CarSeats Analysis

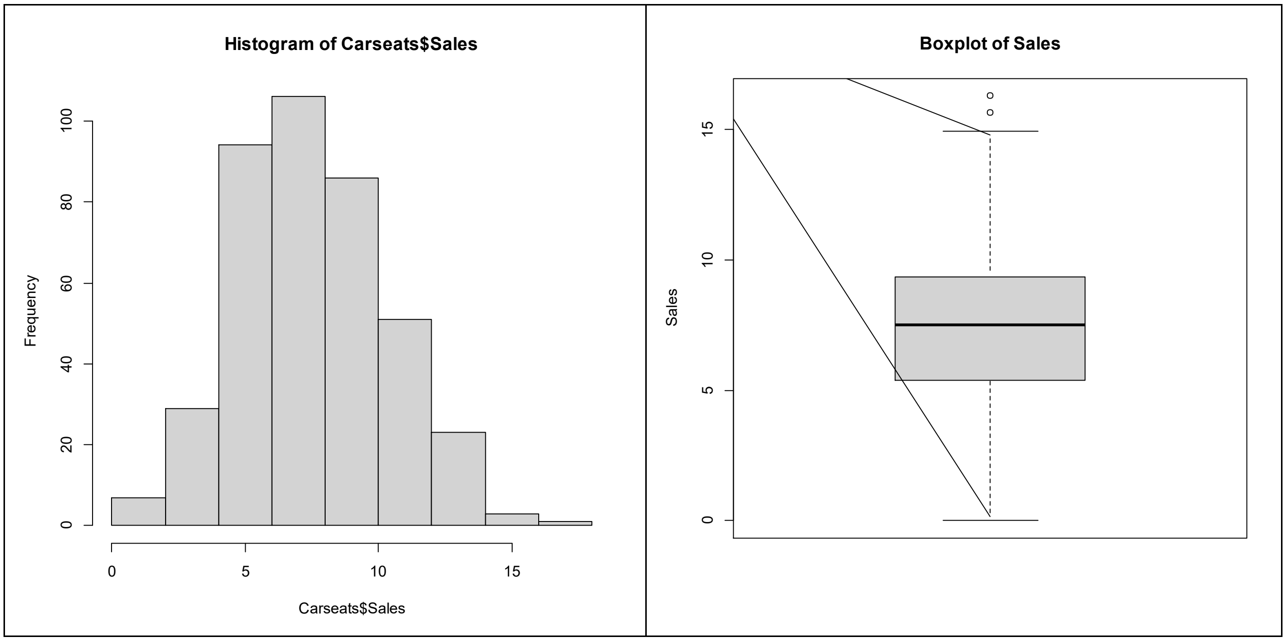
Exploratory Data Analysis

The *Carseats* dataset contains multiple continuous and categorical predictors describing store characteristics, pricing, demographics, and shelf placement quality. Since the same product type is sold across multiple stores, variation in **Sales** reflects both store-level and market-level factors, suggesting that multivariate analysis is appropriate. In this study, **Sales** is treated as the dependent variable. The Y variable for this analysis is **Sales**

Histograms of DV:

A visual inspection of the histogram for **Sales** indicates a slightly right-skewed distribution with no extreme outliers. The variable appears reasonably continuous and approximately bell-shaped, making it well-suited for regression modeling. The distribution does not follow a Poisson-like form, which supports using standard regression and tree-based methods rather than count-data models.

