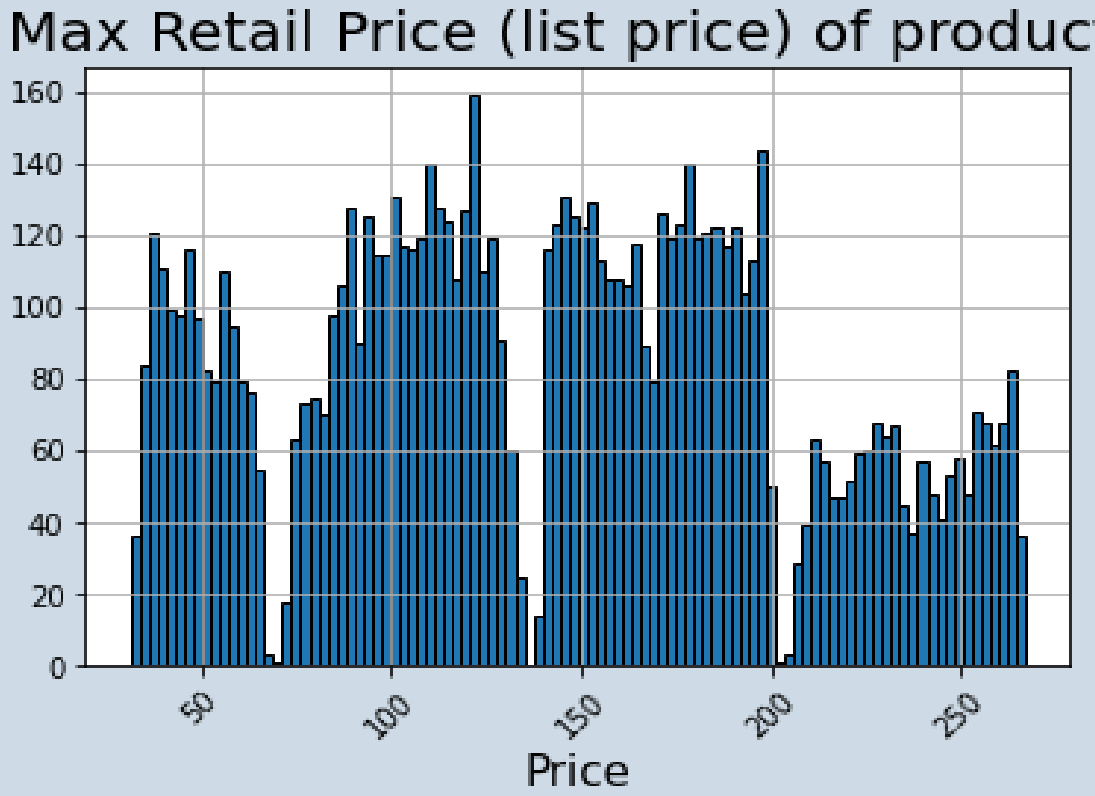
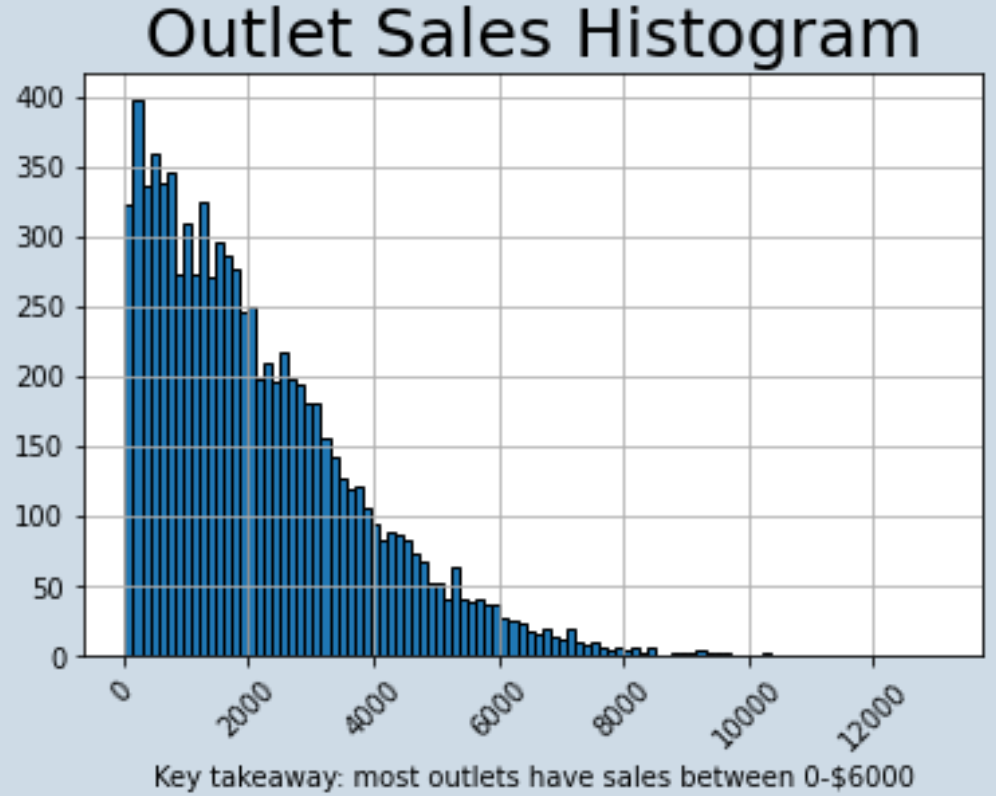
Sales Predictions Project

