Big Mart Sales Practice Problem

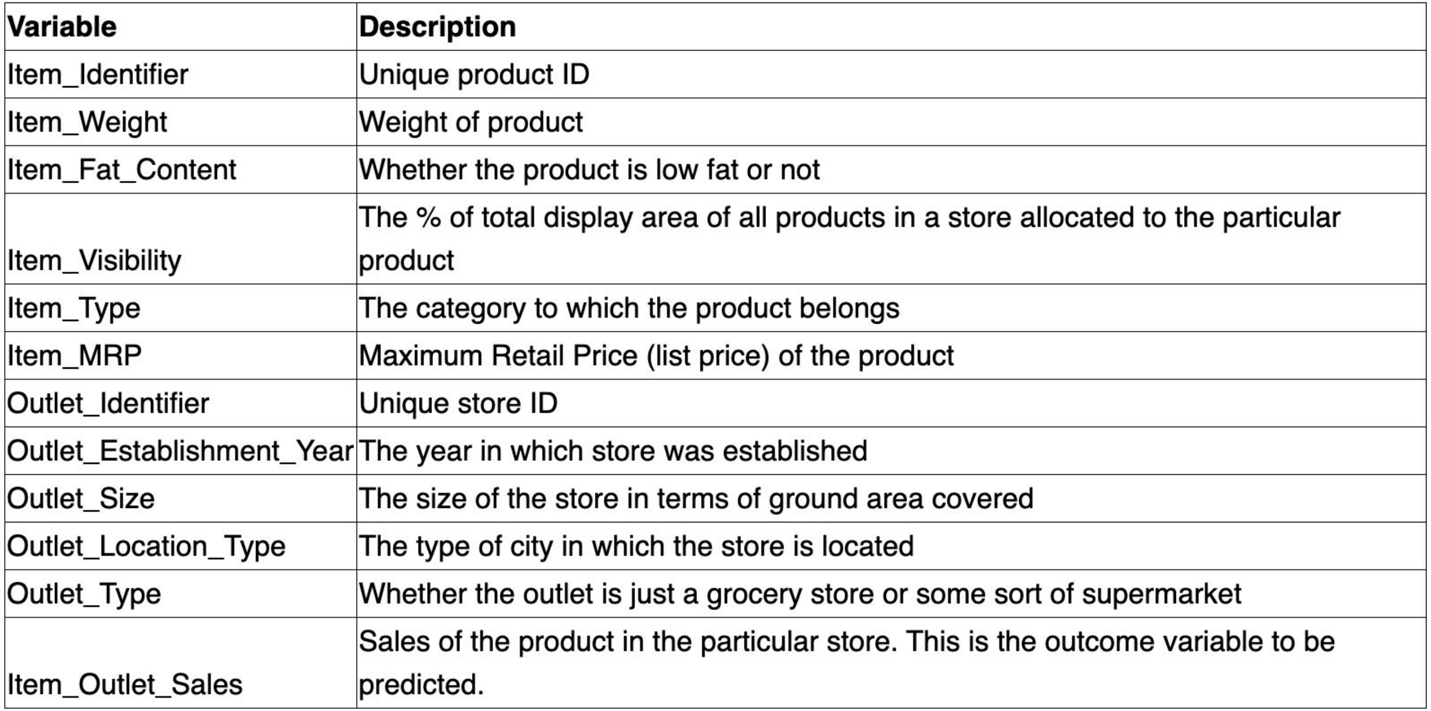
第六組

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

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

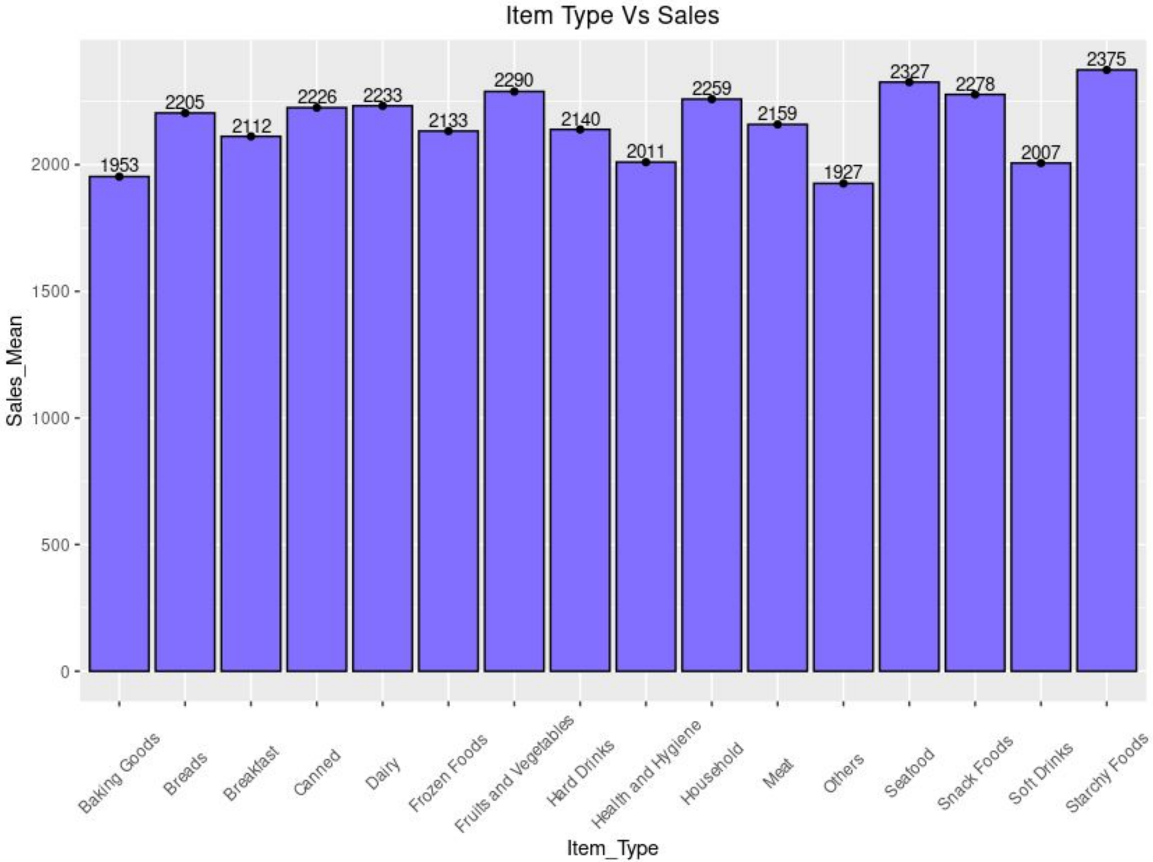
Data Dictionary



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.