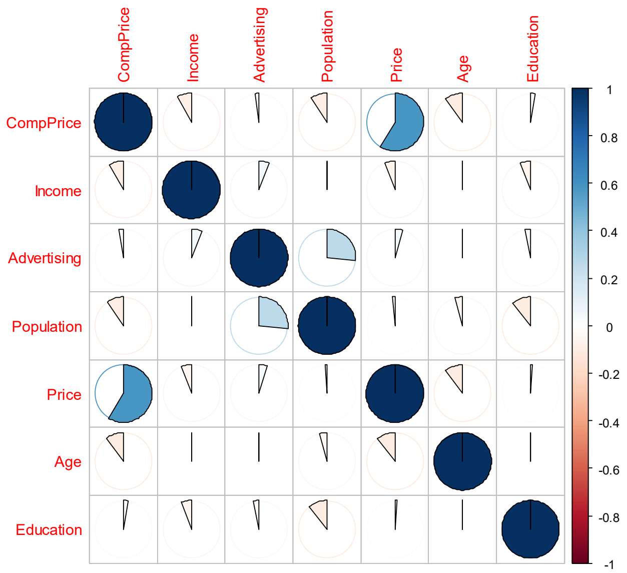
Given these characteristics, tree-based models such as **regression trees, bagging, random forests, and BART** are appropriate for capturing nonlinear relationships, interactions, and complex predictor effects. Further diagnostics will help determine which model offers the best predictive accuracy with minimal violations of model assumptions.



Correlations:

Item sales are Correlated to **Sales,** Auto correlation doesn't exists as the data isn't multilevel nature

```
correlation_matrix <- cor(Carseats[c("CompPrice", "Income","Advertising","Population","Price","Age","Education")])
corrplot(correlation_matrix, method = "pie")
```



The correlation analysis shows mostly weak to moderate relationships among the predictors, with no correlation exceeding 80%. This indicates that the variables are not highly collinear and each contributes unique information for modeling Sales.

Predictor table:

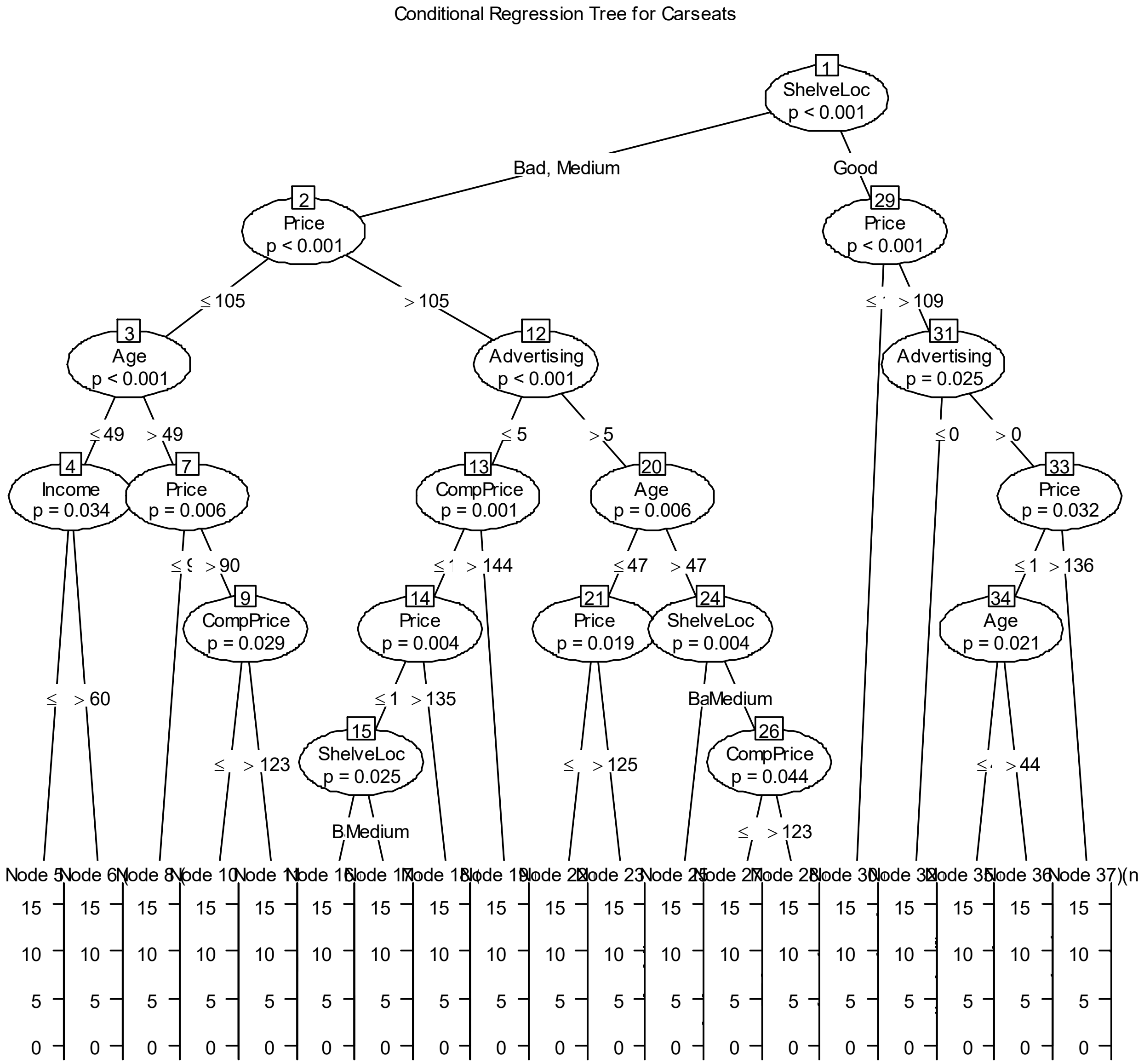
Predictor	Effect	Rationale
DV: sales		
CompPrice	-	Higher competitor prices can reduce price pressure on the store, but if competitors lower prices, store sales may decrease
Income	+	Higher-income markets have more purchasing power, often increasing sales
Advertising	+	More advertising generally increases awareness hence sales
Population	+	Larger populations may lead to higher customer traffic and increased sales opportunities
ShelveLoc	+	Better shelf placement (Good/Medium) increases visibility and accessibility, improving sales
Age	+/-	Older customers may purchase less frequently, but patterns can vary by location and store type
Education	+/-	Higher education levels can influence purchasing preferences, though direction depends on product type
Urban	+	Urban locations generally have higher customer traffic and more consistent sales volume
US	+	U.S. stores may differ in consumer behavior, marketing, and purchasing patterns, contributing to higher sales
Excluded: None		

