Sales Prediction for Big Mart Outlets

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and predict the sales of each product at a particular outlet.

Using this model, BigMart will try to understand the properties of products and outlets which play a key role in increasing sales.

Please note that the data may have missing values as some stores might not report all the data due to technical glitches. Hence, it will be required to treat them accordingly.

**Data Dictionary**

We have train (8523) and test (5681) data set, train data set has both input and output variable(s). You need to predict the sales for test data set.

**Train file:**CSVcontaining the item outlet information with sales value

|  |  |
| --- | --- |
| **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. |

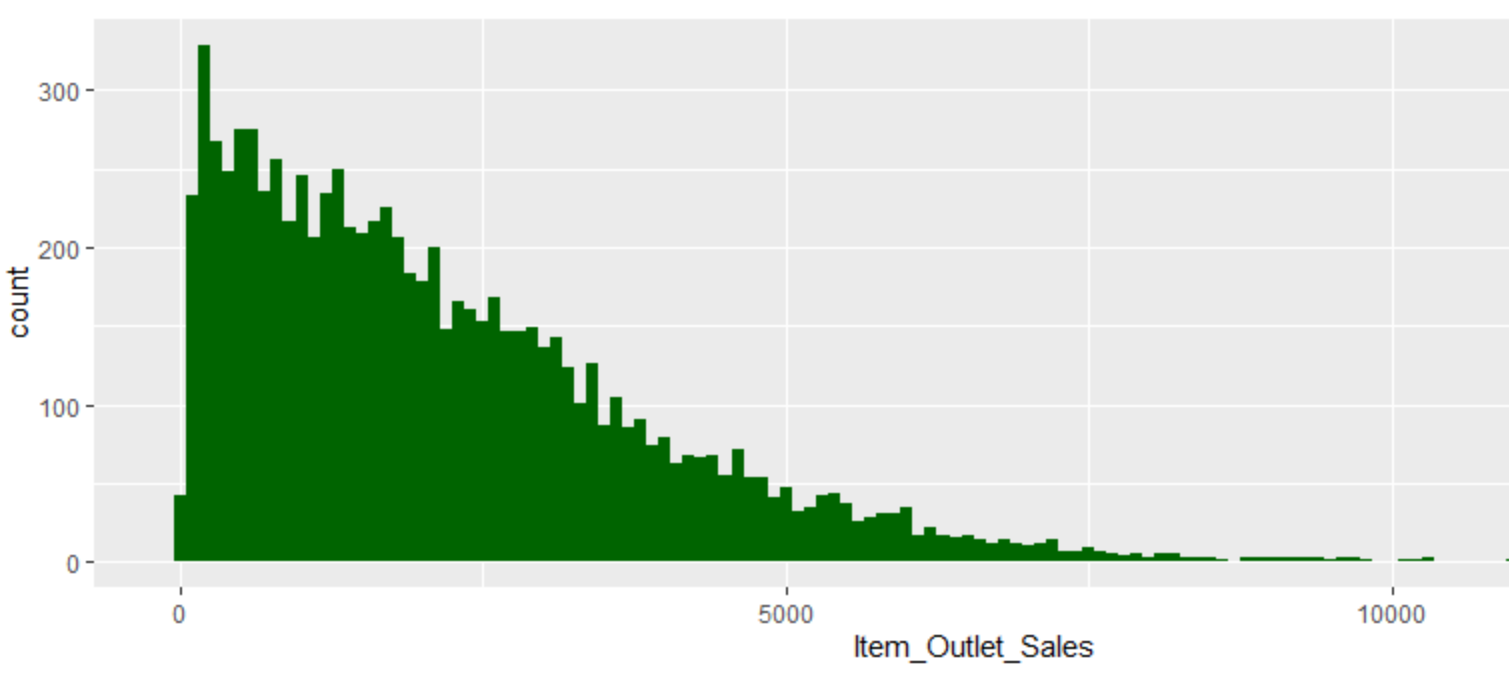
**Test file:**CSV containing item outlet combinations for which sales need to be forecasted

|  |  |
| --- | --- |
| **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 |

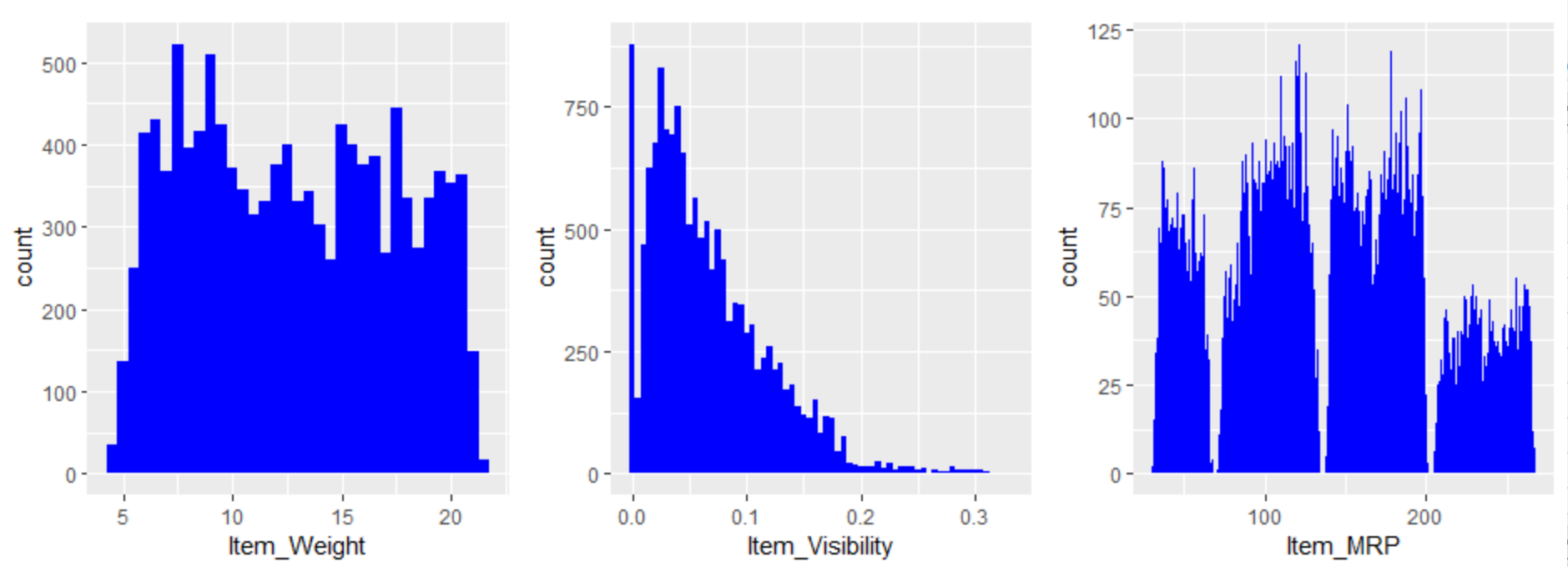
Steps:

