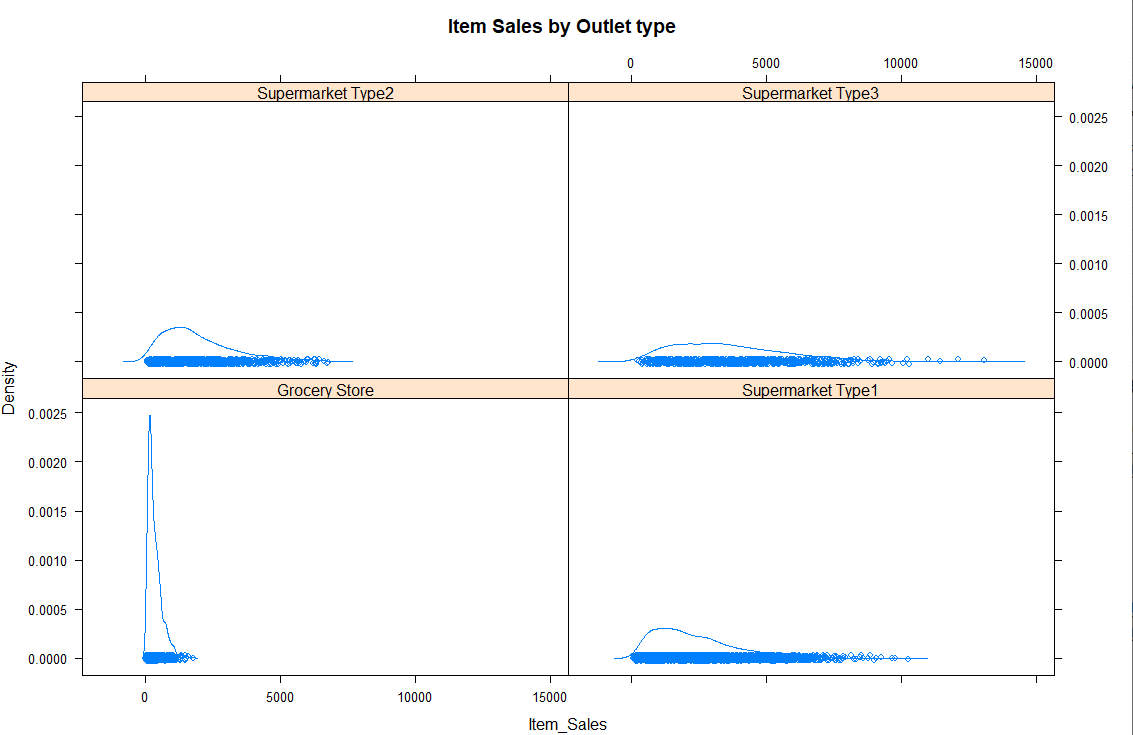
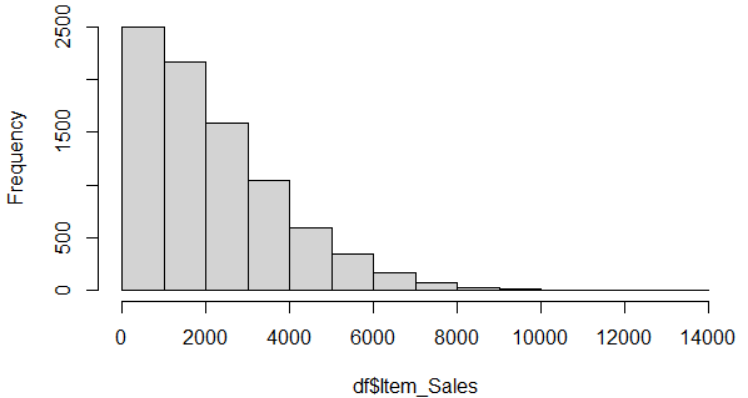
**Big Mart Sales**

