第六組

資科碩二 108753014 廖宇凡
資科碩一 109753128 紀秉杰 109753102 黃渝庭
資科三 107703048 陳子賢 107703004 李元亨

Our goal / Input

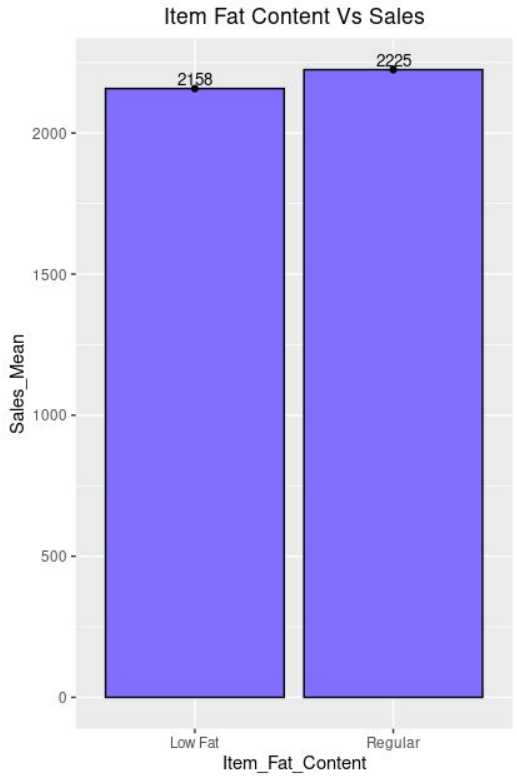
- Data source
  - From **Analytics Vidhya -** Big Mart Sales Prediction Problem
- Input format
  - CSV file
- Any preprocessing
  - Handle Missing Data
  - Handle Redundant Data (Object Identification)