* Read test , train, submission data set separately
* Analyzed the variables in test and train, train had output\_sales(we want to predict for test) but test had only 11 columns
* Combined both datasets
* Performed univariate exploratory analysis

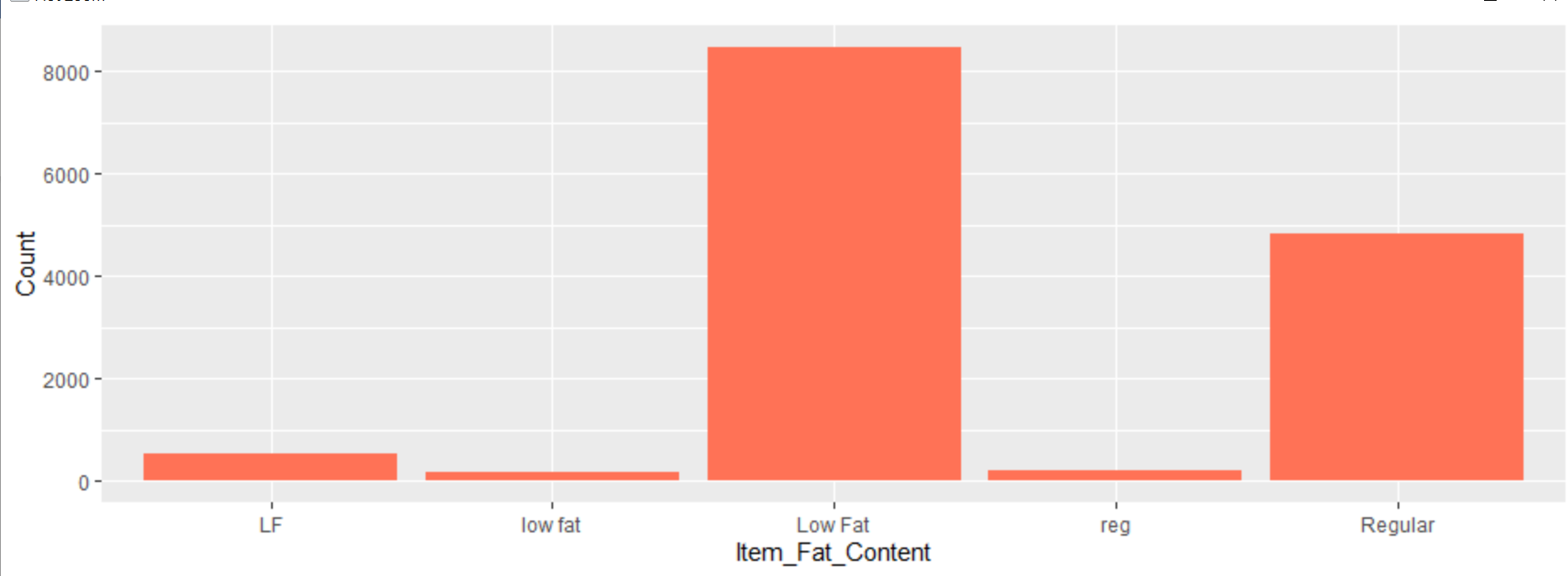
#target variable sales is continous so histogram

