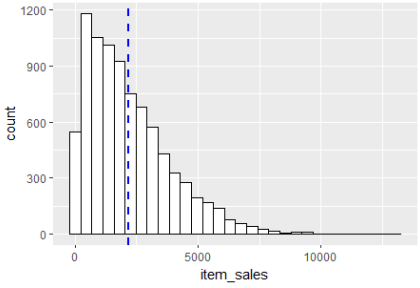
Big Mart Sales

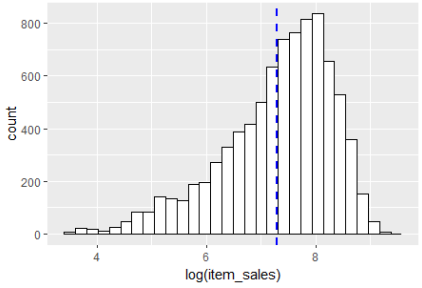
Predictor Table:

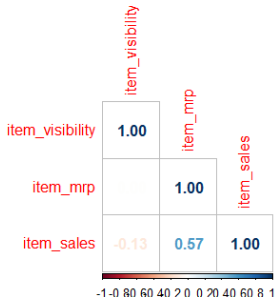
DV: Item_Sales

Predictors	Expected Effect	Rationale
Item_Visibility	+	Total display area of a product, should logically impact the related sales as more the percentage of display, more likely to buy.
Item_MRP	+/-	It's the first thing checked by customers while making decisions. Its effect can be both as it depends on quantity also. More no. of cheaper items and a few of expensive ones eventually might be similar.
Outlet_Size	+	More the size, more space for covering large product range, giving more options to customers, might increase sales.
City_Type	-	Larger the population, larger the economies thereby meaning more sales. As tier increases from 1 to 2, the sales might decrease.
Outlet_Type	?	This would determine the type of products available, which may affect sales as some product categories are required frequently than others. Also, it will be required to tell which type will return best sales.
Outlet_Id	?	To find outlet level differences for answering top 3 highest performing and lowest performing stores, we need this.
Excluded Factors		
Item_ID, Weight, Fat_content	NA	Could have been useful if our target was to figure out variation in sales based on different products, but as of now item details are of no use.
Item_Type	NA	Might be correlated with outlet_type as item_type in a store depends on what type of store it is.
Outlet_year	NA	Might not affect sales as other factors like location, ease of access, demand and population are more important rather than year of opening.

Visualization







As the dependent variable distribution is not normal or negative, we will try to normalize it by transforming it with log. It seems to become somewhat normal. We can try Poisson models by looking at the plots. But it works with non-continuous data, hence we can try using truncated values for sales which can then be considered as count of dollars. Also, mrp and sales will have correlation (<0.7) which has to be neglected as it is the main factor in determining sales.

Transformations

Converting outlet_id, outlet_size, city_type, outlet_type to factor variables as we need to find differences between categories in these variables.

After exploring the data, it was found that, all the grocery or tier2 cities had small outlets, therefore the NA values that corresponded to either grocery or tier2 city_type was replaced with "Small" for outlet_size variable.

Modelling

1.Using base case glm for fixed effects without considering the hierarchical relation between variables and testing for overdispersion:

```
base <- glm(item_sales ~ item_visibility + item_mrp  + outlet_type + city_type
        + outlet_id + outlet_size, data = df, family=poisson(link=log))
```

2.Trying random effects model to test for city level and outlet level variability assuming to be from a randomly distributed population and not fixed. We are considering two models, one as negative binomial also because the dispersion test on the base model resulted in very high value for lambda (~435).

```
m1 <- glmer (item_sales ~ item_visibility + outlet_size + item_mrp + city_type + (1|outlet_id) + (1|outlet_type), data = df,family=poisson(link=log))
m2 <- glmer.nb(item_sales ~ item_visibility + outlet_size + item_mrp + city_type + ( 1 | outlet_type) + (1 | outlet_id) ,data = df)
```