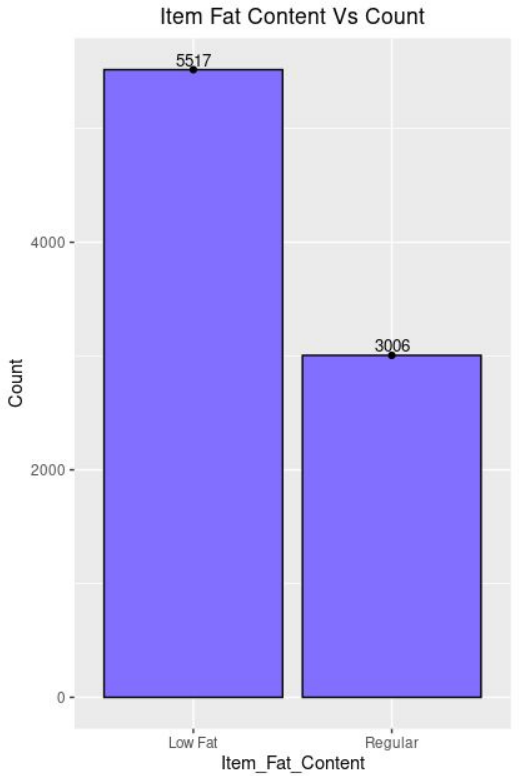
Data Dictionary

Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

Handle Redundant Data - Item_Fat_Content

- Low Fat
  - Low Fat
  - low fat
  - LF
- Regular
  - Regular
  - reg

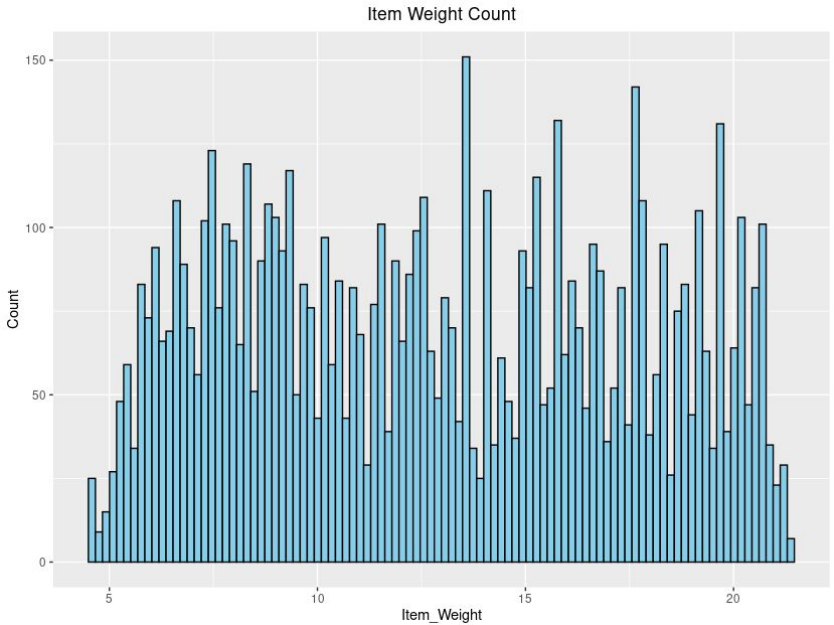




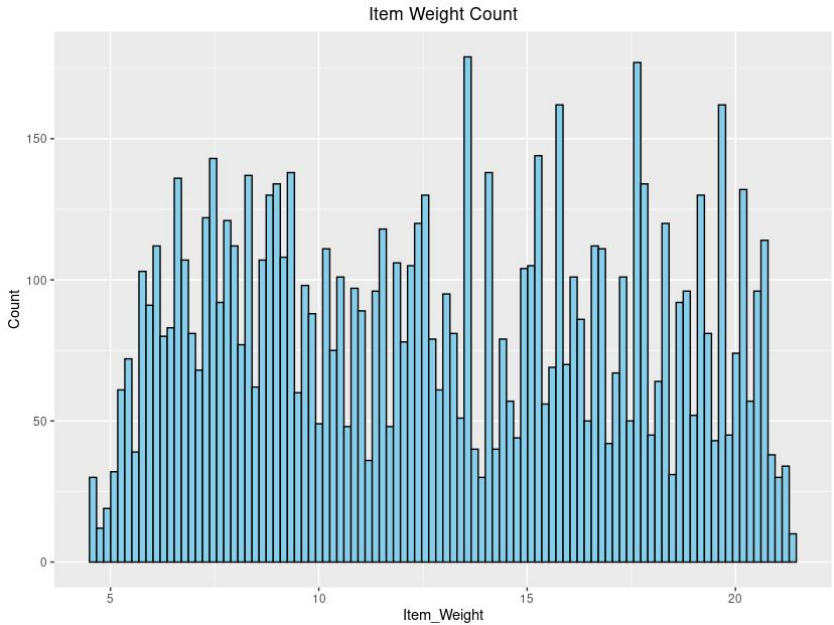
Handle Missing Data - Item_Weight

Fill the missing Item_Weight by same Item_Identifier’s Item_Weight

Before



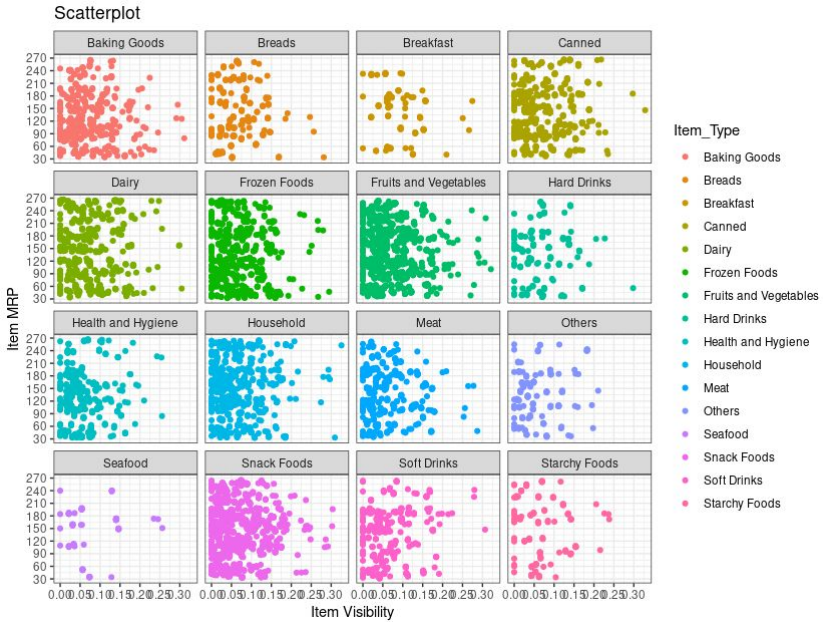
After



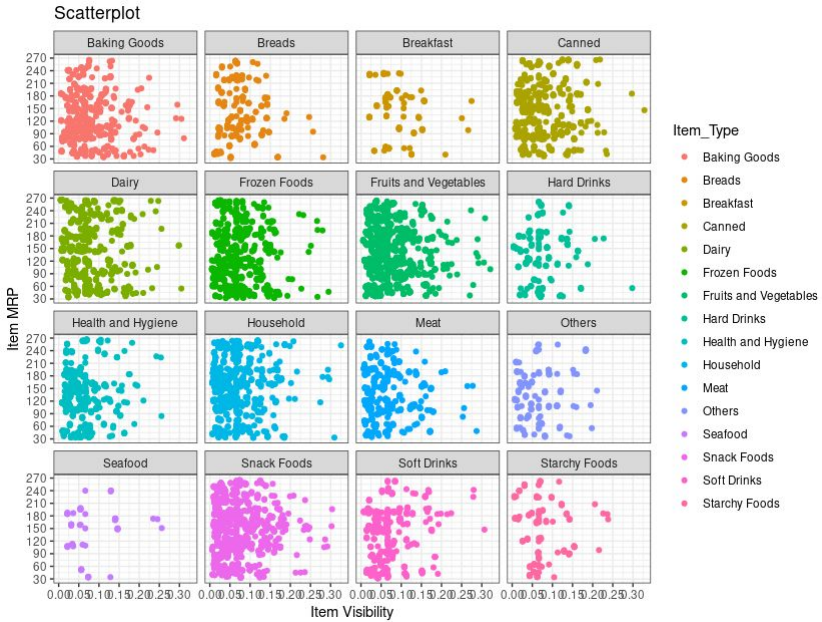
Handle Missing Data - Item_Visibility

Replace 0 Item_Visibility by mean of Item_Visibility

Before

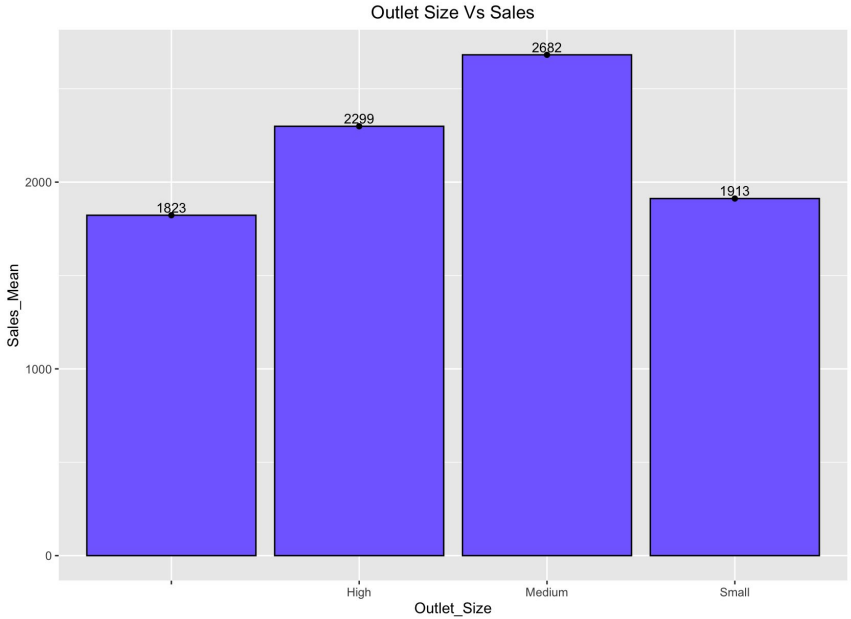


After

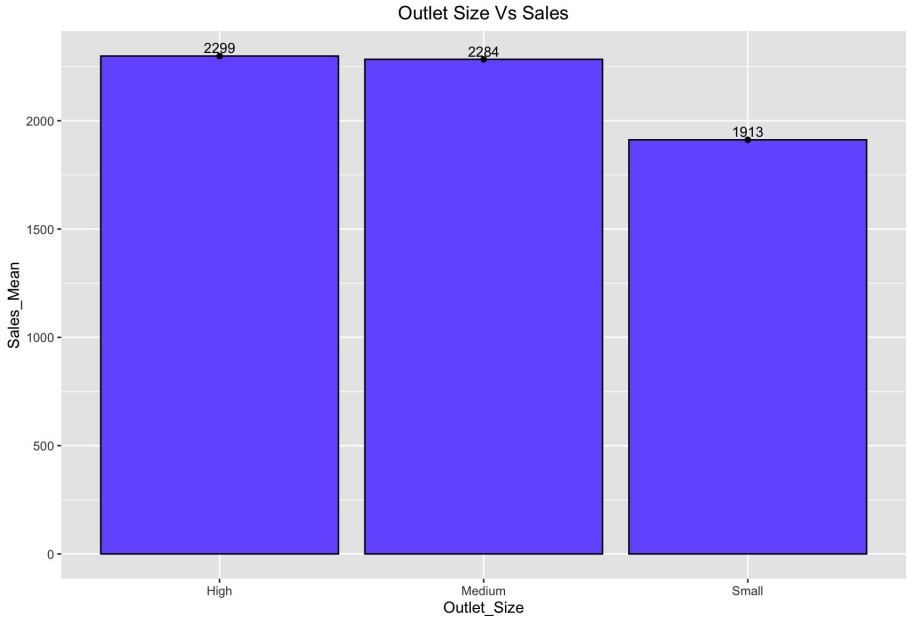


Handle Missing Data - Outlet_Size

Before

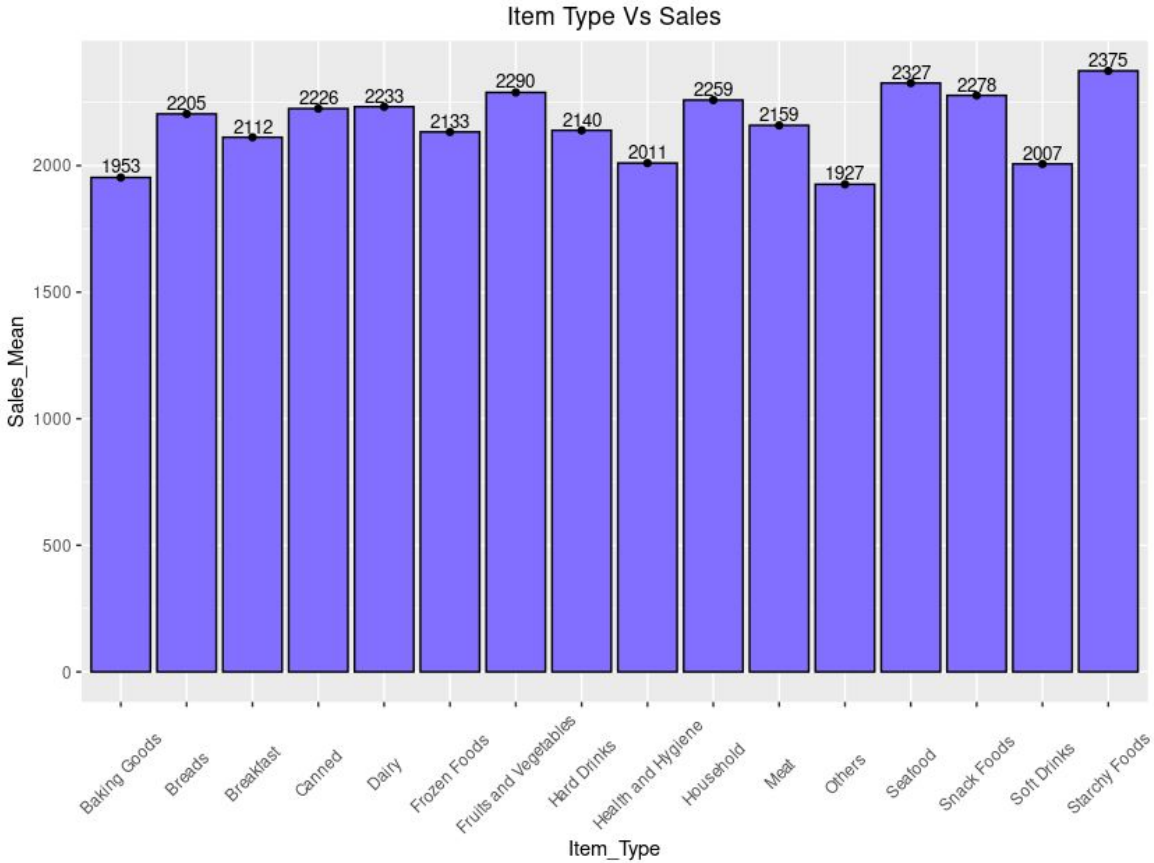


After



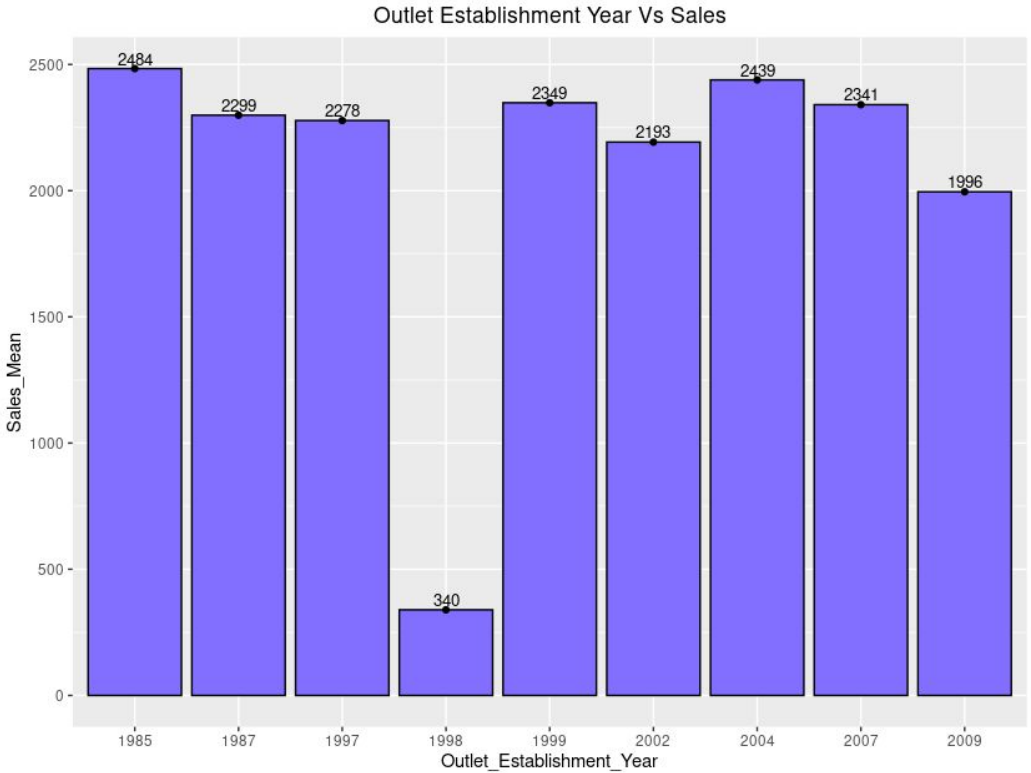
Item_type

16 categories

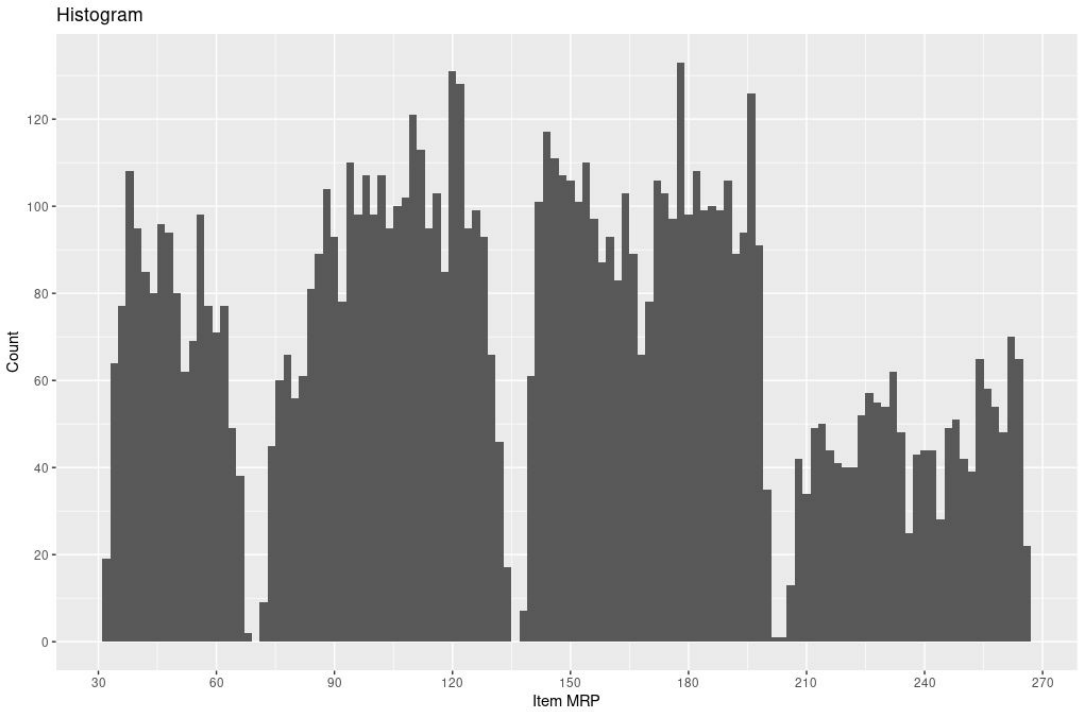


Outlet_Establishment_Year

