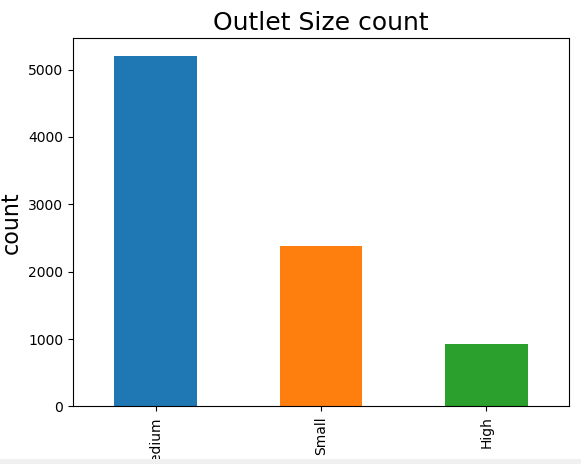
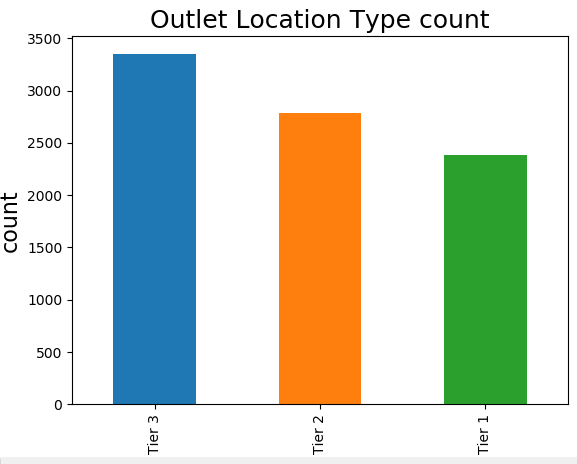
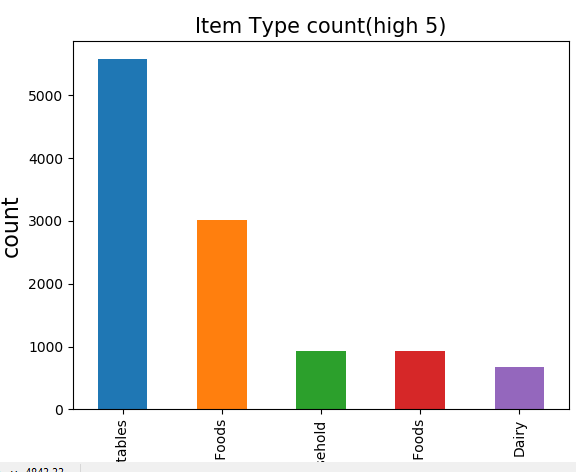
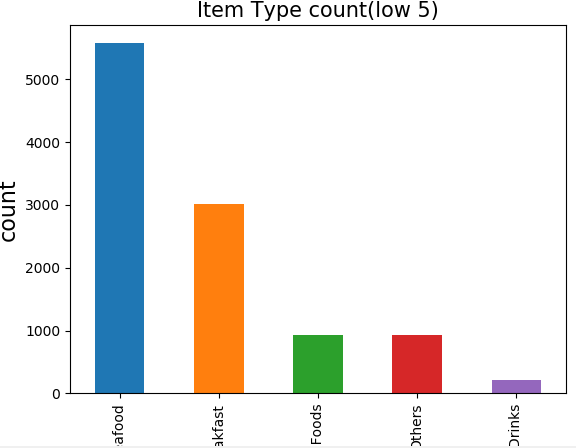
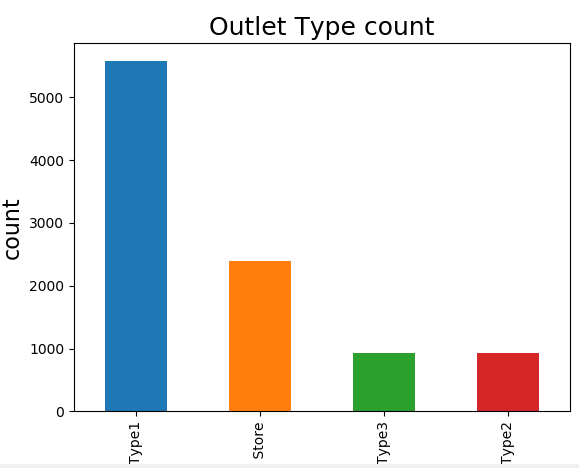
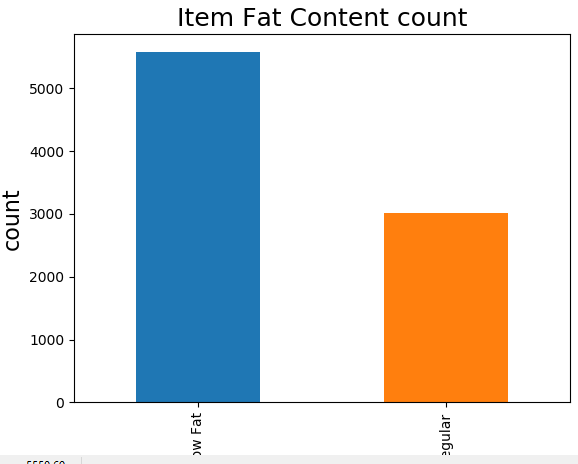
**Regression (store dataset)**

