**SDM A5 Bigmart Sales**

This file contains sales data on different items at multiple outlets of a major retail chain. Note that the data is multi-level. Data definitions are provided on "Data Definition" of the Excel spreadsheet.

Your client is a business entrepreneur considering franchising one or more stores of this retail chain and is looking for the following answers, with adequate justification:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Here the distribution of data is not normal, hence we got log transformation and it is normally distributed with slight negative skewness. It is better to use OLS models.

Here our dependent variable is Y = item\_sales

|  |  |  |
| --- | --- | --- |
| Predictor | Effect | Rationale |
| Item\_weight | + | Item weight can potentially affect item sales, as it may influence factors such as product packaging, shipping costs, and perceived value for the consumer |
| Item\_fat\_content | - | Item\_fat\_content increases, the sales may be decreased as most people don’t prefer items with more fat content |
| Item\_visisbility | + | Items which are mostly visible are more tent to be bought |
| Item\_type | +/- | Item type can potentially affect item sales, as different types of items may have different levels of demand, pricing, and competition in the market |
| Item\_mrp | +/- | Item MRP (Maximum Retail Price) can potentially affect item sales, as it is a commonly used pricing strategy in the retail industry |
| Outlet\_year | - | Outlet year can potentially affect item sales, as the age of the retail outlet may impact factors such as consumer traffic, brand recognition, and competition in the local market |
| Outlet\_size | + | Outlet size can potentially affect item sales, as the size of the retail outlet may impact factors such as the variety and quantity of items available, the store layout and display, and the overall shopping experience for customers |
| City\_type | +/- | City type can potentially affect item sales, as factors such as population size, income levels, and cultural preferences may differ across different types of cities |
| Outlet\_type | +/- | Outlet type can potentially affect item sales, as different types of outlets may have different target markets, pricing strategies, and product offerings |
| Excluded Factors | | |
| Item\_id | n/a | It doesn’t have any significant effect |
| Outlet\_id | n/a | It is correlated with outlet\_size |

Let us do the correlation test to check for autocorrelation.

Diagram

Description automatically generated

There is no significant auto correlation and hence no need to drop any columns. Item\_fat content is correlated with item type and hence needed to drop it

Also let us create new column outlet\_age using outlet\_year and use it as a factor instead of outlet\_year.

Let us categorize different levels of data.

Level 1 – Outlet\_data columns

Level 2 – Item\_data columns

From the data we can see that item weight has 1463 null values and mostly, items with higher price and specific items(item\_type) have higher weight. So its better to omit item\_weight from the model. From the data, it is clear that all the tier\_1 outlets are either small or medium, and tier 2 items are always small. Also tier 3 are always either medium or high.

Let us build models from the remaining predictors.

Models:

*m1 <- lm(log(item\_sales) ~item\_mrp+item\_type+item\_visibility+outlet\_type+city\_type , data=bd)*

*fe <- lm(log(item\_sales) ~ item\_mrp+item\_type+item\_visibility+outlet\_type+city\_type+ outlet\_id , data=bd)*

*re <- lmer(log(item\_sales) ~item\_mrp+item\_type+item\_visibility+outlet\_type+city\_type+ (1|outlet\_id), data=bd, REML=FALSE)*

Graphical user interface, text

Description automatically generated Graphical user interface, text

Description automatically generated Text

Description automatically generated Text

Description automatically generated Text

Description automatically generated

The model with lowest AIC is FE model but where we could not estimate for outlet\_ids which need to be answered. Also there is not much different in the coefficients among the models. Our RE model also passes multicolleniarity tests

Our models are stable in all the three cases. We could not estimate all the coefficients for outlet\_ids in fixed effect model, but could estimate then in random effect models.

In the random effect, we used order id as the higher level (level\_1) and tried to do the hierarchy to estimate the coefficients.

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

Chart, box and whisker chart

Description automatically generated

Outlet\_type supermarket type 3 have the highest sales when compared to groceries. It has 250% more sales than groceries, followed by type 1 super market with 195% more sales than groceries.