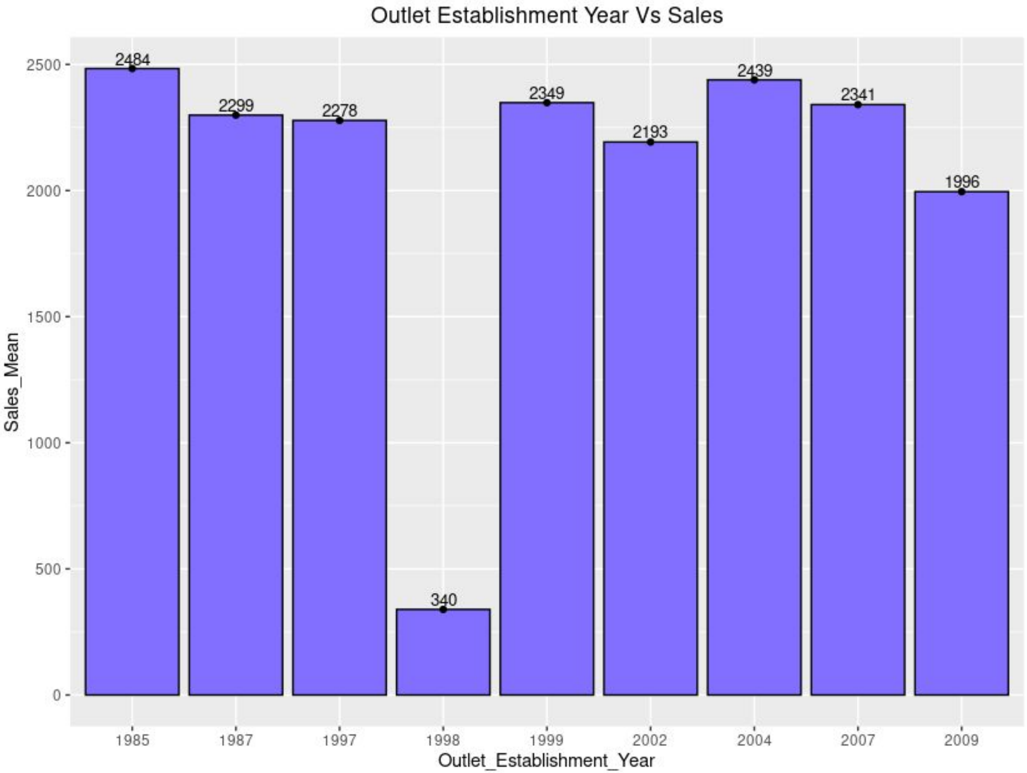
Item_type

16 categories

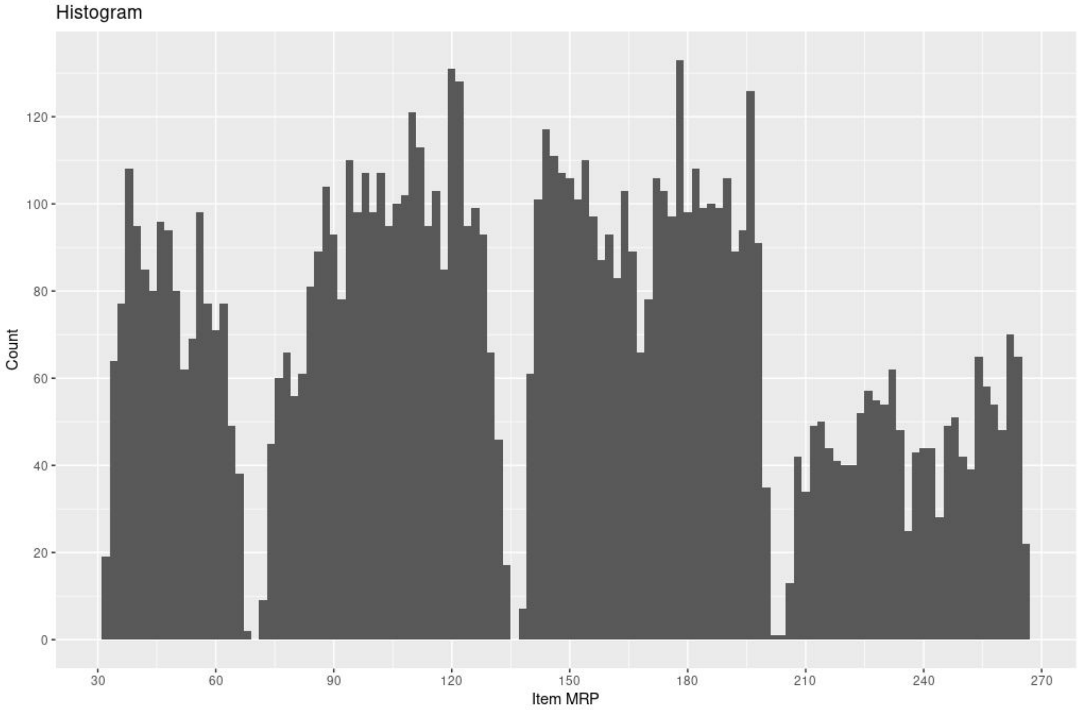


Outlet_Establishment_Year

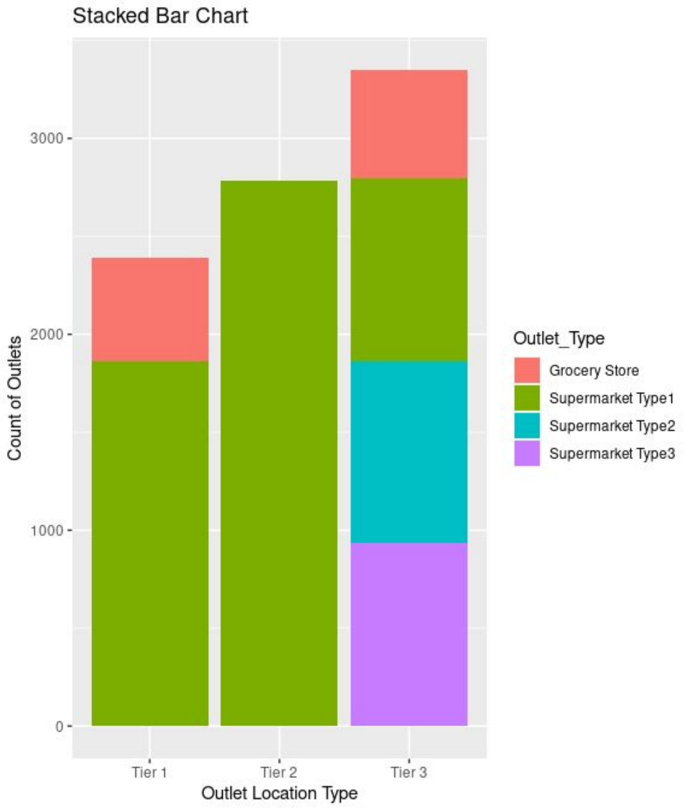
Year = 2013 - Outlet_Establishment_Year

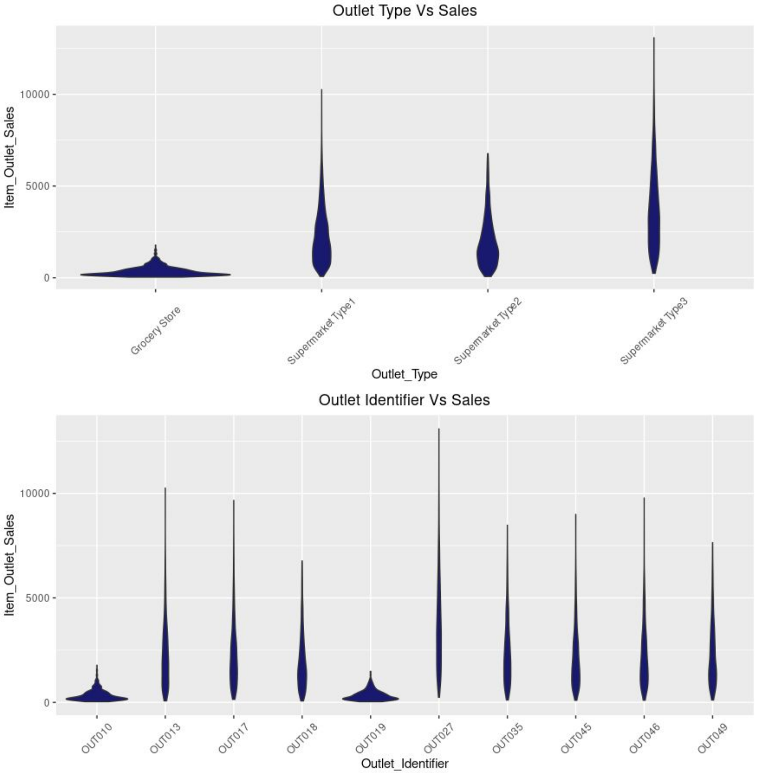


Item_MPR (Maximum Retail Price)