Train and test

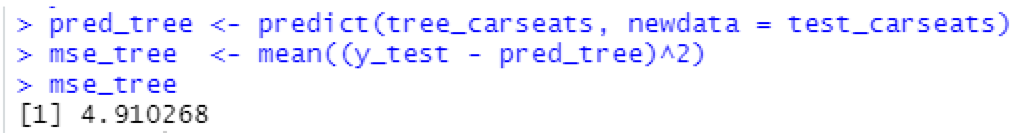
A train–test split is used to evaluate a model's ability to generalize to unseen data. The dataset is divided into a training set, which the model uses to learn underlying patterns, and a test set, which provides an unbiased assessment of predictive performance. Using a 75/25 split ensures the model has sufficient data to learn while still reserving a meaningful portion for validation. Setting a seed makes the split reproducible and consistent across runs.

Regression Tree

```
tree_carseats <- tree(Sales ~ ., data = train_carseats)
summary(tree_carseats)

plot(tree_carseats)
text(tree_carseats, pretty = 0)
```



Results MSE

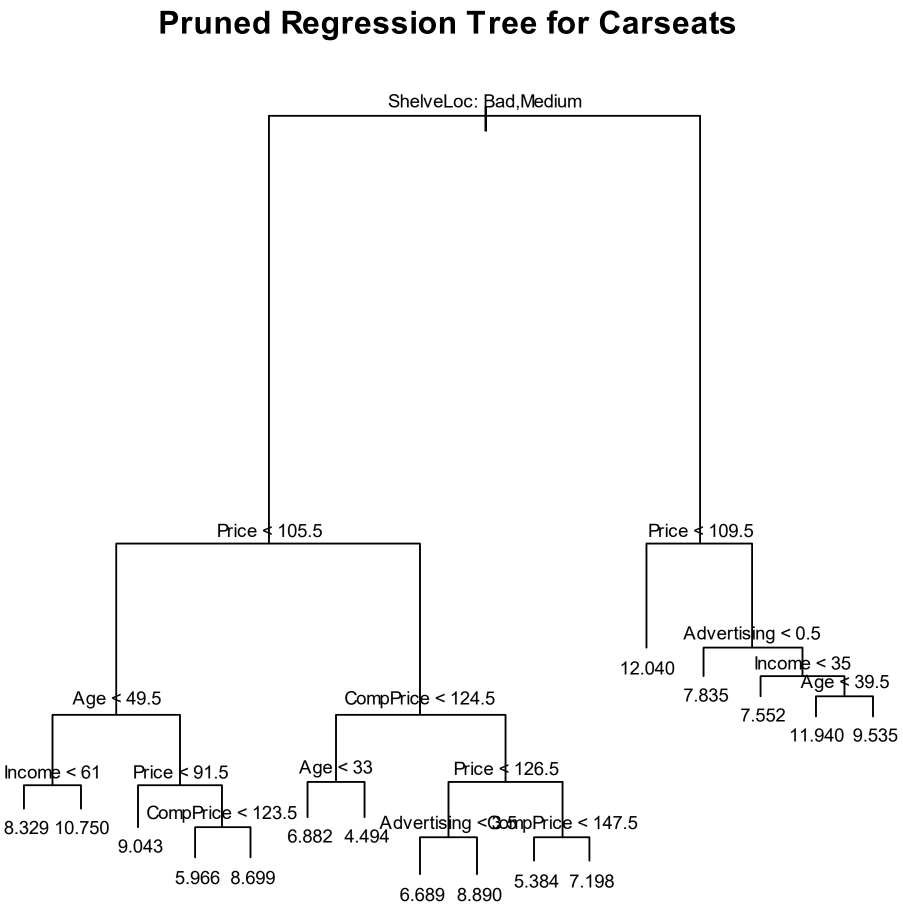


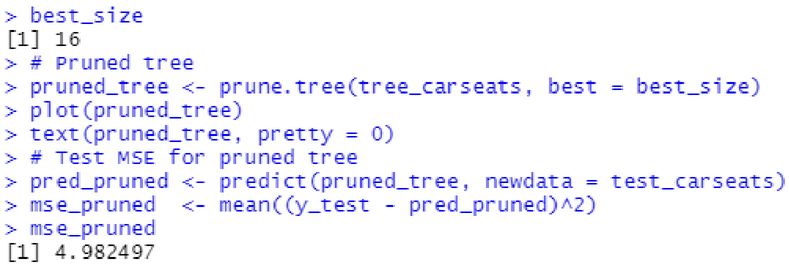
```
> pred_tree <- predict(tree_carseats, newdata = test_carseats)
> mse_tree  <- mean((y_test - pred_tree)^2)
> mse_tree
[1] 4.910268
```

The test MSE of 4.91 indicates that, on average, the regression tree’s predictions differ from the actual Sales values by about √4.91 ≈ 2.22 units. This shows the model has moderate prediction error and may benefit from pruning or ensemble methods for improved accuracy.

Tree Complexity

```
cv_tree <- cv.tree(tree_carseats)  # 10-fold CV by default
cv_tree$size
cv_tree$dev

best_size <- cv_tree$size[which.min(cv_tree$dev)]
best_size
```