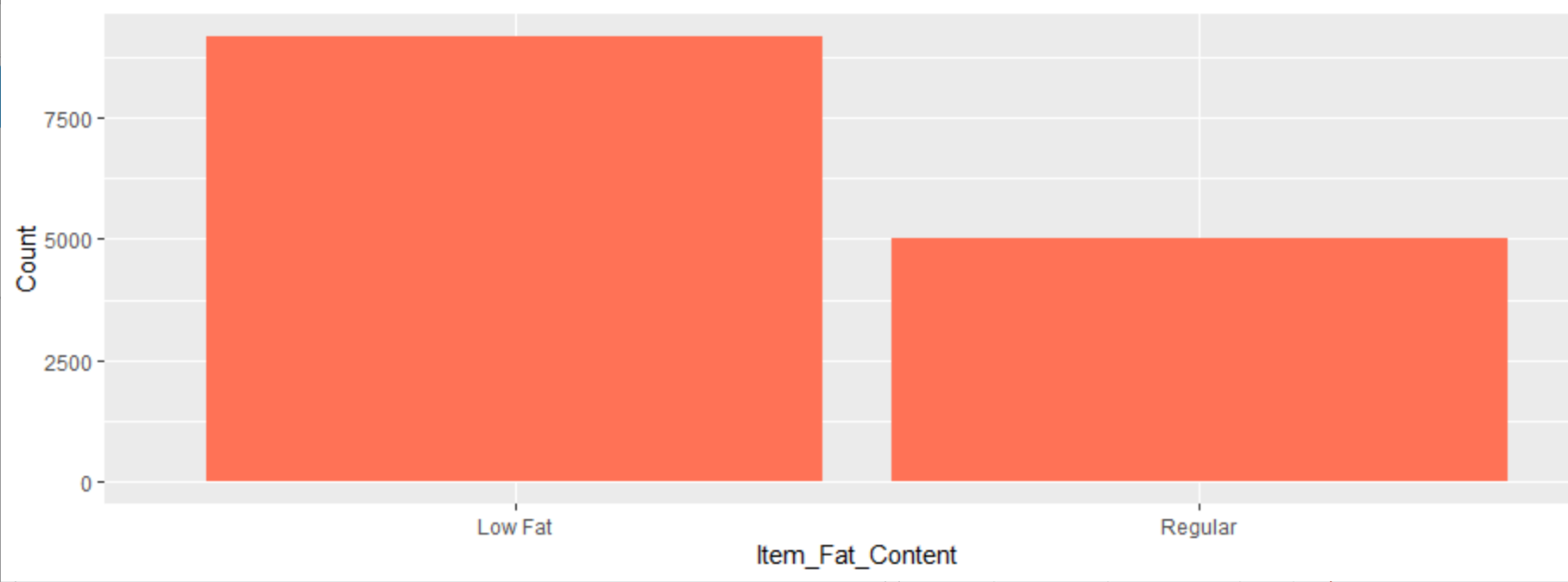


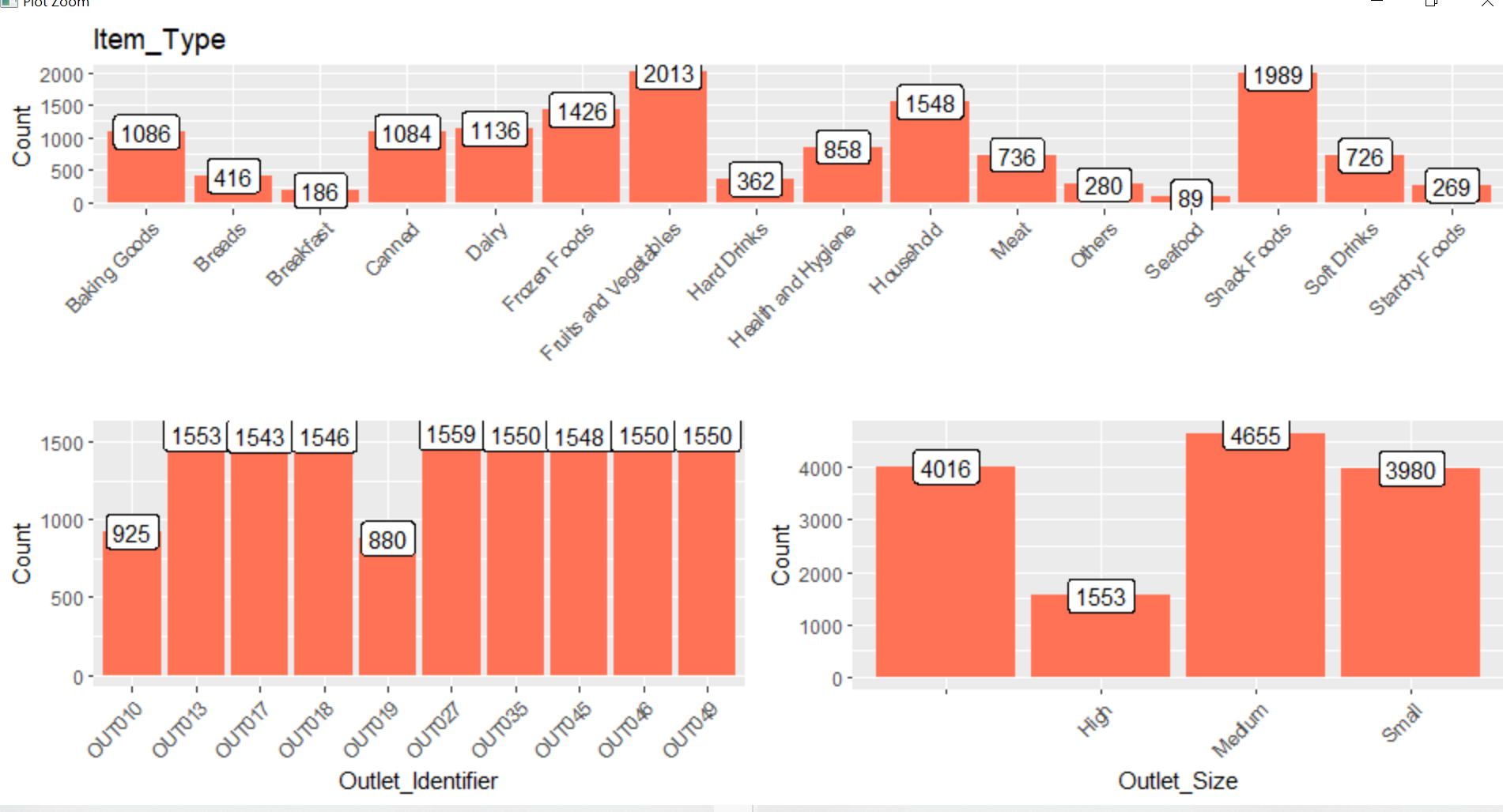
#histogram for independent numeric variables