Irina Nizhnik

Important findings from initial data exploration (1/3)

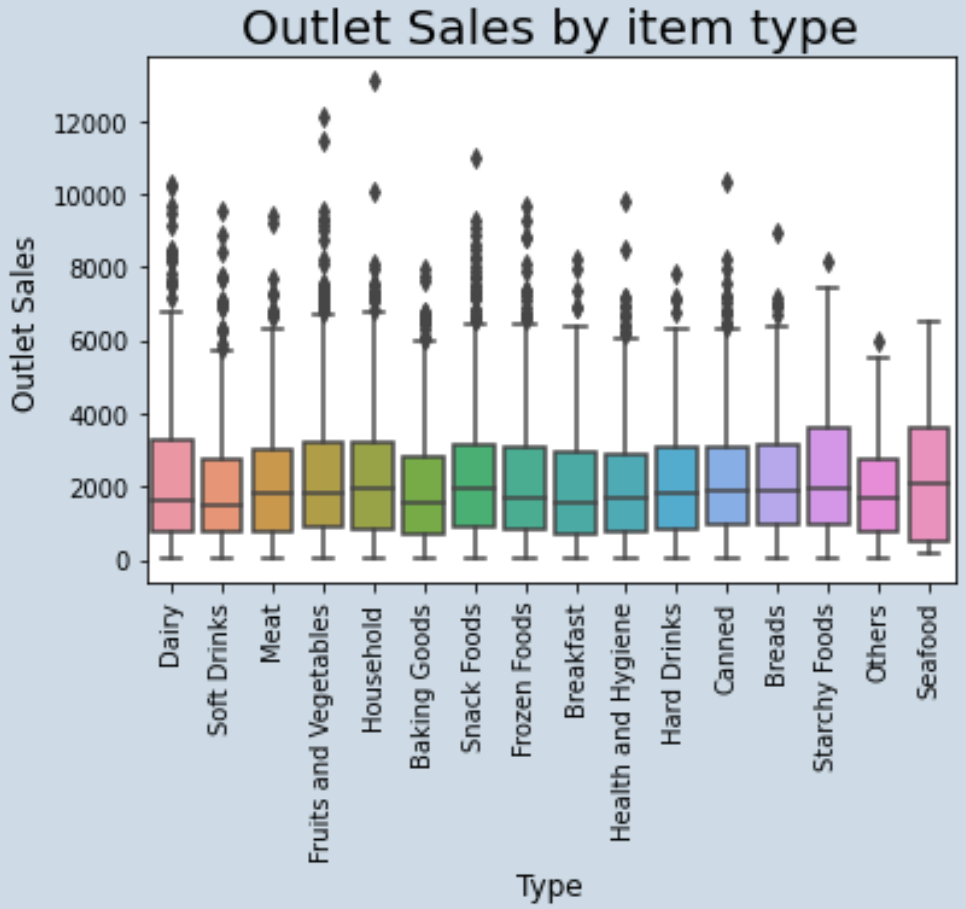


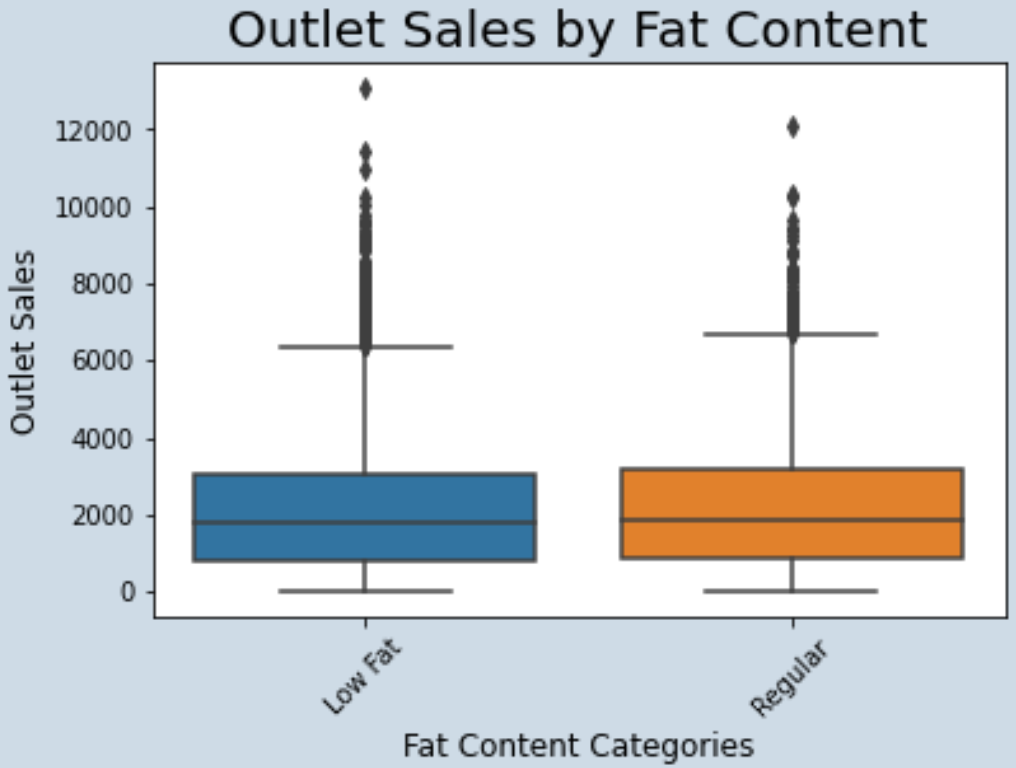


The list price ranges from about $25-$275 with a notable concentration in the $25-$75, $90-$120, and $150-$200 ranges.

As indicated in the graph, most outlets in the dataset have sales between $0-$8,000, with the most significant concentration in <$2,000 sales

Important findings from initial data exploration (2/3)

