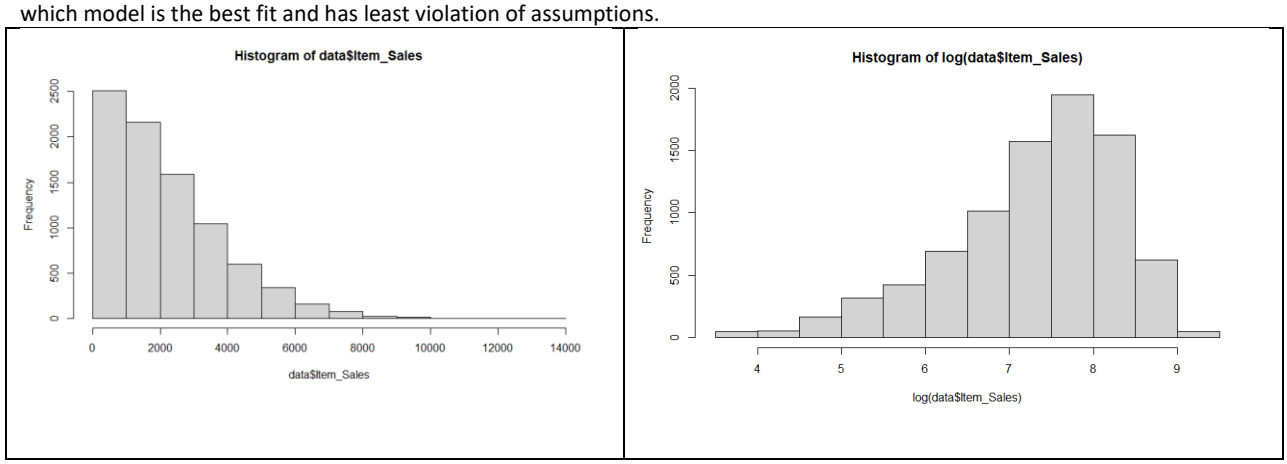
Aryan Sharma
Survival Model
BigMart Sales

Exploratory Data Analysis

The data has multi-layers of items and outlets, the same item can be present at multiple stores. Therefore, multiple level analysis should be used. **Item_sales** is the Y variable of this analysis.

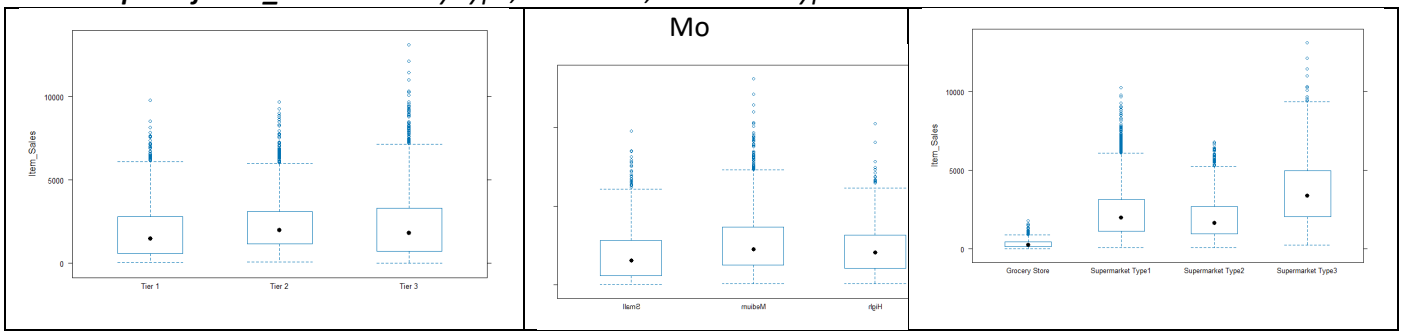
Histograms of DV:

The distributions of **Item_Sales** are right-skewed. The distributions of **Item_Sales** are close to normal, and therefore, more suited for MLS regression. The variable is Poison distribution hence we'll have to use MLS models. Then we can determine which model is the best fit and has least violation of assumptions.