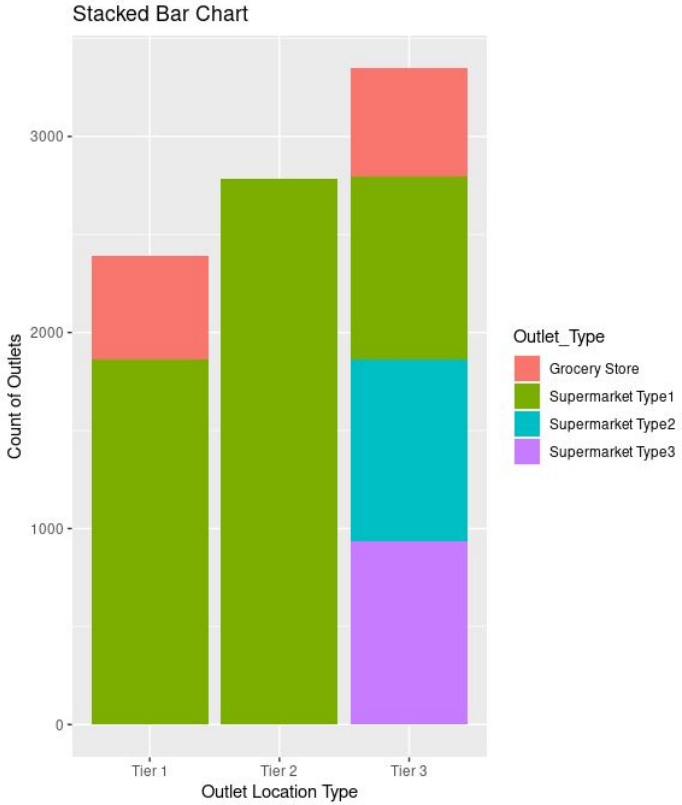
Year = 2013 - Outlet_Establishment_Year

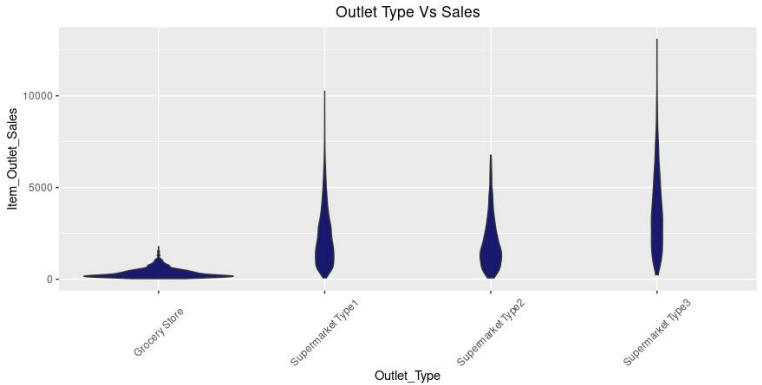


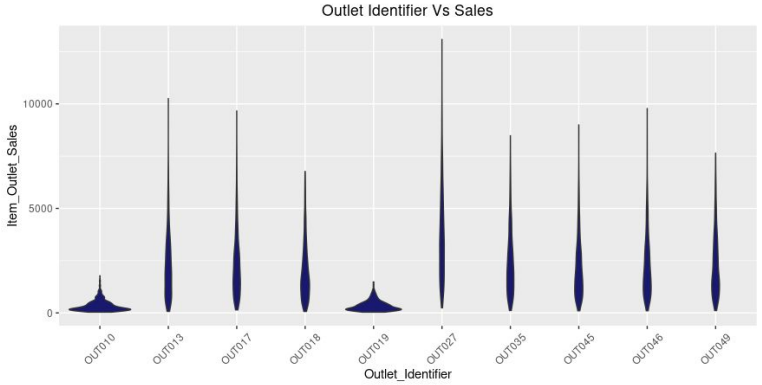
Item_MPR (Maximum Retail Price)