#barplot for categorical variable

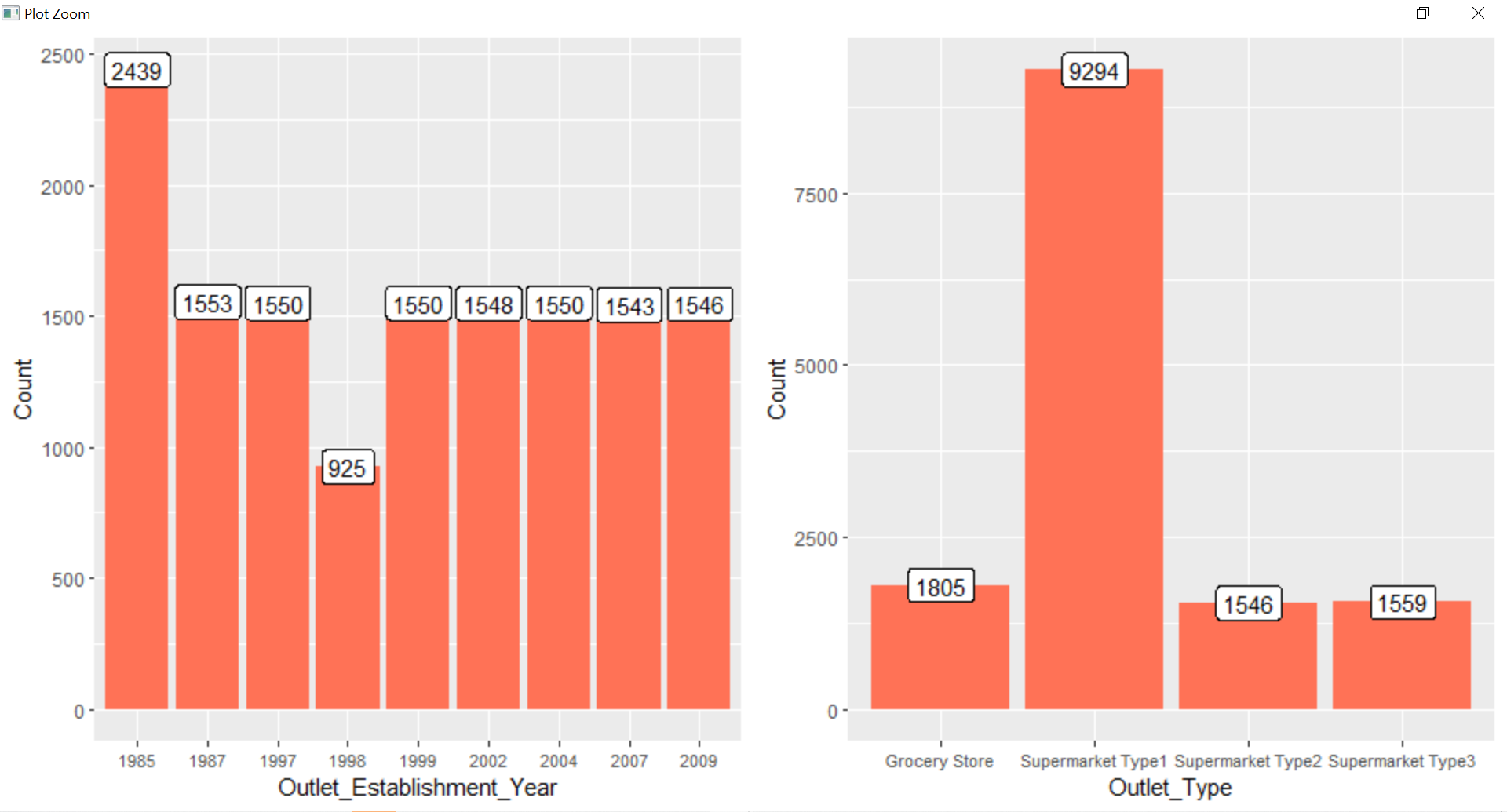






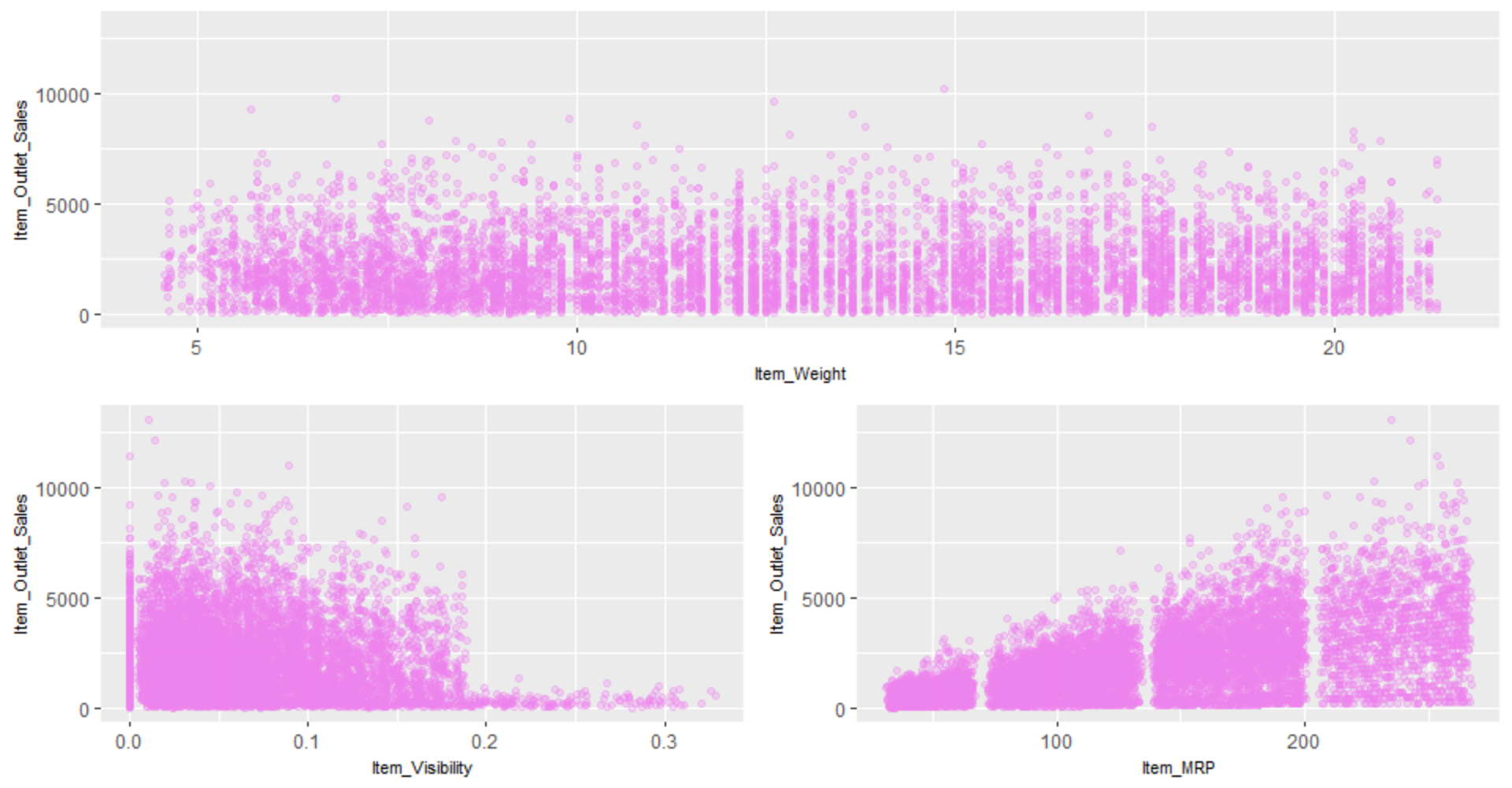
**Observations:**

In Outlet\_Size’s plot, for 4016 observations, Outlet\_Size is blank or missing.We will check for this in the bivariate analysis to substitute the missing values in the Outlet\_Size.

