# DSCI6659 Project

Chiaki Ikeda

2024-04-18

# DSCI 6659 Project Report

## April 18, 2024

## Chiaki Ikeda, Haizhou Yuan, Mauricio Gomez Macedo, Michael Bimal, and Chen Chen

# 1. Introduction

In this project, our objective is to analyze car sales data using methods covered in our class, focusing on data analysis and machine learning techniques. We aim to predict a target of our interests based on various explanatory variables present in the data set obtained from Kaggle (https://www.kaggle.com/datasets/missionjee/car-sales-report (https://www.kaggle.com/datasets/missionjee/car-sales-report)).

The data set includes 23,906 observations over two years, from 2022 to 2023, and includes 16 variables related to car sales transactions.

Our analysis will focus on key aspects following:

1. Customers' Perspective: We will explore how customer income correlates with car sales, identify favorite car companies among customers, and analyze the impact of gender on car sales.
2. Dealers' Perspective: We will examine regional variations in car sales performance to provide insights for dealers about what to expect for the upcoming year.
3. Machine Learning Applications: Using machine learning techniques, we will identify the key factors influencing car prices and sales, contributing to a better understanding of market trends. By addressing these questions, we aim to provide insights into the car sales market.

```
##         Car_id     Date Customer.Name Gender Annual.Income
## 1 C_CND_000001 1/2/2022     Geraldine   Male         13500
## 2 C_CND_000002 1/2/2022           Gia   Male       1480000
## 3 C_CND_000003 1/2/2022        Gianna   Male       1035000
## 4 C_CND_000004 1/2/2022       Giselle   Male         13500
## 5 C_CND_000005 1/2/2022         Grace   Male       1465000
## 6 C_CND_000006 1/2/2022     Guadalupe   Male        850000
##                           Dealer_Name    Company      Model
## 1 Buddy Storbeck's Diesel Service Inc       Ford Expedition
## 2                     C & M Motors Inc      Dodge    Durango
## 3                          Capitol KIA   Cadillac   Eldorado
## 4              Chrysler of Tri-Cities     Toyota     Celica
## 5                   Chrysler Plymouth      Acura         TL
## 6                       Classic Chevy Mitsubishi   Diamante
##                       Engine Transmission      Color Price  Dealer_No
## 1 DoubleÂ Overhead Camshaft         Auto      Black 26000 06457-3834
## 2 DoubleÂ Overhead Camshaft         Auto      Black 19000 60504-7114
## 3          Overhead Camshaft       Manual        Red 31500 38701-8047
## 4          Overhead Camshaft       Manual Pale White 14000 99301-3882
## 5 DoubleÂ Overhead Camshaft         Auto        Red 24500 53546-9427
## 6          Overhead Camshaft       Manual Pale White 12000 85257-3102
##   Body.Style   Phone Dealer_Region
## 1        SUV 8264678    Middletown
## 2        SUV 6848189        Aurora
## 3  Passenger 7298798    Greenville
## 4        SUV 6257557         Pasco
## 5  Hatchback 7081483    Janesville
## 6  Hatchback 7315216    Scottsdale
```

Project contribution:

We have five members in our team, each contributing equally to the project. The task allocations are as follows.

Task 1 (Exploratory Data Analysis (EDA), Preprocessing, and Robust PCA Analysis): Chiaki Ikeda

Task 2 (Customer Segmentation, Profile Analysis, and PCA Regression): Haizhou Yuan

Task 3 (Sales Prediction, Dealership Performance Analysis, and Ensemble Learning) : Mauricio Gomez Macedo

Task 4 (Robust Discriminant Analysis, Sentiment Analysis, and Penalized Linear Discriminant Analysis): Michael Bimal

Task 5 (Visualization, Reporting, Interactive Dashboard Development, and Discriminant Adaptive Nearest Neighbor Classification): Chen Chen

# 2. Data Analysis - Task 1

## 2-1. Data Overview (before data cleaning)

In this data set, 23906 observations with 16 variables are included.There is no missing values detected. Except for “Annual.Income”, “Dealer_No”, and “Price”, the rest of variables are character.

For data analysis, we do not use identification information, such as “Car_id”, “Customer.Name”, “Dealer_Name”,“Dealer_No”, and “Phone”.

```
##  [1] "Car_id"        "Date"          "Customer.Name" "Gender"
##  [5] "Annual.Income" "Dealer_Name"   "Company"       "Model"
##  [9] "Engine"        "Transmission"  "Color"         "Price"
## [13] "Dealer_No"     "Body.Style"    "Phone"         "Dealer_Region"
```

```
## 'data.frame':    23906 obs. of  16 variables:
##  $ Car_id       : chr  "C_CND_000001" "C_CND_000002" "C_CND_000003" "C_CND_000004" ...
##  $ Date         : chr  "1/2/2022" "1/2/2022" "1/2/2022" "1/2/2022" ...
##  $ Customer.Name: chr  "Geraldine" "Gia" "Gianna" "Giselle" ...
##  $ Gender       : chr  "Male" "Male" "Male" "Male" ...
##  $ Annual.Income: int  13500 1480000 1035000 13500 1465000 850000 1600000 13500 815000 13500
...
##  $ Dealer_Name  : chr  "Buddy Storbeck's Diesel Service Inc" "C & M Motors Inc" "Capitol KI
A" "Chrysler of Tri-Cities" ...
##  $ Company      : chr  "Ford" "Dodge" "Cadillac" "Toyota" ...
##  $ Model        : chr  "Expedition" "Durango" "Eldorado" "Celica" ...
##  $ Engine       : chr  "DoubleÂ Overhead Camshaft" "DoubleÂ Overhead Camshaft" "Overhead Cam
shaft" "Overhead Camshaft" ...
##  $ Transmission : chr  "Auto" "Auto" "Manual" "Manual" ...
##  $ Color        : chr  "Black" "Black" "Red" "Pale White" ...
##  $ Price        : int  26000 19000 31500 14000 24500 12000 14000 42000 82000 15000 ...
##  $ Dealer_No    : chr  "06457-3834" "60504-7114" "38701-8047" "99301-3882" ...
##  $ Body.Style   : chr  "SUV" "SUV" "Passenger" "SUV" ...
##  $ Phone        : int  8264678 6848189 7298798 6257557 7081483 7315216 7727879 6206512 71948
57 7836892 ...
##  $ Dealer_Region: chr  "Middletown" "Aurora" "Greenville" "Pasco" ...
```

```
##     Car_id             Date           Customer.Name        Gender
##  Length:23906       Length:23906       Length:23906       Length:23906
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Annual.Income      Dealer_Name          Company            Model
##  Min.    :   10080   Length:23906       Length:23906       Length:23906
##  1st Qu.:  386000   Class :character   Class :character   Class :character
##  Median :  735000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :  830840
##  3rd Qu.: 1175750
##  Max.    :11200000
##     Engine          Transmission          Color              Price
##  Length:23906       Length:23906       Length:23906       Min.    : 1200
##  Class :character   Class :character   Class :character   1st Qu.:18001
##  Mode  :character   Mode  :character   Mode  :character   Median :23000
##                                                           Mean   :28090
##                                                           3rd Qu.:34000
##                                                           Max.    :85800
##   Dealer_No          Body.Style            Phone          Dealer_Region
##  Length:23906       Length:23906       Min.    :6000101   Length:23906
##  Class :character   Class :character   1st Qu.:6746495   Class :character
##  Mode  :character   Mode  :character   Median :7496198   Mode  :character
##                                        Mean   :7497741
##                                        3rd Qu.:8248146
##                                        Max.    :8999579
```

# 2-2. Interpretation of each variable and data preprocessing

We selected variables that are our interests, keeping car_id for the data integration.

```
##     Car_id             Date              Gender          Annual.Income
##  Length:23906       Length:23906       Length:23906       Min.    :   10080
##  Class :character   Class :character   Class :character   1st Qu.:  386000
##  Mode  :character   Mode  :character   Mode  :character   Median :  735000
##                                                           Mean   :  830840
##                                                           3rd Qu.: 1175750
##                                                           Max.    :11200000
##    Company             Model              Engine          Transmission
##  Length:23906       Length:23906       Length:23906       Length:23906
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Color              Price         Body.Style        Dealer_Region
##  Length:23906       Min.    : 1200   Length:23906       Length:23906
##  Class :character   1st Qu.:18001   Class :character   Class :character
##  Mode  :character   Median :23000   Mode  :character   Mode  :character
##                     Mean   :28090
##                     3rd Qu.:34000
##                     Max.    :85800
```

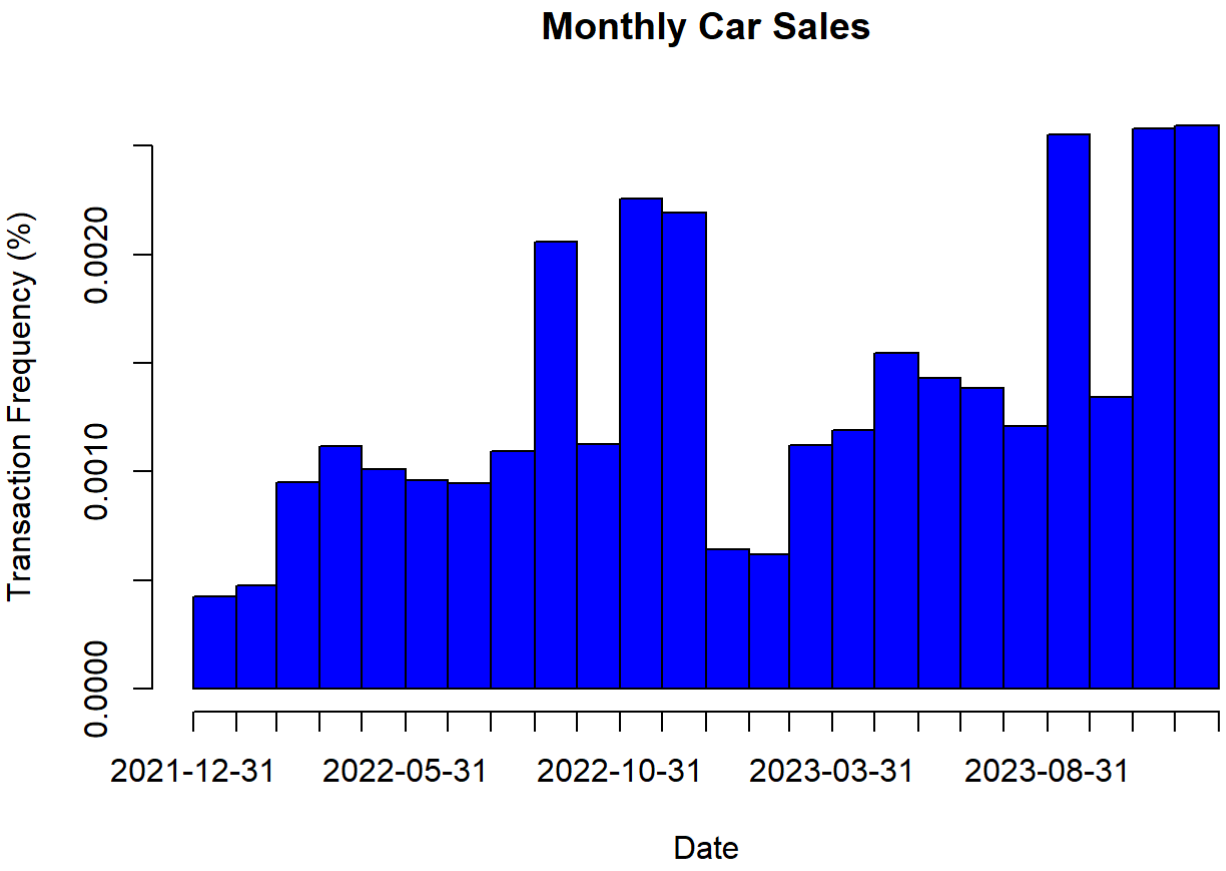
In the selected data, now we have 12 variables including “Car_id”, which refers to unique identifier for each car in the data set.

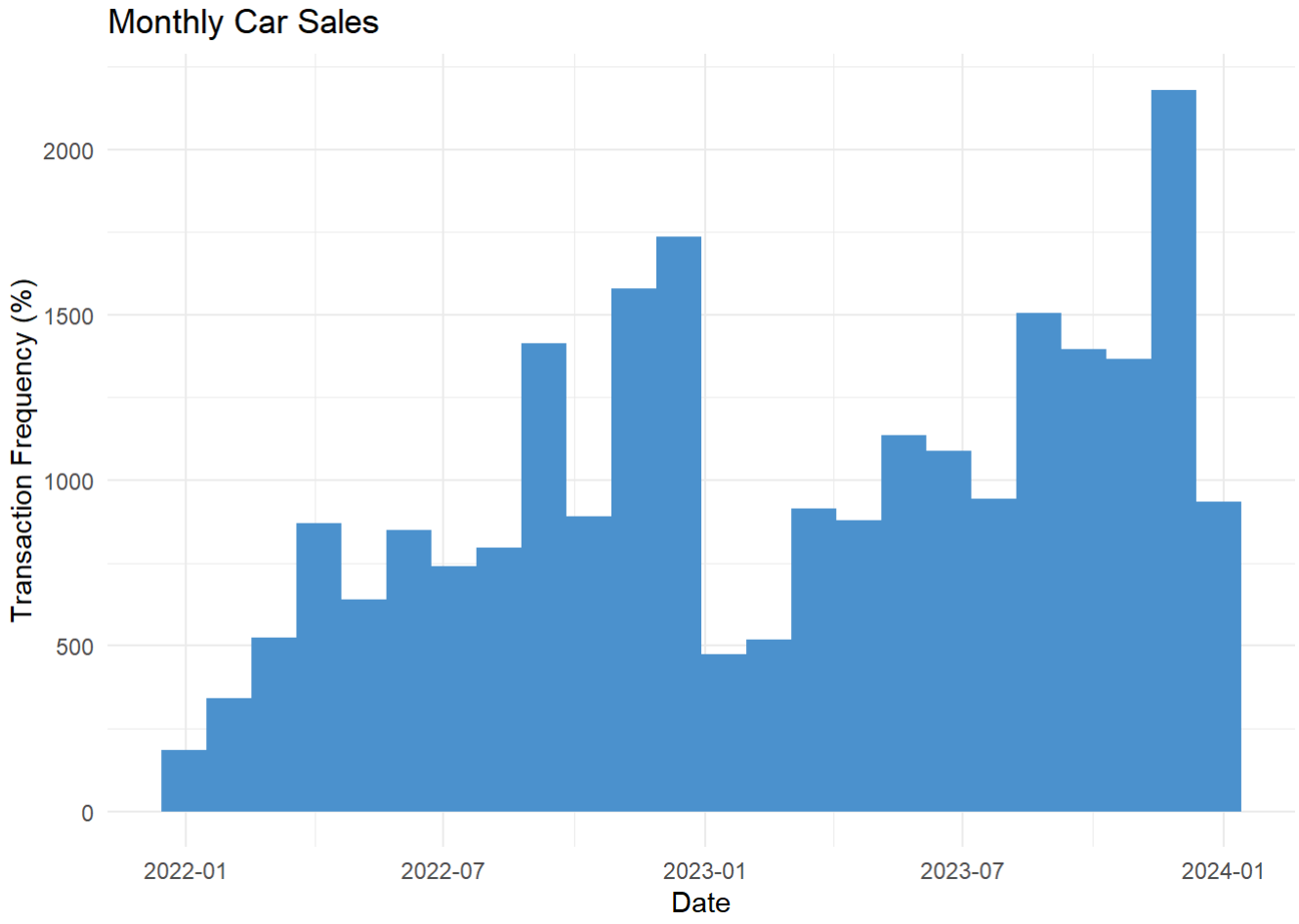
```
##  [1] "Car_id"        "Date"          "Gender"        "Annual.Income"
##  [5] "Company"       "Model"         "Engine"        "Transmission"
##  [9] "Color"         "Price"         "Body.Style"    "Dealer_Region"
```

## V2: “Date” - Date of car sales transaction

This data starts from January 2, 2022 to December 31, 2023, which is the daily car sale transaction over two years. Each bar represents the number of car sales in the month. As shown in the histogram below, car sales transaction increases in September, November, and December both in 2022 and 2023.

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2022-01-02" "2022-09-20" "2023-03-13" "2023-03-01" "2023-09-08" "2023-12-31"
```



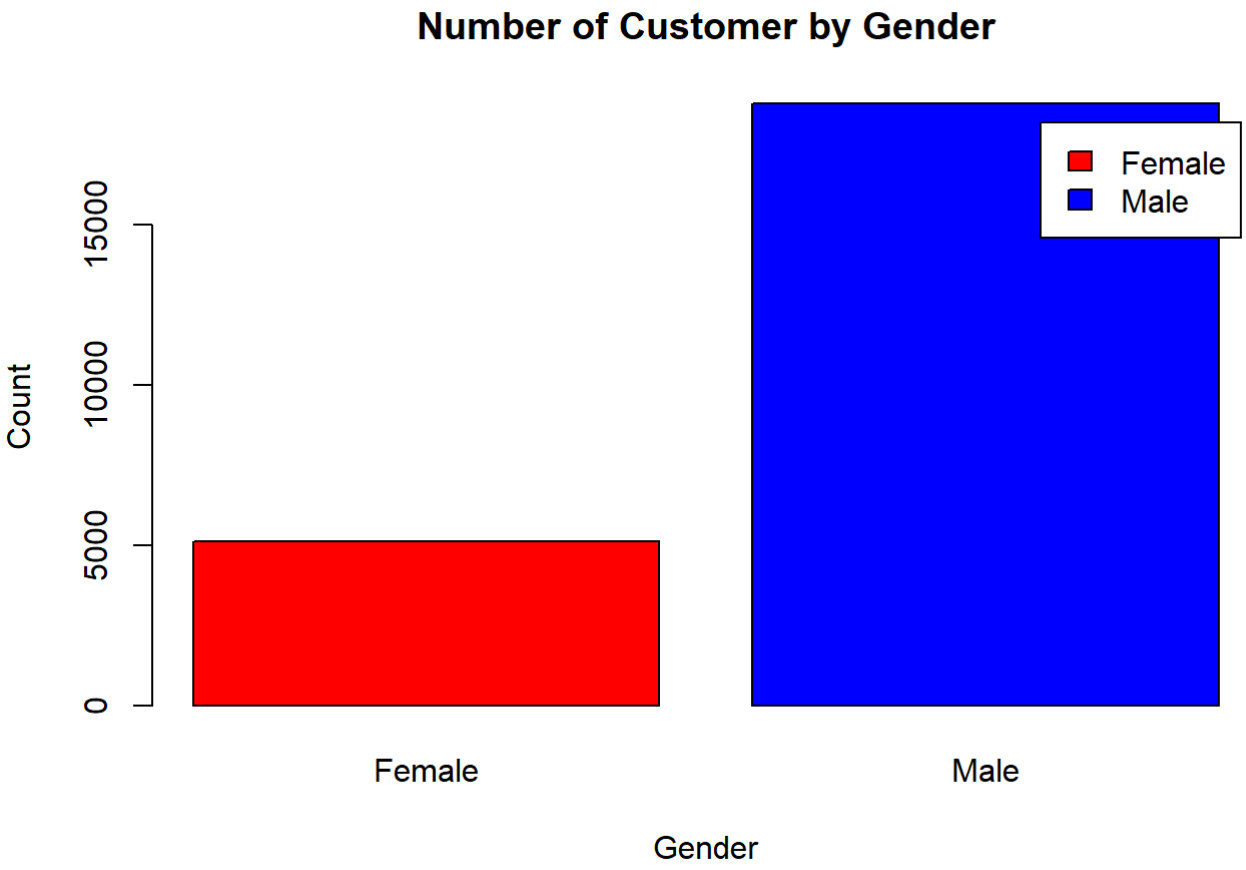

## V3: “Gender” - Gender of the customer, either Male or Female

Gender is a categorical variable including two classes, “Male” and “Female”. In this data set, male customers account for 79% (21% for females) of total car sales transactions.

```
## [1] Male   Female
## Levels: Female Male
```

```
## Female   Male
##   5108  18798
```

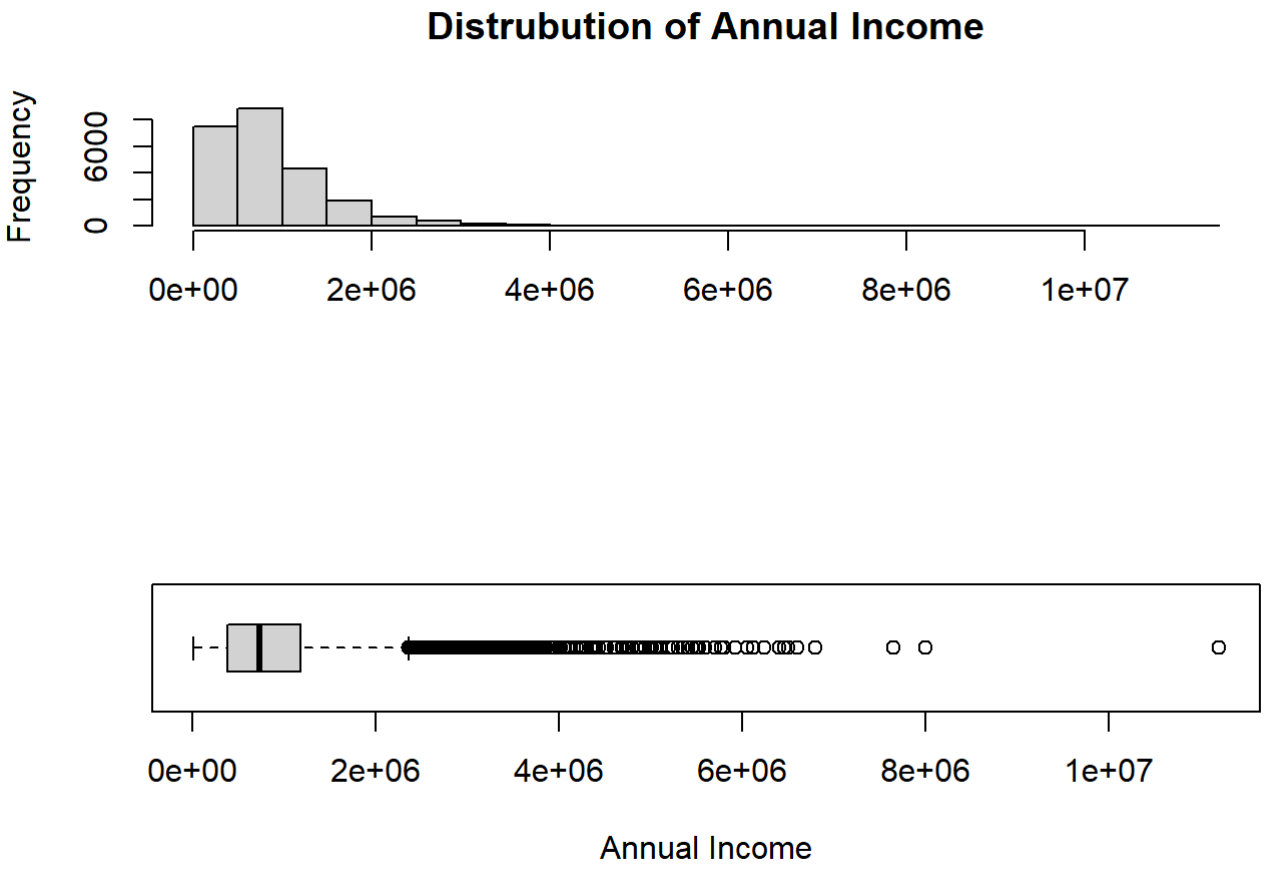
```
## Female %   Male %
## 21.36702 78.63298
```




## V4: “Annual.Income” - Annual Income of Customers ($)

The annual income of customers are in the range from 10.1 k to 11.2 million dollars. Overall, the distribution is right-skewed, suggesting that more observations fall in lower incomes.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    10080   386000   735000   830840  1175750 11200000
```




## V5: “Company” - Company or Brand of the Car Purchased

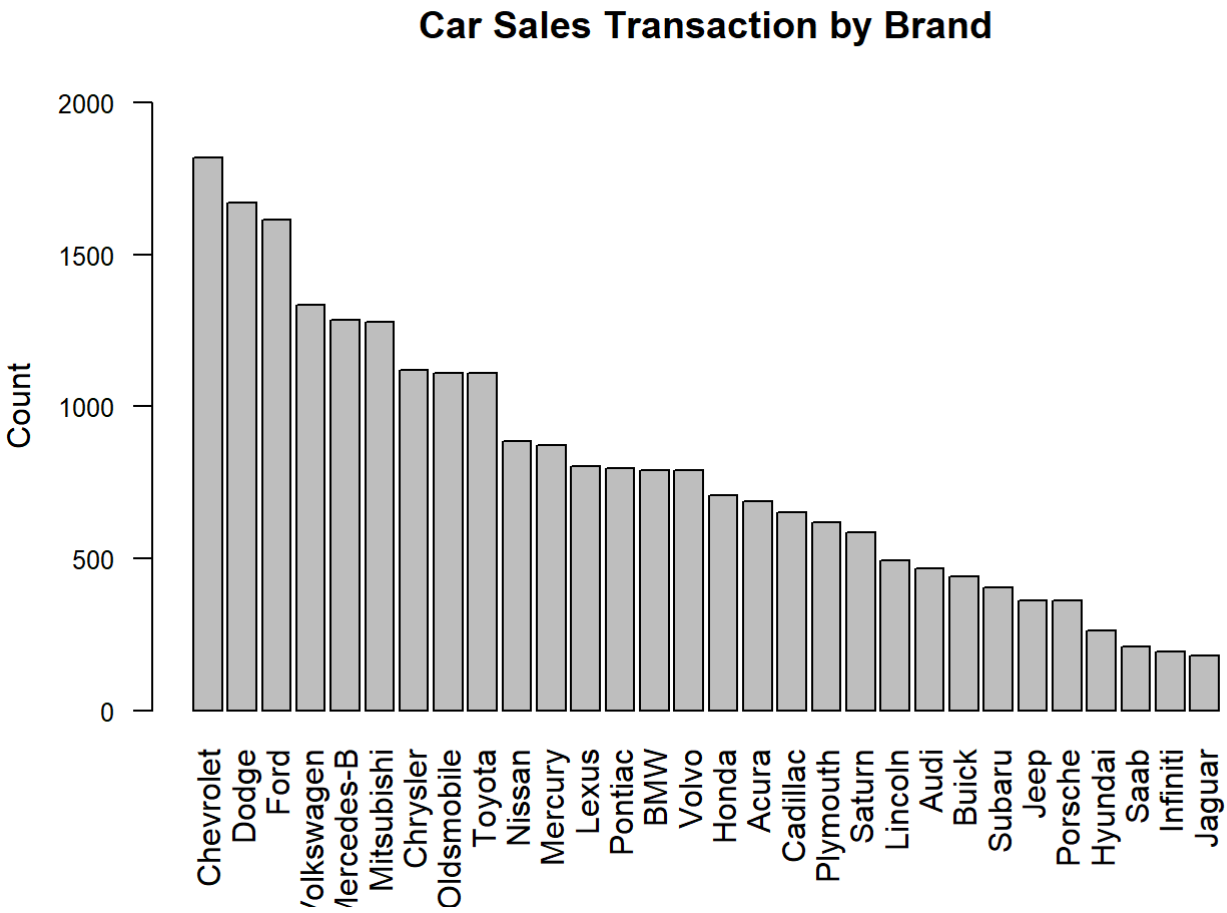
In this variable, we have 30 brands/companies as shown below. Chevrolet seems to be the most popular brand accounting for 7.6%, followed by Dodge (7.0%) and Ford (6.8%).

```
## [1] 30
```

```
##  [1] Ford       Dodge      Cadillac   Toyota     Acura      Mitsubishi
##  [7] Chevrolet  Nissan     Mercury    BMW        Chrysler   Subaru
## [13] Hyundai    Honda      Infiniti   Audi       Porsche    Volkswagen
## [19] Buick      Saturn     Mercedes-B Jaguar     Volvo      Pontiac
## [25] Lincoln    Oldsmobile Lexus      Plymouth   Saab       Jeep
## 30 Levels: Acura Audi BMW Buick Cadillac Chevrolet Chrysler Dodge ... Volvo
```

```
##  Chevrolet      Dodge       Ford Volkswagen Mercedes-B Mitsubishi   Chrysler
##       1819       1671       1614       1333       1285       1277       1120
## Oldsmobile     Toyota     Nissan    Mercury      Lexus    Pontiac        BMW
##       1111       1110        886        874        802        796        790
##      Volvo      Honda      Acura   Cadillac   Plymouth     Saturn    Lincoln
##        789        708        689        652        617        586        492
##       Audi      Buick     Subaru       Jeep    Porsche    Hyundai       Saab
##        468        439        405        363        361        264        210
##   Infiniti     Jaguar
##        195        180
```

```
##  Chevrolet      Dodge       Ford Volkswagen Mercedes-B Mitsubishi    Chrysler
##  7.6089685  6.9898770  6.7514432  5.5760060  5.3752196  5.3417552  4.6850163
## Oldsmobile     Toyota     Nissan    Mercury      Lexus    Pontiac        BMW
##  4.6473689  4.6431858  3.7061825  3.6559859  3.3548063  3.3297080  3.3046097
##       Volvo      Honda      Acura   Cadillac   Plymouth     Saturn    Lincoln
##  3.3004267  2.9615996  2.8821216  2.7273488  2.5809420  2.4512675  2.0580607
##        Audi      Buick     Subaru       Jeep    Porsche    Hyundai       Saab
##  1.9576675  1.8363591  1.6941354  1.5184473  1.5100812  1.1043253  0.8784406
##    Infiniti     Jaguar
##  0.8156948  0.7529491
```

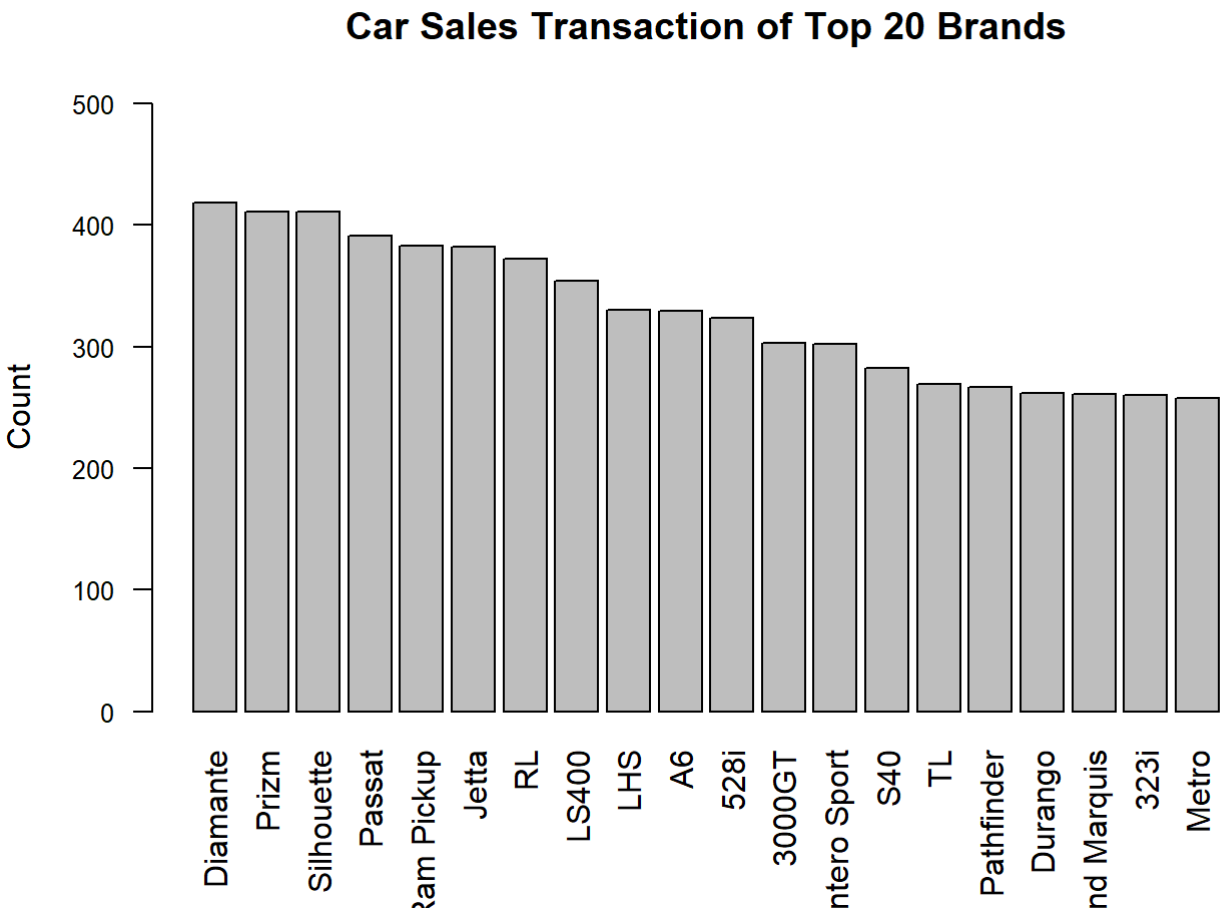



## V6: “Model” - Model Name of the Car

This is a categorical variable of model of the car, including 154 different models. The most popular model is Diamante accounting for about 1.75%, followed by Prizm and Silhouette (1.72% for both).

```
## [1] 154
```

```
##   [1] Expedition     Durango        Eldorado       Celica         TL
##   [6] Diamante       Corolla        Galant         Malibu         Escort
##  [11] RL             Pathfinder     Grand Marquis  323i           Sebring Coupe
##  [16] Forester       Accent         Land Cruiser   Accord         4Runner
##  [21] I30            A4             Carrera Cabrio Jetta          Viper
##  [26] Regal          LHS            LW             3000GT         SLK230
##  [31] Civic          S-Type         S40            Mountaineer    Park Avenue
##  [36] Montero Sport  Sentra         S80            Lumina         Bonneville
##  [41] C-Class        Altima         DeVille        Stratus        Cougar
##  [46] SW             C70            SLK            Tacoma         M-Class
##  [51] A6             Intrepid       Sienna         Eclipse        Contour
##  [56] Town car       Focus          Mustang        Cutlass        Corvette
##  [61] Impala         Cabrio         Dakota         300M           328i
##  [66] Bravada        Maxima         Ram Pickup     Concorde       V70
##  [71] Quest          ES300          SL-Class       Explorer       Prizm
##  [76] Camaro         Outback        Taurus         Cavalier       GS400
##  [81] Monte Carlo    Sonata         Sable          Metro          Voyager
##  [86] Cirrus         Avenger        Odyssey        Intrigue       Silhouette
##  [91] 5-Sep          528i           LS400          Aurora         Breeze
##  [96] Beetle         Elantra        Continental    RAV4           Villager
## [101] S70            LS             Ram Van        S-Class        E-Class
## [106] Grand Am       SC             Passat         Xterra         Frontier
## [111] Crown Victoria Camry          Navigator      CL500          Escalade
## [116] Golf           Ranger         Prowler        Windstar       GTI
## [121] Passport       Boxter         LX470          CR-V           Sunfire
## [126] Caravan        Ram Wagon      Neon           Wrangler       Integra
## [131] Grand Prix     Grand Cherokee F-Series       A8             Mystique
## [136] 3-Sep          Cherokee       Carrera Coupe  Catera         Seville
## [141] CLK Coupe      LeSabre        Sebring Conv.  GS300          Firebird
## [146] V40            Montero        Town & Country SL             Alero
## [151] Mirage         Century        RX300          Avalon
## 154 Levels: 3-Sep 3000GT 300M 323i 328i 4Runner 5-Sep 528i A4 A6 A8 ... Xterra
```




## V7: “Engine” - Specifications of the Car Engine

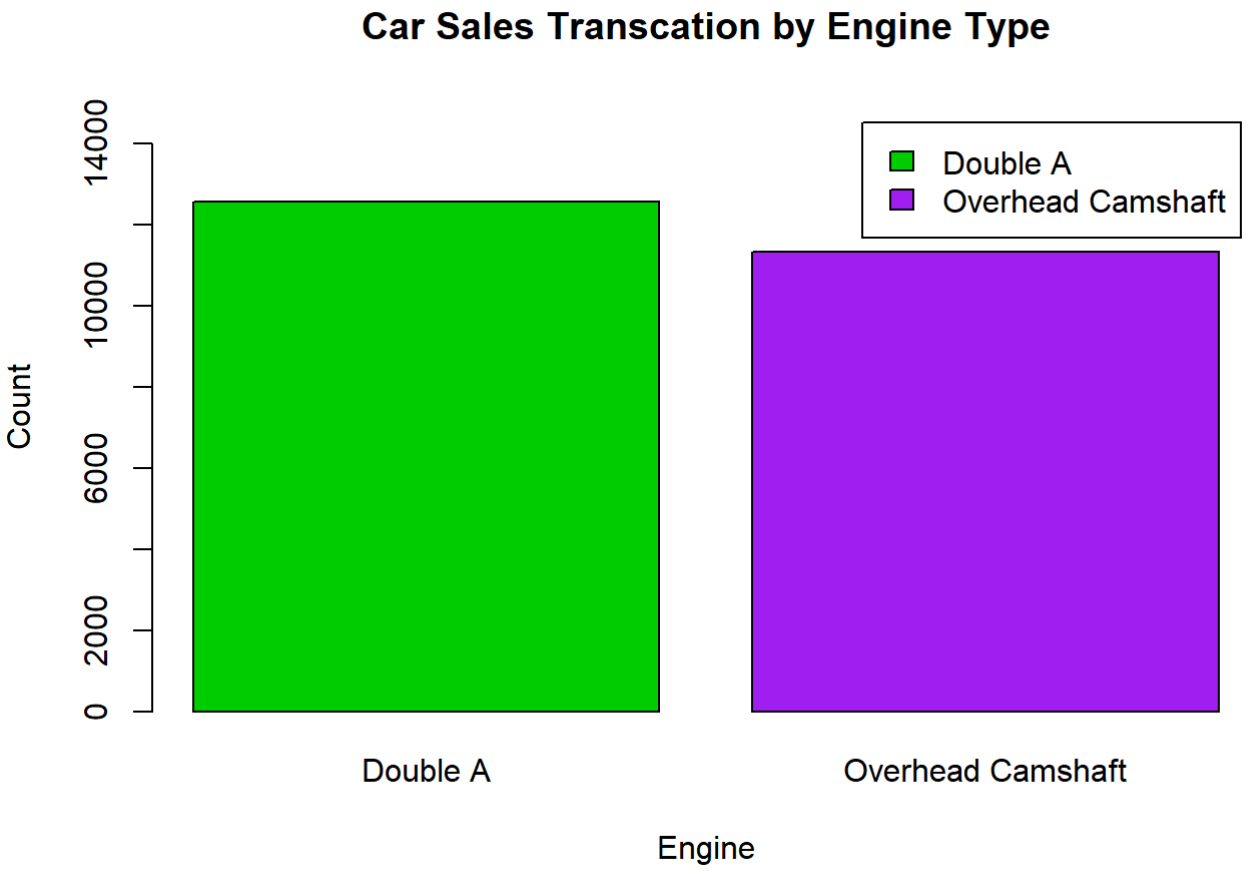
Variable “Engine” is also a categorical variable including 2 classes, either DoubleÃ Overhead Camshaft or Overhead Camshaft. In this modified data set, we call them “Double A” and “Overhead Camshaft”. Among two types of engines, Double A is slightly more popular than the other one. The percentages of Double A and Camshaft are 53% and 47%, respectively.

```
## [1] 2
```

```
##          Double A Overhead Camshaft
##                 1                 1
```

```
##          Double A Overhead Camshaft
##             12571             11335
```

```
##          Double A Overhead Camshaft
##          52.58513          47.41487
```



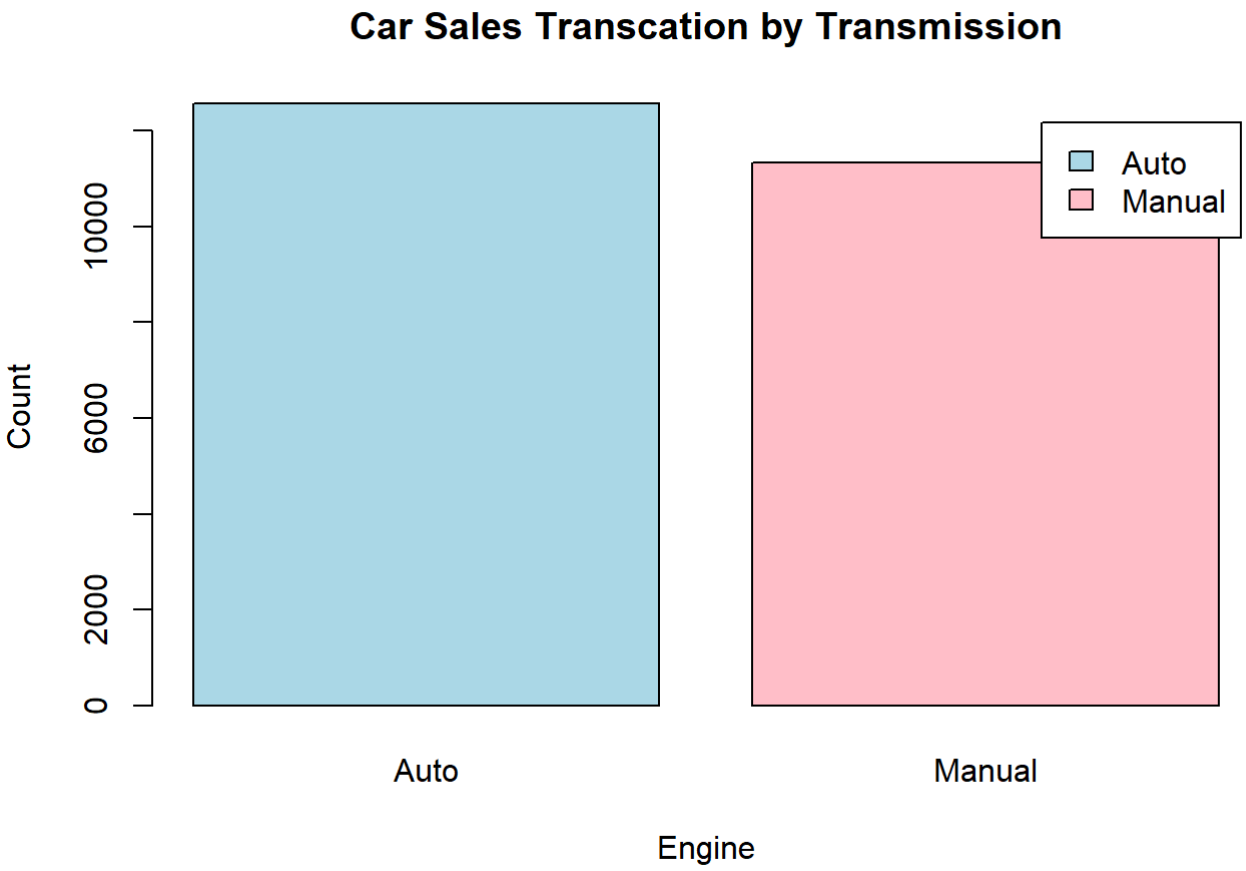

## V8: “Transmission” - Type of Transmission in the Car

Variable “Transmission” has a class of two type of transmission of the car in the data set, either Auto or Manual. Overall, Automatic cars are purchased more than Manual cars, accounting for 53% of the total transaction, while the rest of 47% is manual transmission cars.

```
## [1] "Auto"   "Manual"
```

```
##   Auto Manual
##  12571  11335
```

```
##     Auto   Manual
## 52.58513 47.41487
```



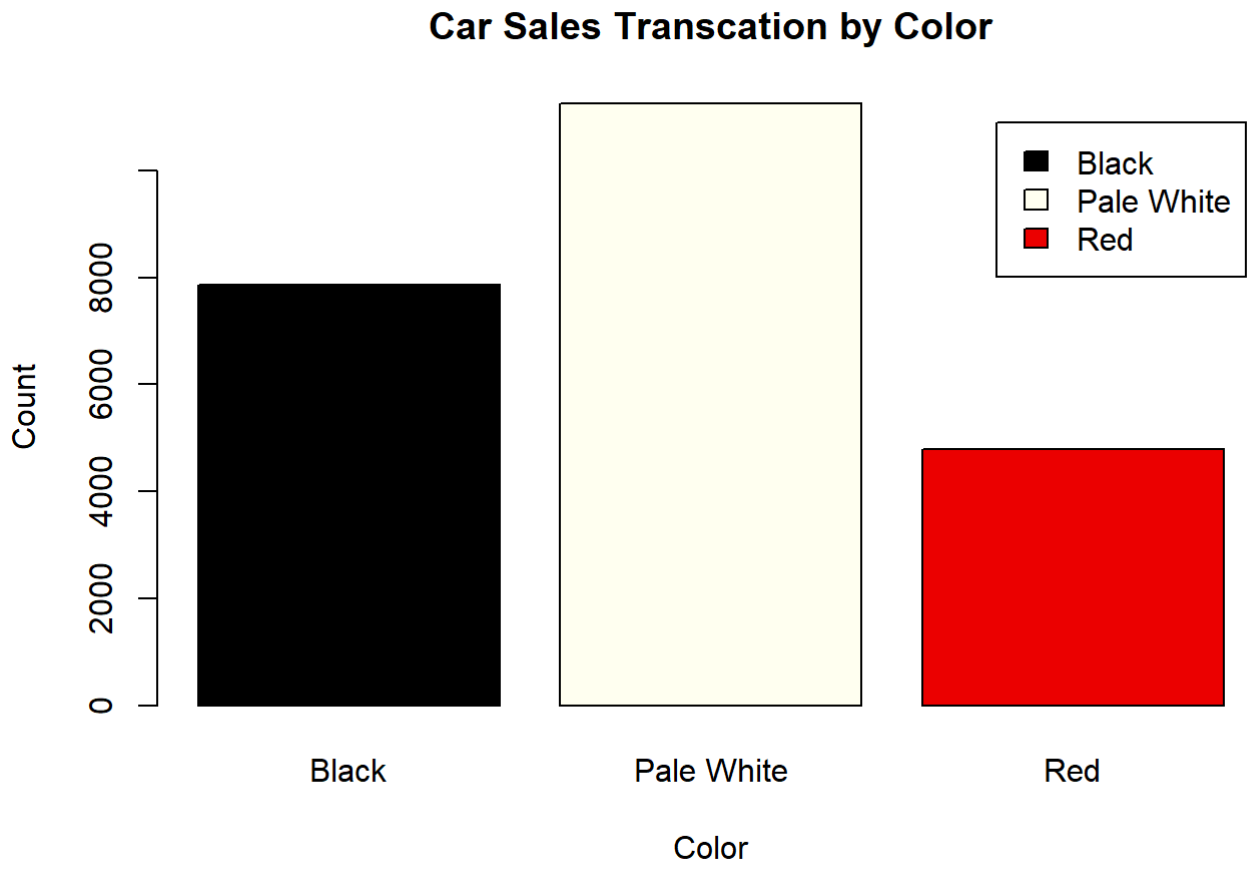

## V9: “Color” - Color of the Exterior of the Car

This variable refers to the color of the car, including 3 classes of black, red, and pale white.

```
## [1] "Black"      "Red"        "Pale White"
```

```
##      Black Pale White        Red
##       7857      11256       4793
```

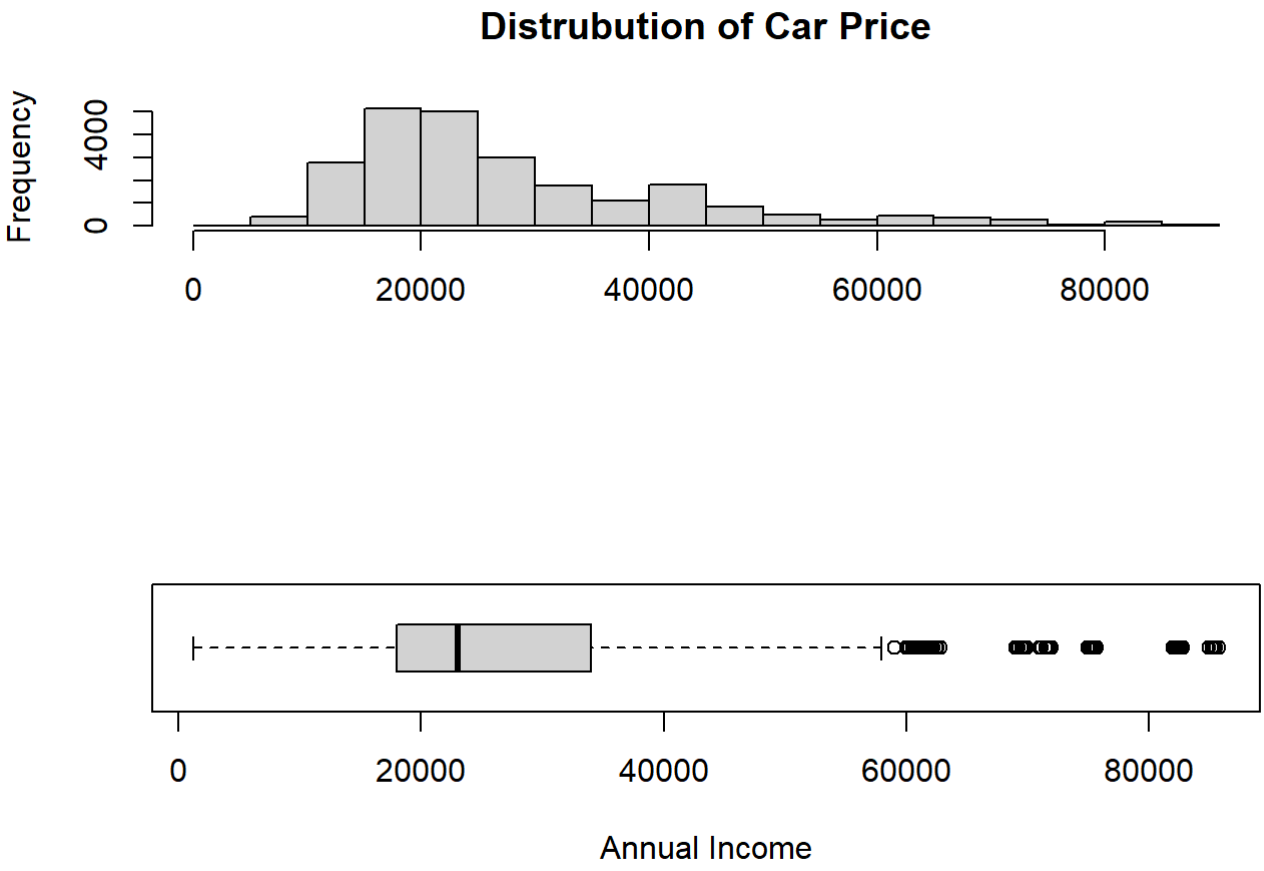
```
##      Black Pale White        Red
##   32.86623   47.08441   20.04936
```




## V10: “Price” - Listed Sales Price of the Car

The price of the car is in the range from 1.2k to 85.8 k dollars. The distribution is right-skewed, which is similar to that of the annual income of customers. The median is 23k and the mean price is 28 k.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1200   18001   23000   28090   34000   85800
```



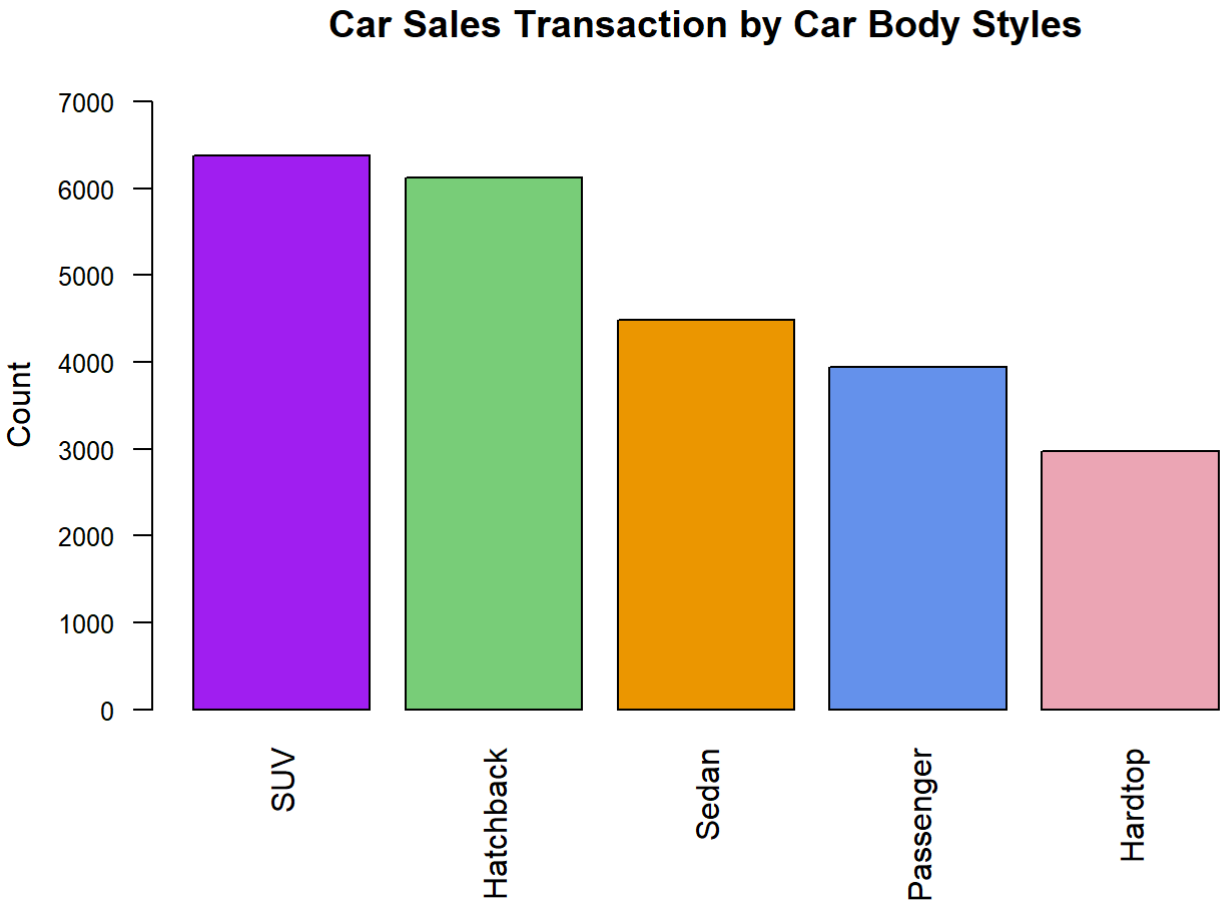

## V11: “Body.Style” - Style or Design of the Car’s Body

This variable represents car’s body style including 5 categories SUV, Passenger, Hatchback, Hardtop, and Sedan. Among these five classes, SUV and Hatchback are two top styles accounting for 27% and 26%, respectively. Passenger cars are the least popular, which are 17% of the total transaction.

```
## [1] "SUV"       "Passenger" "Hatchback" "Hardtop"   "Sedan"
```

```
##   Hardtop Hatchback Passenger     Sedan       SUV
##      2971      6128      3945      4488      6374
```

```
##   Hardtop Hatchback Passenger     Sedan       SUV
##  12.42784  25.63373  16.50213  18.77353  26.66276
```



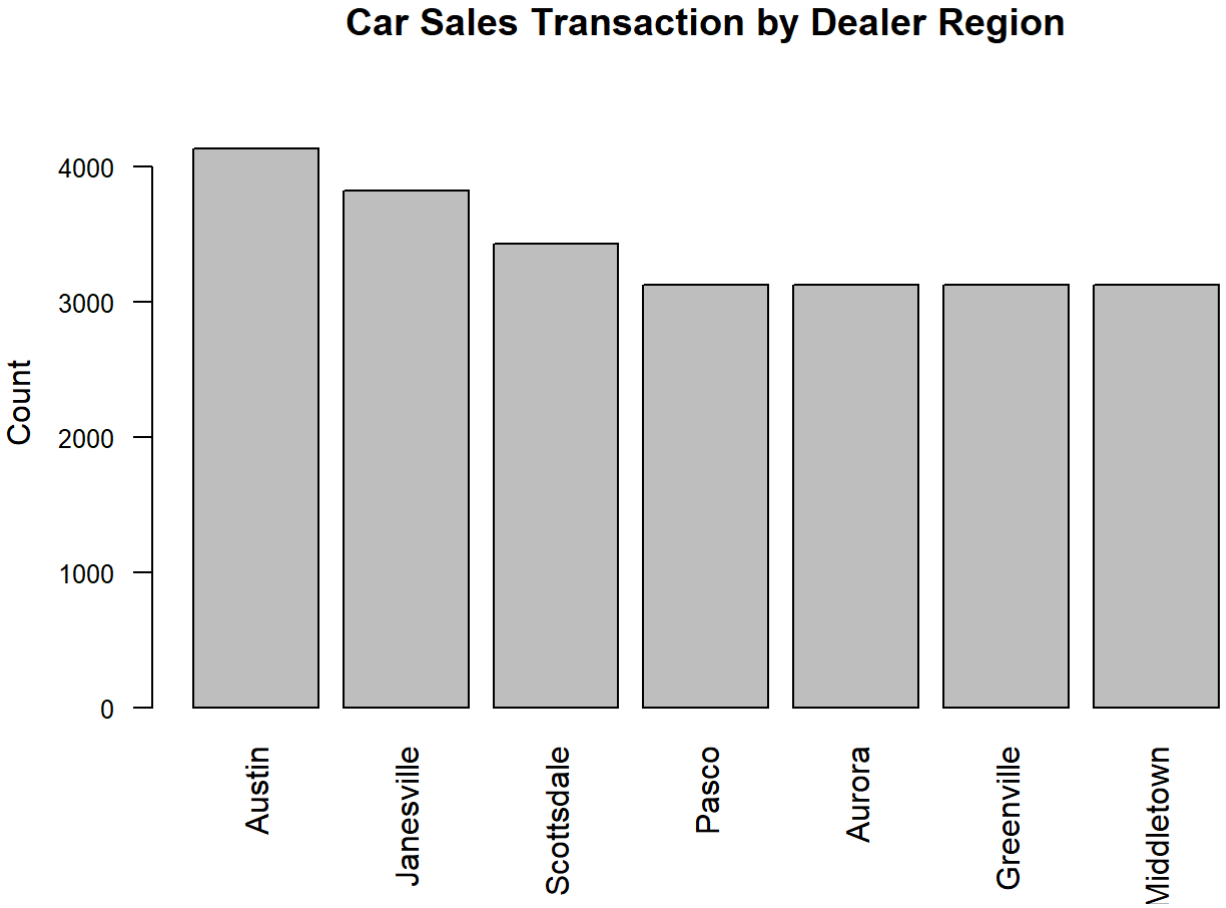

## V12: “Dealer_Region” - Geographic Region of the Car Dealer

The region includes seven areas including Middletown, Aurora, Greenville, Pasco, Janeville, Scottsdale, and Austin. The most common region is Austin accounting for 17%, followed by Janesville (16%) and the rest of regional rates are in the range from 13-14%.

```
## [1] "Middletown" "Aurora"     "Greenville" "Pasco"      "Janesville"
## [6] "Scottsdale" "Austin"
```

```
##     Aurora     Austin Greenville Janesville Middletown      Pasco Scottsdale
##       3130       4135       3128       3821       3128       3131       3433
```

```
##     Aurora     Austin Greenville Janesville Middletown      Pasco Scottsdale
##   13.09295   17.29691   13.08458   15.98344   13.08458   13.09713   14.36041
```




## Summary

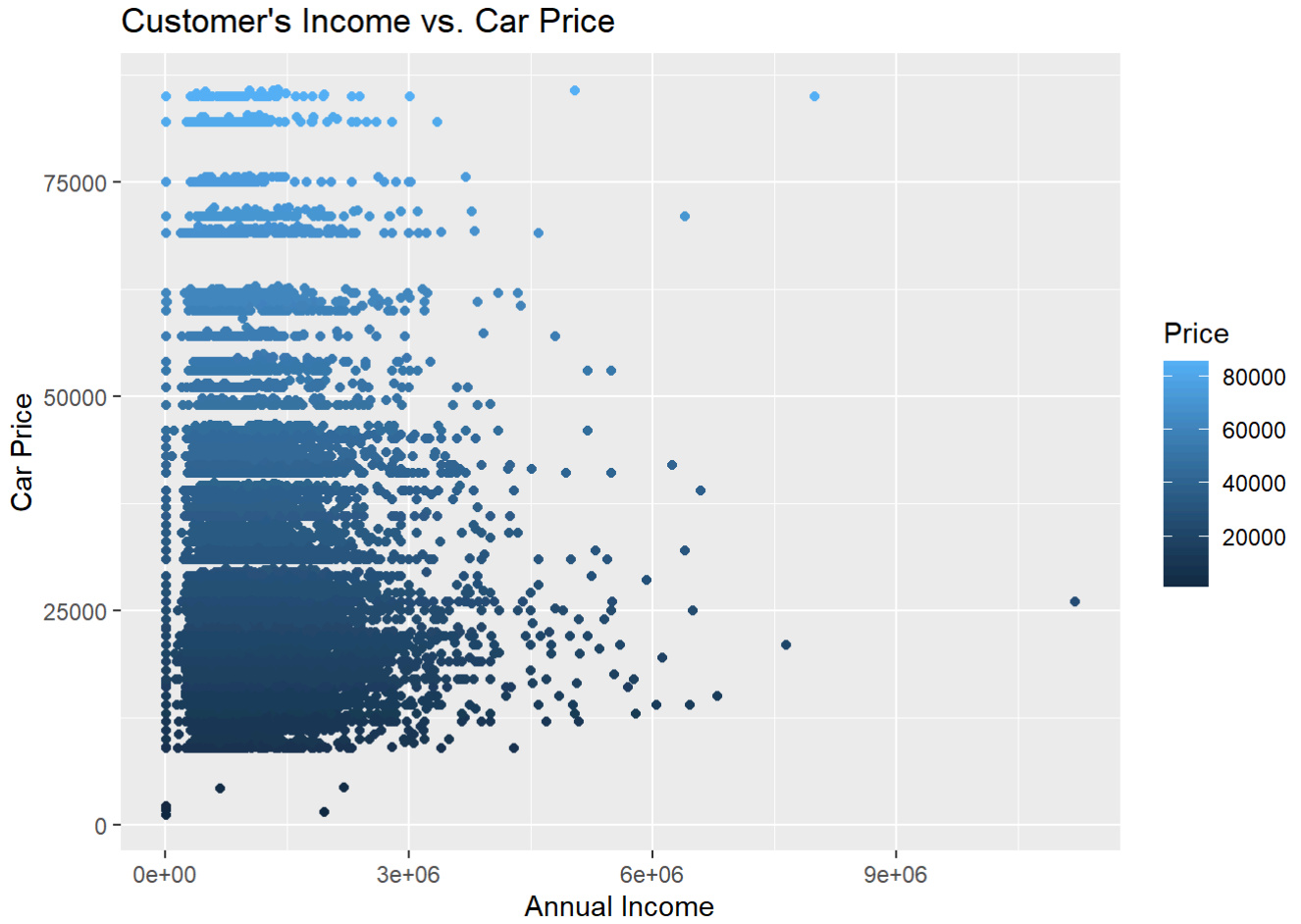
After the transformation of data type, the summary of the new data set is as shown below.