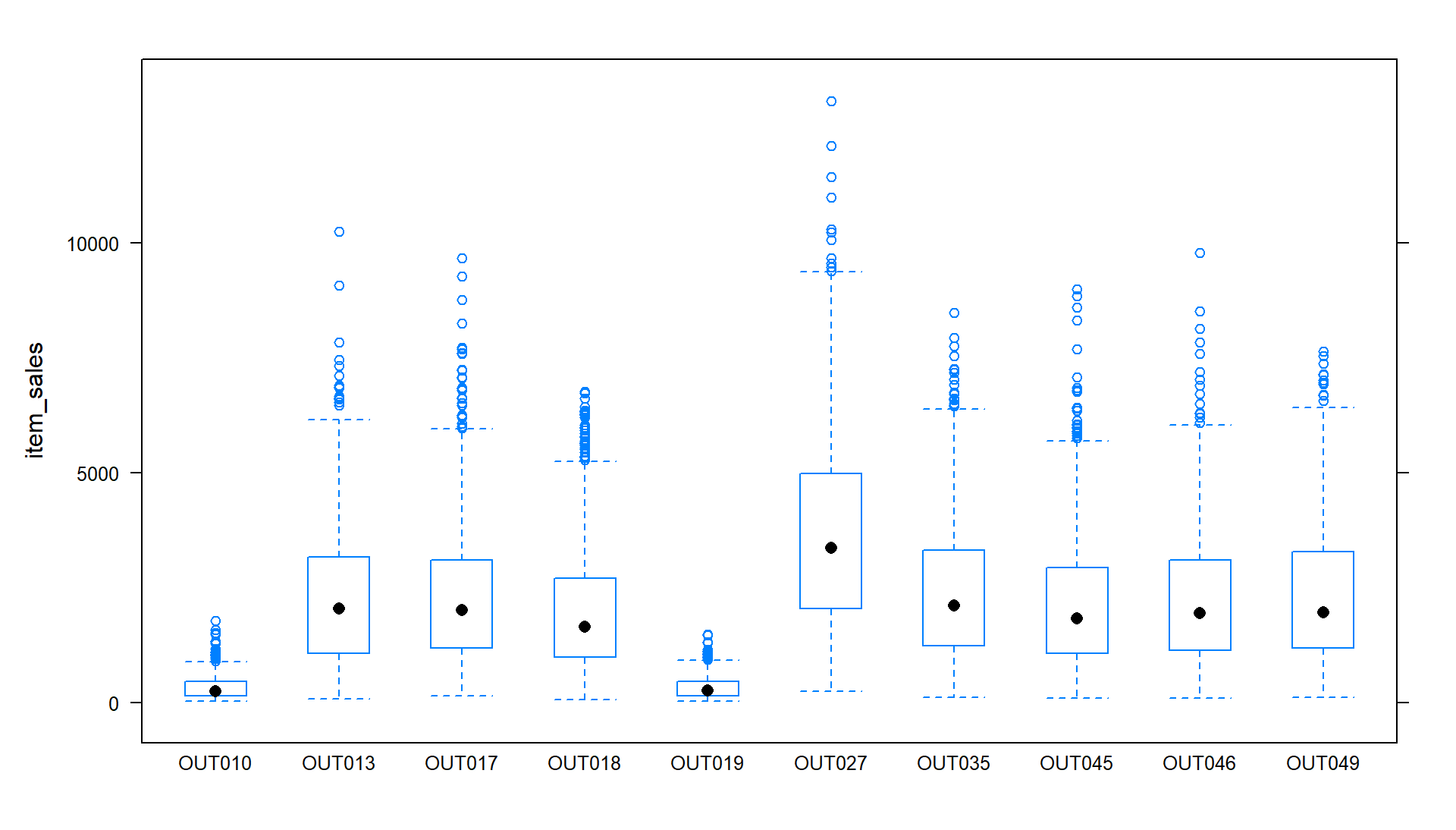
* What type of city will return him the best sales: Tier 1, 2 or 3.

Chart, box and whisker chart

Description automatically generated

City tier 1 has the best sales average when compared to the other types. Sales in tier 2 and tier 3 are lower by 1.5% and 3.3% respectively, when compared to tier 1 cities.

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



Outlet 35, 49 and 17 are the best performing stores and 10,45,46 are the lowest performing stores when compared to other stores.