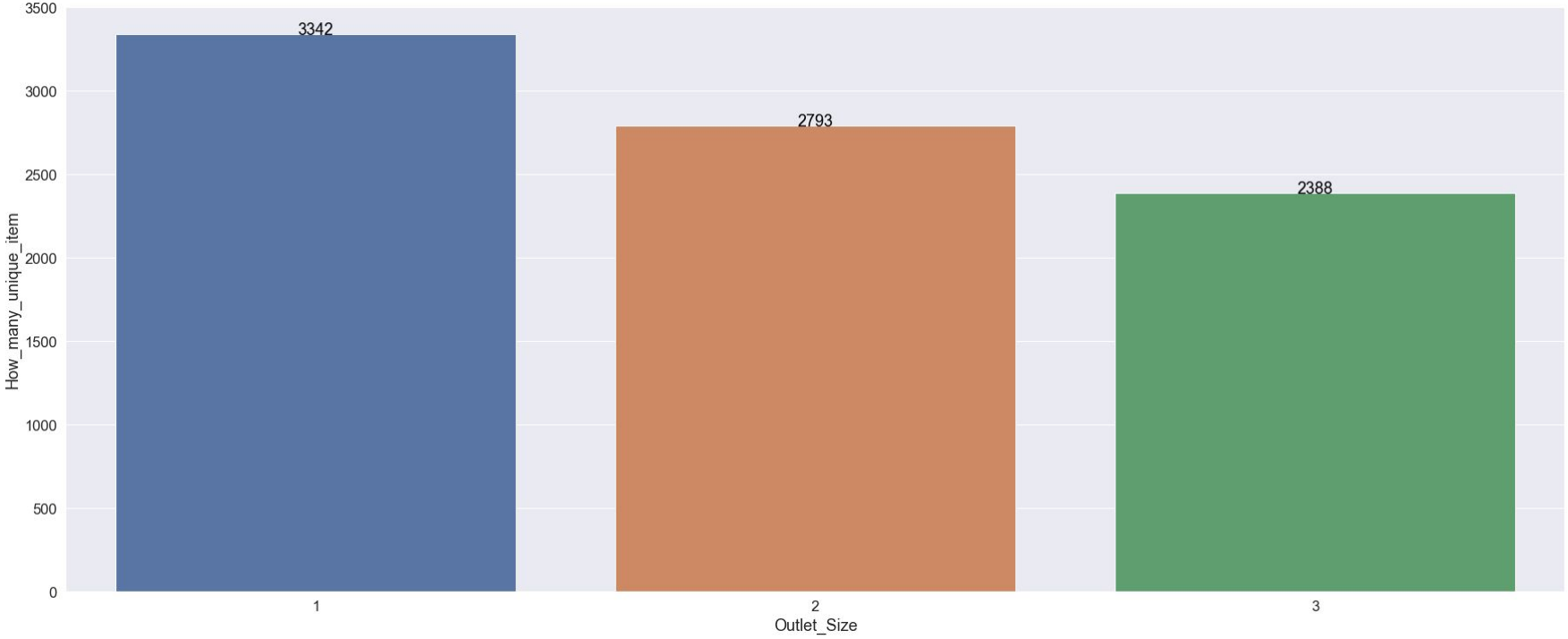
Outlet_Location & Outlet_Type & Outlet_Identifier







Outlet Size with number of kinds of product



New Feature !!! Item_Volume ?

New Feature !!! Item_Volume ?

```
## Create Item Volume Sold.
Train_Base$Item_Volume = Train_Base$Item_Outlet_Sales/Train_Base$Item_MRP
Train_Base$Item_Volume = round(Train_Base$Item_Volume)
Train_Base$Item_Outlet_Sales = NULL
```

With Volume :

Without Volume :

Feature Enginnering

- One-Hot Encoding

Modeling

- Which method do we use?
  - Decision Tree
  - Random Forest ✓ (The Best)
  - LM
  - XGBoost
- What is a null model for comparison?
- How do our perform evaluation?
  - Cross-validation, or extra separated data

Output

- precision, recall, R-square
- Is your improvement significant?

Demo

Reference

- https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/
- https://github.com/Param-Trivedi/Big-Mart-Sales-Data-Prediction?fbclid=IwAR3vso5dZWupAZPELKsMs_bDh1buYPoDiT_t-SdKZQrXN4iHVRUJmgCHSsA
- https://www.kaggle.com/usamakhan8199/big-mart-prediction-top-100-with-optimisation
- https://www.kaggle.com/bgsumanth/plots-in-r
- https://rpubs.com/prateekjoshi565/381886?fbclid=IwAR3G67crQULEmecWedgaIysWx4OuA9DzWdY8S2Km96xv5wf7IW2gN7z2Z2Q