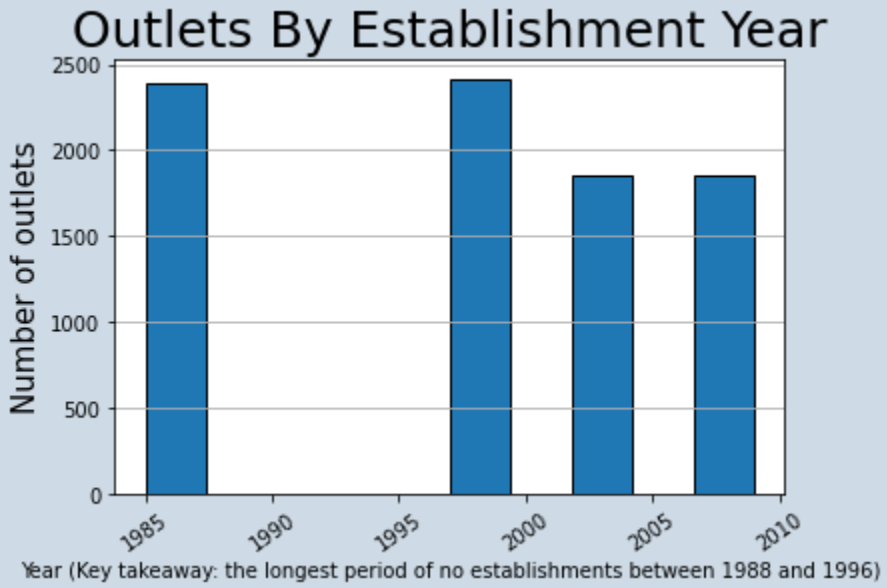


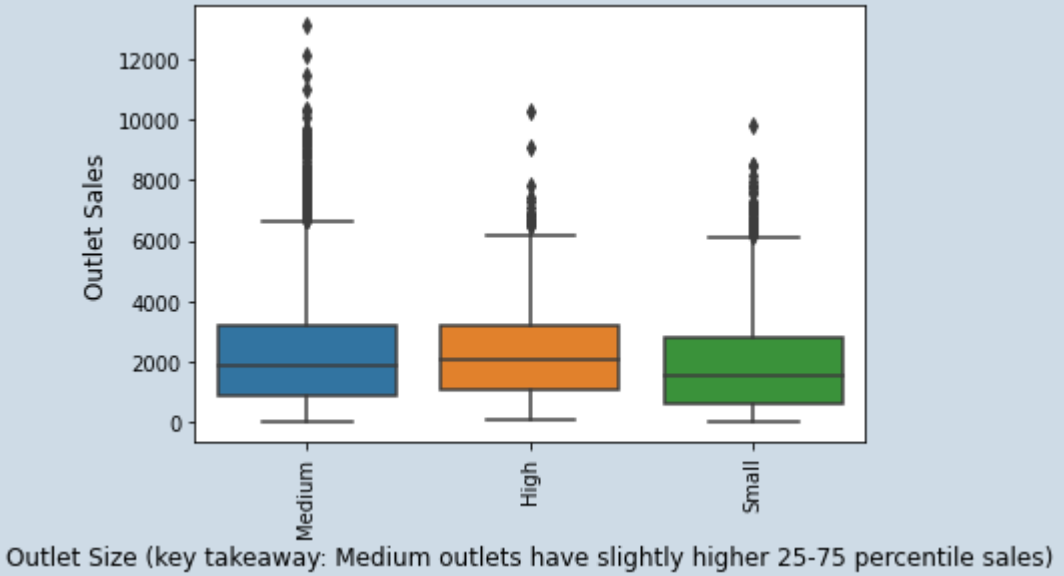


There are several notable item types that are instrumental in driving sales in outlets: seafood, starchy foods, frozen foods, snack foods, household, fruits and veg, and dairy

The fat content of low fat vs regular fat do not appear to have significate impact on the sales differentiation.

Important findings from initial data exploration (3/3)





A relatively large number of outlet establishments have been during 1997-2009 with zero establishments tween 1988 and 1996.

The size of outlets appear to potentially play a somewhat slight impact on sales. Overall, medium and high size establishments have about 25-75% higher sales than small establishments.

We used 4 models to predict the variability in outlet sales with varying degrees of explainability

	OLS Regression	Regular Decision Tree	Bagged Trees	Random Forest
R^2 in training	56.1%	61%	92%	62%
R^2 in test	56.7%	59%	52%	60%
Additional details	Training RMSE: 1139.1 Testing RMSE: 1092.8	Best result achieved with 5 depth	Best result achieved with 5 max depth	Best result achieved with 5 max depth and n-estimators of 200

Training and Testing Dataset

Several models were run to train and test the dataset: 1) regression 2) decision tree 3) bagged tree 4) random forest tree

The dataset was split into 75% for training and 25% was testing

Outlet sales in the dataset was the selected as a target vector that was tried to be predicted for the analysis.