**Observations**

* Lesser number of observations in the data for the outlets established in the year 1998 as compared to the other years.
* Supermarket Type 1 seems to be the most popular category of Outlet\_Type.

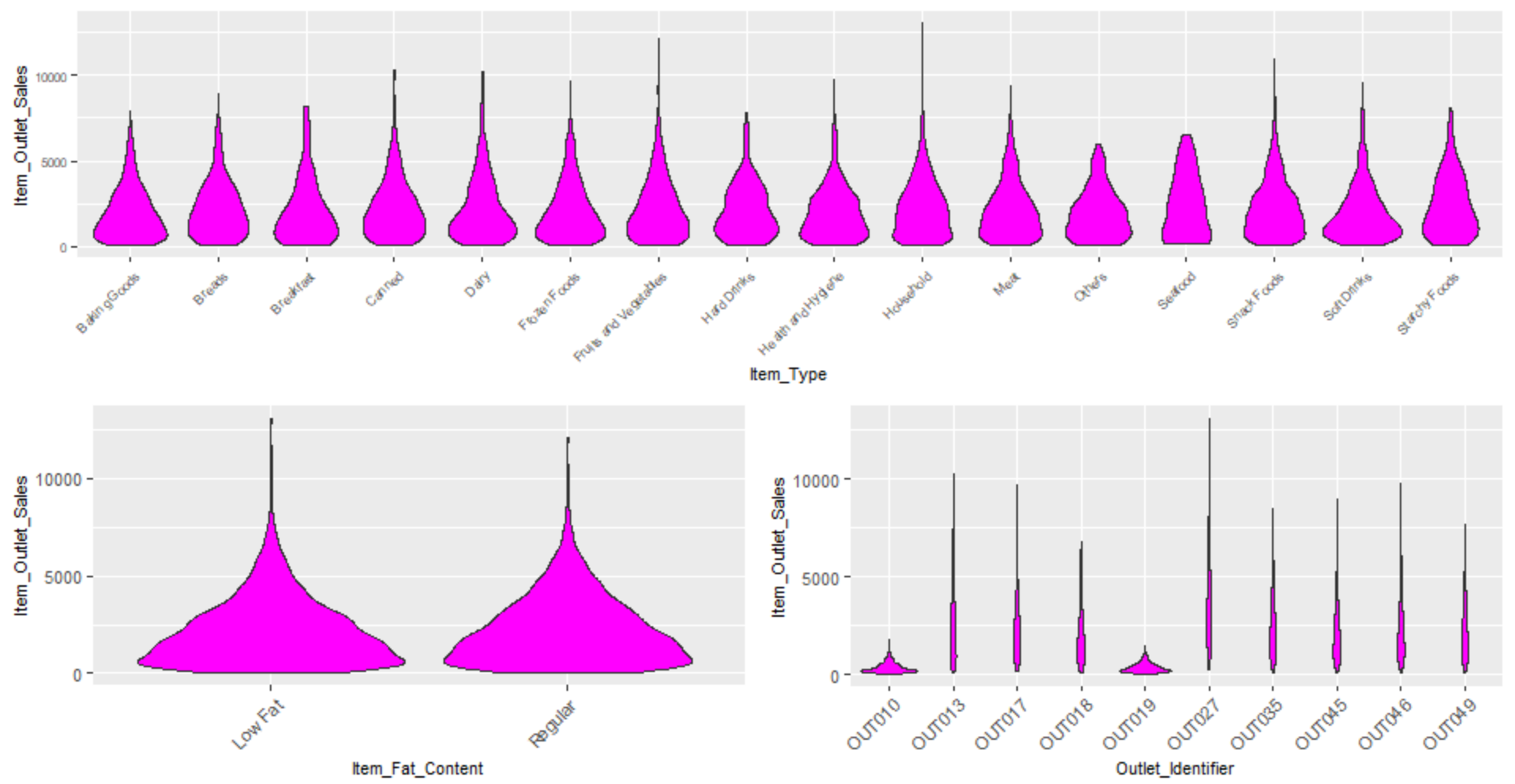
#Target Variable vs Independent Numerical Variables