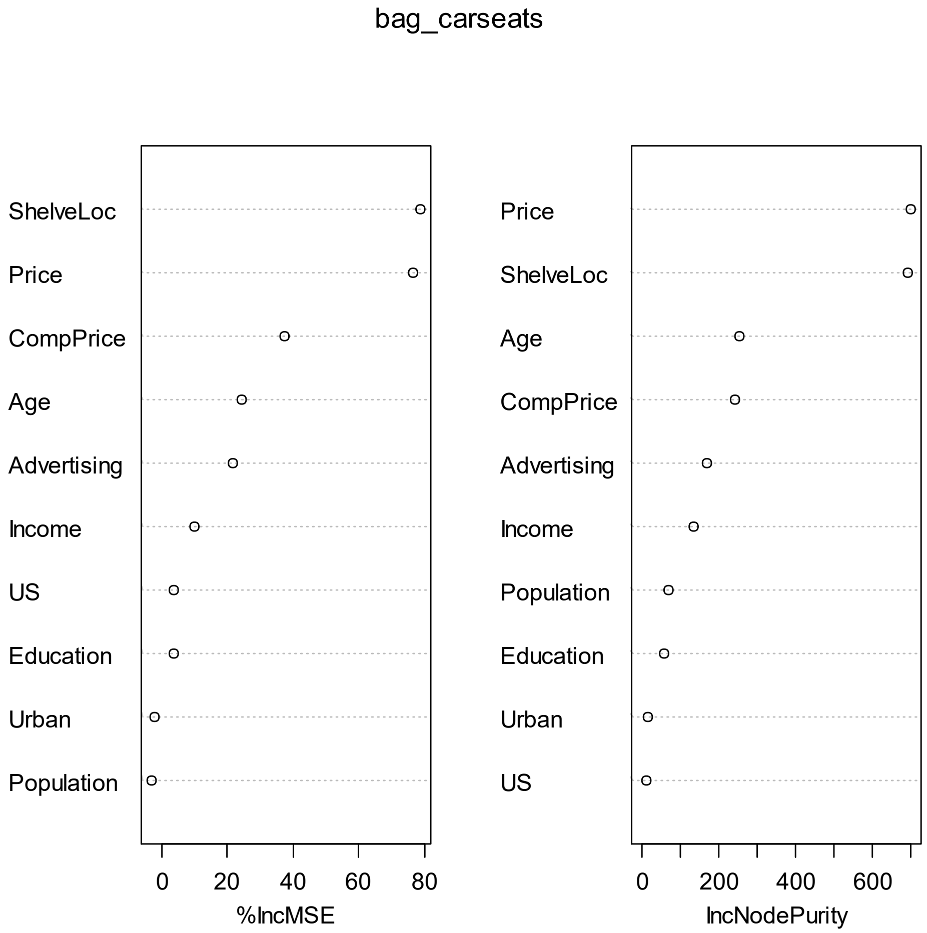

Pruned Regression Tree for Carseats



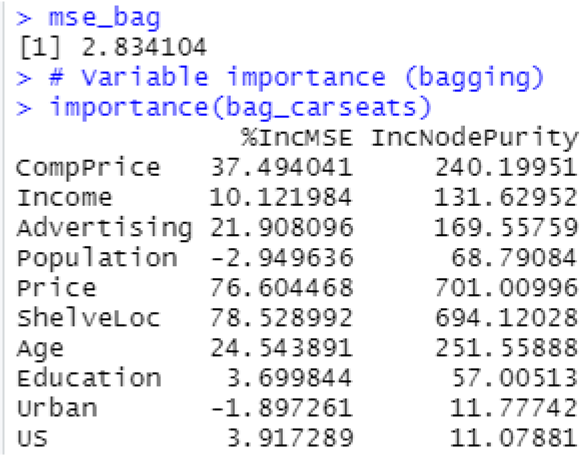
```
> best_size
[1] 16
> # Pruned tree
> pruned_tree <- prune.tree(tree_carseats, best = best_size)
> plot(pruned_tree)
> text(pruned_tree, pretty = 0)
> # Test MSE for pruned tree
> pred_pruned <- predict(pruned_tree, newdata = test_carseats)
> mse_pruned  <- mean((y_test - pred_pruned)^2)
> mse_pruned
[1] 4.982497
```

The pruned tree’s test MSE of 4.98 indicates that, on average, its Sales predictions differ from the actual values by about √4.98 ≈ 2.23 units. This suggests that pruning simplifies the model without significantly increasing prediction error. The model still exhibits moderate accuracy, and further improvements could be achieved using ensemble methods such as bagging or random forests.