Boxplots:

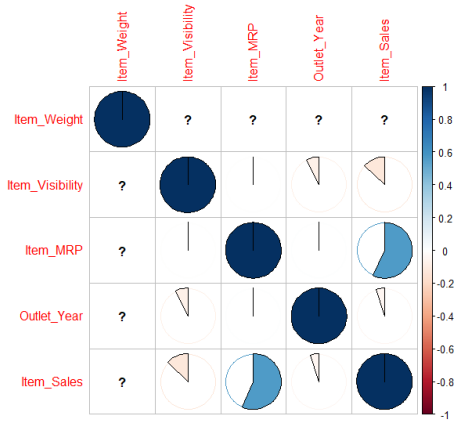
Boxplot of Item_Sales *with City type, outlet size, and outlet type*



Correlations:

Item sales are Correlated to **Item_MRP,** Auto correlation exists as the data is multilevel nature

```
correlation_matrix <- cor(data[c("Item_Weight", "Item_Visibility","Item_MRP","Outlet_Year","Item_Sales")])
corrplot(correlation_matrix, method = "pie")
```



Predictor table:

Predictor	Effect	Rationale
DV: Item_sales		
Item_Visibility	+	If the product is visible more on the aisle, then the chances of the selling increases
Item_Type	+	Past Spending pattern may be helpful in determining the spend
Item_MRP	+	The selling price can affect the number of sales
Outlet_ID	+/-	To identify which outlet is doing the sale
Outlet_Type	+	Type of outlet can be driving factor as different type of outlet may have more listed items
City_Type	+	Tier 1 cities may tend to have more sales
Outlet_Age	+/-	If an outlet is old people in vicinity may know its old and can age can help in driving sales
Excluded: Item_ID as it is just a unique identifier for product, Item_weight shouldn't be a significant factor in our analysis and item_fat_content. Outlet_size can't be a major contributor when the customer is in need of a product.		

Regression Analysis

```
    model_1 <- lm(log(Item_Sales)~ Item_Visibility + Outlet_Size + Item_Type + log(Item_MRP) + Outlet_ID +
    Outlet_Type + City_Type + yearsofoutlet,data=data)
    model_2 <- lmer(log(Item_Sales)~ Item_Visibility + Item_Type + log(Item_MRP) + Outlet_Type + City_Type
    + yearsofoutlet +(1|Outlet_ID),data=data,REML=FALSE)
model_3 <- lmer(log(Item_Sales)~ Item_Visibility + Item_Type + log(Item_MRP) + Outlet_Type + City_Type +
yearsofoutlet +(1|Outlet_ID/City_Type),data=data,REML=FALSE)
```