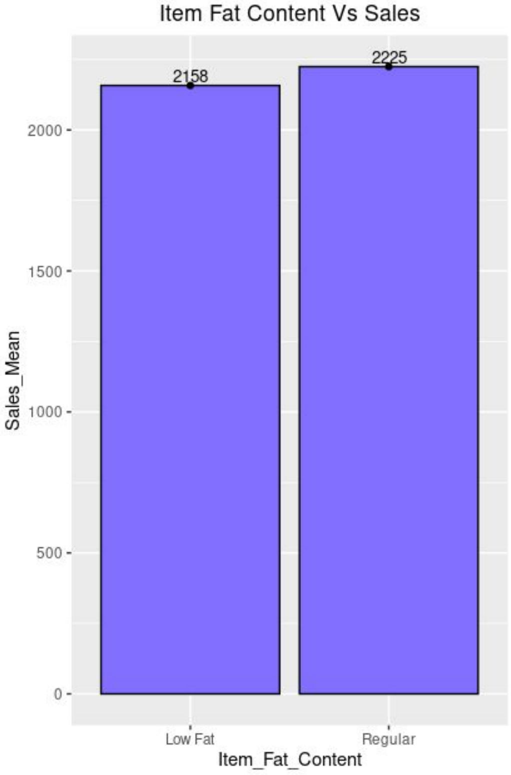
Outlet_Location & Outlet_Type & Outlet_Identifier

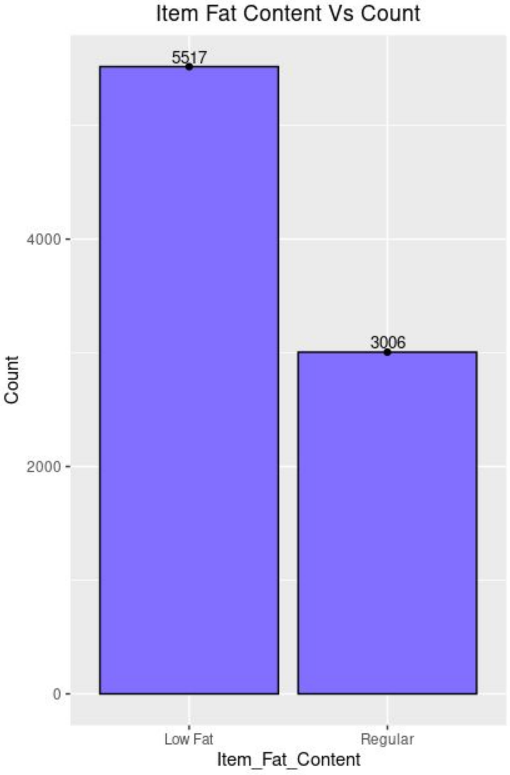




Handle Redundant Data - Item_Fat_Content

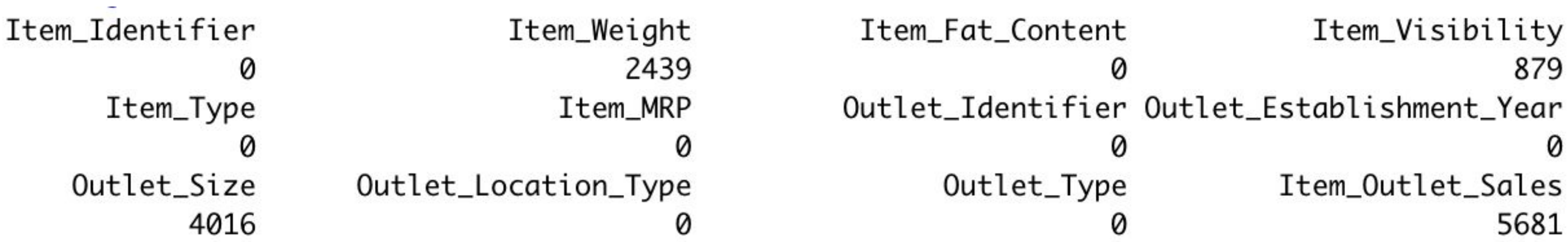
- Low Fat
  - Low Fat
  - low fat
  - LF
- Regular
  - Regular
  - reg