**Observations**

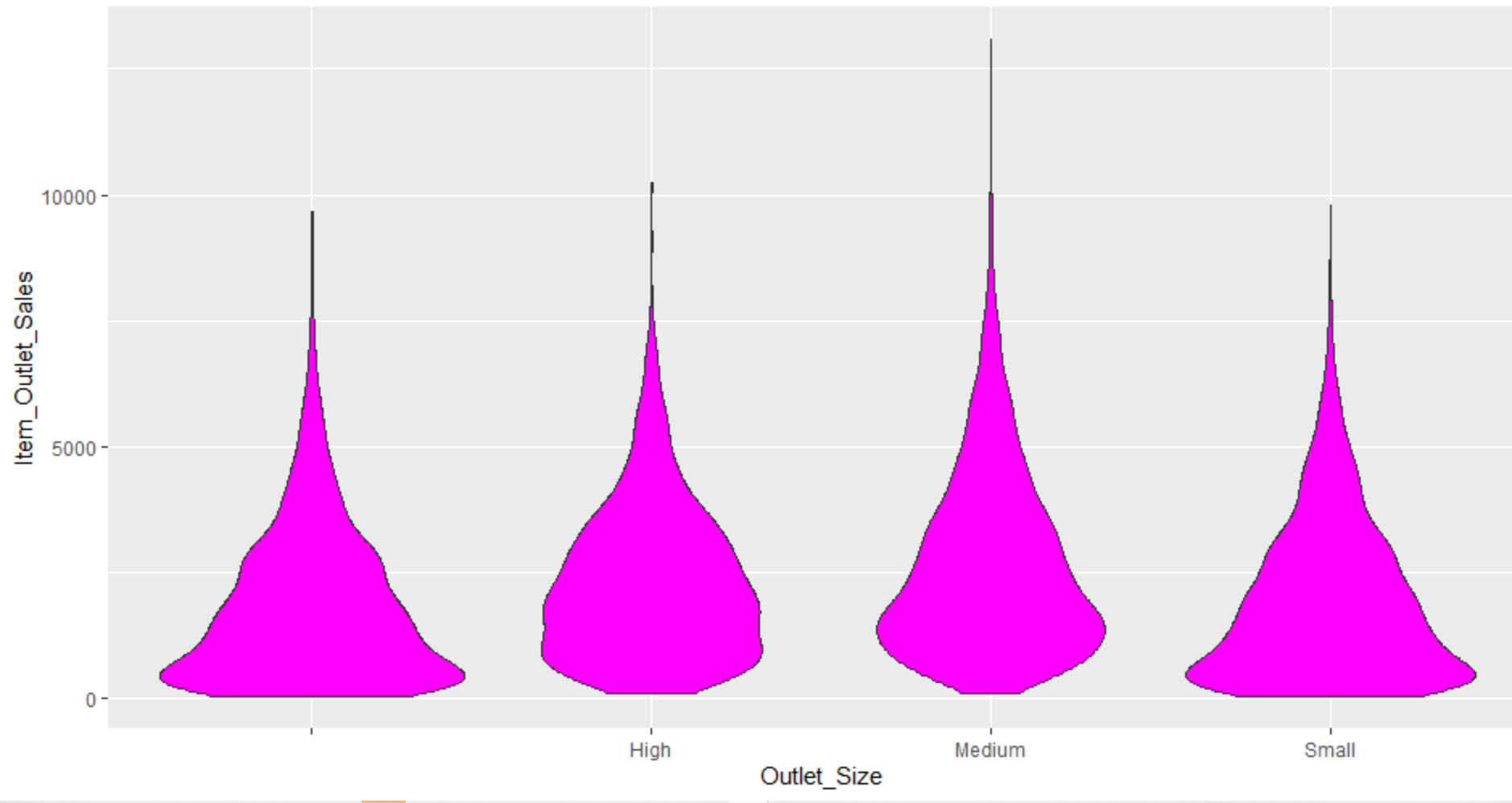
* Item\_Outlet\_Sales is spread well across the entire range of the Item\_Weight without any obvious pattern.
* In Item\_Visibility vs Item\_Outlet\_Sales, there is a string of points at Item\_Visibility = 0.0 which seems strange as item visibility cannot be completely zero.
* In the third plot of Item\_MRP vs Item\_Outlet\_Sales, we can clearly see 4 segments of prices that can be used in feature engineering to create a new variable.

