -1.283
item_typeBreads  -19.20     98.12  -0.196
item_typeBreakfast 156.54    136.02   1.151
item_typeCanned   96.42     79.23   1.217
item_typeDairy    225.52     72.38   3.116
item_typeFrozen Foods 94.16     68.63   1.372
item_typeFruits and Vegetables 224.00     65.96   3.502
item_typeHard Drinks -35.71     105.25  -0.339
item_typeHealth and Hygiene 74.74     79.32   0.942
item_typeHousehold 171.44     69.73   2.459
item_typeMeat     135.74     82.38   1.648
item_typeOthers   131.44    114.99   1.143
item_typeSeafood  210.30    172.72   1.218
item_typeSnack Foods 192.54     64.31   2.994
item_typeSoft Drinks -41.77     81.88  -0.510
item_typeStarchy Foods 310.11    120.06   2.583
outlet_sizeMedium -254.53    1037.37  -0.245
outlet_sizeSmall  -668.23    1296.73  -0.515
city_typeTier 2    465.82     778.04   0.599
city_typeTier 3     72.66     898.43   0.081
```

```
> ranef(strel3)
$outlet_id
(Intercept)
OUT010 -1.700818e+03
OUT013  4.576404e-09
OUT017 -9.866009e+01
OUT018 -7.487187e+01
OUT019 -1.065414e+03
OUT027  1.775689e+03
OUT035  3.740663e+02
OUT045 -2.754062e+02
OUT046  6.913481e+02
OUT049  3.740663e+02
```

AIC	BIC	logLik	deviance	df.resid
146744.9	146921.1	-73347.4	146694.9	8498

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.7247	-0.6373	-0.1366	0.4540	6.5762

#### Interpretations and Recommendations.

- Looking at the random effect coefficients we can infer that Outlets **OUT027, OUT046, OUT035** are top 3 performing outlets with sales **1775.68, 691.34 and 374.06** more than the mean of the random effect variable.
- Outlets **OUT010, OUT019, OUT045** are least performing outlets with sales **1700.83, 1065.414 and 275.4** less than the mean of the random effect variable.