Handle Missing Data

Item_Weight with **NA** value, Item_Visibility with **0** value and Outlet_Size with **“” (Null String)**

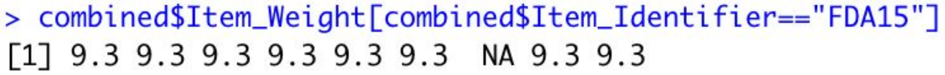


```
  Item_Identifier               Item_Weight          Item_Fat_Content           Item_Visibility
                0                      2439                         0                       879
        Item_Type                  Item_MRP         Outlet_Identifier Outlet_Establishment_Year
                0                         0                         0                         0
      Outlet_Size      Outlet_Location_Type               Outlet_Type         Item_Outlet_Sales
             4016                         0                         0                      5681
```

Handle Missing Data - Item_Weight

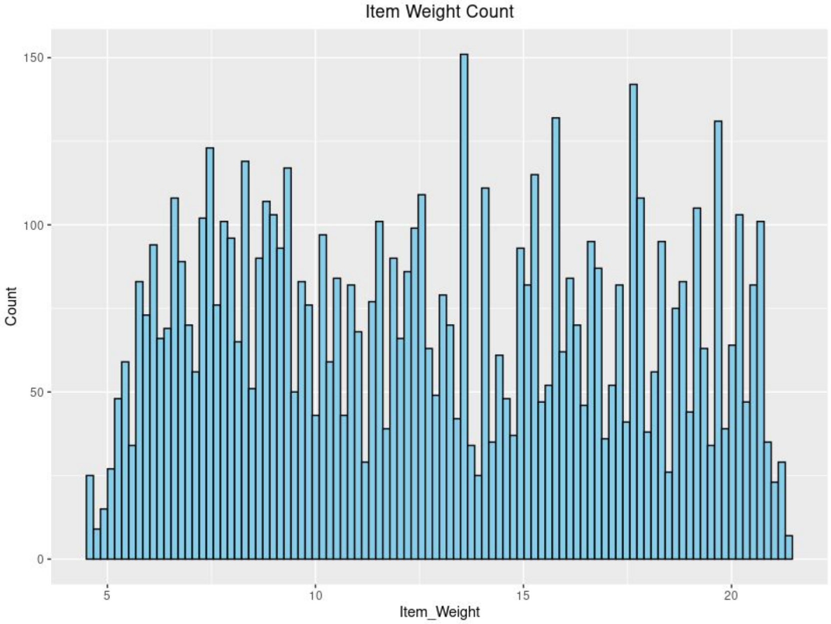
Fill the missing Item_Weight by same Item_Identifier’s Item_Weight

ex:

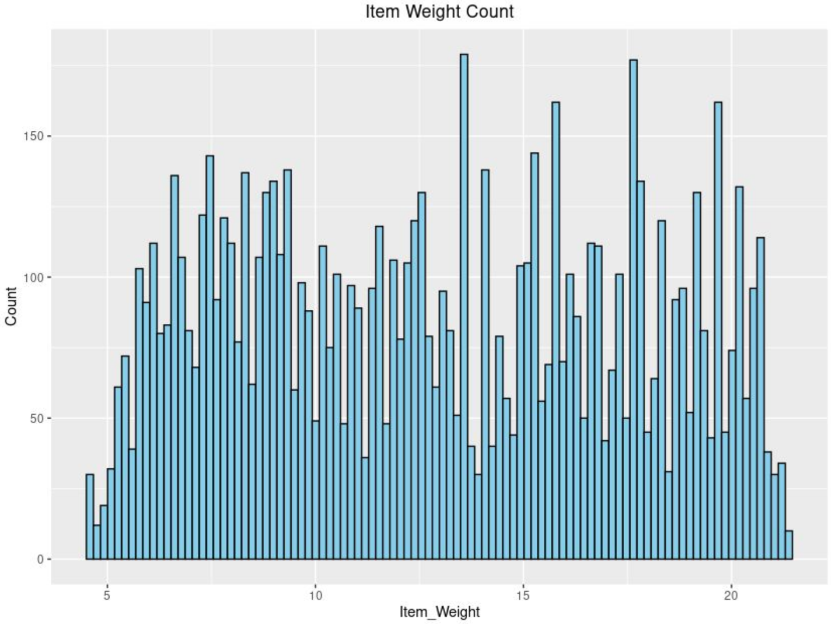


```
> combined$Item_Weight[combined$Item_Identifier=="FDA15"]
[1] 9.3 9.3 9.3 9.3 9.3 9.3  NA 9.3 9.3
```