	(1)	(2)	(3)
Item_Visibility	-0.020 (0.137)	0.021 (0.114)	0.021 (0.114)
Outlet_SizeMedium	0.060** (0.024)		
Outlet_SizeSmall	0.020 (0.024)		
Item_TypeBreads	0.061 (0.045)	0.041 (0.038)	0.041 (0.038)
Item_TypeBreakfast	-0.026 (0.063)	-0.065 (0.053)	-0.065 (0.053)
Item_TypeCanned	0.023 (0.034)	0.023 (0.029)	0.023 (0.029)
Item_TypeDairy	-0.029 (0.033)	-0.030 (0.028)	-0.030 (0.028)
Item_TypeFrozen Foods	-0.017 (0.032)	-0.021 (0.027)	-0.021 (0.027)
Item_TypeFruits and Vegetables	0.004 (0.030)	-0.001 (0.025)	-0.001 (0.025)
Item_TypeHard Drinks	0.008 (0.049)	-0.003 (0.041)	-0.003 (0.041)
Item_TypeHealth and Hygiene	0.014 (0.036)	0.015 (0.030)	0.015 (0.030)
Item_TypeHousehold	-0.009 (0.031)	-0.016 (0.027)	-0.016 (0.027)
Item_TypeMeat	0.024 (0.038)	0.021 (0.032)	0.021 (0.032)
Item_TypeOthers	-0.016 (0.052)	-0.005 (0.045)	-0.005 (0.045)
Item_TypeSeafood	0.051 (0.080)	0.028 (0.068)	0.028 (0.068)
Item_TypeSnack Foods	0.003 (0.030)	-0.006 (0.025)	-0.006 (0.025)
Item_TypeSoft Drinks	-0.018 (0.038)	-0.003 (0.032)	-0.003 (0.032)
Item_TypeStarchy Foods	0.011 (0.056)	-0.001 (0.047)	-0.001 (0.047)
log(Item_MRP)	1.026*** (0.013)	1.023*** (0.011)	1.023*** (0.011)
Outlet_IDOUT018	-0.207*** (0.024)		
Outlet_IDOUT019	-1.928*** (0.029)		
Outlet_IDOUT027	0.501*** (0.024)		
Outlet_IDOUT035	0.058** (0.024)		
Outlet_IDOUT046			
Outlet_IDOUT049			
Outlet_TypeSupermarket Type1		1.938*** (0.025)	1.938*** (0.025)
Outlet_TypeSupermarket Type2		1.756*** (0.049)	1.756*** (0.049)
Outlet_TypeSupermarket Type3		2.511*** (0.035)	2.511*** (0.035)
City_TypeTier 2		-0.016 (0.027)	-0.016 (0.027)
City_TypeTier 3		-0.030 (0.024)	-0.030 (0.024)
yearsofoutlet		-0.002 (0.002)	-0.002 (0.002)
Constant	2.522*** (0.067)	0.693*** (0.085)	0.693*** (0.085)
Observations	6,113	8,523	8,523
R2	0.712		
Adjusted R2	0.711		
Log Likelihood		-6,471.138	-6,471.138
Akaike Inf. Crit.		12,994.280	12,996.280
Bayesian Inf. Crit.		13,177.590	13,186.640
Residual Std. Error	0.515 (df = 6089)		
F Statistic	653.786*** (df = 23; 6089)		
Note:			*p<0.1; **p<0.05; ***p<0.01

Model1: is a linear model and we have used log transformation because the y variable is skewed distribution, as it helps to stabilize variance and make the relationship between predictors and the response more linear.
Model 2: Model_2 utilizes a linear mixed-effects model (LMM) to account for potential correlation within groups (Outlets) and variability between groups
Model 3: Model_3 extends Model_2 by including a nested random effect structure, where Outlet_ID is nested within City_Type. This measures for the hierarchical structure of the data, where outlets are within cities.

Results of AIC And BIC

```
> AIC(model_1,model_2,model_3)
        df       AIC
model_1 25  9268.591
model_2 26 12994.276
model_3 27 12996.276
> BIC(model_1,model_2,model_3)
        df       BIC
model_1 25  9436.545
model_2 26 13177.590
model_3 27 13186.640
```

Model 3 best fit out of the three models.

Interpretation

1. In accordance to outlet type, With base as Grocery store, Supermarket type III has the maximum number of sales with 251% more sales. Super Mart Type II will have 175% more sales than the grocery store and Supermarket Type I will have 193% more sales than grocery store.
2. Tier 1 cities as a base. When we compare tier 1 cities with tier II and tier III cities, the sales in tier I cities is the highest. Tier II has 1.6% less sales than tier 1 and tier III has 3% less sales.
3. Best performing stores are:
   Outlet_IDOUT027 which is performing 50.1% better than the baseline store, Outlet_IDOUT035 is performing 5.8% better than baseline, Outlet_IDOUT049 is performing 1% better. Worst performing stores are: 1. Outlet_IDOUT019 is performing the worst with 192% worse than the baseline, Outlet_IDOUT018 is second worst with 20% less sales and outlet_IDOUT046 being the least worst.

Recommendations:

like model 3 we can build upon the hierarchical structure of the data, Model 4 can incorporate random effects to account for variability at different levels. As we know the outlets are situated within the cities. This accounts for clustering of observations and potential correlations within groups.