Bagging



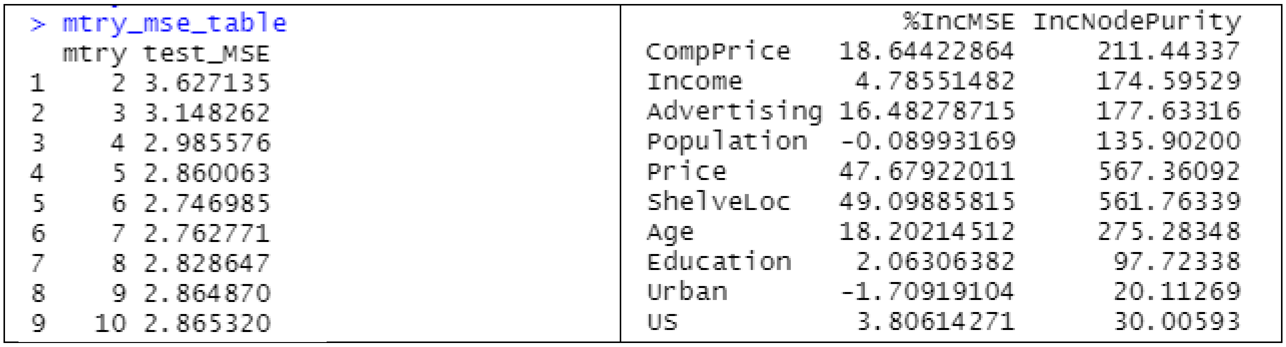
Bagging reveals that Price and ShelveLoc are the dominant drivers of Sales, while variables like Urban and Population contribute little to prediction accuracy. Bagging stabilizes predictions by averaging many trees, which is why it dramatically reduces test MSE



```
> mse_bag
[1] 2.834104
> # variable importance (bagging)
> importance(bag_carseats)
              %IncMSE IncNodePurity
CompPrice   37.494041     240.19951
Income      10.121984     131.62952
Advertising 21.908096     169.55759
Population  -2.949636      68.79084
Price       76.604468     701.00996
ShelveLoc   78.528992     694.12028
Age         24.543891     251.55888
Education    3.699844      57.00513
Urban       -1.897261      11.77742
US           3.917289      11.07881
```

The bagging model achieves a test MSE of 2.83, which is substantially lower than both the unpruned tree (≈ 4.91) and the pruned tree (≈ 4.98). This indicates that bagging improves predictive accuracy by reducing variance—its predictions differ from actual Sales by about $\sqrt{\mathbf{2.83}} \approx \mathbf{1.68}$ units showing a strong improvement over single-tree models.

Random Forrest

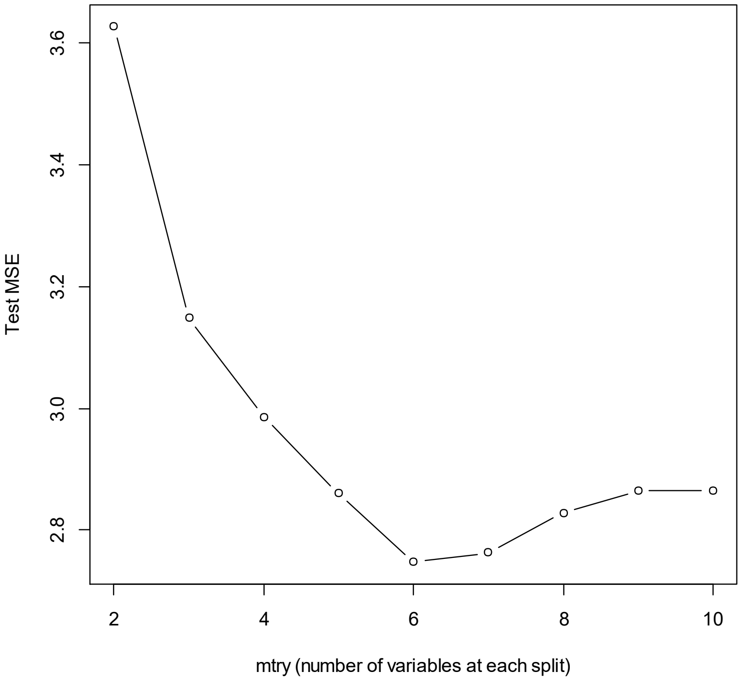


```
> mtry_mse_table
  mtry test_MSE
1    2 3.627135
2    3 3.148262
3    4 2.985576
4    5 2.860063
5    6 2.746985
6    7 2.762771
7    8 2.828647
8    9 2.864870
9   10 2.865320
```

```
              %IncMSE IncNodePurity
CompPrice   18.64422864     211.44337
Income       4.78551482     174.59529
Advertising 16.48278715     177.63316
Population  -0.08993169     135.90200
Price       47.67922011     567.36092
ShelveLoc   49.09885815     561.76339
Age         18.20214512     275.28348
Education    2.06306382      97.72338
Urban       -1.70919104      20.11269
US           3.80614271      30.00593
```