Before



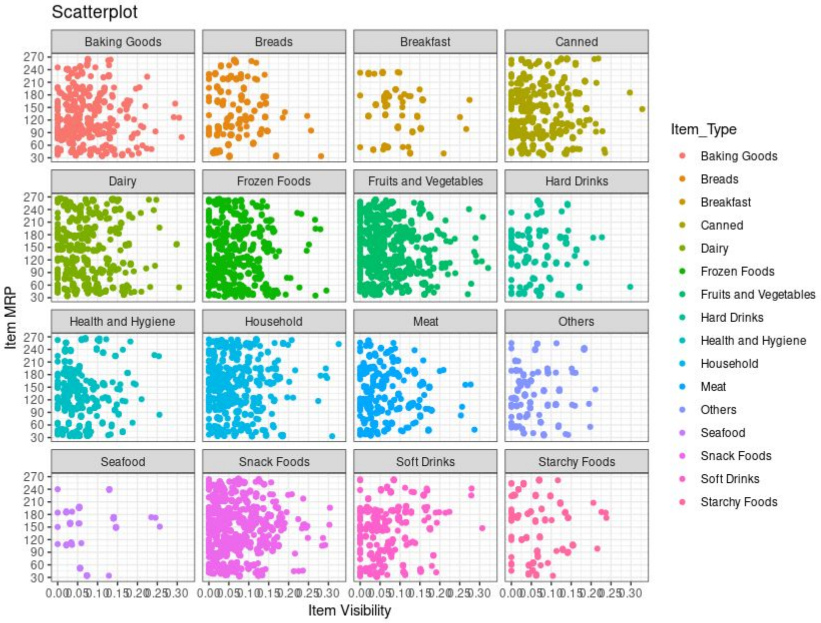
After



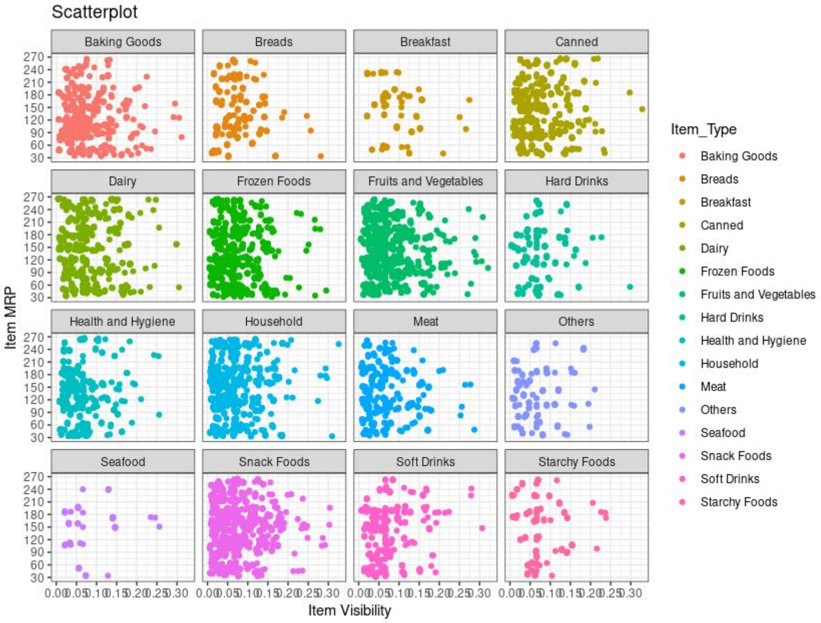
Handle Missing Data - Item_Visibility

Replace 0 Item_Visibility by mean of Item_Visibility

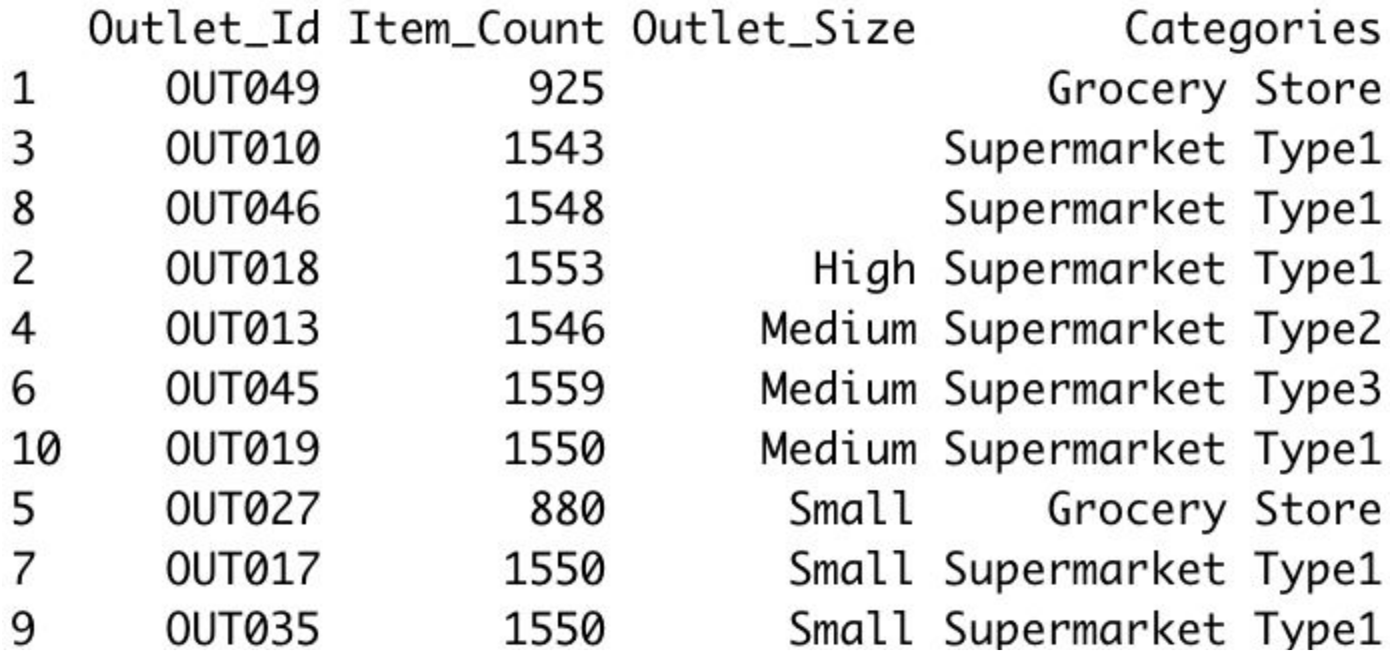
Before



After



Handle Missing Data - Outlet_Size

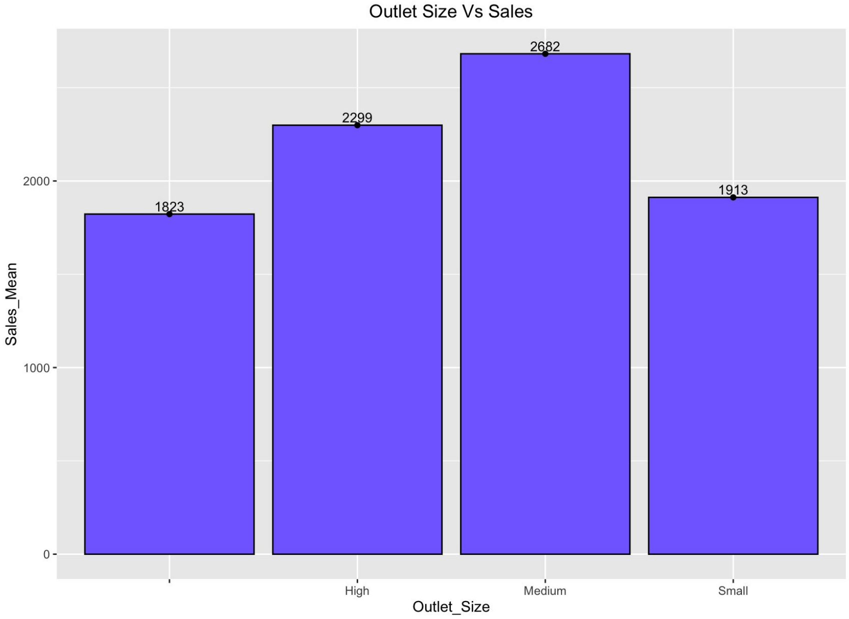


```
   Outlet_Id Item_Count Outlet_Size        Categories
1     OUT049        925                 Grocery Store
3     OUT010       1543             Supermarket Type1
8     OUT046       1548             Supermarket Type1
2     OUT018       1553        High Supermarket Type1
4     OUT013       1546      Medium Supermarket Type2
6     OUT045       1559      Medium Supermarket Type3
10    OUT019       1550      Medium Supermarket Type1
5     OUT027        880       Small     Grocery Store
7     OUT017       1550       Small Supermarket Type1
9     OUT035       1550       Small Supermarket Type1
```

