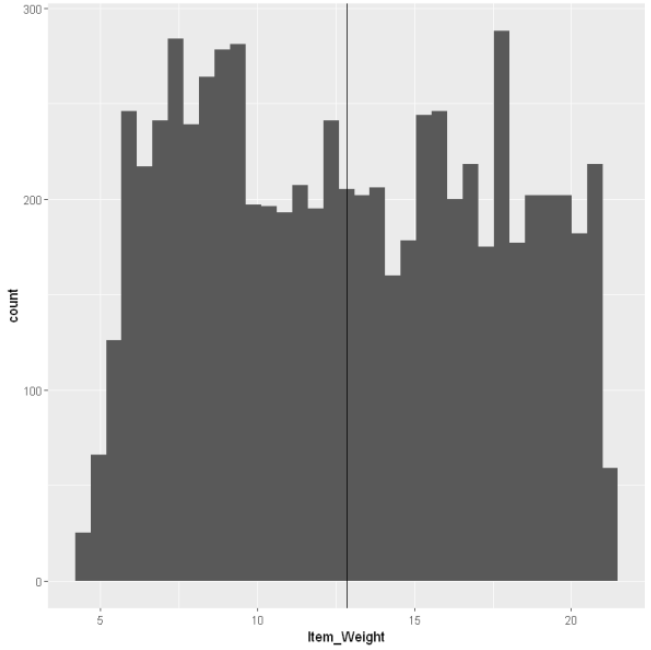
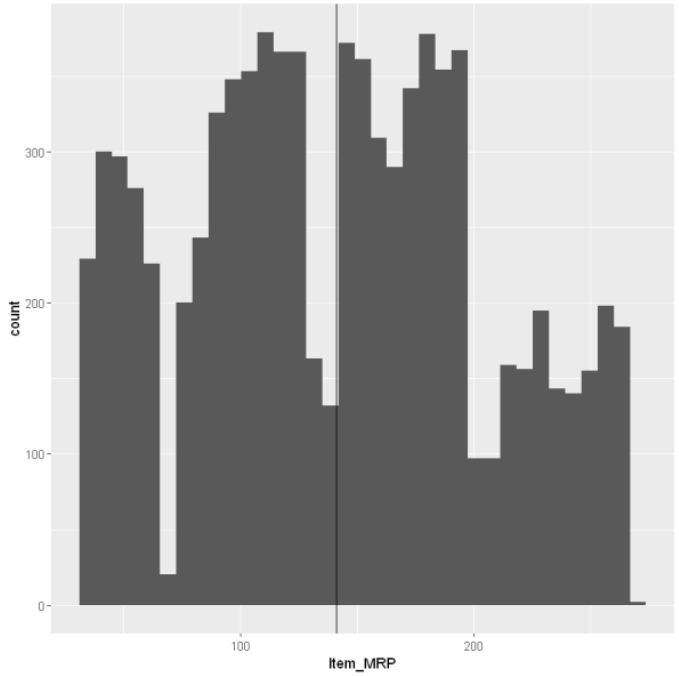
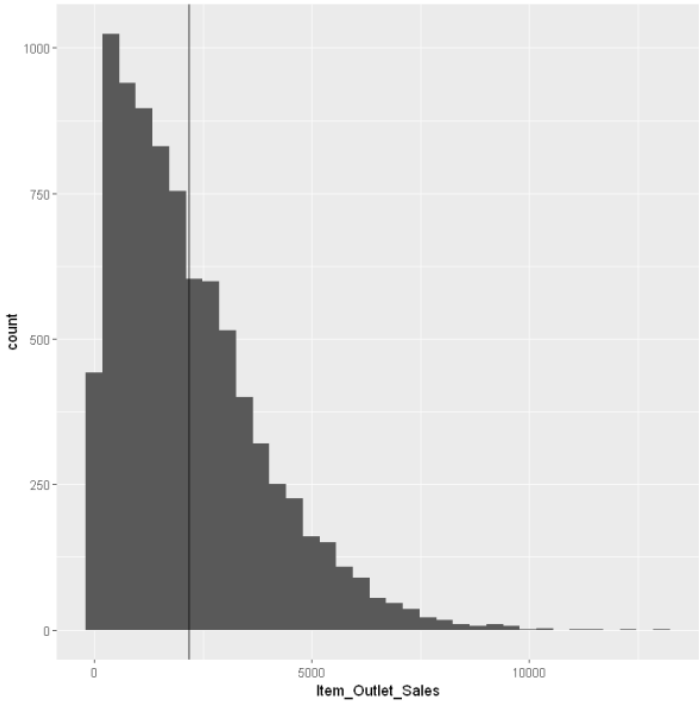
BigMart Project—Product Analytics and EDA