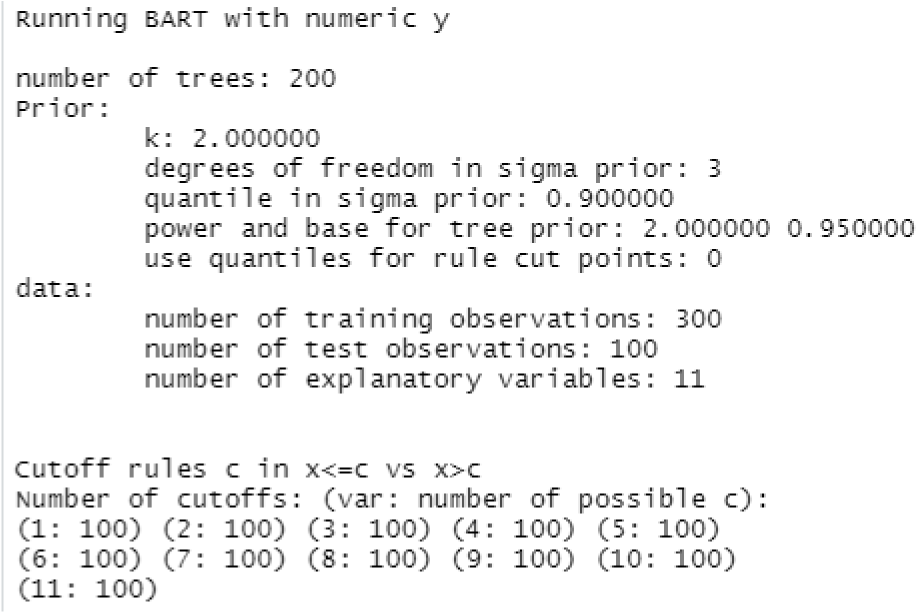
The mtry tuning results show that the test MSE decreases steadily as mtry increases, reaching its lowest value at mtry = 6, after which the error begins to rise again. This suggests that using around six randomly selected predictors at each split gives the most accurate model, balancing randomness and predictive strength.

The variable importance results indicate that ShelveLoc and Price are the strongest predictors of Sales, contributing the largest increases in MSE when permuted and showing the highest node purity. Other meaningful predictors include Age, CompPrice, and Advertising, which also play important roles in splitting the data effectively. Variables such as Population, Urban, and Education show very low or even negative importance, suggesting they add little value to the model's predictive performance. Overall, the random forest highlights that store display quality (ShelveLoc) and product pricing are the dominant drivers of Sales.