#Target Variable vs Independent Categorical Variables

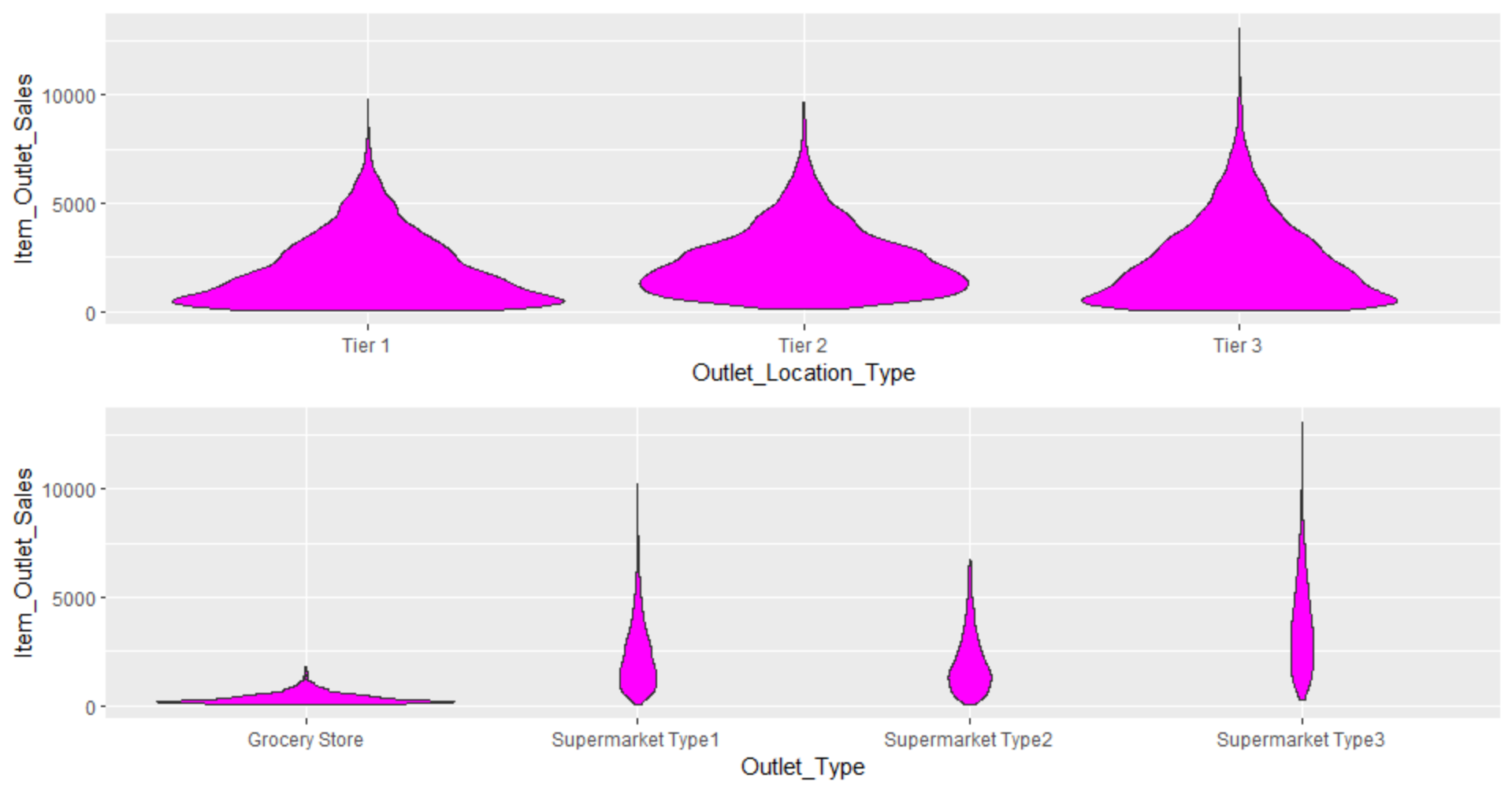


**Observations**

* Distribution of Item\_Outlet\_Sales across the categories of Item\_Type is not very distinct and same is the case with Item\_Fat\_Content.
* The distribution for OUT010 and OUT019 categories of Outlet\_Identifier are quite similar and very much different from the rest of the categories of Outlet\_Identifier.



The distribution of ‘Small’ Outlet\_Size is almost identical to the distribution of the blank category (first vioin) of Outlet\_Size. So, we can substitute the blanks in Outlet\_Size with ‘Small’.