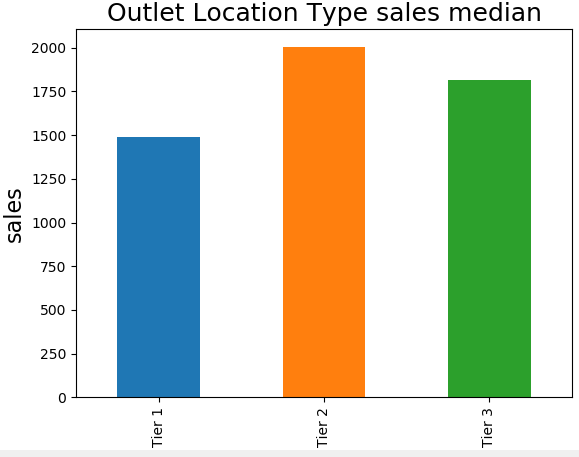
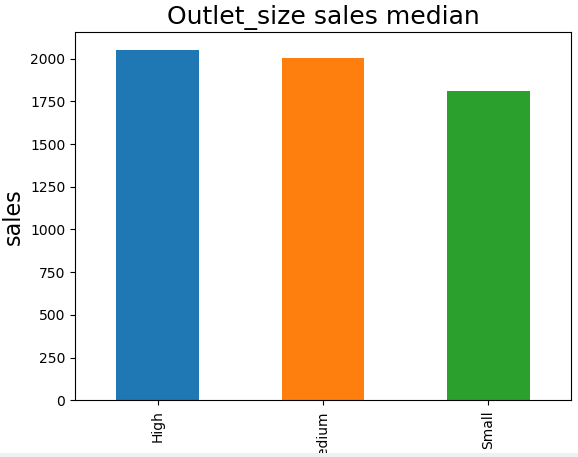
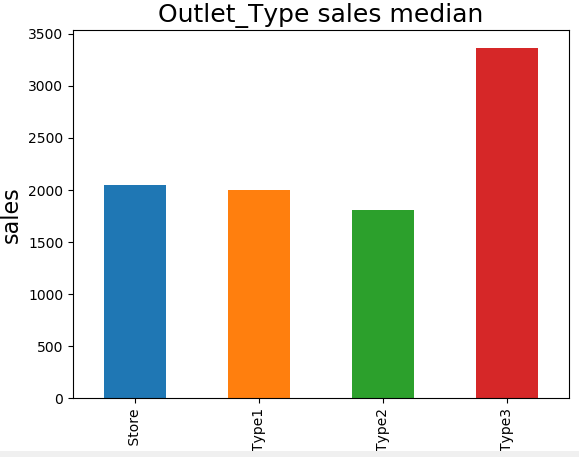
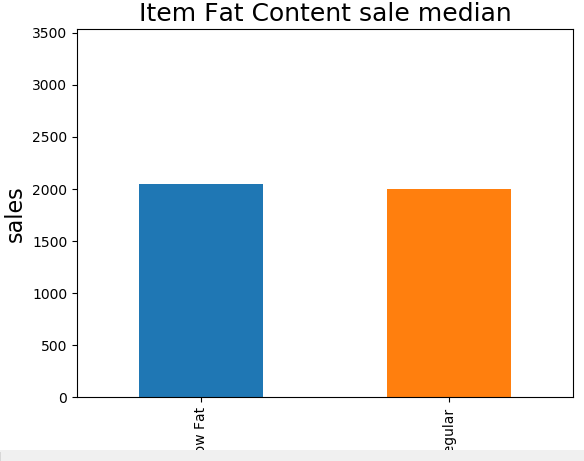
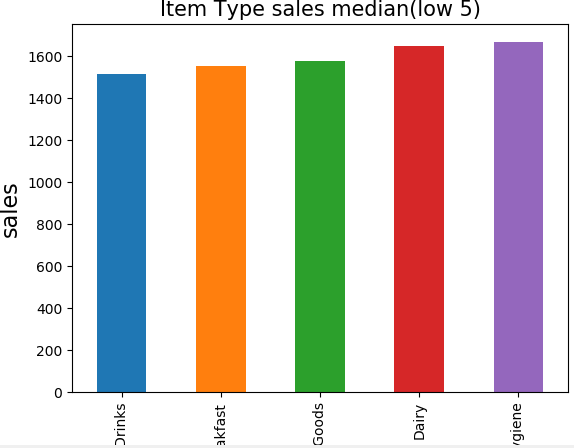
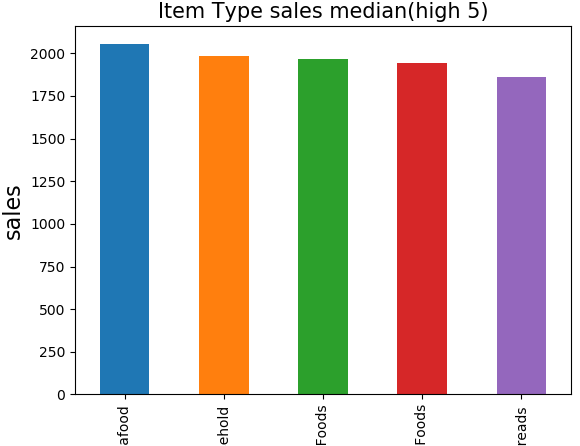
The dataset- Store\_Train\_Forecast\_Class contains 8523 observations, including seemingly useful features: (**Item Weight,** **Item Fat Content, Item Visibility, Item Type, Item MRP, Outlet Size, Outlet Location Type, Outlet Type, Item Outlet Sales** (response variable)), and seemingly useless features (**Item Identifier, Outlet Identifier, Outlet Establishment Year**) that are dropped in the later analysis.