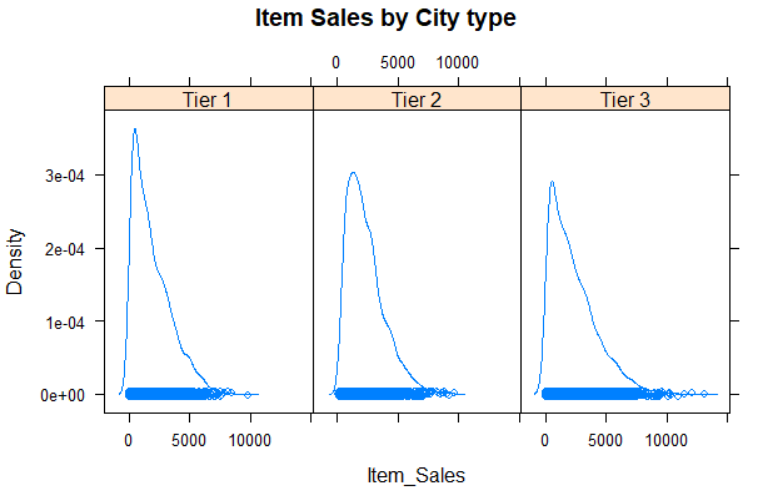
1. **Create a table of predictors for our dependent variable, listing all relevant predictors, the sign of their hypothesized effects, and a short 1-sentence rationale for each effect.**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Sign** **of effect** | **Rationale** |
| Item\_Fat\_Content | +/- | There might be a increasing preference for low fat items compared to regular fat items. |
| Item\_Visibiility | + | Items with more display area in a store would likely be sold more. |
| Item\_MRP | + | Items with higher price could contribute more to sales than items with lower price. |
| Outlet\_Size | + | Bigger outlets could attract more customers resulting in better sales. |
| Outlet\_ID | +/- | There are outlets which do better and there are some which don’t, so I will be using it as random variable assuming these outlets are independent of each other. |
| Outlet\_Age (2013 – Outlet\_Year) | + | Outlets which are old would have established themselves and could have more loyal customers compared to new ones contributing to better sales. |
| Outlet\_Type |  | A particular outlet type could sell its own unique set of products which could contribute to sales |
| City\_Type |  | Cities which are bigger could have more competition that can lead to lesser sales, it could also go the other way. |
| Item\_Type |  | Certain types of items like groceries would lead to more item sales because these are basic necessities. |

I have not used Item\_ID Item\_weight since I felt they were not relevant.

**Visualizations:**

* Item sales follows a Poisson distribution, this could be attributed to the underlying distribution which follows a Poisson distribution ie number of items sold.



**3 Best models:**

* 1. glmer\_m2 <- glmer(round(Item\_Sales) ~ Item\_Fat\_Content + Item\_MRP + Item\_Visibility + Item\_Type + Outlet\_Size + Outlet\_yrsold +(1|Outlet\_ID) + Outlet\_Type + City\_Type,family = poisson(link = "log"))
  2. lmer6 <- lmer(Item\_Sales ~ Item\_Fat\_Content + Item\_Visibility + Item\_MRP + Item\_Type + Outlet\_yrsold + (1|Outlet\_ID) + (1|Outlet\_Type) + City\_Type,REML = FALSE) ) – **BEST**
  3. lmer7 <- lmer(Item\_Sales ~ Item\_Fat\_Content + Item\_Visibility + Item\_MRP + Item\_Type + Outlet\_yrsold + Outlet\_Size + (1|Outlet\_ID) + (1|Outlet\_Type) + City\_Type ,REML = FALS)

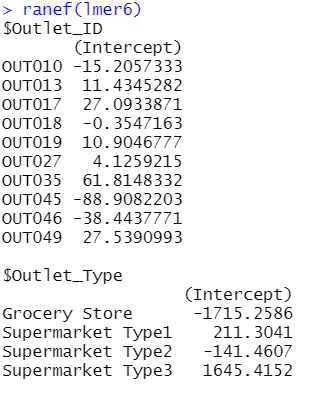
**Model Justification**:

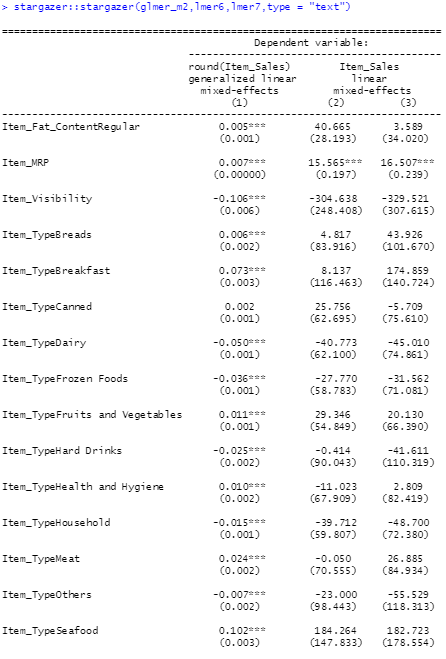
This is clearly a multi level problem as the datastructure clearly shows that there is item level, outlet level, City level data. So, I decided to use a multi-level model

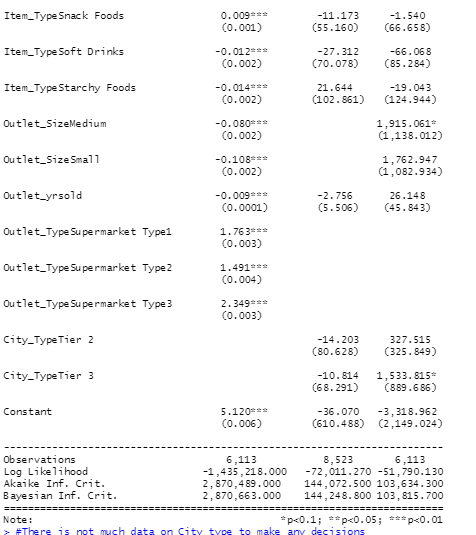
I tried using glmer models initially because item sales was following a poisson distribution. But, the glmer model was failing to converge. Hence, I decided to go with lmer models.

I used outlet ID and outlet types as random effects, as I assumed them to have independent random effects. Also, it gives more degrees of freedom. Also, there are fixed level 1 and level 2 predictors in my model. I decided to use model b as my best model because model c was giving random intercepts of 0 variance. This was due to the outlet size variable, even removing null values did not work. Hence, I went with model c (lmer7) as my best model.

Adding the random effects of the best model below. Please find the stargazer output in the next page.







* **What type of outlet will return the best sales?**

Looking at the random effect coefficients we can see that the outlet type ‘**supermarket type 3**’ has **1645 more sales** than the mean estimate of the random variable.

**Recommendation**: I would recommend to start a new franchise of ‘supermarket type 3’ type of outlets as their sales are much higher compared to the rest. While investing in grocery stores leads to **1715 dollars less in sales,** so avoiding franchising grocery stores is better.

* **What type of city will return the best sales?**

The estimate of city type suggest that for **tier** **2** and **tier** **3**, sales goes **down** by **10.8** and **14.20 dollars when compared to tier 1.**

**Recommendation:** From the model estimates, I would recommend to look at Tier 1 cities as they seem to be more profitable. The data does not have different outlet types in tier 1 and tier 2. So further analysis must be done to see on what type of outlet could be started here.

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

Outlets **35,49 and 17** are the best performing outlets as they contribute **61.8,27.5 and 27** dollars **more in sales** compared to the average estimate of outlets.

Outlets **45,46 and 10** are the worst performing outlets as they contribute **88.9,38.4 and 15.2** dollars **less in sales** compared to the average estimate of outlets.

**Recommendation:** I would advice the entrepreneur to look at what makes Outlet 35 stand out as the highest contributor to sales. He can also look into why outlets 45,46 and 10 are doing bad. Learning yhis would help him in making his new franchise successful.