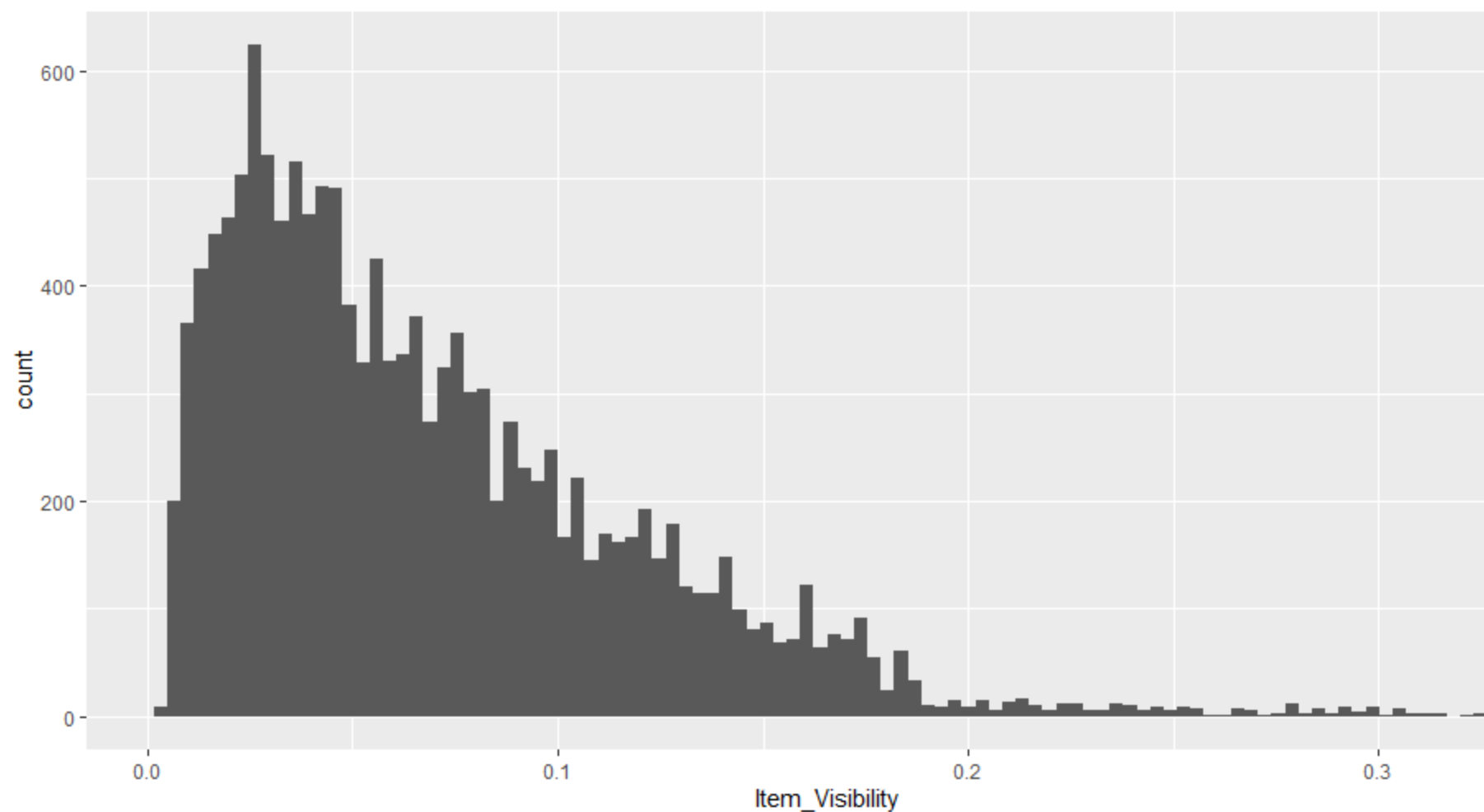
**Observations**

* Tier 1 and Tier 3 locations of Outlet\_Location\_Type look similar.
* In the Outlet\_Type plot, Grocery Store has most of its data points around the lower sales values as compared to the other categories.

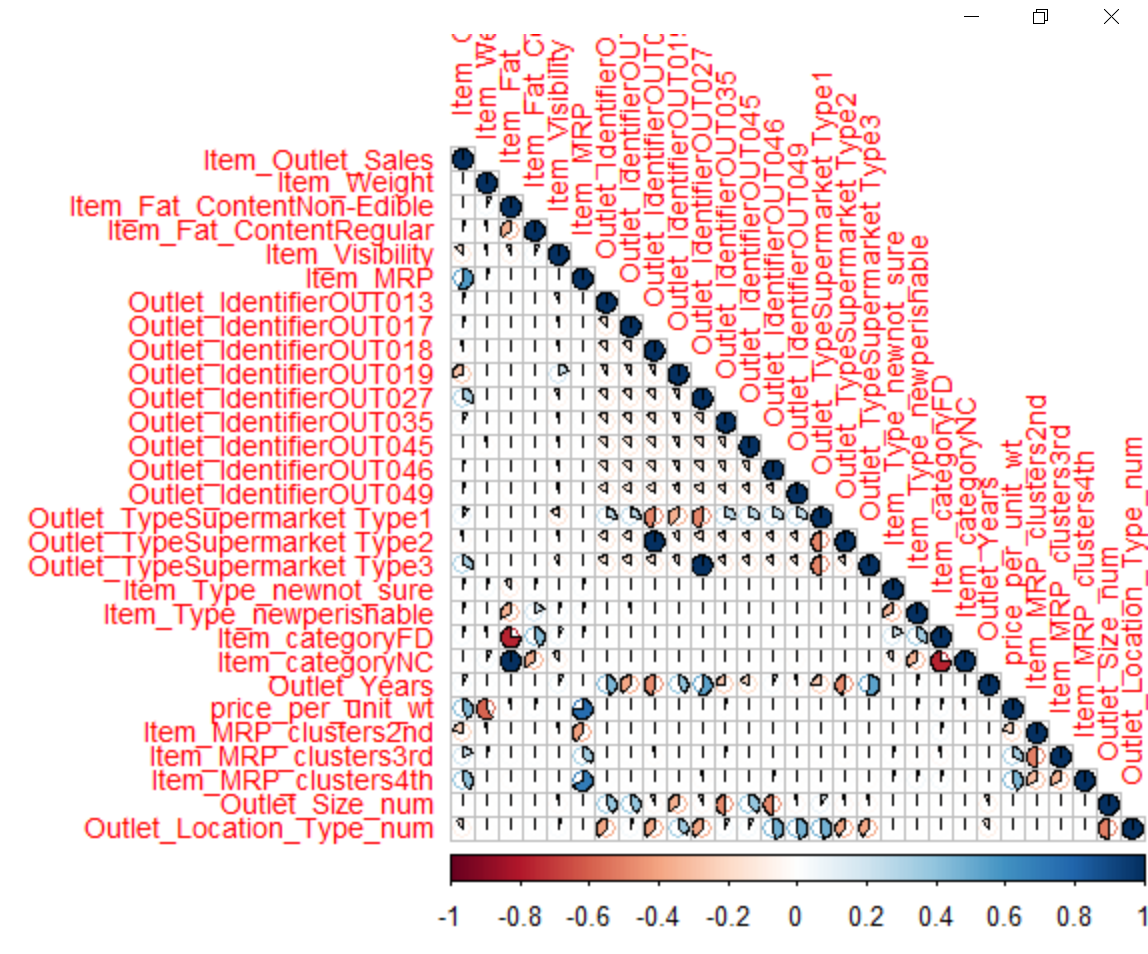
**Imputing Missing Value**

impute Item\_Weight with mean weight based on the Item\_Identifier variable

Replacing 0 in item\_visibility



Correlated variables



As Amee mentioned we are watched every single sec. [Today we all give away our data all day long while aiming to maintaining our privacy.](https://www.quotemaster.org/q9b85fb14e850f140989a694b5b44cacb) which is shown in this spyderware dashboard of Redmorph extension As many of us are unaware of these malicious practices Thus the main objective of this project was to create user awareness about cyber security and privacy issues. A lot of such cybercrimes take place with students as they spend much of their time on web and social media. So we focused on educating the student population through different means to help in protecting their data and privacy. There is no question that our property, personal belongings are precious to us but so should be our digital life and it should have the same degree of security and protection. Even the law bodies like US federal and NIST are taking it as a serious crime and embedding laws related to cybersecurity and user rights to his/ her privacy along with providing guidelines for the users on how can they protect their device and data.But it is very important for us as individual to take up responsibility and ensure our safety and think before you click anything on internet. Now Arzoo will take it ahead from here