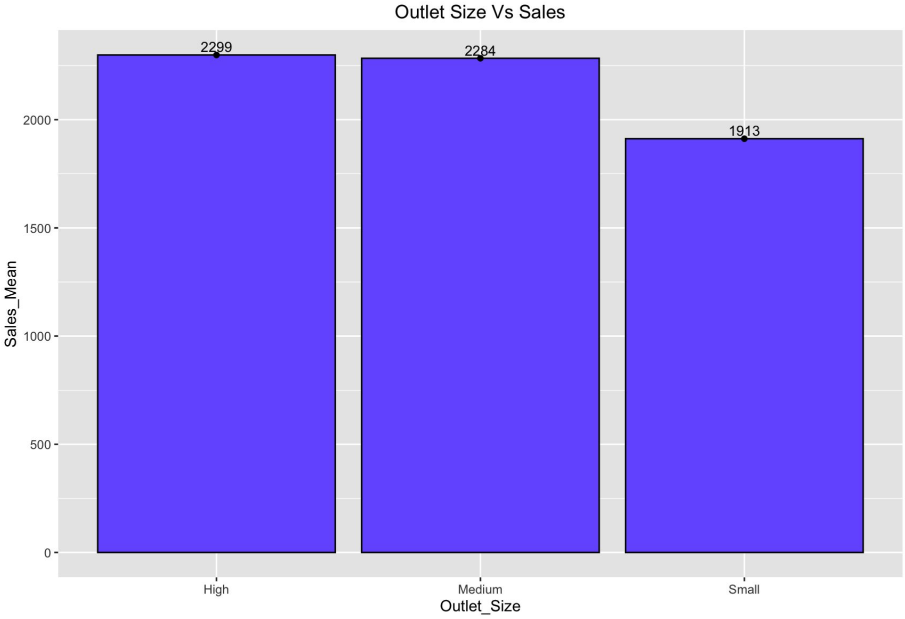
Handle Missing Data - Outlet_Size

Fill the missing Outlet_Size by mode Outlet_Size

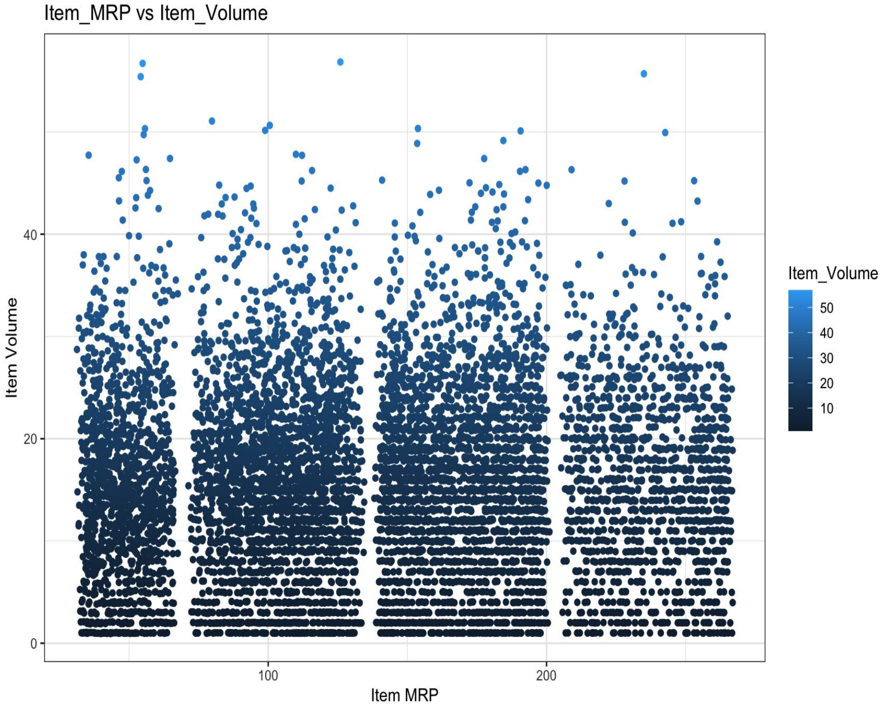
Before

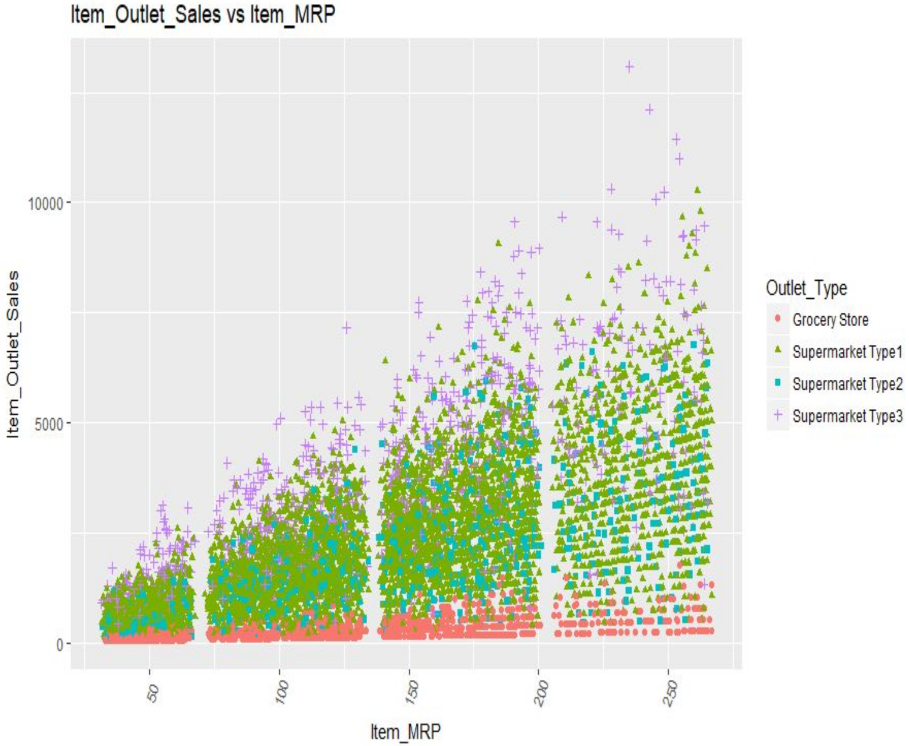


After

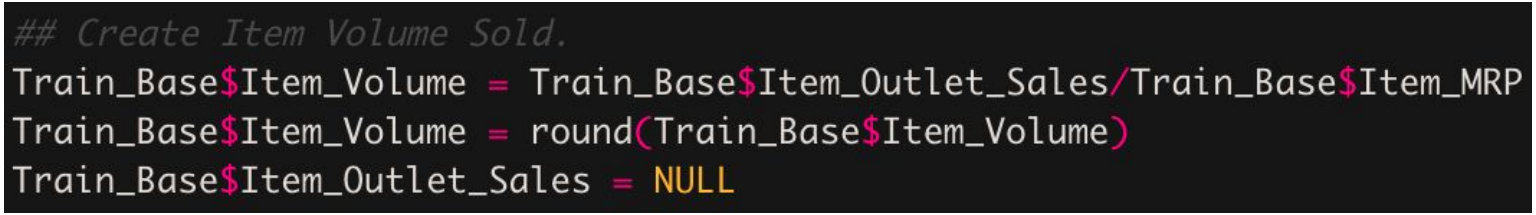


New Feature !!! Item_Volume ?





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 Item_Outlet_Sales :
1142.60973123	1148.27349433

Feature Enginnering