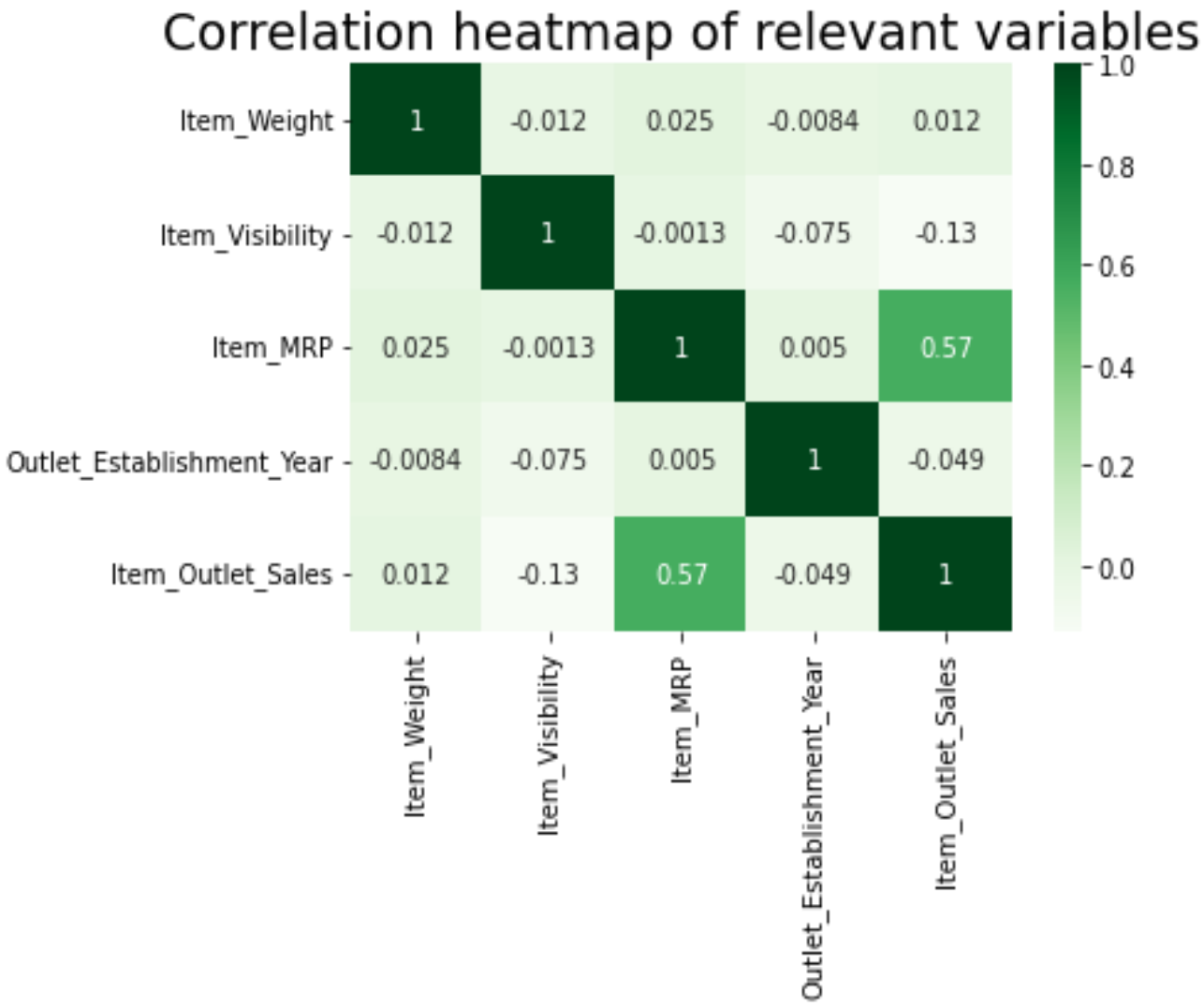
First, in the regression model, the r2 of 56% also known as the coefficient of the variation in the outlet sales or y can be explained by the features. Furthermore, on average, the model illustrated to be incorrect by about 1139 dollars in trained set and 1092 dollars in tested.

Second, the decision trees model by default was outfitting and needed several tries to find the best score in testing. It improved with tuning max depth to 5 where the dataset trained data performed its best in tested dataset (i.e., 0.61 trained and 0.59 tested).

Third, by introducing more models to improve the score in testing such as a bagged trees model, the output showed some improvement in training dataset predictability (0.91), but slightly reduced predictability in test dataset (0.51).

Finally, the random forest showed improvement in testing from 0.51 in bagged trees to 0.60. In this model, trained score was 0.62 and 0.60 in testing resulting in best output.

Appendix – Interesting Analyzed Observations from the Dataset



Key takeaway:

- Visibility does not play important role in driving the sale per outlet.
- However, as expected, a price of an product does play role in driving the overall sale for that item.