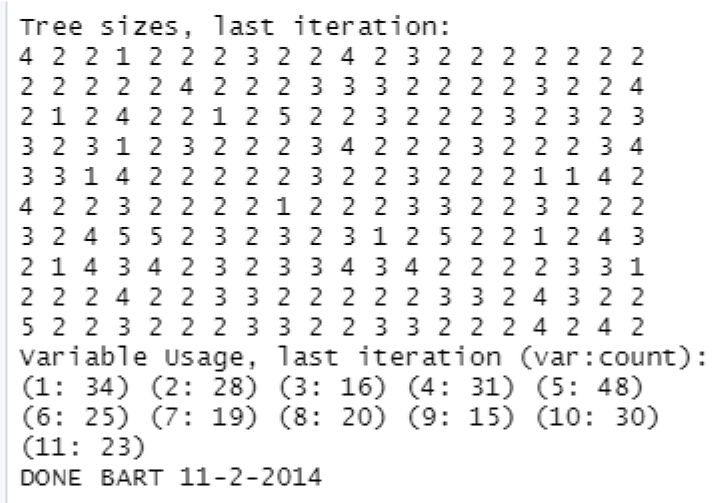


The test MSE curve shows that model performance improves as mtry increases, reaching its minimum at mtry = 6, after which the error begins to rise again. This pattern indicates that selecting six predictors at each split provides the best balance between randomness and predictive accuracy, making it the optimal mtry value for this random forest model

BART

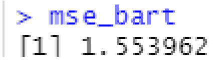


```
Running BART with numeric y

number of trees: 200
Prior:
        k: 2.000000
        degrees of freedom in sigma prior: 3
        quantile in sigma prior: 0.900000
        power and base for tree prior: 2.000000 0.950000
        use quantiles for rule cut points: 0
data:
        number of training observations: 300
        number of test observations: 100
        number of explanatory variables: 11


Cutoff rules c in x<=c vs x>c
Number of cutoffs: (var: number of possible c):
(1: 100) (2: 100) (3: 100) (4: 100) (5: 100)
(6: 100) (7: 100) (8: 100) (9: 100) (10: 100)
(11: 100)
```



```
Tree sizes, last iteration:
4 2 2 1 2 2 2 3 2 2 4 2 3 2 2 2 2 2 2 2
2 2 2 2 2 4 2 2 2 3 3 3 2 2 2 2 3 2 2 4
2 1 2 4 2 2 1 2 5 2 2 3 2 2 2 3 2 3 2 3
3 2 3 1 2 3 2 2 2 3 4 2 2 2 3 2 2 2 3 4
3 3 1 4 2 2 2 2 2 3 2 2 3 2 2 2 1 1 4 2
4 2 2 3 2 2 2 2 1 2 2 2 3 3 2 2 3 2 2 2
3 2 4 5 5 2 3 2 3 2 3 1 2 5 2 2 1 2 4 3
2 1 4 3 4 2 3 2 3 3 4 3 4 2 2 2 2 3 3 1
2 2 2 4 2 2 3 3 2 2 2 2 2 3 3 2 4 3 2 2
5 2 2 3 2 2 2 3 3 2 2 3 3 2 2 2 4 2 4 2
Variable Usage, last iteration (var:count):
(1: 34) (2: 28) (3: 16) (4: 31) (5: 48)
(6: 25) (7: 19) (8: 20) (9: 15) (10: 30)
(11: 23)
DONE BART 11-2-2014
```

The BART model mostly builds small trees and uses variables like Price, ShelveLoc, and CompPrice most frequently, indicating they are the strongest predictors of Sales. Less frequent usage of Urban and US shows they contribute little. Overall, BART confirms the same key drivers identified by earlier models.



```
> mse_bart
[1] 1.553962
```

The BART model achieves a test MSE of 1.55, which is substantially lower than all previous methods (trees, pruning, bagging, and random forests). This indicates that BART provides the most accurate predictions for Sales, capturing nonlinear relationships and interactions more effectively than the other models.

Recommendations:

Across all models, **Price and ShelveLoc consistently emerge as the strongest drivers of Sales**, highlighting that competitive pricing and better shelf placement are critical levers for improving store performance. While individual regression trees provide a basic understanding of decision structure, they show moderate accuracy due to high variance. Bagging and Random Forests significantly enhance performance by stabilizing predictions and revealing deeper variable interactions, with Random Forests ranking ShelveLoc and Price as the top predictors. Ultimately, **BART outperforms all models with the lowest test MSE (1.55)**, demonstrating its ability to capture complex nonlinear patterns and subtle interactions that traditional tree-based methods miss.

Overall, the analysis suggests that **advanced ensemble methods—especially BART—offer the most reliable predictive accuracy**, and business strategies should prioritize optimizing pricing and enhancing product visibility (ShelveLoc) to maximize sales.