EDA - Numeric



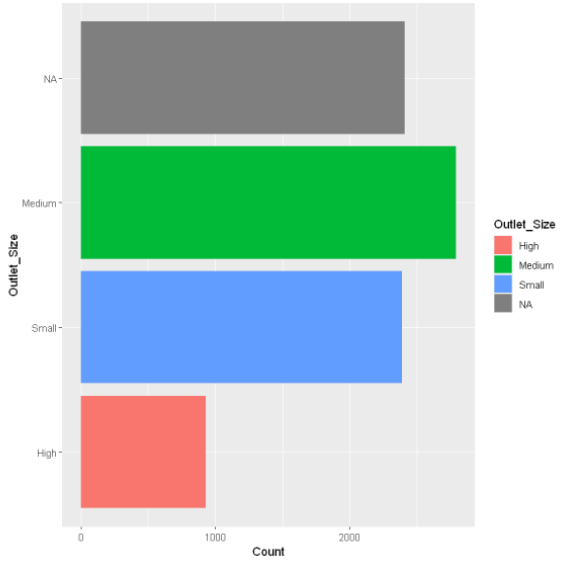


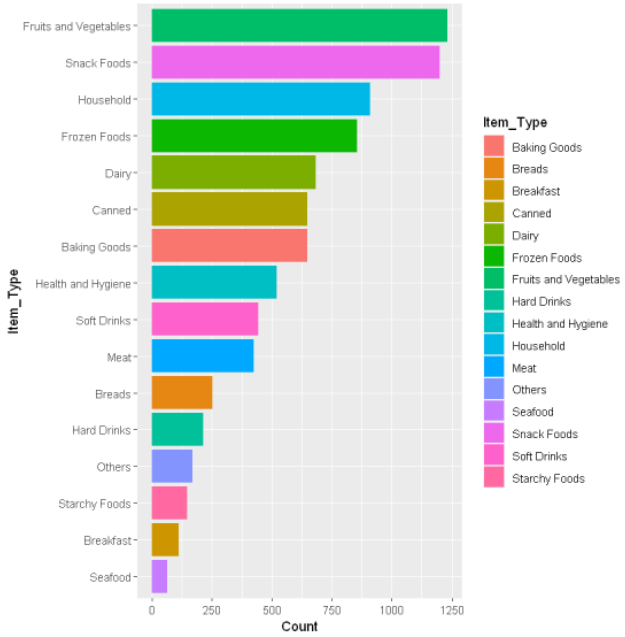


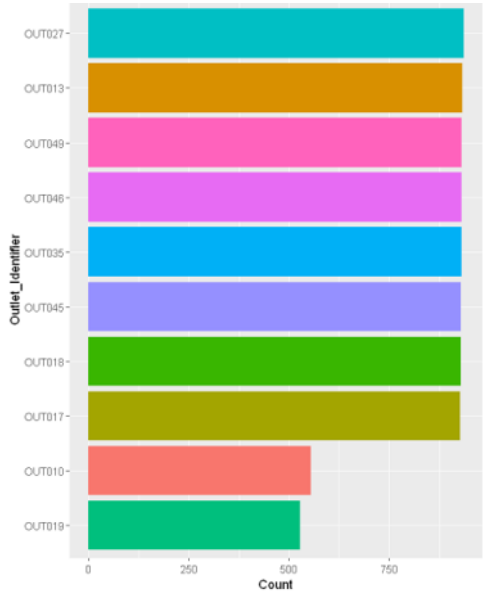
Dataset contains:

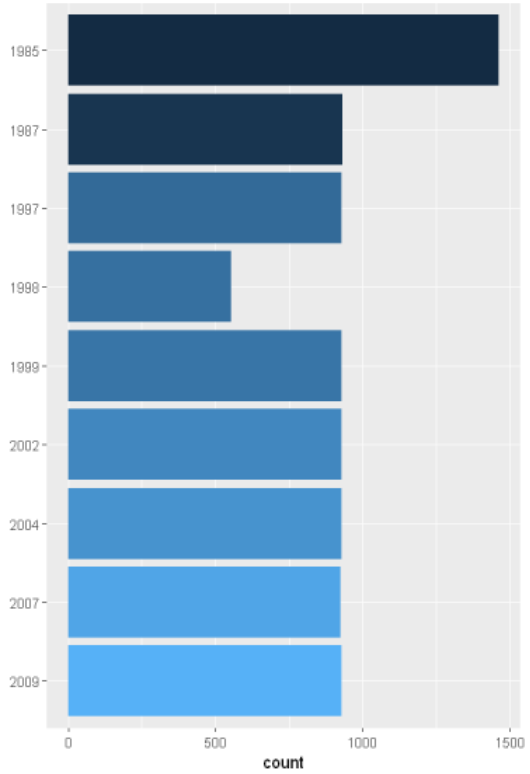
- 5 numeric columns
- Item_Weight has 1463 missing records

EDA - Categorical









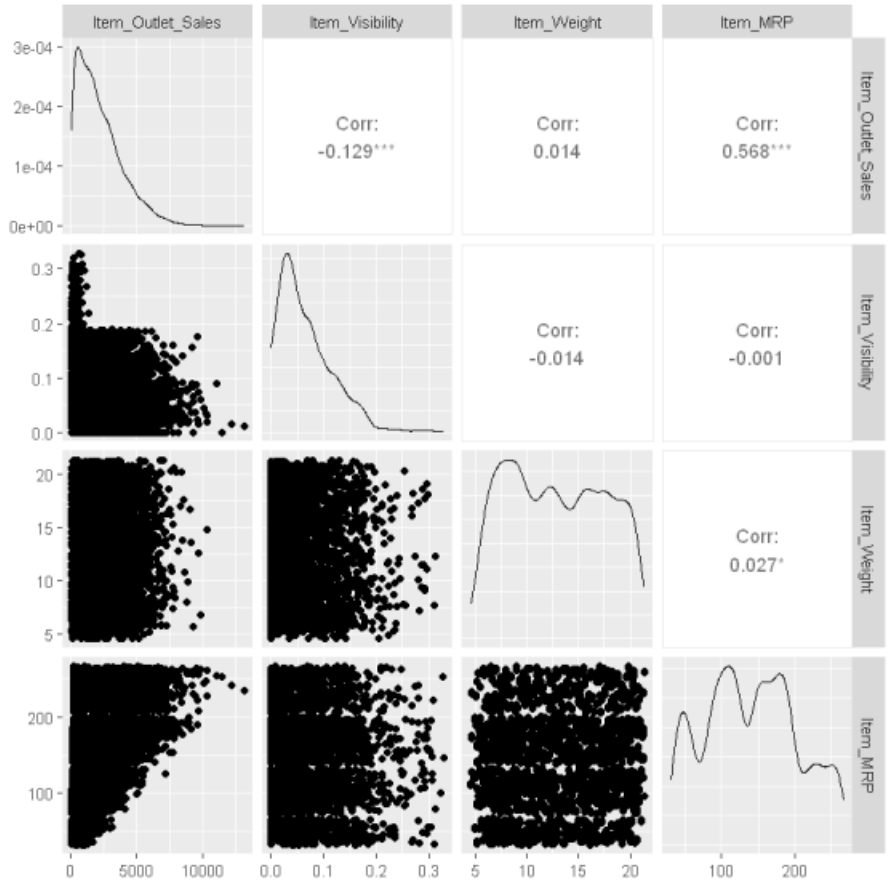
Dataset contains:

- 7 categorical columns
- Outlet size contains 2410 missing records
- Low Fat records IN Item_Fat_Content are represented differently and this has to be cleaned. Low Fat is represented as low fat and LF. And also, replace reg with Regular

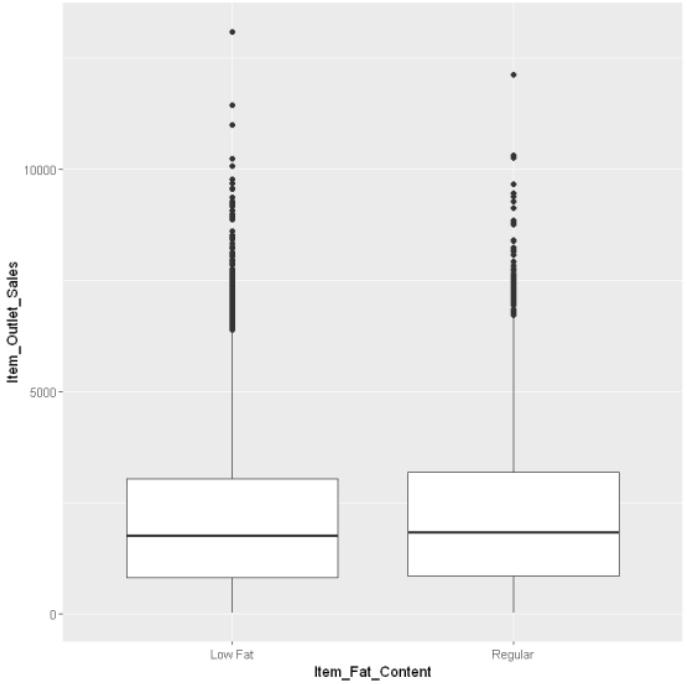
EDA – Bivariate Analysis

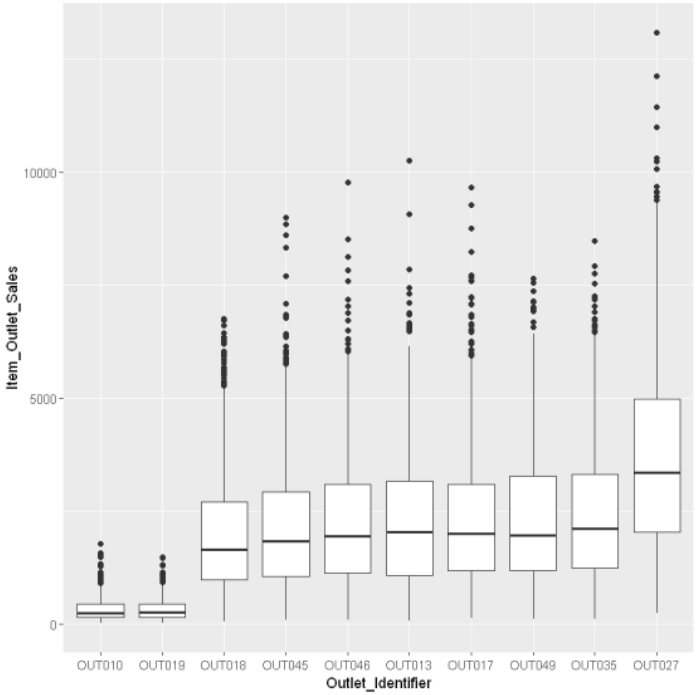
1. No clear linear relationships exist b/n Item Outlet sales and Item_Visibility, Item_Weight.

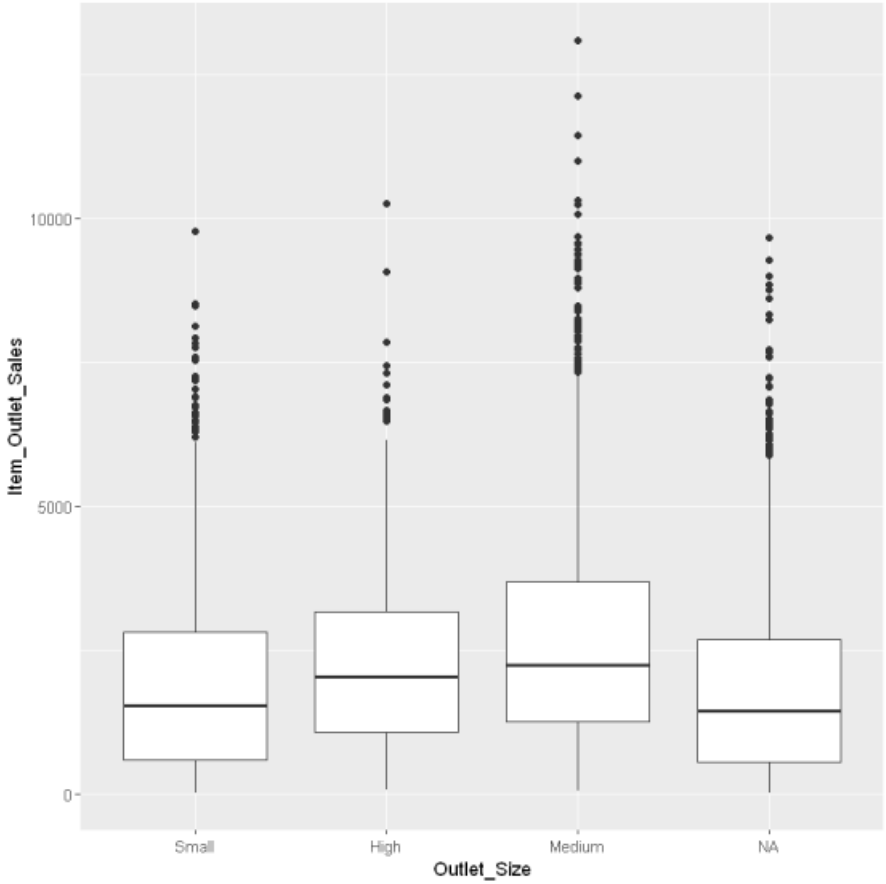
2. A linear relationship exists between Item_Sales_Outlet and Item_MRP