In the preprocessing stage, 2 variables (Item Weight, Outlet Size) contained missing values, and they were then fixed by filling the nulls of Item Weight with mean (12.8576451841), and filling the nulls of Outlet Size with the value ‘Medium’. Further more, the variable - Item Fat Content contained some messy values (eg: ‘low fat’, ‘Low Fat’ and ‘LF’ are the same thing), and this is fixed simply by bring them into conformity.

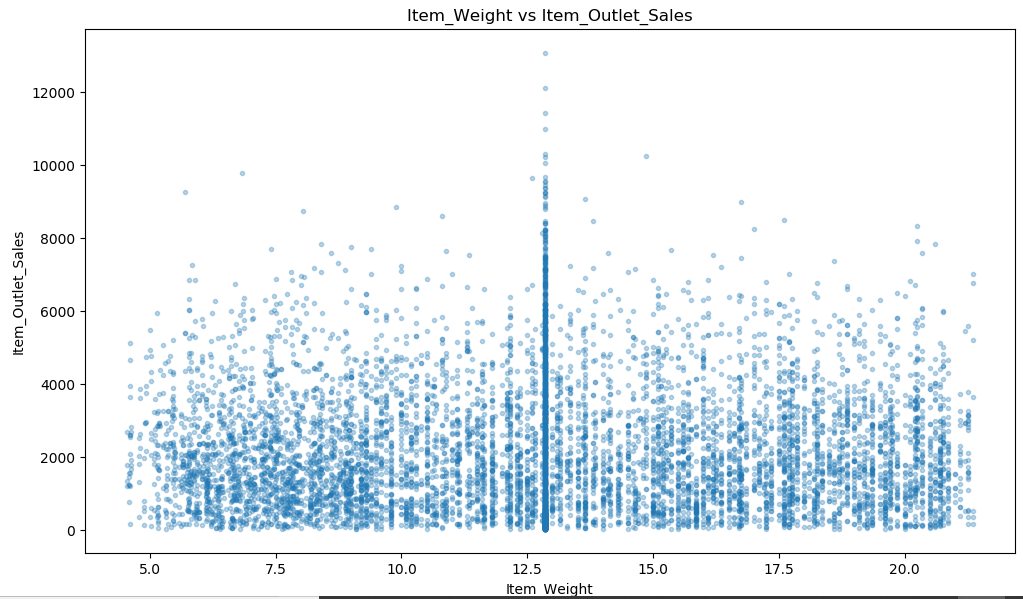
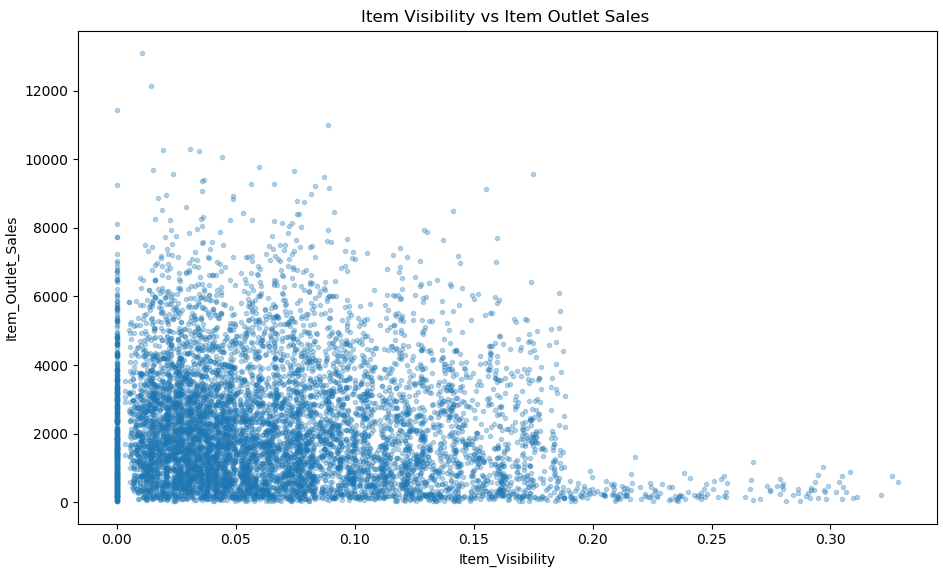
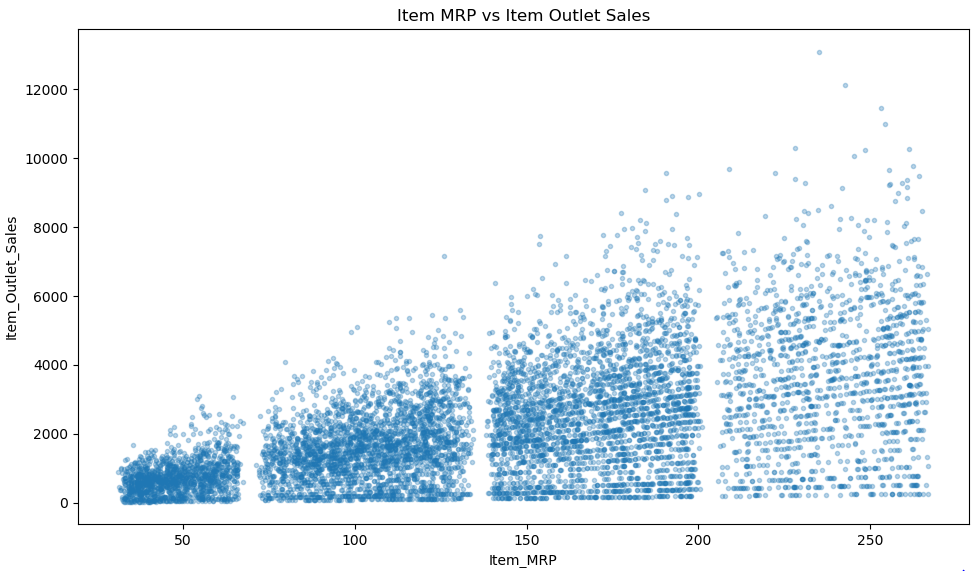
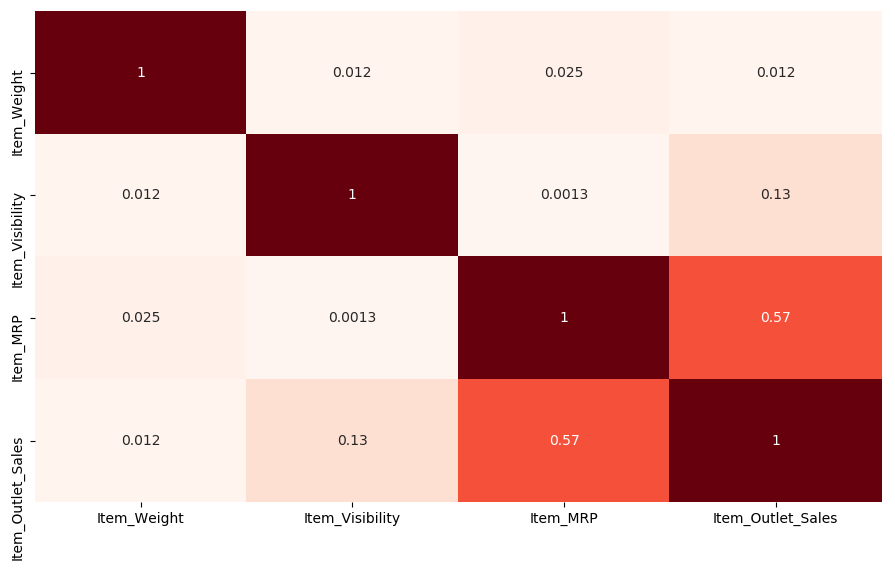
**Plot section 1: (categorical variables vs count)**