- One-Hot Encoding

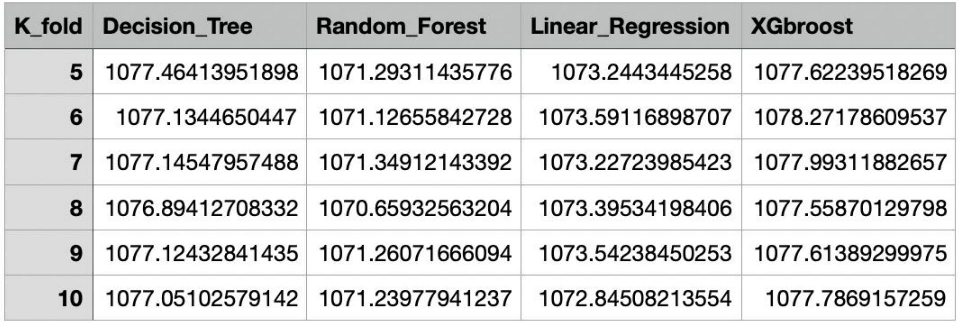
Feature Name	轉變特徵數
Item_Fat_Content	2
Item_Type	16
Outlet_Identifier	10
Outlet_Location_Type	3
Outlet_Type	4

Modeling

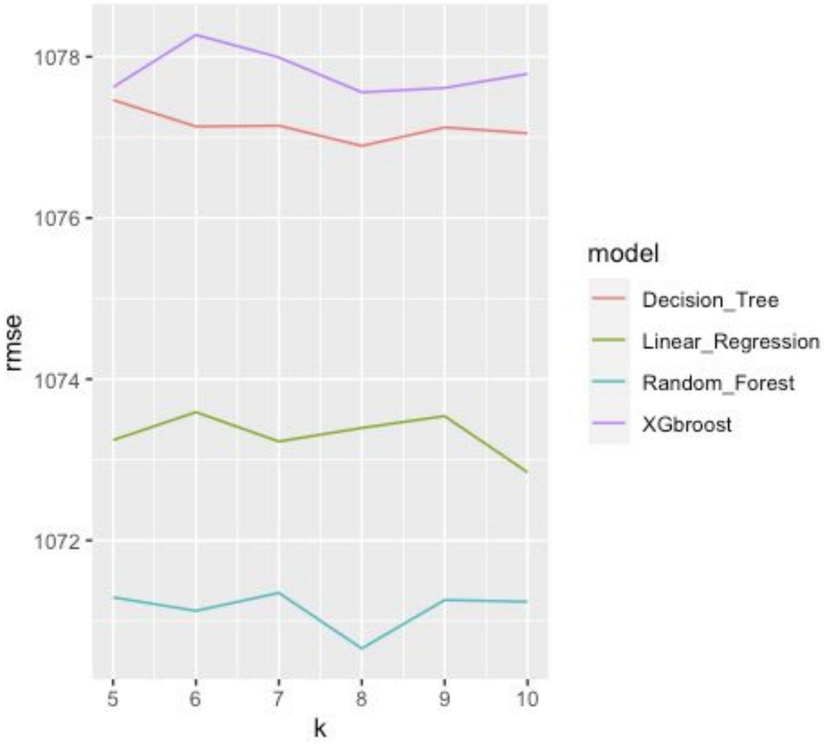
- Which method do we use?
  - Decision Tree
  - Random Forest ✓ (The Best)
  - LM
  - XGBoost
- What is a null model for comparison?
- How do our perform evaluation?
  - Cross-validation