3. Stargazer and conditional variance for the selected model

	Dependent variable:		
	item_sales		
	Poisson	generalized linear mixed-effects	
	(1)	(2)	(3)
item_visibility	-0.136*** (0.005)	-0.136*** (0.005)	-0.067 (0.108)
item_mrp	0.007*** (0.00000)	0.007*** (0.00000)	0.008*** (0.0001)
outlet_typeSupermarket Type1	1.921*** (0.002)		
outlet_typeSupermarket Type2	1.757*** (0.002)		
outlet_typeSupermarket Type3	2.391*** (0.002)		
city_typeTier 2	-0.075*** (0.001)	0.030 (0.033)	0.015 (0.032)
city_typeTier 3	-0.014*** (0.003)	-0.013 (0.041)	-0.040 (0.043)
outlet_idOUT013	-0.019*** (0.003)		
outlet_idOUT017	0.075*** (0.001)		
outlet_idOUT018			
outlet_idOUT019			
outlet_idOUT027			
outlet_idOUT035	0.094*** (0.001)		
outlet_idOUT045			
outlet_idOUT046	-0.047*** (0.001)		
outlet_idOUT049			
outlet_sizeMedium		0.021 (0.057)	0.002 (0.058)
outlet_sizeSmall		-0.029 (0.058)	-0.031 (0.058)
Constant	4.749*** (0.003)	6.257*** (0.387)	6.115*** (0.457)
Observations	8,523	8,523	8,523
Log Likelihood	-1,928,688.000	-1,928,743.000	-68,651.920
Akaike Inf. Crit.	3,857,399.000	3,857,503.000	137,323.800
Bayesian Inf. Crit.		3,857,567.000	137,394.300

```
> ranef(m2)
$outlet_id
       (Intercept)
OUT010 -6.694138e-04
OUT013  6.657479e-08
OUT017  1.067180e-02
OUT018  1.405033e-04
OUT019 -2.260222e-04
OUT027  5.289889e-04
OUT035  2.330351e-02
OUT045 -3.397496e-02
OUT046  8.952956e-04
OUT049 -6.697740e-04

$outlet_type
                (Intercept)
Grocery Store     -1.5005847
Supermarket Type1  0.3786345
Supermarket Type2  0.2354575
Supermarket Type3  0.8864873

with conditional variances for "outlet_id" "outlet_type"
```

Assumptions

We know that glm models are robust to Linearity, normality and heteroscedasticity assumptions, so we will test for dispersion, multicollinearity and independence.

```
> vif(m2)
                    GVIF Df GVIF^(1/(2*Df))
item_visibility 1.000137  1        1.000068
outlet_size     5.362428  2        1.521740
item_mrp        1.000127  1        1.000064
city_type       5.363063  2        1.521785
```

Passes Multicollinearity as vif values are small <10.

```
> durbinWatsonTest(resid(m2))
[1] 1.962972
>
```

Passes DurbinWatson test for autocorrelation/ independence as value is near 2.

Negative binomial models are good at handling overdispersion so considering the m2 model, it will be robust to overdispersion.

Best model

Using the m2 model for interpretations and answering the questions.

Interpretations

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

>Random effect model assumes that the effect is with respect to a hypothesized mean store type. Looking at the random effect coefficients for various outlet_types we can say that supermarket type 3 will return best sales which will be ~88% better than the average.

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

>Considering fixed effects for city_type, there seems not much of a difference. But still if we want to suggest based on the selected model, tier 2 city will have 1% better sales. So, I believe, Tier1/2 could be targeted for best sales as Tier 3 is consistent with not proving to be good across all 3 models.

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

>The three stores with highest performance are OUT035, OUT046, and OUT017 when compared to average outlet.

The three stores with lowest performance are OUT049, OUT010, and OUT019 with respect to the mean outlet.