From the above bar plots, we can get a sense of the consumers’ behaviors, it seems that consumers prefer more shopping at Tier 3 location and type 1 outlet type, and buy medium sized, and low-fat products, specifically items like Fruits and Vegetables and Snack Foods.

**Plot section 2: (categorical variables vs outlet sales median)**

From the above bar plots, we can observe that, tier 2 and type 3 outlet would generate more sales median. In terms of the products per se, items’ fat content and size don’t make much difference in sales median. Furthermore, items like seafood, Household and Starchy Foods are relatively more expensive.

**Plot section 3: (numerical variables vs outlet sales)**

From the above scatter plots, we can observe that, although not so obvious, outlet sales and item MRP have a positive association, and the correlation matrix actually proved it- correlation coefficient (p)=0.57. On the other hand, item visibility and weight, respectively, are not really associated with generating more outlet sales.

Before modeling, dummy variables are created for all the categorical variables, and then using feature selection, some insignificant variables and dummy variables are dropped to generate the highest possible R2 score. (variable: **Item MRP**, and dummy variables: **Item Fat Content Regular, Outlet Type Supermarket Type1, Outlet Type Supermarket Type2, Outlet Type Supermarket Type3, Item Type Fruits and Vegetables** and **Item Type Seafood** are kept for later regression analysis). At the end, standard scaling was applied, which marks the last step of preprocessing.

In the modelling section, **linear regression, KNN, SVM** and **random forest** were applied to the training sets and test sets. As a result (see the table below), random forest is the best algorithm for the prediction of this model. Relatively, it gives the highest adjusted-R2 score of 0.59, and the lowest RMSE value of 1099 (low prediction error). According to the overall R2 score and RMSE value, we can conclude that the independent/ predictive variables used for regression modelling are not bad (moderate) predictors to predict item outlet sales.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear regression | KNN | SVM | ***random forest*** |
| R2 | 0.5616 | 0.5780 | 0.3546 | ***0.5971*** |
| Adjusted R2 | 0.5562 | 0.5728 | 0.3467 | ***0.5921*** |
| RMSE | 1146 | 1124 | 1391 | ***1099*** |