The first type of analysis allowed us to test the originally generated hypotheses. We then explored several variations for the baseline model. Our strategy was to compare whether a specific set of additional terms or interaction terms improve the current best predictive model using Extra-Sum-of-Squares F test. We fitted a total of 7 linear regression models (details in Results section), which we visually examined model assumptions by a residual vs. fitted scatter plot and a QQ plot. We also looked at the R-squared improvement with each model we fitted. For the best predictive regression model (as defined by the best R-squared and AIC performance and estimated by cross-validation MSE), we used bootstrap to perform inference on the fitted coefficients permuting residuals, as we did not trusted the normality test for residuals. Finally, we fitted lasso, ridge and elastic net models in attempt to reduce overfitting and improve model predictive ability.

3.1 Hypothesis Generation

Before beginning our analysis, we have thought of some important hypothesis that could be interesting to test. In this study, we have managed to analyze some of them, whilst the rest are here for further reflection and improvement. In order to avoid any sort of data dredging as well as leave some space for unexpected results, we have stated two types of hypotheses the first related to the features of the stores and the 2nd related to the features of the product.

Store related Hypotheses:

1. Location type: Retail stores located in urban areas(represented as Tier 1 in our data) would generally generate more sales on average because of a higher influx of customers.
2. Store Capacity: Stores of Size Large (Outlet.Size) are expected to benefit from higher sales, for the same reason why stores in more populated areas do. We also expect an interaction between store size and store location, given that larger stores populate more urban areas in general.
3. Neighbourhood: Stores located within very close to other big marketplaces might suffer from lower sales, given that there is increased competition. If our data included geographical coordinates, it would allow us to study how location and competitors interact in association with sales, using a machine-learning based approach and measuring a competitiveness factor, for instance.
4. Ambiance: Stores which are well-maintained and managed by polite and humble people are expected to have higher footfall and thus higher sales.

Product Related Hypotheses:

1. Marketing Index: Branded Items are more prone to be attractive to customers. In our data, we used visibility as a marker of marketing.

2. Branding: The customers are more willing to buy items of the same type which have more colorful labels, a better design, a better smell.

3. Utility: Daily use products, such as milk, bread, fruit and vegetables have higher sales and replacement. This is reflected under the Item.Type in our data, and specifically diary, fruits and vegetables

4. Ambiance: New and well designed stores (reflected probably in the Outlet.Establishment.Year variable) are expected to have higher sales.

3.1.1 Exploratory Data Analysis (EDA)

We first checked the distribution of the response variable (Item.Sales) to determine if there is any skewness to be corrected with transformation. As shown in **Figure 6**, the distribution of Sales. Sales is fairly right skewed and such right-skewness is usually corrected by applying the log transformation. However, in our case log-transformation was not the best solution, so we found that cube-root is the right transformation for our response **(Figure 7)**. Note that we decided to sacrifice some interpretability by $(Sales)^{1/3}$ over some other transformations for a more symmetrically distributed response variable. We then inspected the association of numeric and categorical predictor variables with the transformed response in order to choose the best type of OLS (linear or polynomial) **(Figure 10)**.

Specifically, we found that there is linear association between all the numerical variables and our response, Sales. The association with Item Visibility is approximately flat, probably because of the extreme small values of the visibility factor (a one-to-one simple model, would be appropriate to test this association). Additionally, for the categorical variables, the boxplots show enough variation in their distribution, more prominent for Item.Type and Outlet.Identifier.

3.1.2 Multiple Linear Regression Model Performance, Interpretation, Assumptions and Comparisons

Based on our hypotheses related to the store features, we were expecting to see an association between `Outlet_Type` and Sales, given that stores of higher capacity have a higher volume of sales and also higher prices compared to grocery stores, so we fitted a model predicting sales from `Item_MRP`+`Outlet_Type`(**Model 2**) and then another one which has the interaction between these two predictors (**Model 2b**). The associations were significant and positive for all predictors and interaction terms. The second model, which included the interaction terms had a 1 percent higher R-squared (from **68** to **69**) and can be trusted more, since there is significant interaction between prices and outlet types (level 1 stores have more expensive products compared to small grocery stores). The association between price and Sales can be visualized using the plot: