In-Depth Analysis of Car Auctions in USA

MTH443 Course Project Report

Submitted To - Prof. Amit Mitra

Submitted by - Siddharth Pathak (Roll no. - 211034)

Table of contents

Data Understanding

The data set is about car auctions in US by the company Carvana in the year 2009-2010. In this report, We outline the steps to analyse the data and understand its quality. Additionally, we even add ways to deal with missing values.

Data Semantics

The semantics mentioned here are the most important columns used. The columns which are not mentioned are either removed or there name is too obvious to understand.

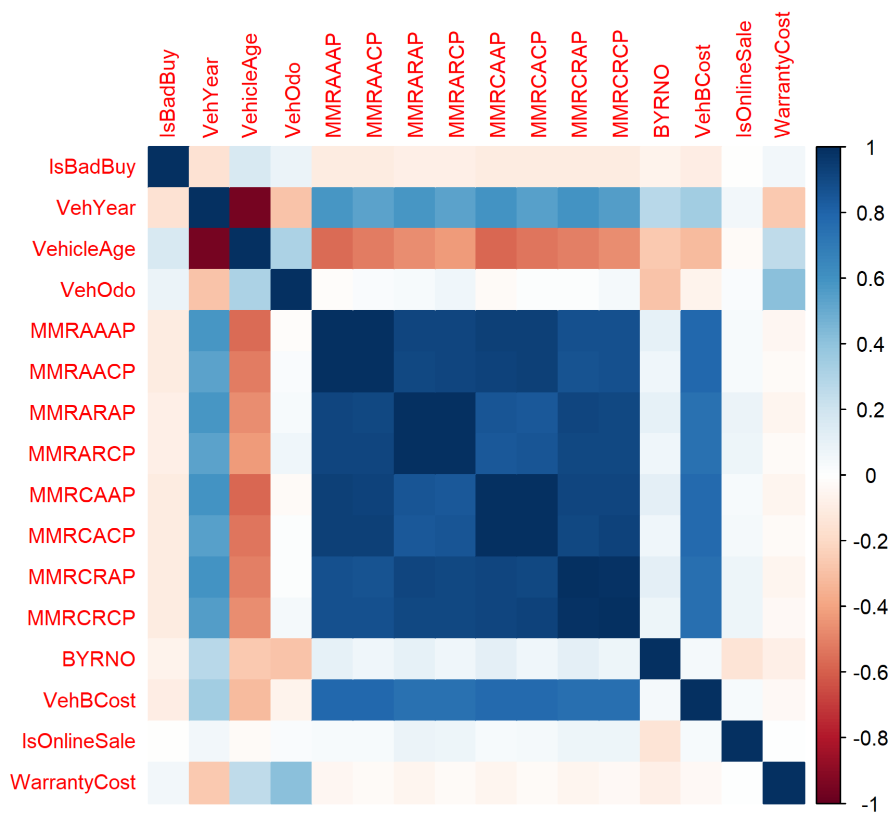
Table 1: Car Auction Dataset Field Descriptions

Field_Name	Definition
IsBadBuy	Indicates whether buying the vehicle was a mistake.
Auction	The auction provider where the vehicle was purchased.
VehYear	The year when the vehicle was manufarured.
VehicleAge	Current year - VehYear
Make	The Company of the Vehicle.
Model	The specific model of the vehicle.
Color	The color of the vehicle.
Transmission	The type of transmission in the vehicle (e.g., Automatic, Manual).
WheelType	The vehicle's wheel type (eg: Alloys, Cover, Special)
VehOdo	The vehicle's odometer reading at the time of purchase.
MMRAAAP	The average acquisition price in average condition at the time of purchase.
MMRAACP	The acquisition price in above-average condition at the time of purchase.
MMRARAP	The retail market acquisition price in average condition at the time of purchase.
MMRARCP	The retail market acquisition price in above-average condition at the time of purchase.
MMRCAAP	The current day acquisition price in average condition at auction.
MMRCACP	The current day acquisition price in above-average condition at auction.
MMRCRAP	The current retail market acquisition price in average condition.
MMRCRCP	The current retail market acquisition price in above-average condition.
BYRNO	A unique identifier assigned to the buyer who purchased the vehicle.
VNST	The state where the vehicle was purchased.
VehBCost	The acquisition cost paid at the time of purchase.
IsOnlineSale	Indicates if the vehicle was originally purchased online.
WarrantyCost	The cost of a warranty with a term of 36 months and 36,000 miles.

Data Cleaning

- **RefId**: This is just a unique identifier and does not provide any meaningful information.
- **PurchDate**: This is redundant information.
- **WheelTypeID**: Information regarding this is already contained in *WheelType*.
- **SubModel** and **Trim**: Contains too much information, but it is not useful for any data mining activity.
- **TopThreeAmericanName**, **PRIMEUNIT**, **AUCGUART**: These columns contained too much missing data.
- **VNZIP1**: The information is already included in *VNST*.

To simplify the dataset, the eight **MMR** columns were renamed using short forms.



This correlation plot shows that all the MMRs are strongly correlated (close to 0.9). We used this note-worthy information in the upcoming sections.

Data Quality :

- Consistency checks were applied, and rows containing `NA` values or missing data were removed.

- Clear outliers were deleted. For example, the column `VehBCost` contained an outlier with a value of $10.

Exploratory Data Analysis

To assess the relationship between categorical features and the target variable `IsBadBuy`, we performed chi-squared tests. This analysis helps identify significant associations that can guide feature selection for predictive modeling.

Table 2: Chi-Squared Test Results for Categorical Features

	Variable	P_Value	Interpretation
Auction	Auction	0.0005	Significant
Make	Make	0.0005	Significant
Model	Model	0.0005	Significant
Color	Color	0.0005	Significant
Transmission	Transmission	0.7826	Not Significant
WheelType	WheelType	0.0005	Significant
Nationality	Nationality	0.0125	Significant
Size	Size	0.0005	Significant
VNST	VNST	0.0005	Significant

We found that *Transmission* does not show significant relationship (p-value = 0.7921)

To evaluate the relationship between numeric features and the target variable `IsBadBuy`, we conducted two-sample t-tests for each numeric column. This helps identify whether there are statistically significant differences in the means of numeric variables across the levels of `IsBadBuy`.

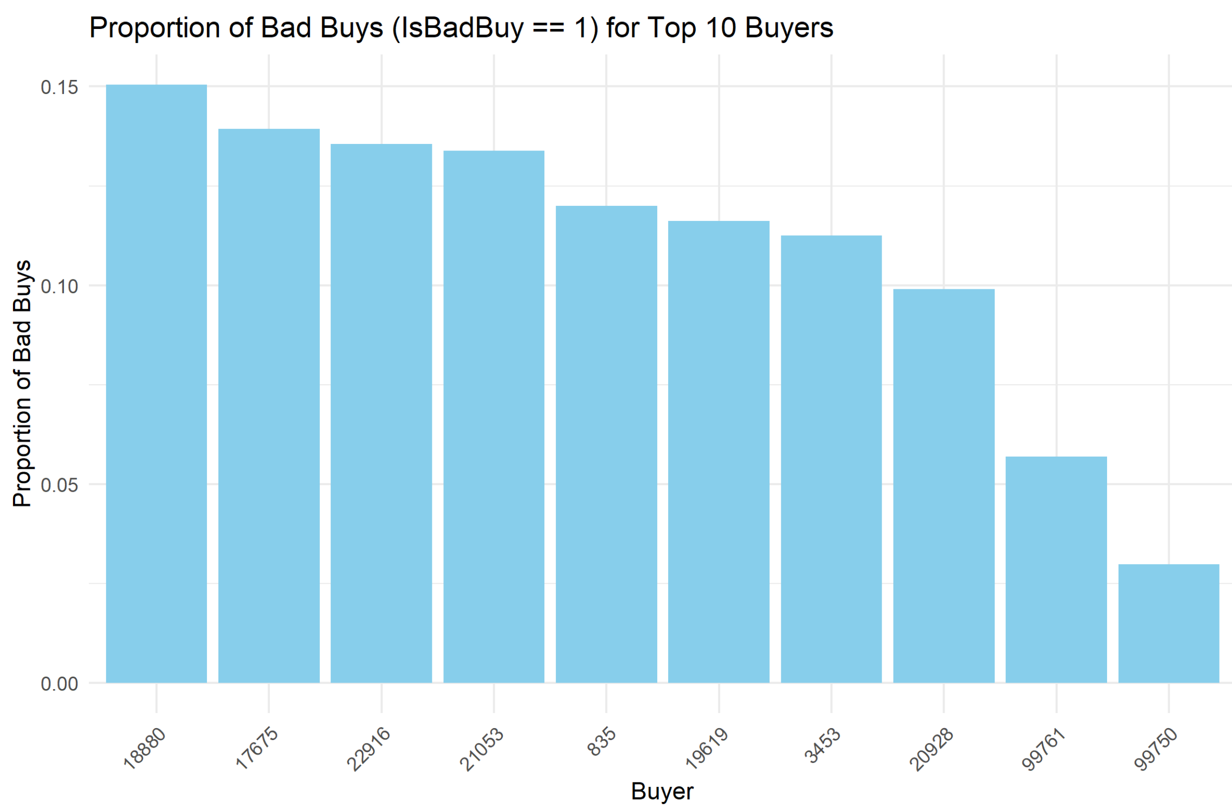
Table 3: T-Test Results for Numeric Features

	Variable	P_Value	Interpretation
IsBadBuy	IsBadBuy	NA	NA
VehYear	VehYear	0.0000	Significant
VehicleAge	VehicleAge	0.0000	Significant
VehOdo	VehOdo	0.0000	Significant
MMRAAAP	MMRAAAP	0.0000	Significant
MMRAACP	MMRAACP	0.0000	Significant
MMRARAP	MMRARAP	0.0000	Significant
MMRARCP	MMRARCP	0.0000	Significant

	Variable	P_Value	Interpretation
MMRCAAP	MMRCAAP	0.0000	Significant
MMRCACP	MMRCACP	0.0000	Significant
MMRCRAP	MMRCRAP	0.0000	Significant
MMRCRCP	MMRCRCP	0.0000	Significant
BYRNO	BYRNO	0.0000	Significant
VehBCost	VehBCost	0.0000	Significant
IsOnlineSale	IsOnlineSale	0.3034	Not Significant
WarrantyCost	WarrantyCost	0.0000	Significant

We found *IsOnlineSale* is not having significant relationship

Next, we look for bar chart for top 10 buyers ranked by highest proportion of bad purchases. Some buyers have relatively higher proportion of bad purchases compared to others. We analyse the trend of these bad buyers.



To understand the attribute VSNT, we plotted the map of USA with this variable.

Interestingly, we found that major auctions took place in only few states. There were many states that had 0 auctions (for eg: Montana).

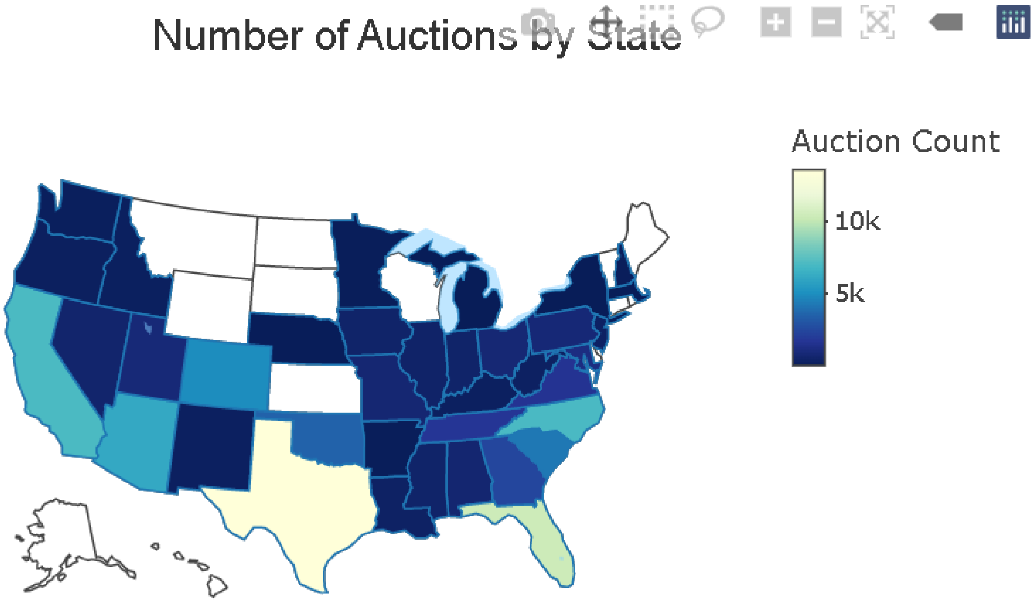
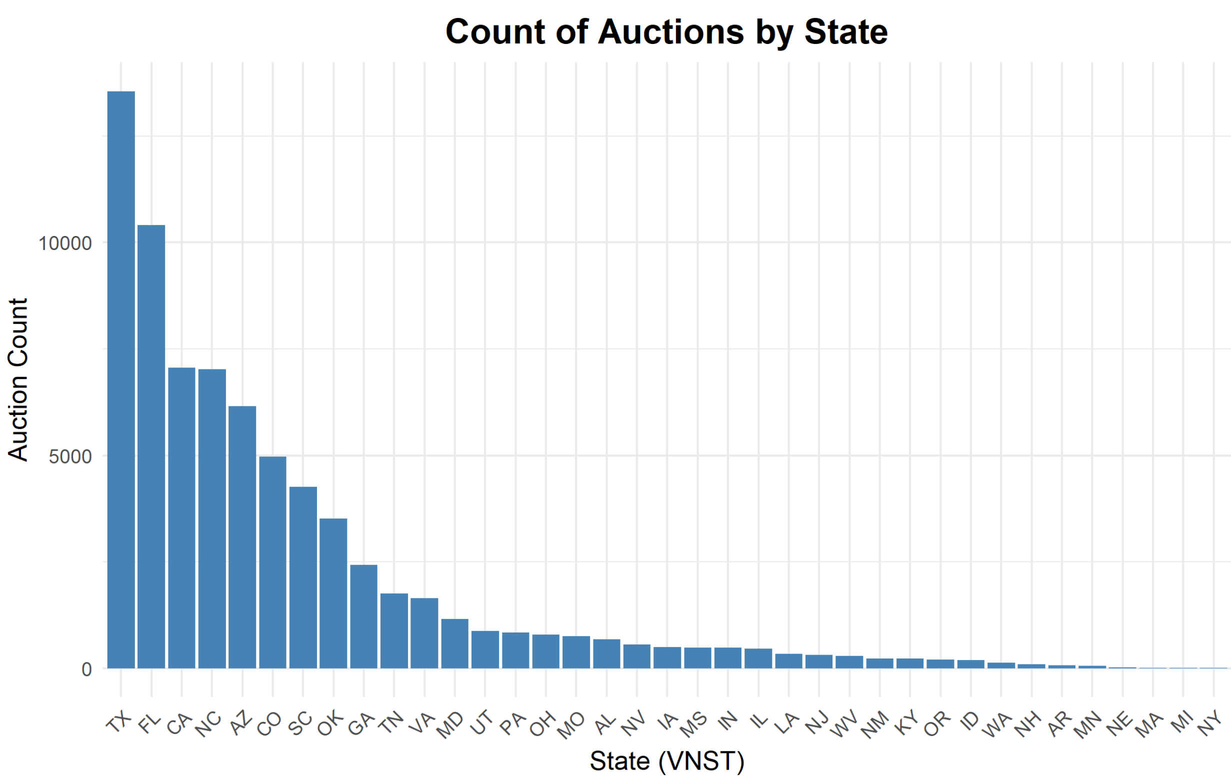
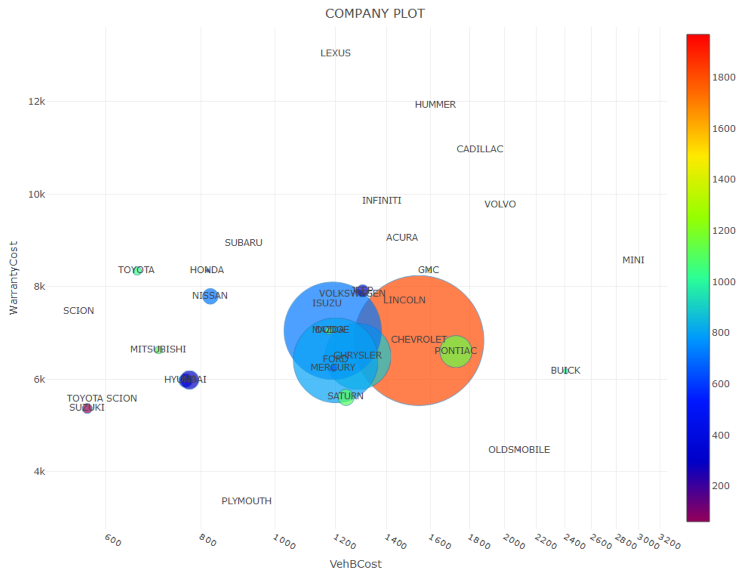


Figure 1: Plot of Auction Data

To check this further we created a bar chart of states v/s auctions. We We found the most of the auctions took place in Texas, Florida and California.

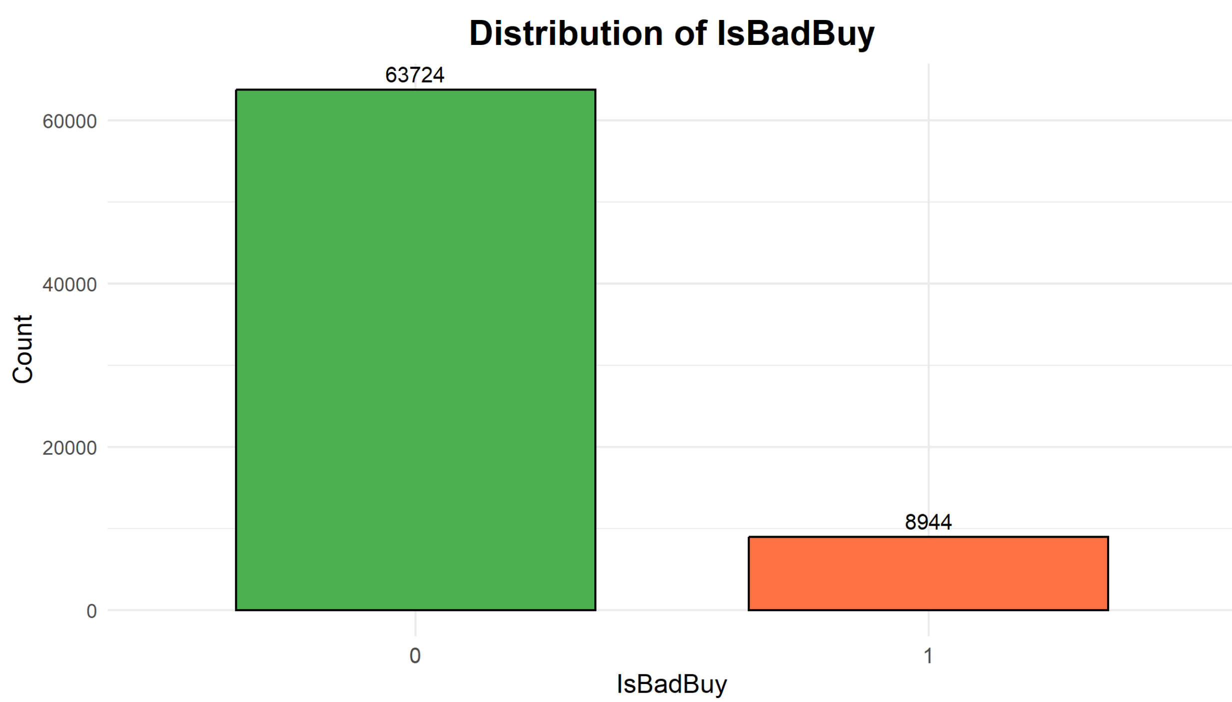


Then, We explored about the *VehBcost, WarrantyCost, Make.*



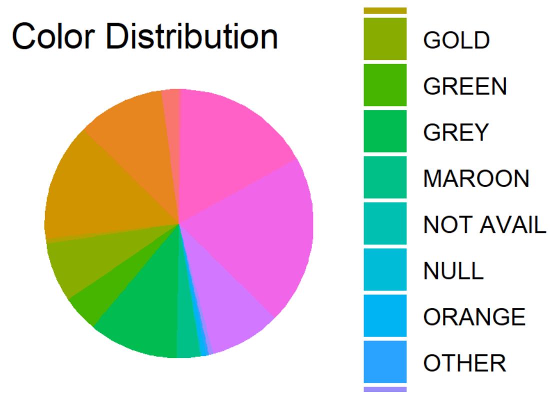
We see 2 clusters, one around low *VehBcost* and other around mid -range *VehBcost*. Economical brands like **Hyundai**, **Toyota**, and **Mitsubishi** offer vehicles with lower acquisition costs and moderate warranty costs, catering to budget-conscious buyers. Luxury brands like **Cadillac** and **Lexus** are associated with higher acquisition and warranty costs, suggesting their premium status. Brands like **Chevrolet** and **Pontiac** dominate the mid-range segment, likely due to higher vehicle sales or activity.

Next, we explored about the most important variable *IsBadBuy*.

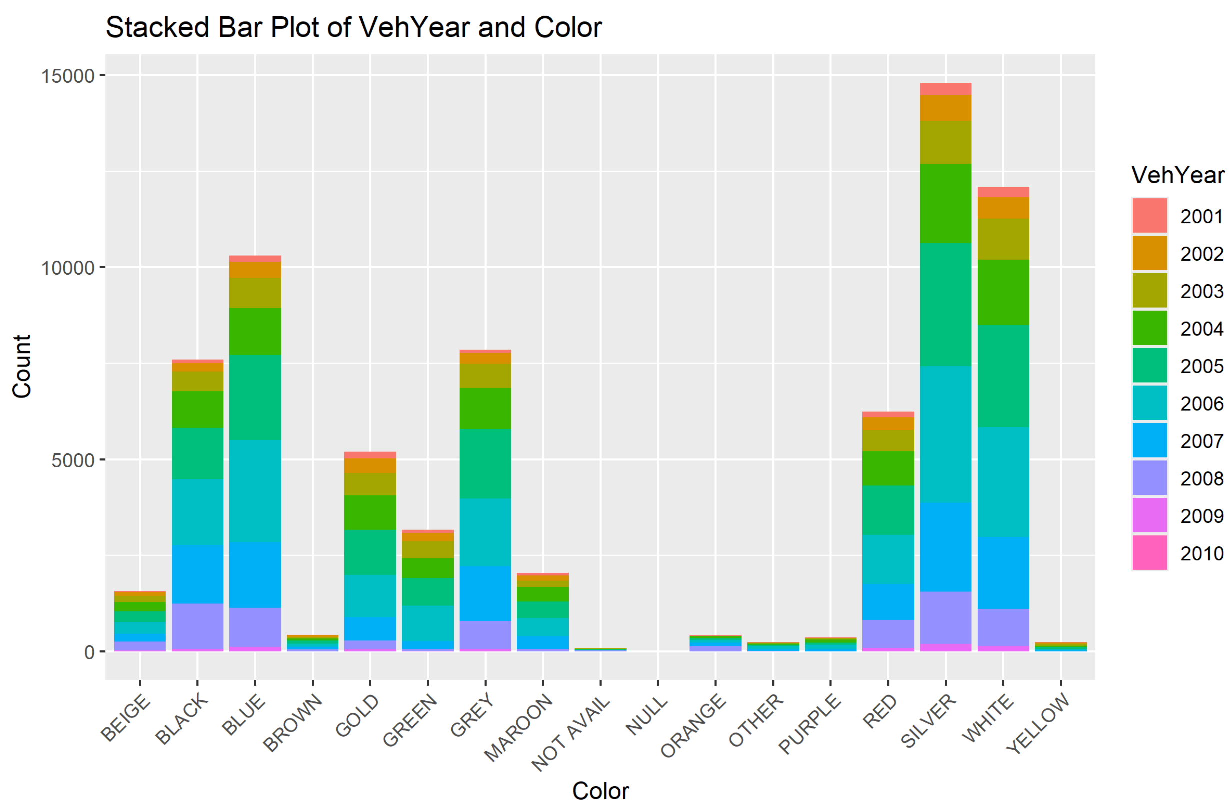


The plot clearly shows the significant class imbalance in the *IsBadBuy*. This may make ML models to show biasness towards predicting the majority class. (We used techniques like SMOTE to tackle this problem)

Next we check the attribute *color*. Starting with the distribution of color.



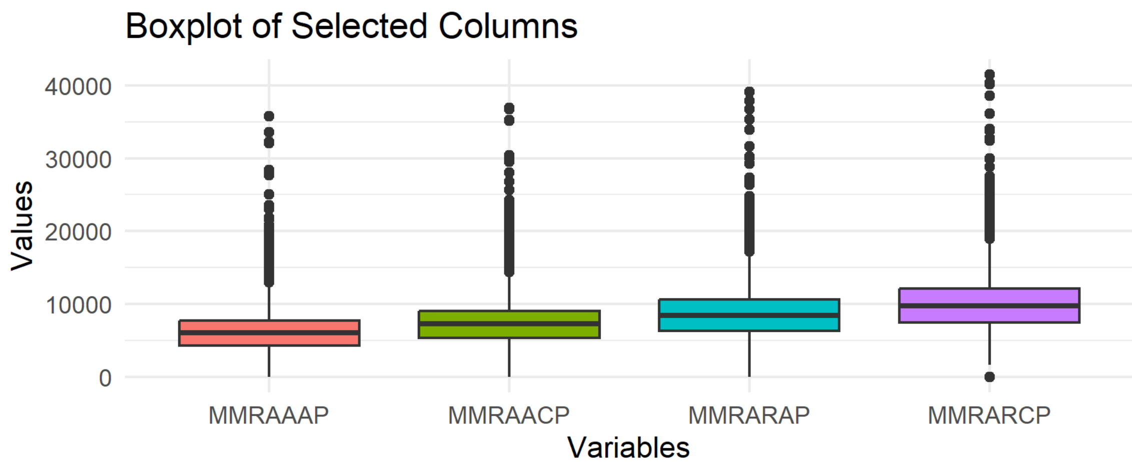
As this could be an important indicator, we look for more plots for *color* attribute.



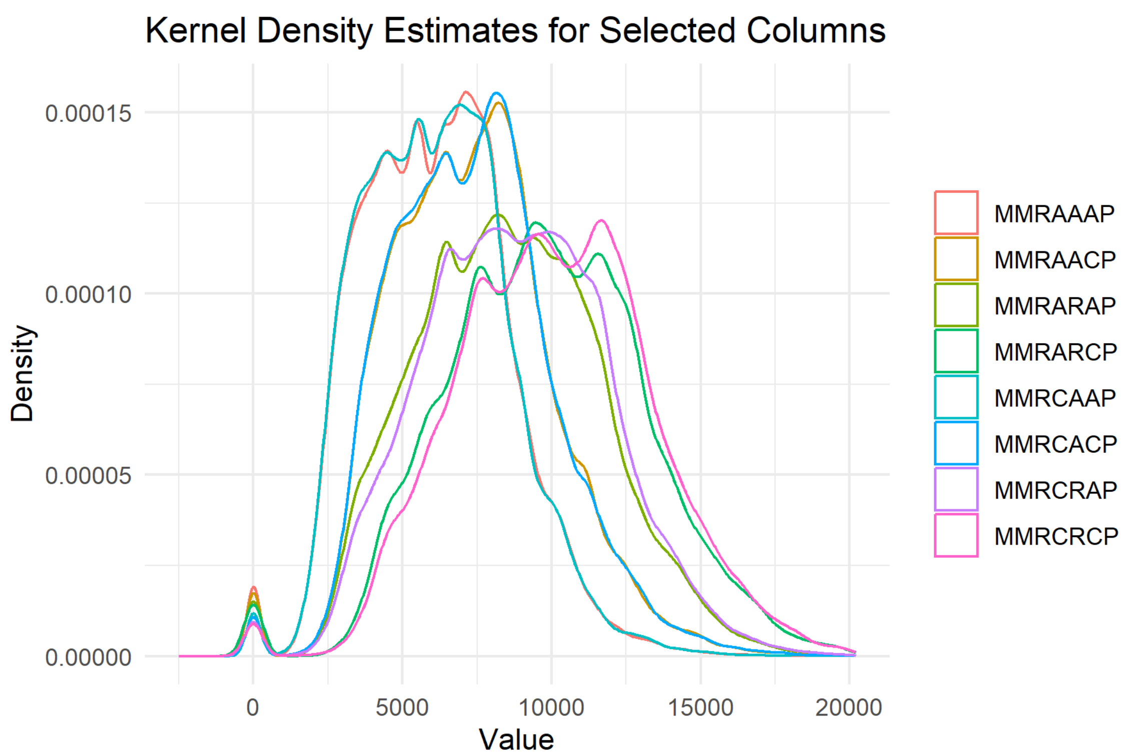
The stacked bar plot shows that **White**, **Silver**, and **Black** are the most dominant vehicle colors, consistently popular across all years (2001–2010). Bright colors like **Yellow**, **Orange**, and **Purple** are rare, indicating limited demand or production for these colors. Older vehicles (2001–2003) contribute less overall, reflecting fewer auctions for these models. Some categories, such as **Not**

Available and **NULL**, highlight missing or incomplete data. The consistent popularity of neutral colors suggests strong customer preferences for classic, universally appealing vehicle tones.

Exploring the MMR variables

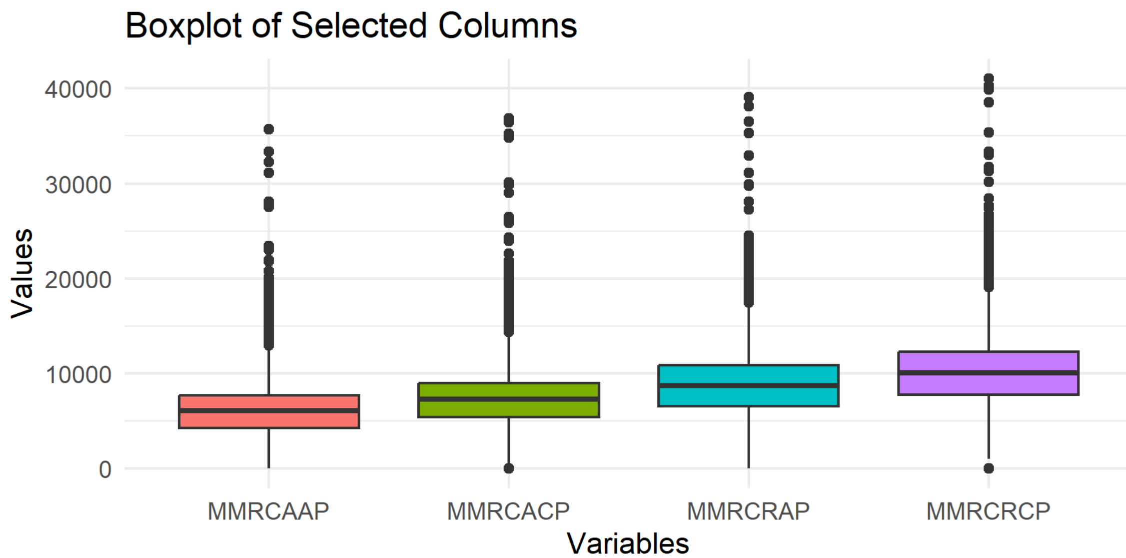


The boxplot reveals that the selected variables (**MMRAAAP**, **MMRAACP**, **MMRARAP**, **MMRARCP**) have similar distributions, with most values concentrated around the lower range but with significant outliers extending beyond 30,000.



We found MMR attributes are very correlated, and the correlation is even higher when consider these pair wise. Further, we noticed before the cleaning, the presence of a peak at 0. We will treat this value as missing value.

Checking the same trend in MMR current day prices.



We see a similar trend in MMR current prices and MMR acquisition price. Even this shows a correlation between these 2 variables.

Clustering

We use **k-means clustering** and **hierarchical clustering** to group observations based on their similarities in selected features.

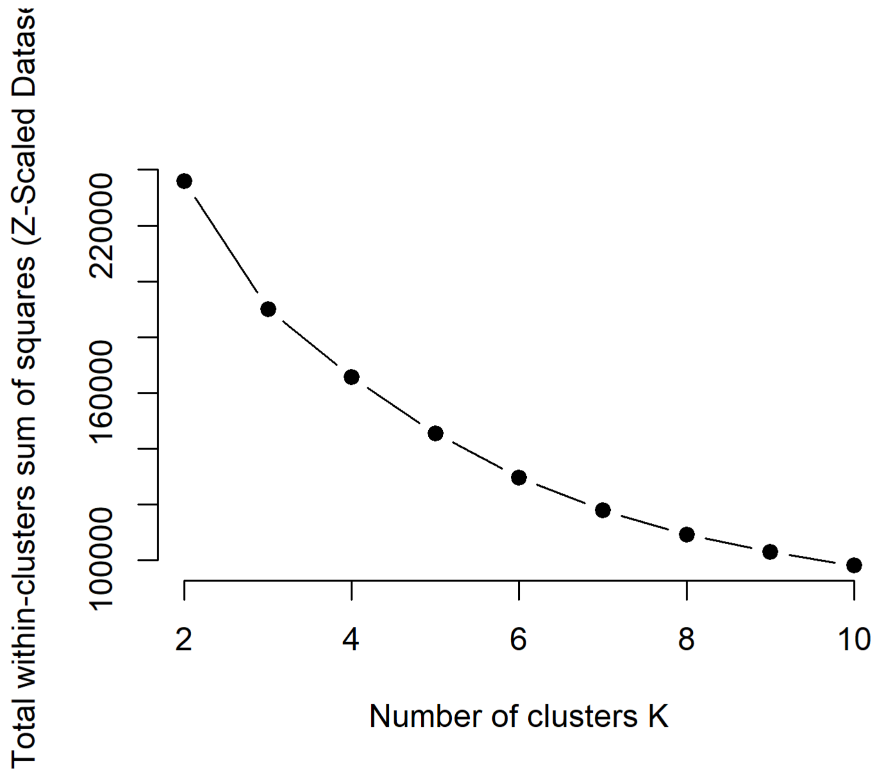
The clustering is based on the following columns:

- VehOdo
- VehBCost
- WarrantyCost
- MMRAcquisitionAuctionAveragePrice
- MMRAcquisitionRetailAveragePrice

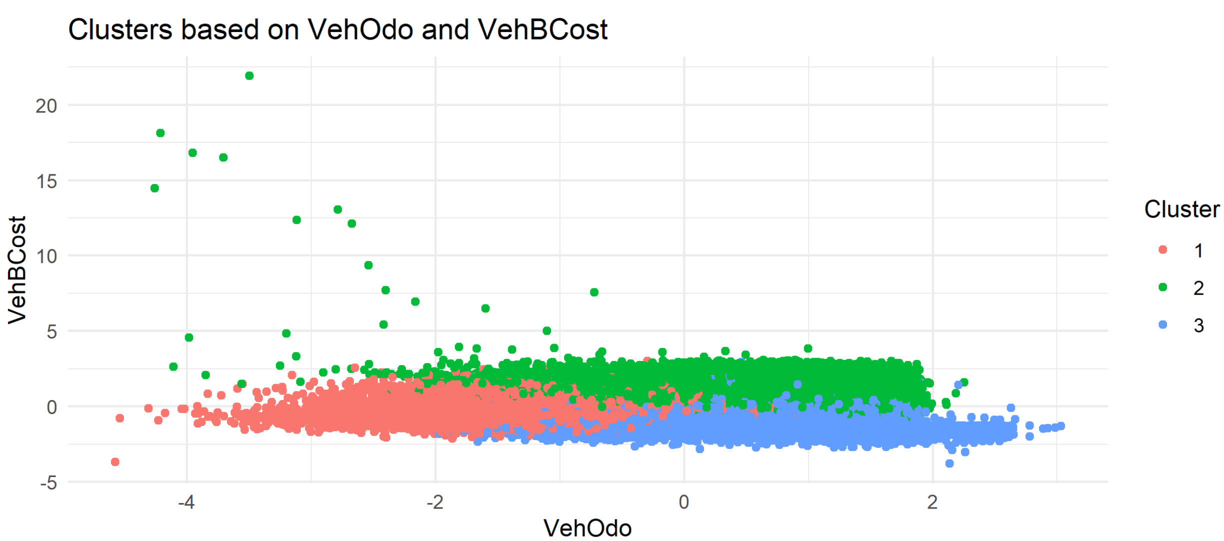
The data is also standardized (z-scores) before performing clustering to obtain consistent and reasonable inference from the formed clusters.

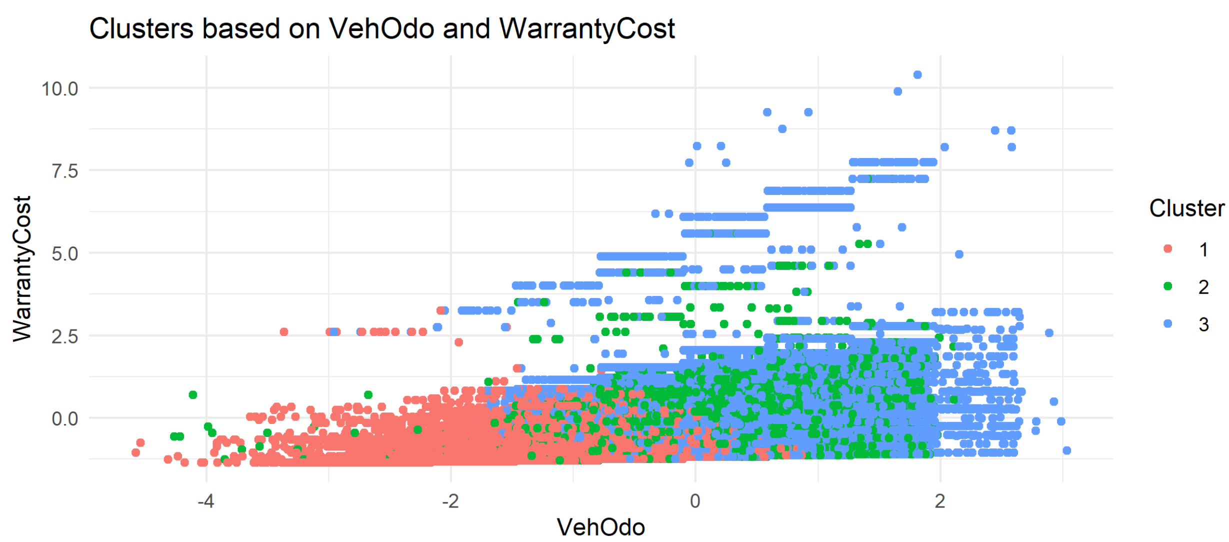
K-Means Clustering

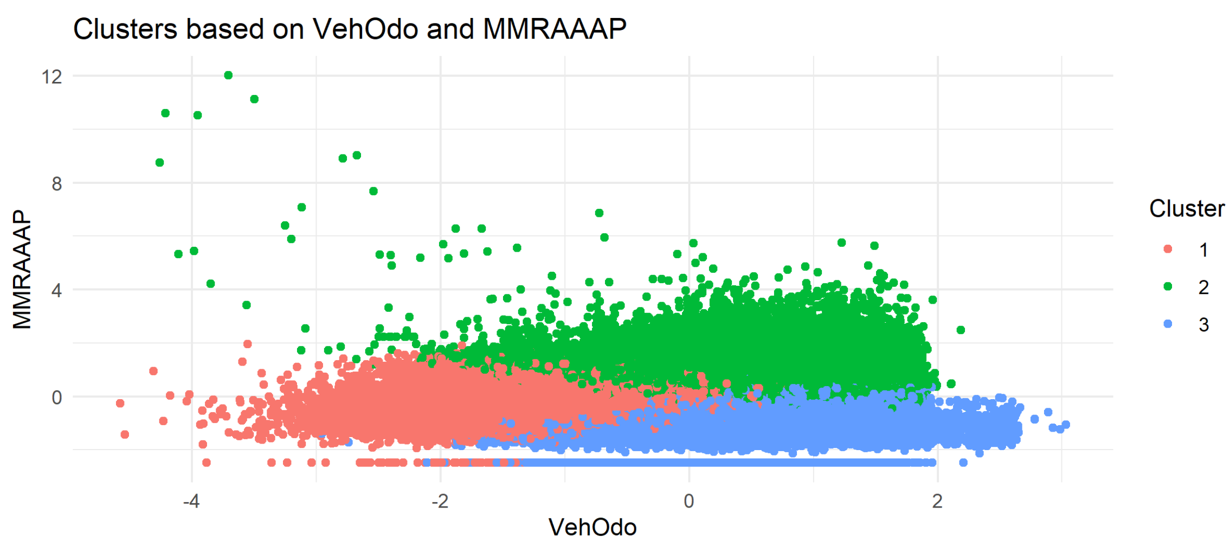
The ideal number of clusters (k) is evaluated using the elbow plot where we plot the total within cluster sum of squares (WSS) against k. The clustering is performed for $k = 2 - 10$. As a result, we see an

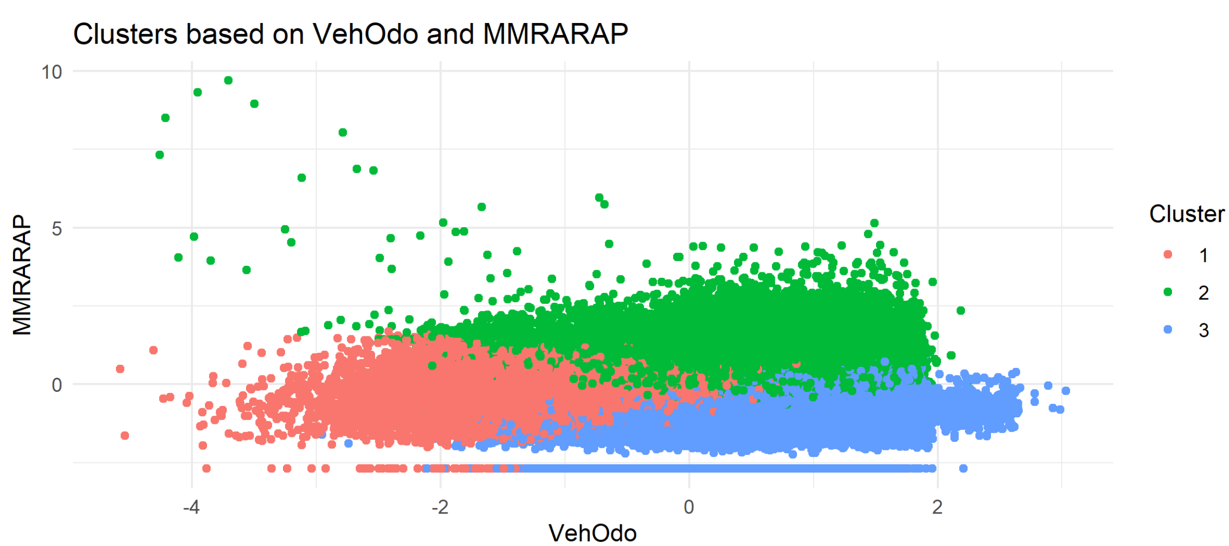


From the plot, we identify $k = 3$ as the best number of clusters. Based on this, we create clusters and plot them as a function of 2 parameters at a time.

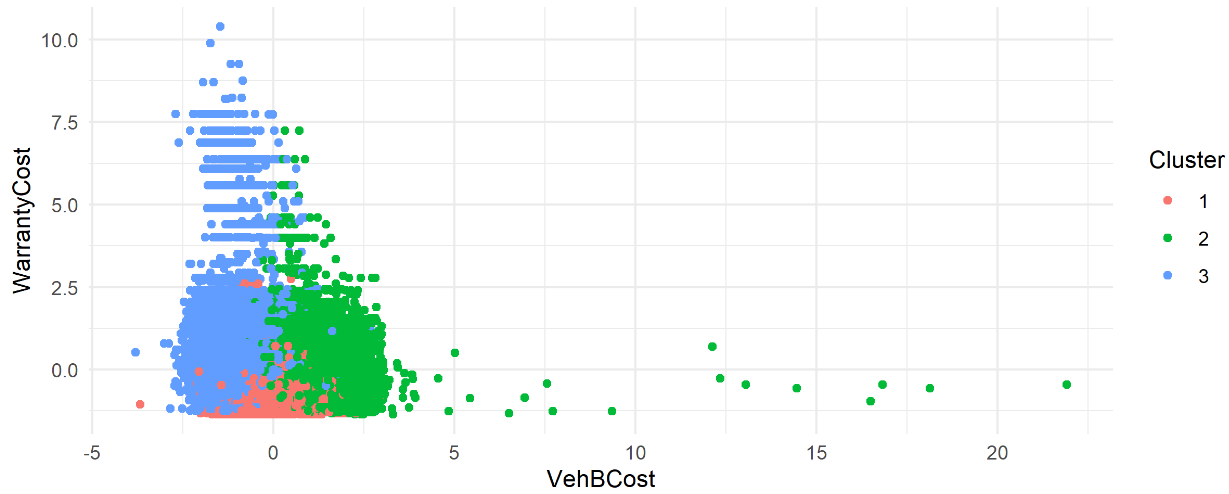



Clusters based on VehOdo and WarrantyCost
WarrantyCost
10.0
7.5
5.0
2.5
0.0
-4
-2
0
2
VehOdo
Cluster
1
2
3

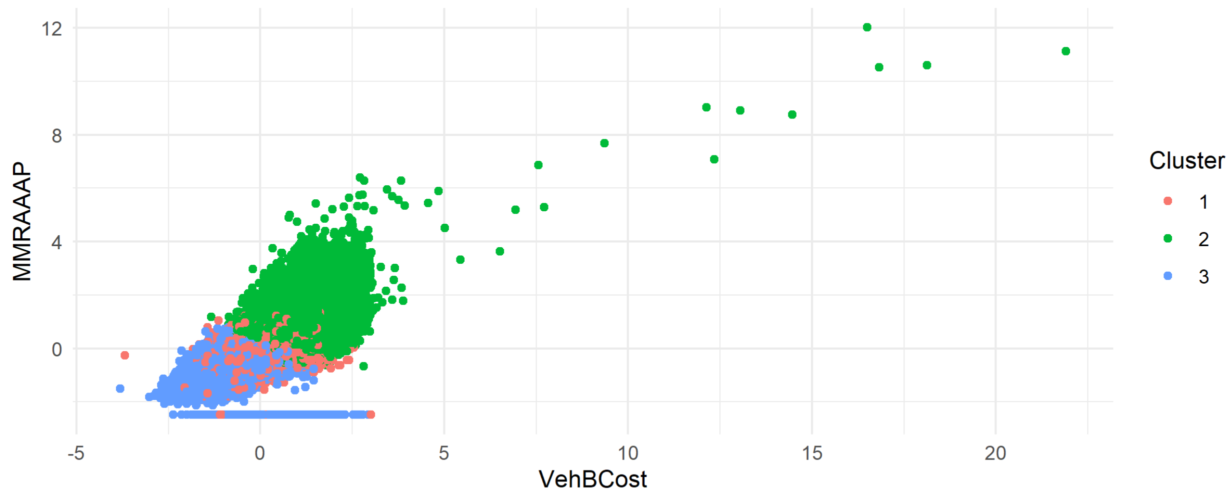

Clusters based on VehOdo and MMRAAAP
MMRAAAP
12
8
4
0
-4
-2
0
2
VehOdo
Cluster
1
2
3


Clusters based on VehOdo and MMRARAP
MMRARAP
10
5
0
-4
-2
0
2
VehOdo
Cluster
1
2
3

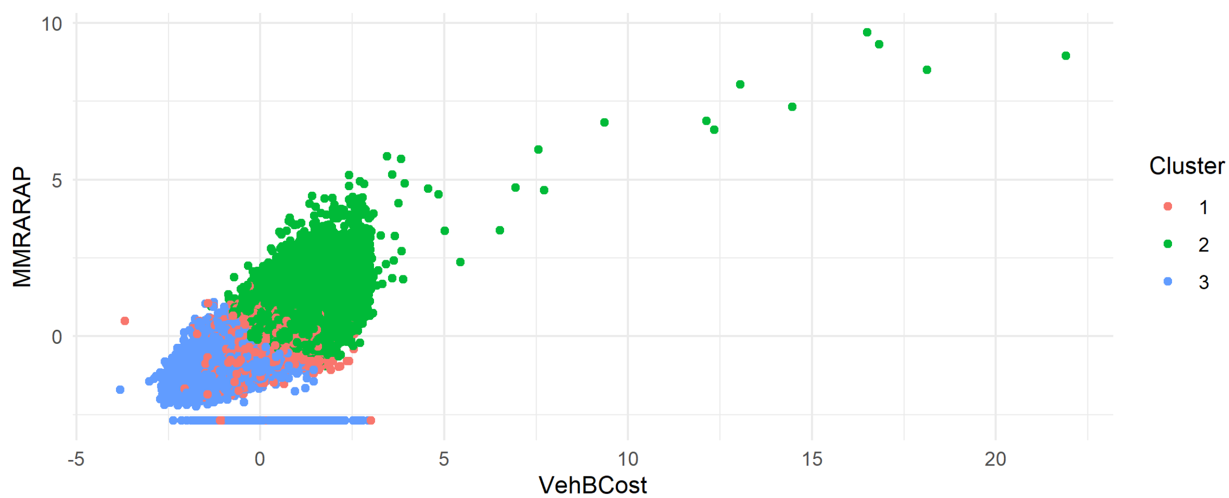
Clusters based on VehBCost and WarrantyCost

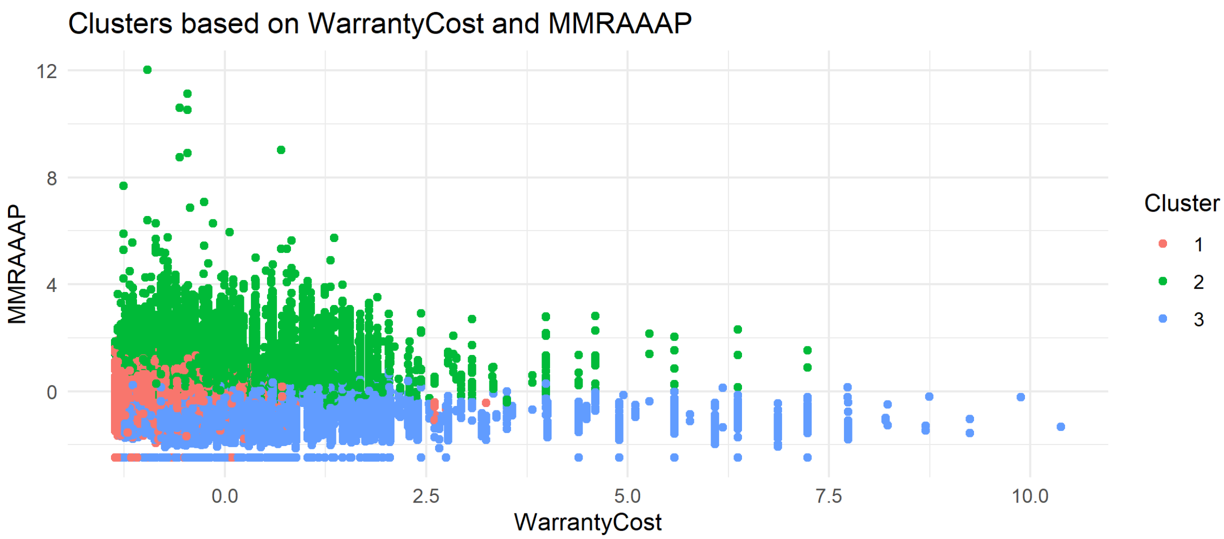


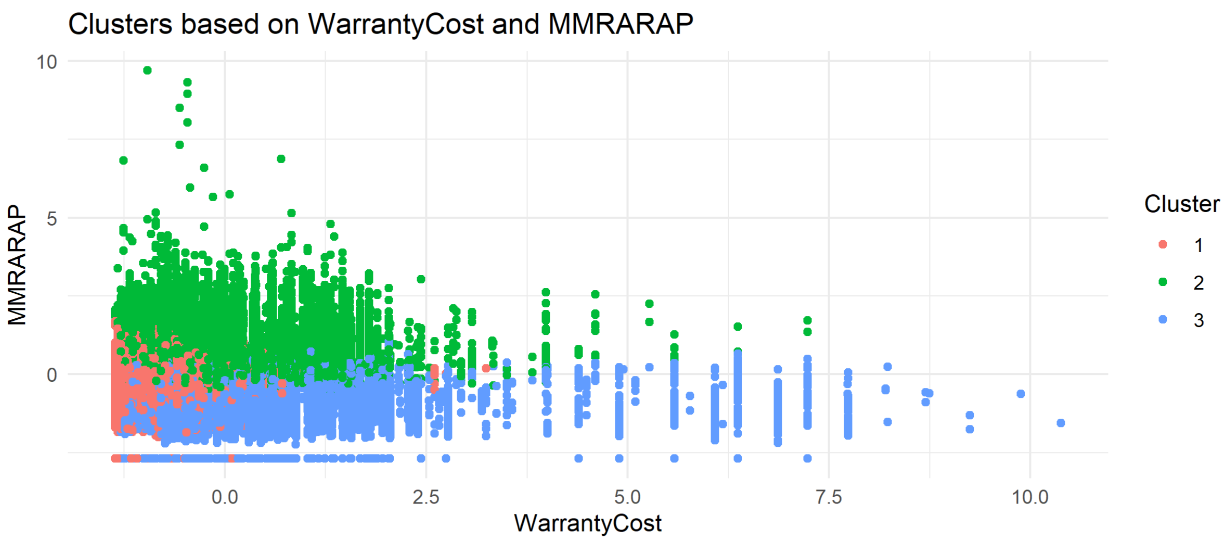
Clusters based on VehBCost and MMRAAAP

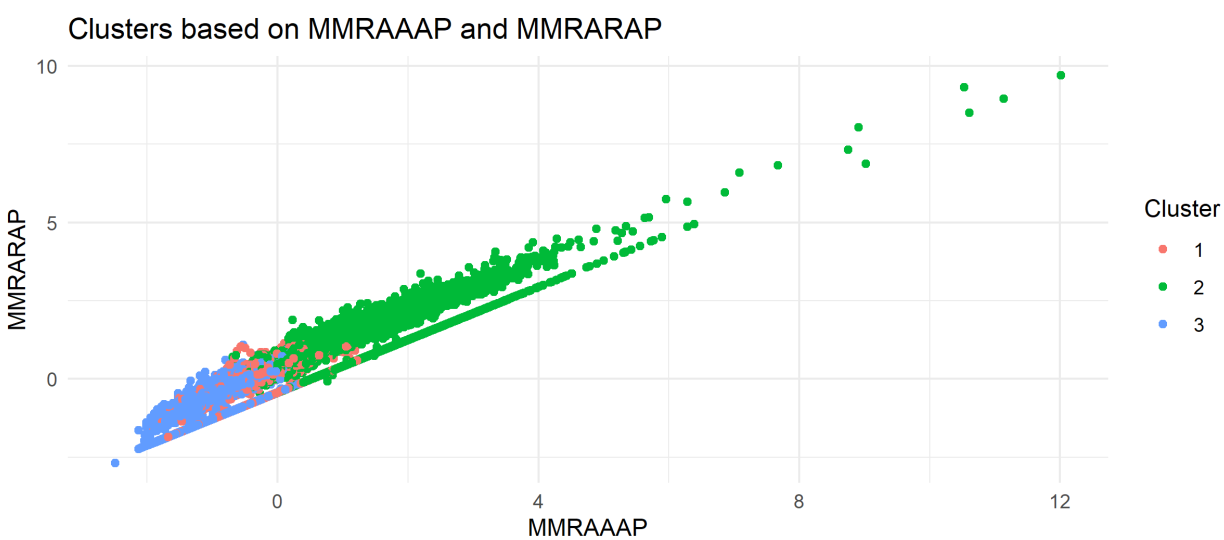


Clusters based on VehBCost and MMRARAP




Clusters based on WarrantyCost and MMRAAAP
MMRAAAP
WarrantyCost
Cluster
1
2
3
12
8
4
0
0.0
2.5
5.0
7.5
10.0


Clusters based on WarrantyCost and MMRARAP
MMRARAP
WarrantyCost
Cluster
1
2
3
10
5
0
0.0
2.5
5.0
7.5
10.0


Clusters based on MMRAAAP and MMRARAP
MMRARAP
MMRAAAP
Cluster
1
2
3
10
5
0
0
4
8
12

Hierarchical Clustering

Since the data is very large, the distance matrix for the original dataset requires a lot of memory and time to run. Instead, we consider 1000 data points sampled from the dataset to try and evaluate any interesting features.

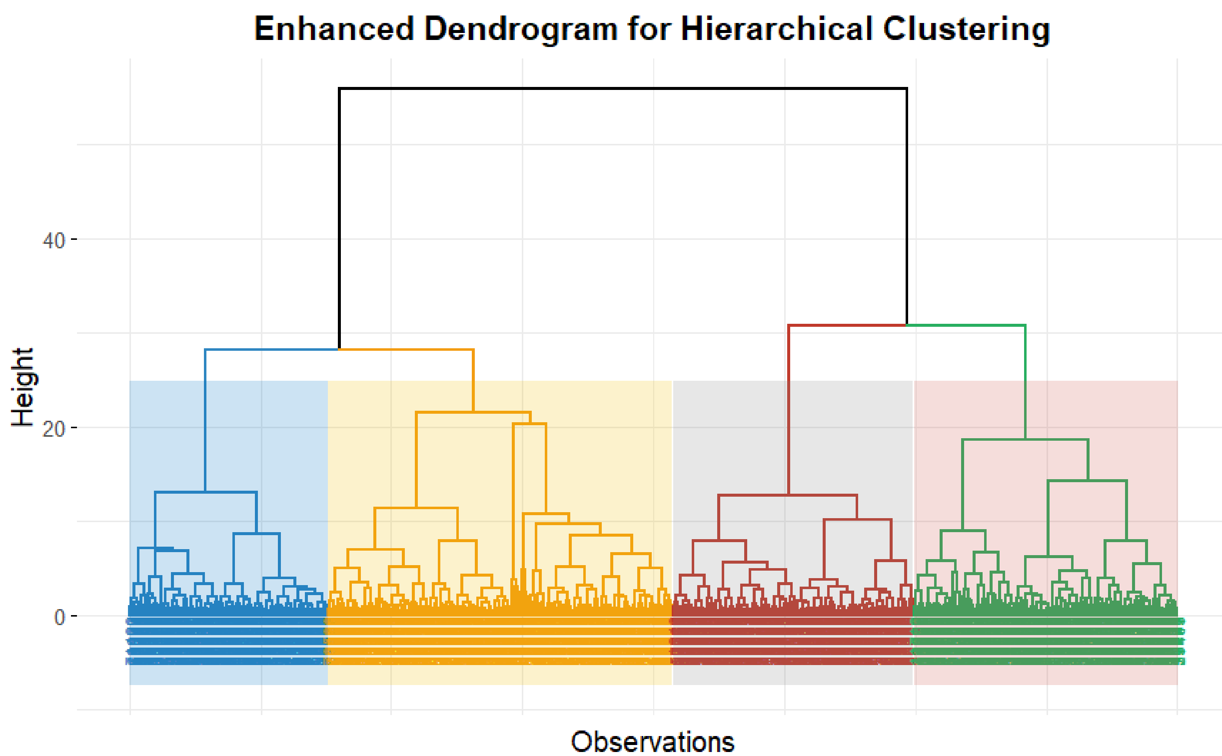


Figure 2: Hierarchical Clustering Dendogram

Classification

Data Processing

The data was standardized using Z-score normalization and PCA was applied to reduce dimensionality. The first principal component (PC1) was extracted to represent the combined effect of MMRA and MMRC-related variables, simplifying the dataset while retaining key information.

Model Training and Evaluation

The `train_and_evaluate()` function is a comprehensive utility for training and assessing the performance of classification models. It employs **5-fold Cross-Validation** (CV), a robust evaluation technique that divides the training data into five subsets (folds) to ensure the model is validated on multiple segments of the dataset, reducing the risk of overfitting. The function utilizes the `caret::train()` method to train the specified model, using `ROC` as the evaluation metric for model selection. It then generates probabilistic predictions on the test set and converts them to binary class labels using a threshold of 0.5. The evaluation step involves computing a confusion matrix, from which key metrics are derived: overall accuracy, Good Buy accuracy (for non-defective cars), and Bad Buy accuracy (for defective cars). This function standardizes the model evaluation process, enabling consistent and interpretable comparisons across different models and data resampling techniques.

LDA

LDA aims to find a linear combination of features that best separates the classes by maximizing the distance between the class means while minimizing the variance within each class.

```
# LDA for normal, upsampled, and downsampled data
lda_normal <- train_and_evaluate("lda", data.frame(X_train, Class = y_train),
                                 X_test, y_test)
```

```
          Reference
Prediction    No   Yes
       No  12525  1347
       Yes   219   441
```

```
lda_up <- train_and_evaluate("lda", up_train, X_test, y_test)
```

```
          Reference
Prediction   No  Yes
       No  9304  755
       Yes 3440 1033
```

```
lda_down <- train_and_evaluate("lda", down_train, X_test, y_test)
```

```
          Reference
Prediction   No  Yes
       No  9235  753
       Yes 3509 1035
```

QDA

QDA is similar to LDA but relaxes the assumption of equal covariance matrices for each class. It allows each class to have its own covariance matrix, resulting in a quadratic decision boundary.

```
# QDA for normal, upsampled, and downsampled data
qda_normal <- train_and_evaluate("qda", data.frame(X_train, Class = y_train),
                                 X_test, y_test)
```

```
          Reference
Prediction    No   Yes
       No  12315  1288
       Yes   429   500
```

```
qda_up <- train_and_evaluate("qda", up_train, X_test, y_test)
```

```
          Reference
Prediction    No   Yes
       No  10032   855
       Yes  2712   933
```

```
qda_down <- train_and_evaluate("qda", down_train, X_test, y_test)
```

```
          Reference
Prediction   No  Yes
       No  9987  844
       Yes 2757  944
```

Decision Tree

A Decision Tree is a non-parametric, tree-structured model that splits the data into subsets based on feature values. At each node, the tree selects the feature and threshold that best separates the classes using a criterion like **Gini impurity** or **information gain**. The tree continues splitting until a stopping condition is met, making it interpretable and useful for both classification and regression tasks.

```r
# Decision Tree for normal, upsampled, and downsampled data
dt_normal <- train_and_evaluate("rpart", data.frame(X_train, Class = y_train),
                                X_test, y_test)
```

```
          Reference
Prediction    No   Yes
       No  12673  1388
       Yes    71   400
```

```r
dt_up <- train_and_evaluate("rpart", up_train, X_test, y_test)
```

```
          Reference
Prediction   No  Yes
       No  7925  517
       Yes 4819 1271
```

```r
dt_down <- train_and_evaluate("rpart", down_train, X_test, y_test)
```

```
          Reference
Prediction   No  Yes
       No  7925  517
       Yes 4819 1271
```

Results Summary

Table 4: Class-wise Accuracy Metrics

Model	Accuracy	Good_Buy_Accuracy	Bad_Buy_Accuracy
Decision Tree - Normal	0.8996009	0.9944288	0.2237136
Decision Tree - Upsampled	0.6328103	0.6218613	0.7108501
Decision Tree - Downsampled	0.6328103	0.6218613	0.7108501
LDA - Normal	0.8922378	0.9828154	0.2466443
LDA - Upsampled	0.7113267	0.7300691	0.5777405
LDA - Downsampled	0.7067162	0.7246547	0.5788591
QDA - Normal	0.8818470	0.9663371	0.2796421
QDA - Upsampled	0.7545417	0.7871940	0.5218121

Model	Accuracy	Good_Buy_Accuracy	Bad_Buy_Accuracy
QDA - Downsampled	0.7522020	0.7836629	0.5279642

Key Takeaways:

- The models generally performed well in predicting “Good Buy” (non-defective cars), with **Decision Tree (Normal)** achieving the highest Good Buy accuracy of **99.3%**
- Predicting “Bad Buy” (defective cars) was challenging across all models, with the highest accuracy observed in **Decision Tree (Upsampled)** and **Decision Tree (Downsampled)**, achieving around **69%**.
- **Upsampling** and **downsampling** led to significant improvements in Bad Buy accuracy at the expense of slightly reduced Good Buy accuracy.
- If the primary focus is on identifying non-defective cars (Good Buys), then **Decision Tree (Normal)** is the best choice.
- If equal importance is to be given to both Bad Buy and Good Buys then upsampled LDA or QDA will be better.

Association Rule Minning

Data Preprocessing

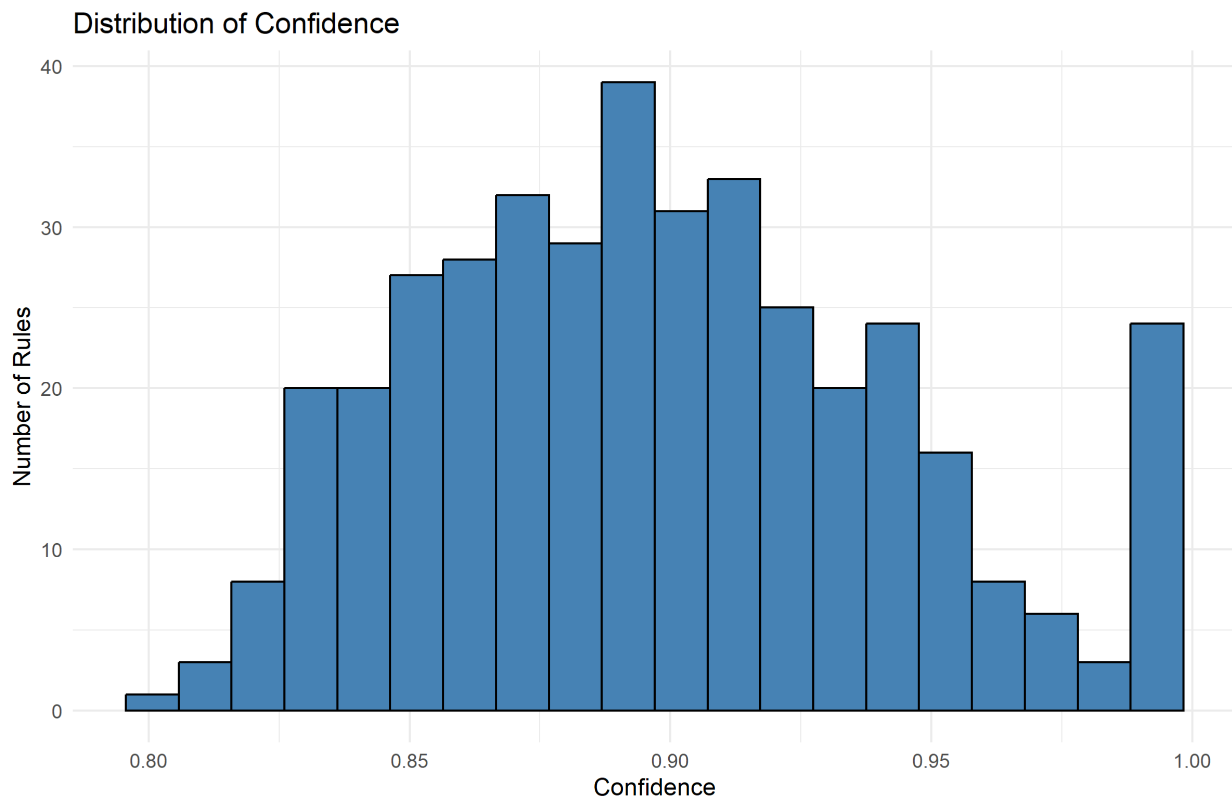
The data was standardized using Z-score normalization and PCA was applied to reduce dimensionality. The first principal component (PC1) was extracted to represent the combined effect of MMRA and MMRC-related variables, simplifying the dataset while retaining key information.

```
Importance of components:
                          PC1     PC2     PC3     PC4    PC5     PC6     PC7
Standard deviation     2.7154 0.54243 0.50019 0.20913 0.1697 0.07745 0.04834
Proportion of Variance 0.9217 0.03678 0.03127 0.00547 0.0036 0.00075 0.00029
Cumulative Proportion  0.9217 0.95847 0.98974 0.99521 0.9988 0.99956 0.99985
                           PC8
Standard deviation     0.03453
Proportion of Variance 0.00015
Cumulative Proportion  1.00000
```

Continuous variables such as PC1, VehOdo, VehBCost, and WarrantyCost were discretized into bins using the cut() function to create categorical data, a requirement for association rule mining.

```
transactions as itemMatrix in sparse format with
 72668 rows (elements/itemsets/transactions) and
 1257 columns (items) and a density of 0.009546539

most frequent items:
  VehBCost=(-44.5,9.09e+03]                  IsBadBuy=0
                      66016                       63724
WarrantyCost=(455,1.87e+03]          PC1=(-7.16,0.304]
                      61714                       38230
 VehOdo=(7.14e+04,9.35e+04]                     (Other)
                      37318                      605014

element (itemset/transaction) length distribution:
sizes
   12
72668

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     12      12      12      12      12      12

includes extended item information - examples:
        labels  variables levels
1 VehicleAge=0 VehicleAge      0
2 VehicleAge=1 VehicleAge      1
3 VehicleAge=2 VehicleAge      2

includes extended transaction information - examples:
  transactionID
1             1
2             2
3             3
```

All relevant columns were converted to factors to ensure compatibility with the arules package.

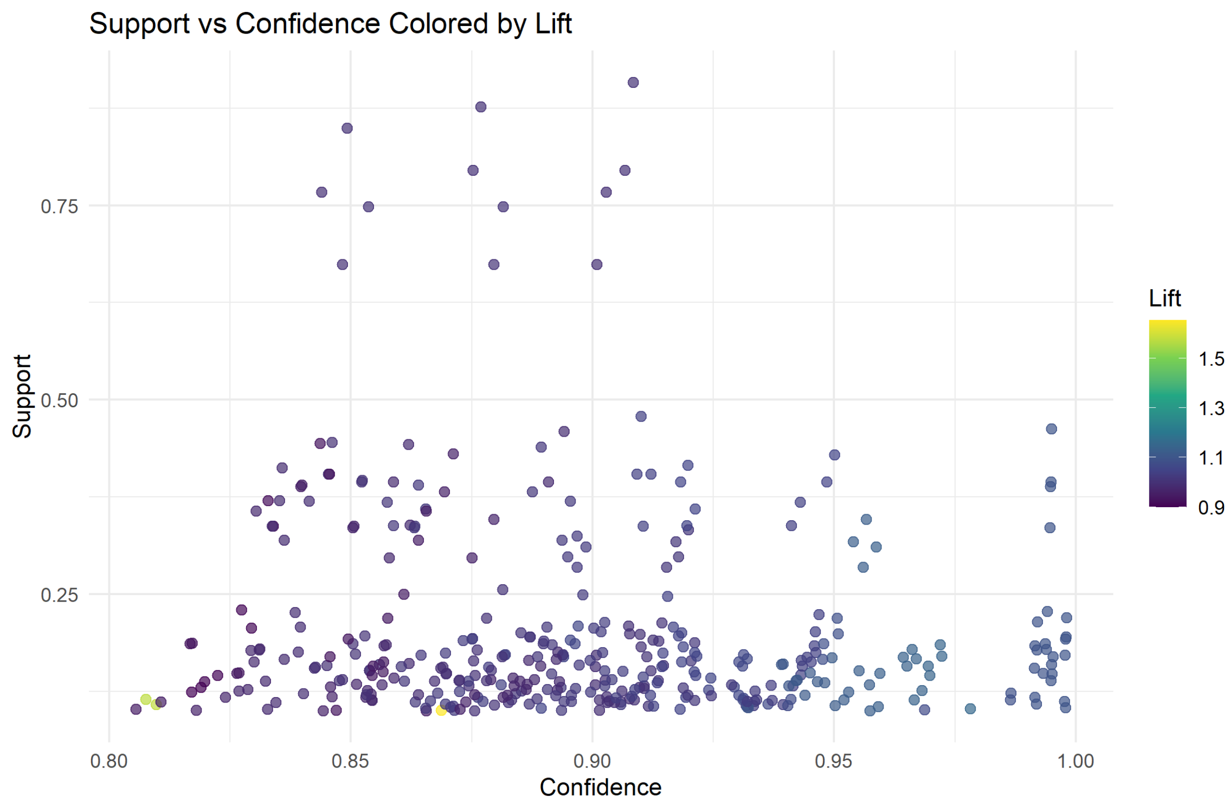


No Bad Buy Rules Found

- The dataset contained no strong rules predicting `{IsBadBuy=1}`. This suggests that the "bad buy" patterns are either too rare or not strongly associated with specific attribute combinations.

Good Buy Patterns

- Vehicles with **WheelType = Covers**, high **PC1 values**, and very low **VehBCost** are strongly associated with being good buys. For example:
  - `{WheelType=Covers, PC1=High, VehBCost=Very Low} → {IsBadBuy=0}`
    Confidence = 0.95, Lift = 1.08.
- Vehicles with **newer ages** (e.g., `VehicleAge = 2 or 3`) and specific makes (e.g., **CHEVROLET**) with attributes like `WheelType = Covers` tend to be good buys:
  - `{VehicleAge=2} → {IsBadBuy=0}`
    Confidence = 0.936, Lift = 1.067.



Impact of PC1

High **PC1 values**, representing combined MMRA and MMRC factors, strongly correlate with good buys: - `{PC1=High} → {IsBadBuy=0}`
Support = 0.478, Confidence = 0.91, Lift = 1.037.

Warranty Cost Insights

Vehicles with very low **warranty costs** are more likely to be good buys, suggesting lower anticipated maintenance costs: - `{WheelType=Covers, WarrantyCost=Very Low} → {IsBadBuy=0}`
Confidence = 0.92, Lift = 1.05.

Role of Odometer Reading

Vehicles with **medium odometer readings** are often associated with lower warranty costs, reflecting a sweet spot where the vehicles are neither too new nor too old: - `{VehOdo=Medium, WarrantyCost=Very Low} → {IsBadBuy=0}`
Confidence = 0.898, Lift = 1.024.

Reference

- Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.
- T. Hastie, R. Tibshirani and J, Friedman: The elements of statistical learning: Data Mining, Inference and Prediction; Springer Series in Statistics, Springer.
- MTH443 Lecture Notes, by Prof. Amit Mitra.