Modeling

- 比賽是要用 RMSE 做評估指標
- Cross-validation 用訓練資料集的銷售額與驗證的銷售額做RMSE



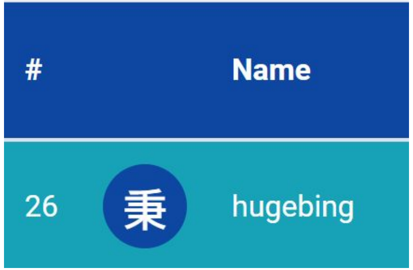
K_fold	Decision_Tree	Random_Forest	Linear_Regression	XGbroost
5	1077.46413951898	1071.29311435776	1073.2443445258	1077.62239518269
6	1077.1344650447	1071.12655842728	1073.59116898707	1078.27178609537
7	1077.14547957488	1071.34912143392	1073.22723985423	1077.99311882657
8	1076.89412708332	1070.65932563204	1073.39534198406	1077.55870129798
9	1077.12432841435	1071.26071666094	1073.54238450253	1077.61389299975
10	1077.05102579142	1071.23977941237	1072.84508213554	1077.7869157259



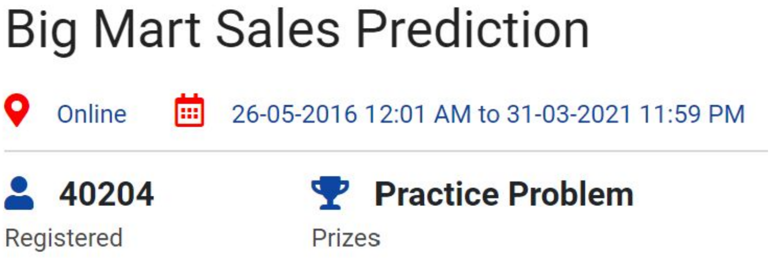
Result

排名26/40204

百分比0.065%



#		Name
26	秉	hugebing

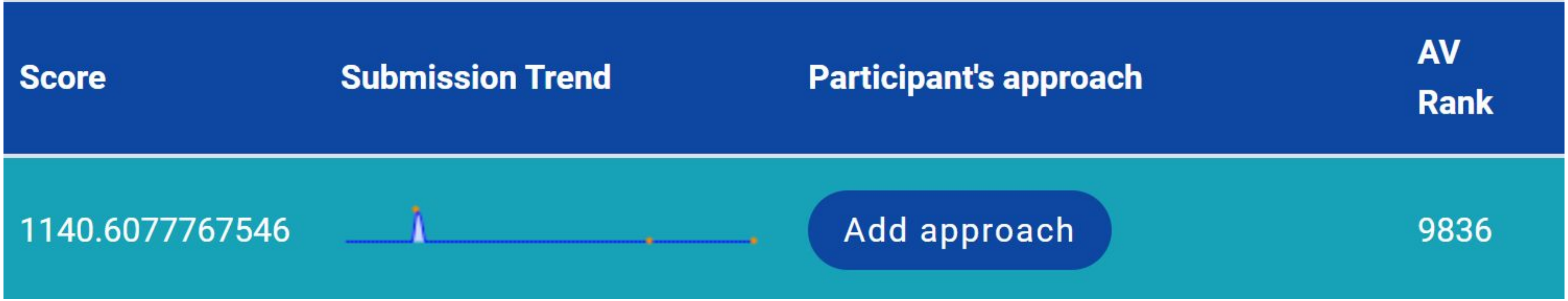


Big Mart Sales Prediction

Online 26-05-2016 12:01 AM to 31-03-2021 11:59 PM

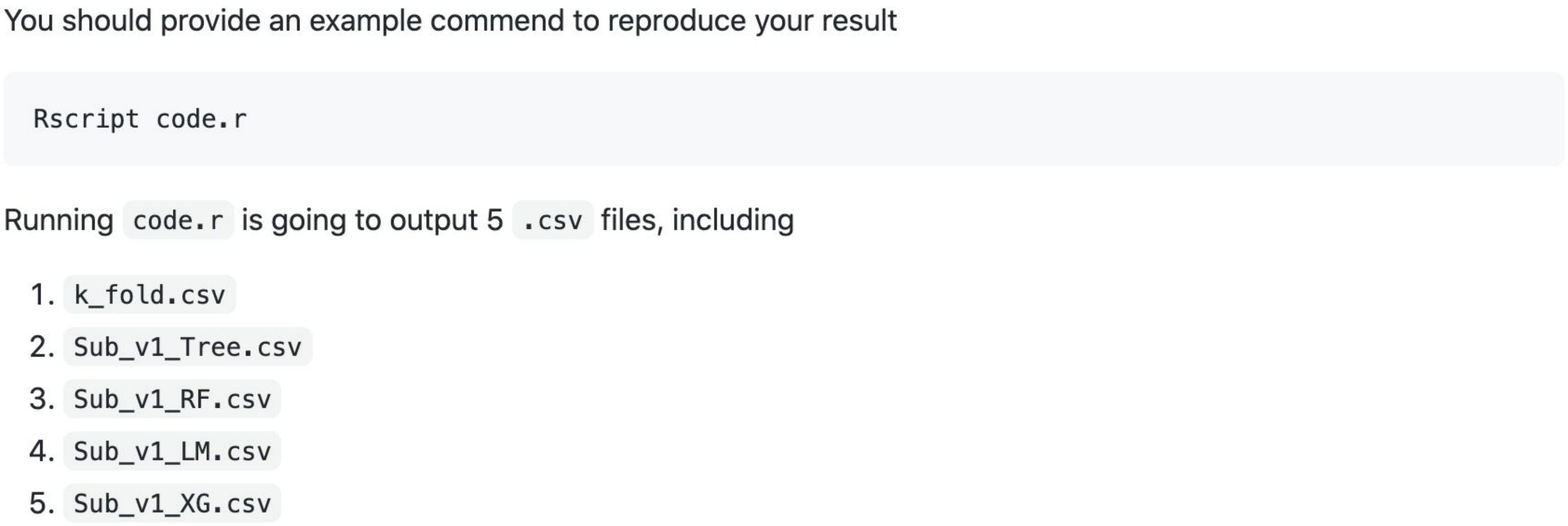
40204 Registered

Practice Problem Prizes



Score	Submission Trend	Participant's approach	AV Rank
1140.6077767546		Add approach	9836

Demo



You should provide an example commend to reproduce your result

```
Rscript code.r
```

Running `code.r` is going to output 5 `.csv` files, including

1. `k_fold.csv`
2. `Sub_v1_Tree.csv`
3. `Sub_v1_RF.csv`
4. `Sub_v1_LM.csv`
5. `Sub_v1_XG.csv`

Reference

- https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/
- 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
- https://github.com/Param-Trivedi/Big-Mart-Sales-Data-Prediction