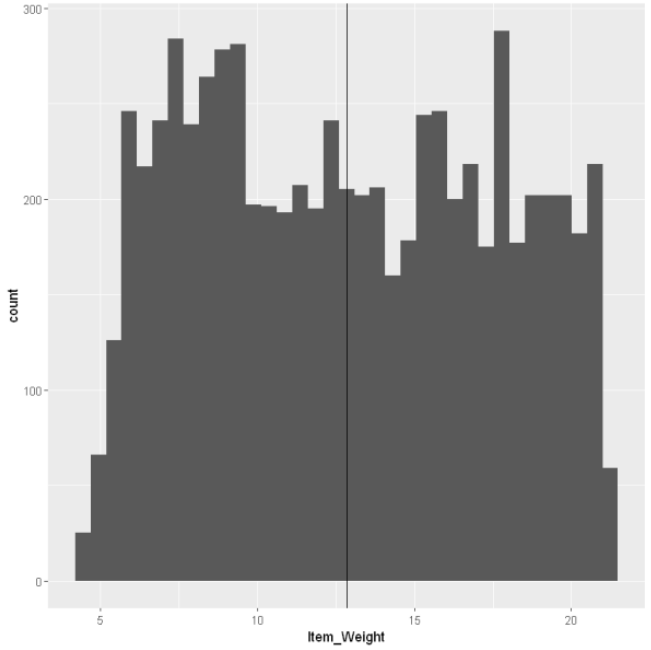
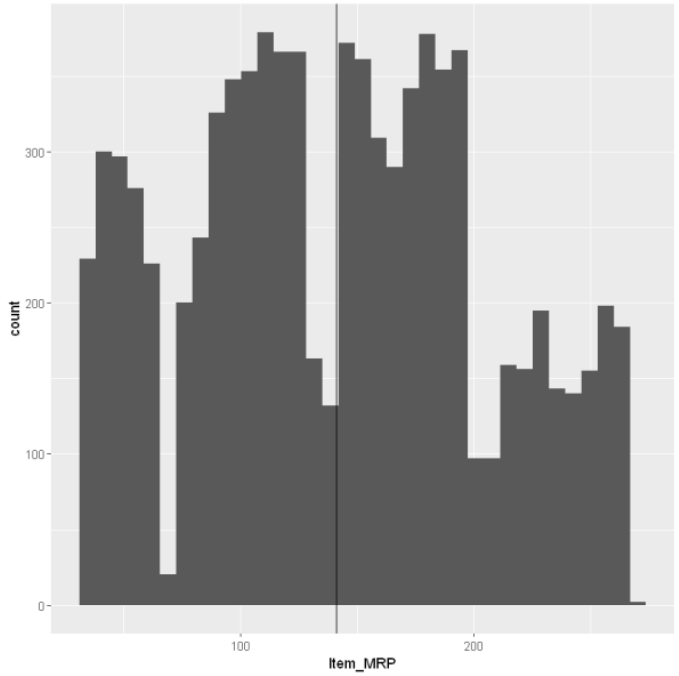
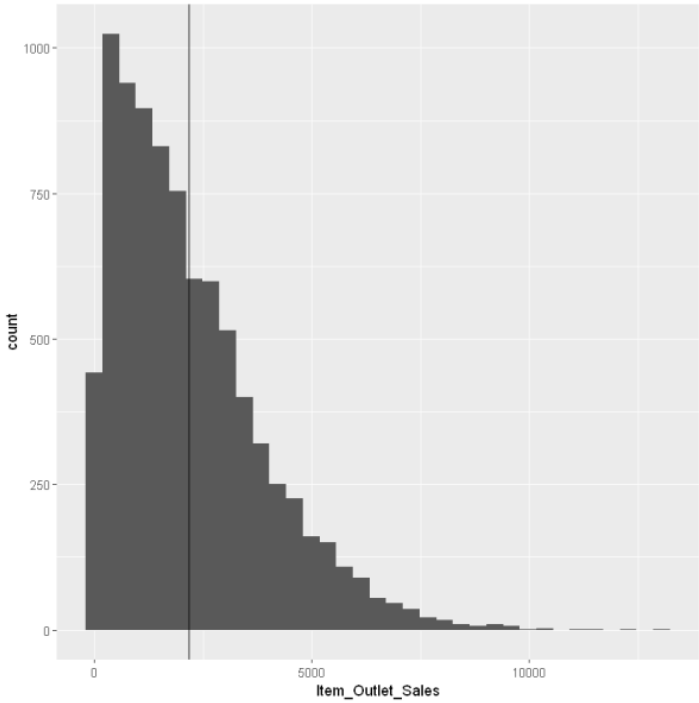
# BigMart Project—Product Analytics and EDA