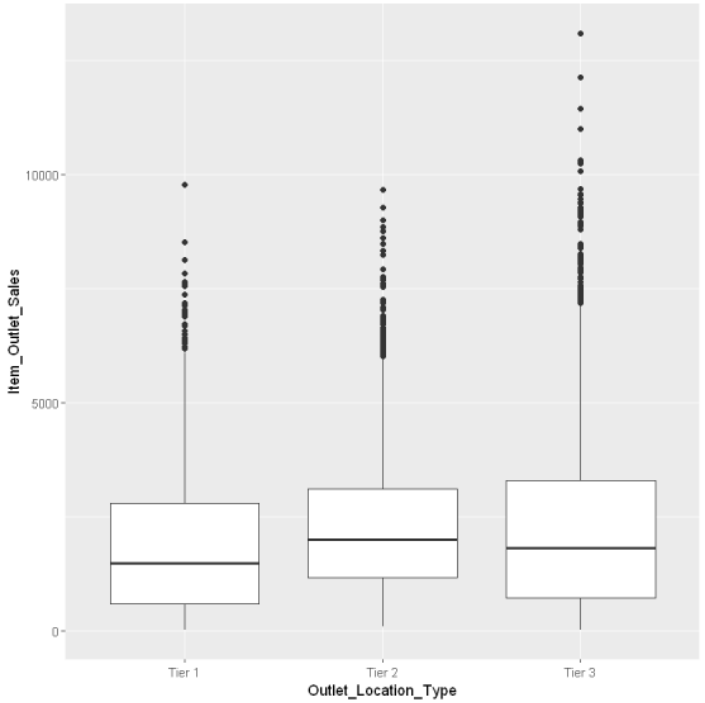
EDA - Bivariate

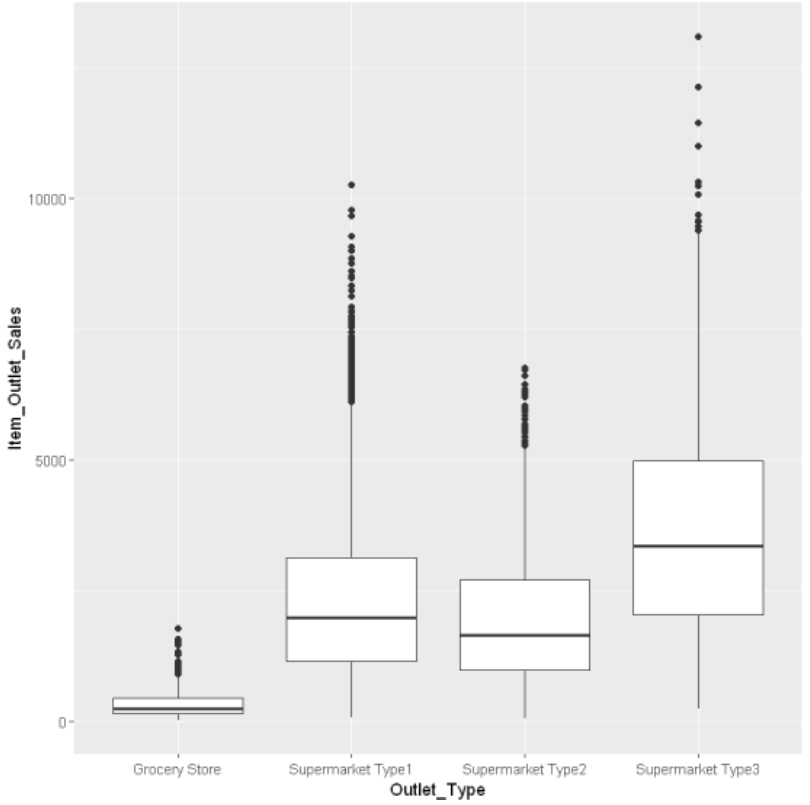


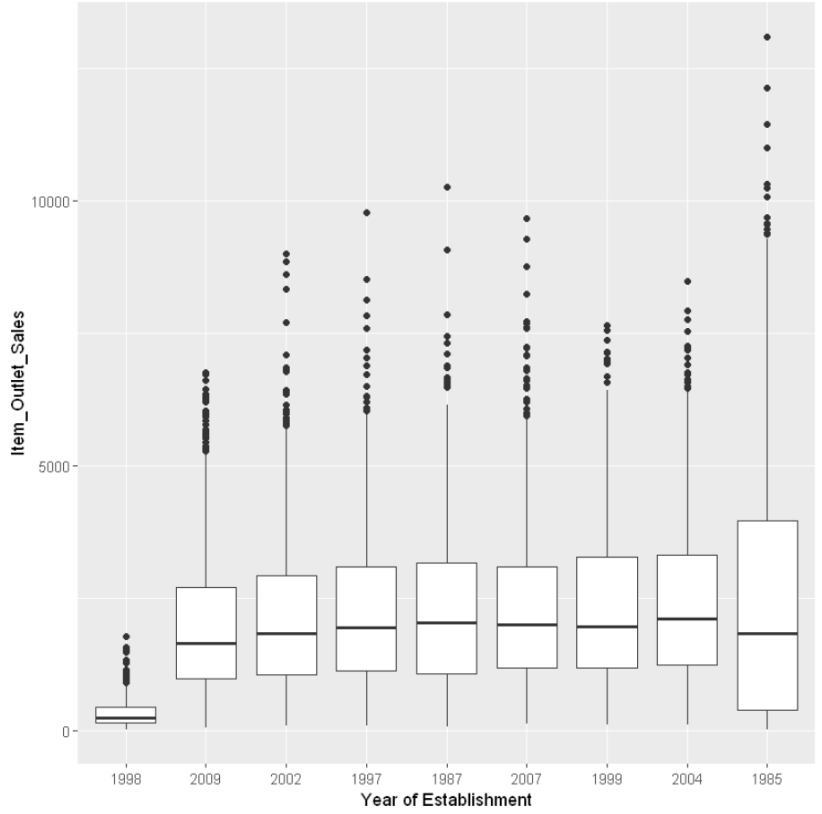




EDA - Bivariate







Product Questions and Methods

The product will be a **report** that details the influence of item characteristics and store profile on item_sales.

In addition, the product will contain a **user-friendly application** that will provide business teams the ability to predict/forecast the sales of an item based on the item characteristics and store profile

Methods I will use include

- Descriptive Statistics:
- Univariate and Bivariate analysis:
- Correlation Analysis:
- Modelling: Here, I will build predictive models that best fit the data and evaluate each model using the adjusted R-squared metric.

More Data Questions

The analysis will be validated in two ways:

1. We will **document domain knowledge within the organization** i.e. interview various stores to understand their beliefs of how item characteristics and store profile affects sales. After this, we will test whether our model validates or contradicts their beliefs.

2. We will request for new data to evaluate/validate our model and check whether our model test performance (measured by RMSE/MAE or Adjusted-R-squared), is at least better than our training performance (from our analysis).

Audience:

The audience of this report are in 2 broad folds:

1. Upper-level management: A comprehensive report detailing the influences/factors that affect item_outlet_sales will be submitted to this audience. This report will significantly guide strategic decision-making.
2. Supply Chain and Sales teams: A prediction/forecasting tool (or application) will be delivered to this audience to enable planning and sales forecast.

Other Considerations

The product will be used in the following ways:

1. Report: This will be used to enable strategic decision-making. For instance, this will inform decisions on what outlet locations and types influence item sales.
2. Application: This tool will be used by mid-level staff for tactical and operational decision-making, such as forecasting sales for a given product in a given store.

How much detail to include?

1. Report: This should be high-level information and not include so much detail on modelling approaches and others.

2. Application: This should contain enough detail for the application to be used to pre dict/forecast sales.

Thank you