# EDA - Numeric



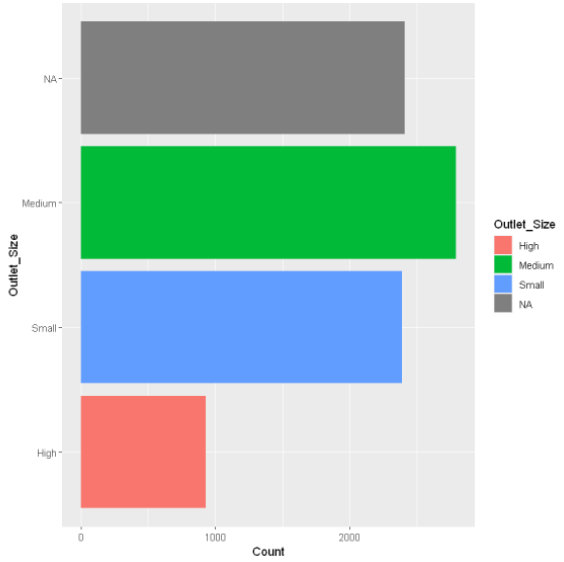



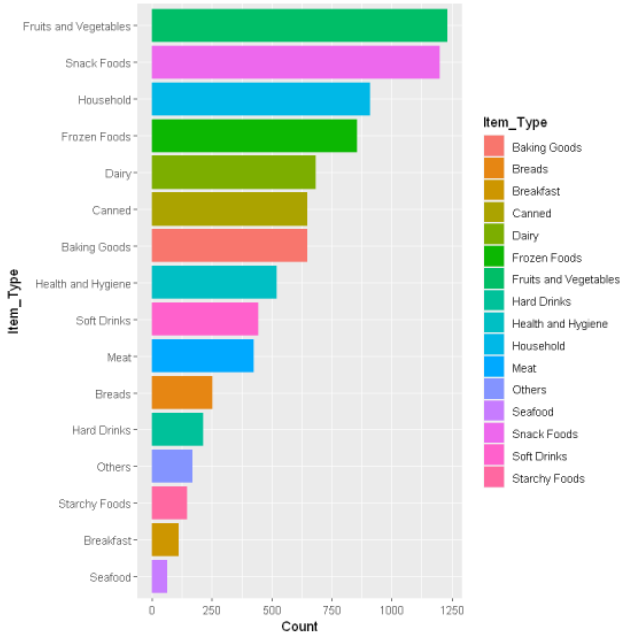



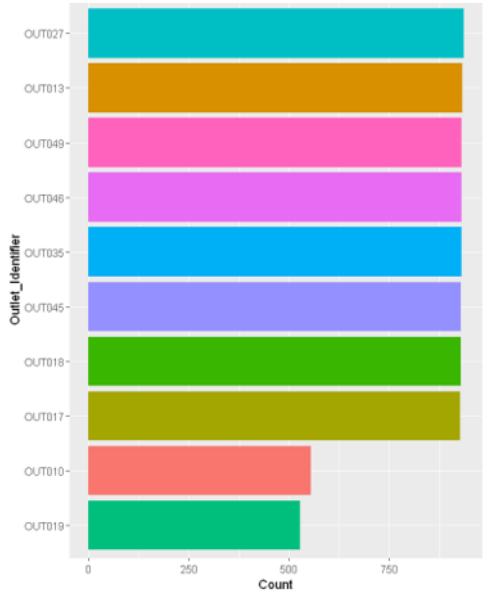

Dataset contains:

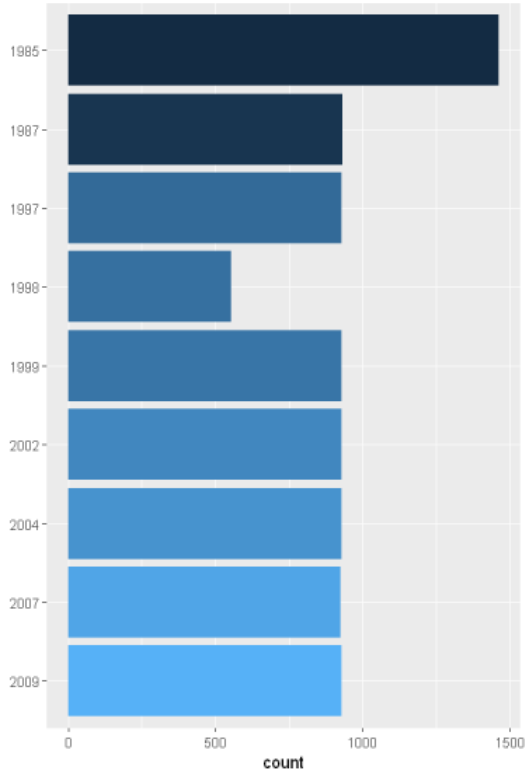
- 5 numeric columns
- Item_Weight has 1463 missing records

# EDA - Categorical



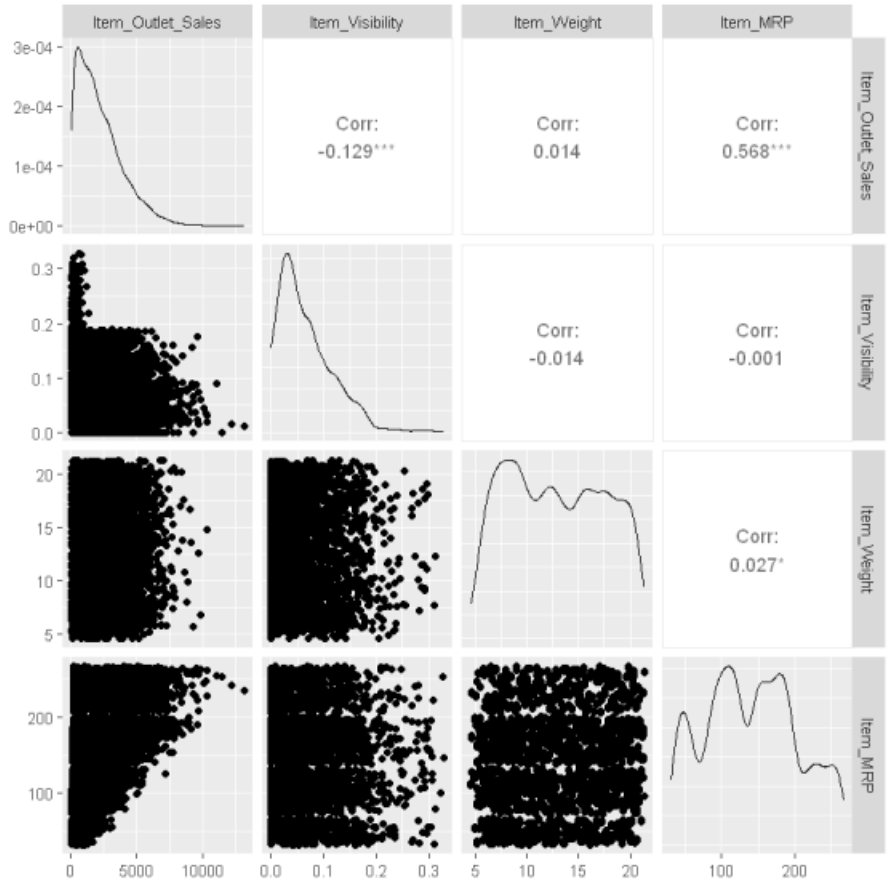

Dataset contains:

- 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

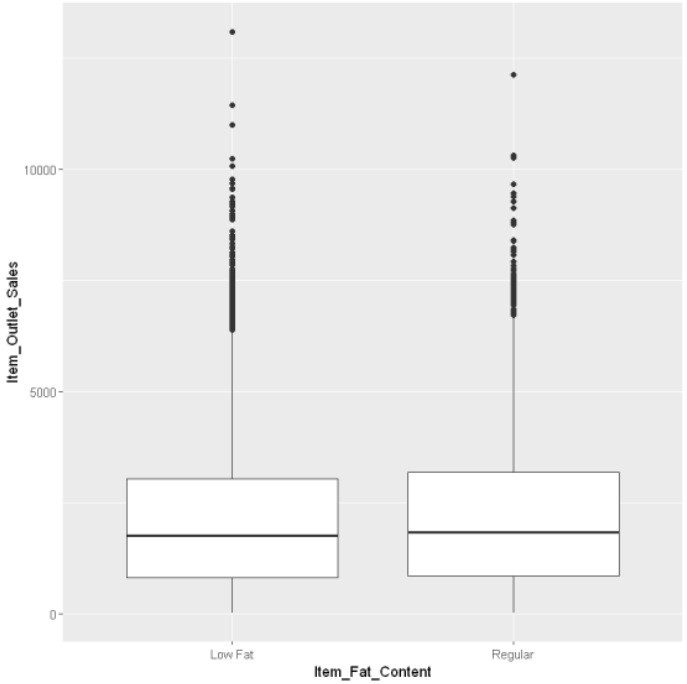
# EDA – Bivariate Analysis

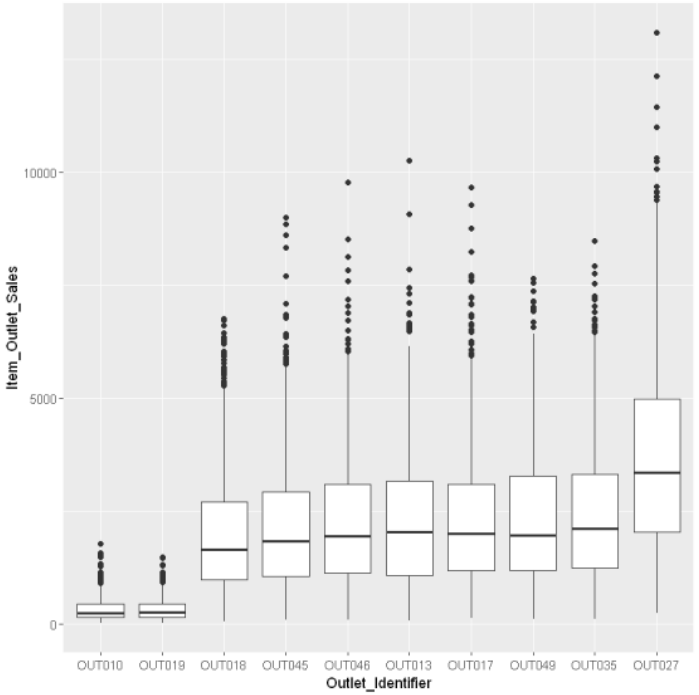
1. No clear linear relationships exist b/n Item Outlet sales and Item_Visibility, Item_Weight.

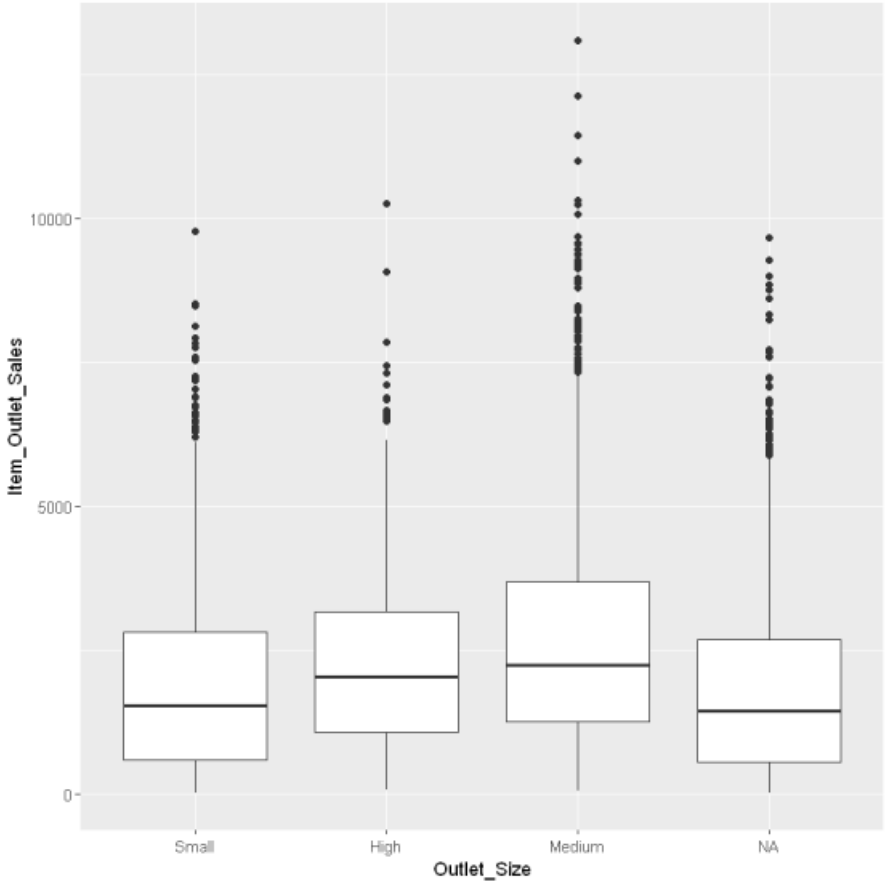
2. A linear relationship exists between Item_Sales_Outlet and Item_MRP



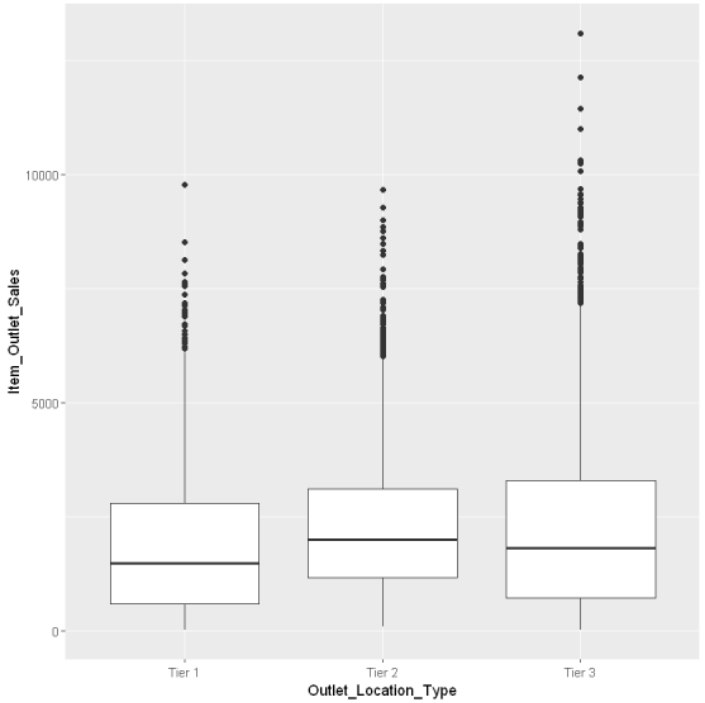
# EDA - Bivariate

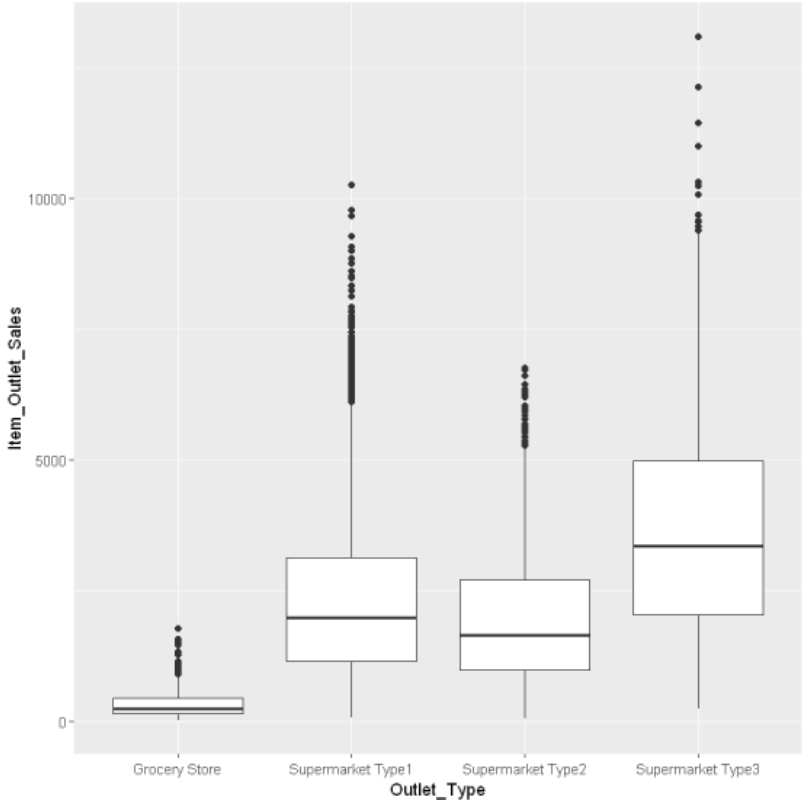


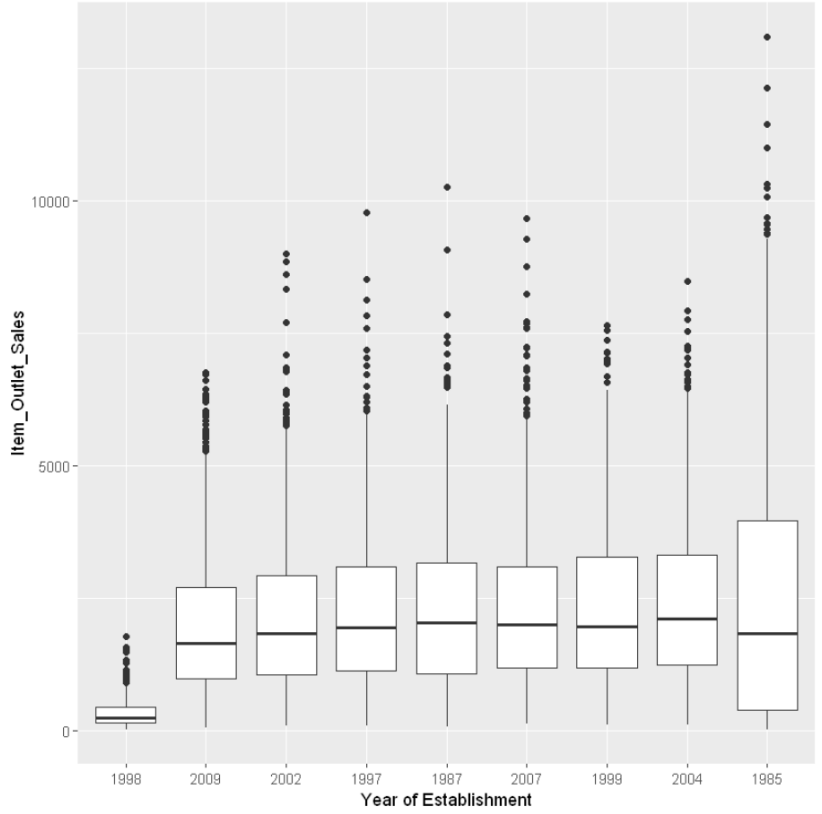



# EDA - Bivariate



# 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 predict/forecast sales.

# Handling Missing Data – Item-Weight

A closer look at the Item Weight column revealed that all the missing records were from a single year, i.e. 1985, for outlets OUT027 and OUT019.

Given that all the missing Item Weight records are from 1985, performing a listwise deletion will mean I lose insightful data that will be invaluable in this analysis.

I used correlation analysis to inspect the relationship between the item weight and other predictors to prevent this data loss from listwise deletion.

The result revealed low correlations between Item_Weight and the other columns. As a result of the low correlations discovered, I decided to exclude the item weight column from my predictive analysis.

| | Item_Outlet_Sales | Item_Visibility | Item_Weight | Item_MRP |
|---|---|---|---|---|
| **Item_Outlet_Sales** | 1.00000000 | -0.085334041 | 0.01412274 | 0.620961316 |
| **Item_Visibility** | -0.08533404 | 1.000000000 | -0.01404773 | -0.006061148 |
| **Item_Weight** | 0.01412274 | -0.014047726 | 1.00000000 | 0.027141154 |
| **Item_MRP** | 0.62096132 | -0.006061148 | 0.02714115 | 1.000000000 |

# Handling Missing – Outlet Size

A closer look at the missing values from the dataset reveals that most missing values from the outlet location type Tier 2 with 1855 missing records and Tier 3 with 555 missing records.

All Tier 3 missing records are from outlet OUT010, while the rest are from outlets OUT017 and OUT045 in Tier 2. I then studied the sales profile from the missing records, discovering that the average sales per outlet size and location type followed no definite pattern, hence not allowing for logical imputation. Based on the above, I decided not to impute the missing records and exclude this column from my analysis.

**Decision: I felt confident that the outlet identifier column implicitly revealed information about the outlet size.**

| Outlet_Size | Outlet_Location_Type | Outlet_Identifier | null_count | avg_sales |
|---|---|---|---|---|
| <chr> | <chr> | <chr> | <int> | <dbl> |
| Small | Tier 1 | OUT019 | 0 | 340.3297 |
| Small | Tier 1 | OUT046 | 0 | 2277.8443 |
| Medium | Tier 1 | OUT049 | 0 | 2348.3546 |
| NA | Tier 2 | OUT017 | 926 | 2340.6753 |
| Small | Tier 2 | OUT035 | 0 | 2438.8419 |
| NA | Tier 2 | OUT045 | 929 | 2192.3848 |
| NA | Tier 3 | OUT010 | 555 | 339.3517 |
| High | Tier 3 | OUT013 | 0 | 2298.9953 |
| Medium | Tier 3 | OUT018 | 0 | 1995.4987 |
| Medium | Tier 3 | OUT027 | 0 | 3694.0386 |

A tibble: 10 × 14

| Item_Identifier | Item_Weight | Item_Fat_Content | Item_Visibility | Item_Type | Item_MRP | Outlet_Identifier | Outlet_Establishment_Year | Outlet_Size | Outlet_Location_Type | Outlet_Type | Item_Outlet_Sales | sales_contribution | rolling_sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <chr> | <dbl> | <chr> | <dbl> | <chr> | <dbl> | <chr> | <dbl> | <chr> | <chr> | <chr> | <dbl> | <dbl> | <dbl> |
| NCE42 | NA | Low Fat | 0.01055095 | Household | 234.9958 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 13086.965 | 0.0007039361 | 0.0007039361 |
| FDQ19 | NA | Regular | 0.01429556 | Fruits and Vegetables | 242.6512 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 12117.560 | 0.0006517927 | 0.0013557288 |
| FDZ20 | NA | Low Fat | 0.00000000 | Fruits and Vegetables | 253.0356 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 11445.102 | 0.0006156218 | 0.0019713506 |
| FDP33 | NA | Low Fat | 0.08883995 | Snack Foods | 254.2672 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 10993.690 | 0.0005913407 | 0.0025626914 |
| FDI50 | NA | Regular | 0.03069331 | Canned | 228.0352 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 10306.584 | 0.0005543819 | 0.0031170733 |
| FDF39 | 14.850 | Regular | 0.01949505 | Dairy | 261.2910 | OUT013 | 1987 | High | Tier 3 | Supermarket Type1 | 10256.649 | 0.0005516960 | 0.0036687693 |
| FDU14 | NA | Low Fat | 0.03458436 | Dairy | 248.3750 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 10236.675 | 0.0005506216 | 0.0042193908 |
| NCH18 | NA | Low Fat | 0.04444496 | Household | 245.2802 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 10072.888 | 0.0005418116 | 0.0047612025 |
| NCM05 | 6.825 | Low Fat | 0.05984697 | Health and Hygiene | 262.5226 | OUT046 | 1997 | Small | Tier 1 | Supermarket Type1 | 9779.936 | 0.0005260540 | 0.0052872565 |
| FDC17 | NA | Low Fat | 0.01538586 | Frozen Foods | 208.9928 | OUT027 | 1985 | Medium | Tier 3 | Supermarket Type3 | 9678.069 | 0.0005205747 | 0.0058078312 |

The top 10 items listed below contribute to about 0.6% of the total item outlet sales in the dataset.

# Correlation Analysis

| | Item_Outlet_Sales | Item_Visibility | Item_Weight | Item_MRP |
|---|---|---|---|---|
| Item_Outlet_Sales | 1 | | | |
| Item_Visibility | -0.085 | 1 | | |
| Item_Weight | 0.0141 | -0.014 | 1 | |
| Item_MRP | 0.621 | -0.006 | 0.027 | 1 |

Item Outlet Sales

| Variable | Correlation | Statistic | P-Value |
|---|---|---|---|
| Item_Fat_Content | 0.019 | 1.728 | 0.084 |
| Outlet_Identifier | 0.162 | 15.185 | 0.000 |
| Outlet Size | -0.129 | -10.175 | 0.000 |
| Outlet Location Type | 0.089 | 8.283 | 0.000 |
| Item Identifier | 0.003 | 0.265 | 0.791 |
| Outlet Type | 0.401 | 40.470 | 0.000 |

# Predictive Model

The following models were trained:

- Linear Regression
- Support Vector Machine
- Classification and Regression Trees
- Xgboost model

| Model | Test - RMSE |
|---|---|
| Support Vector Machine | 0.64 |
| XGBoost | 0.651 |
| Bagged CART Model | 0.658 |
| Linear Regression | 0.662 |
| Random Forest Model | 0.664 |

# Thank you