```
##     Car_id              Date               Gender      Annual.Income
##  Length:23906       Min.   :2022-01-02   Female: 5108   Min.   :   10080
##  Class :character   1st Qu.:2022-09-20   Male  :18798   1st Qu.:  386000
##  Mode  :character   Median :2023-03-13                  Median :  735000
##                     Mean   :2023-03-01                  Mean   :  830840
##                     3rd Qu.:2023-09-08                  3rd Qu.: 1175750
##                     Max.   :2023-12-31                  Max.   :11200000
##
##        Company             Model                      Engine      Transmission
##  Chevrolet : 1819   Diamante  :  418   Double A         :12571   Auto  :12571
##  Dodge     : 1671   Prizm     :  411   Overhead Camshaft:11335   Manual:11335
##  Ford      : 1614   Silhouette:  411
##  Volkswagen: 1333   Passat    :  391
##  Mercedes-B: 1285   Ram Pickup:  383
##  Mitsubishi: 1277   Jetta     :  382
##  (Other)   :14907   (Other)   :21510
##         Color           Price          Body.Style     Dealer_Region
##  Black     : 7857   Min.   : 1200   Hardtop  :2971   Aurora    :3130
##  Pale White:11256   1st Qu.:18001   Hatchback:6128   Austin    :4135
##  Red       : 4793   Median :23000   Passenger:3945   Greenville:3128
##                     Mean   :28090   Sedan    :4488   Janesville:3821
##                     3rd Qu.:34000   SUV      :6374   Middletown:3128
##                     Max.   :85800                    Pasco     :3131
##                                                      Scottsdale:3433
```

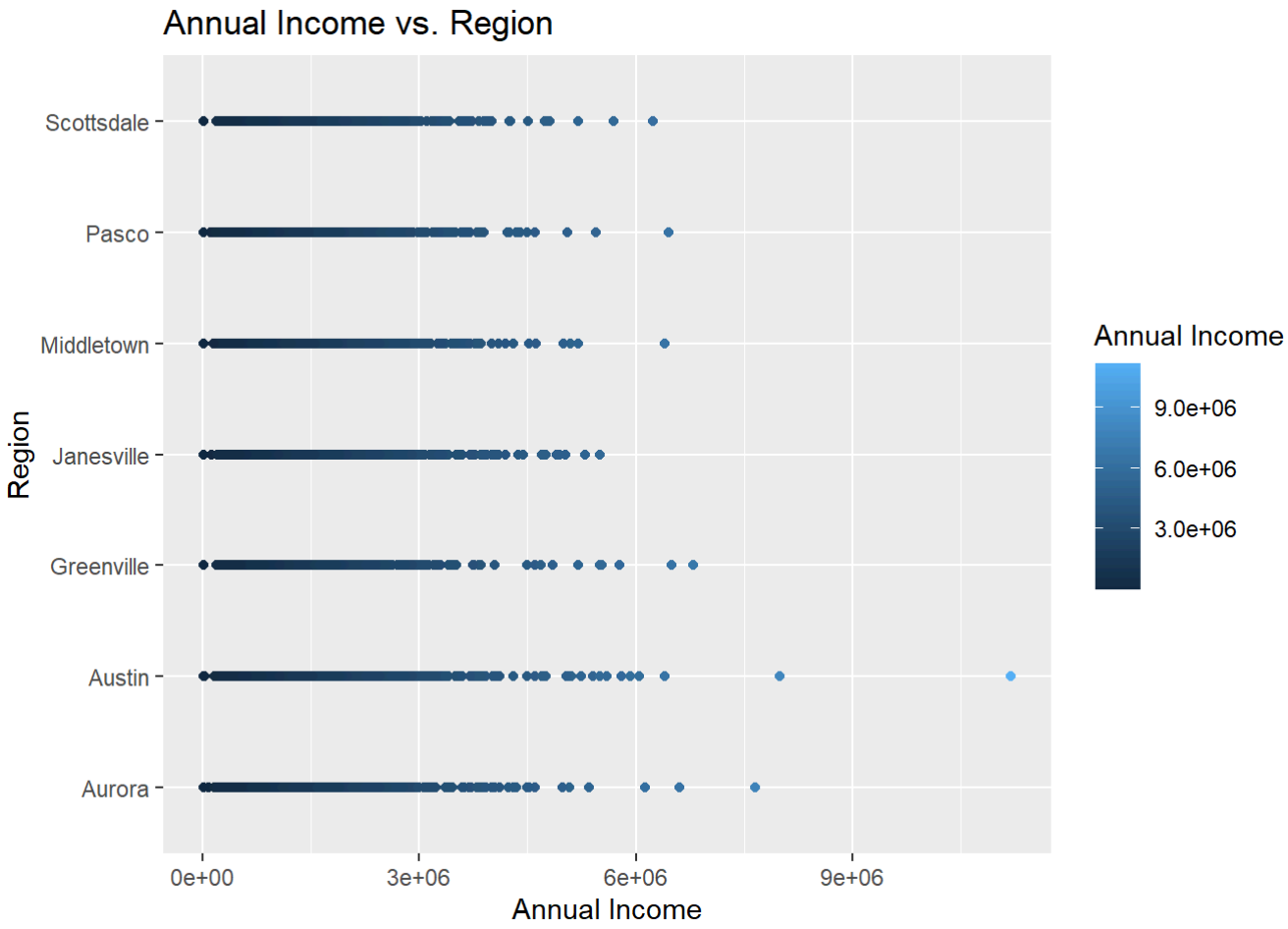
To sum up, cars are sold the most in September, November, and December by mostly male customers. Chevrolet, Dodge, and Ford are well-selling brands, and customers prefer SUV or Hatchback, with pale white color. Diamante by Mitsubishi, Prizm by Chevrolet and Silhouette by Oldsmobile are the most popular models. There is no huge difference between transmission types and engine types. The car price and customer's annual income showed similar trends, with the right-skewed distributions. Car dealers are most commonly located in Austin and Janeville in this data set, that may be associated with the income distribution as customers in this region are richer than others to afford to purchase a car.

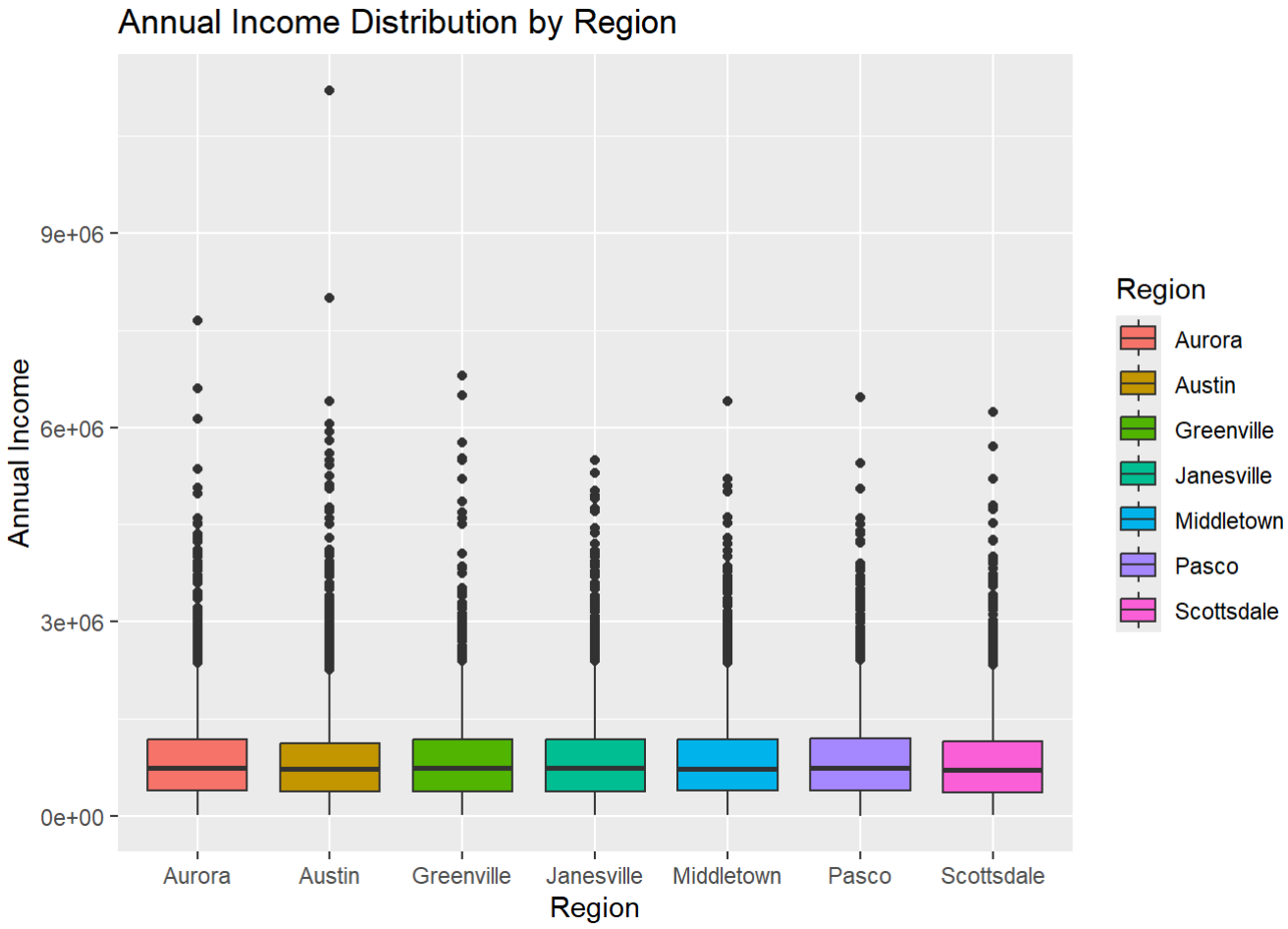
# 2-3. Relationship between variables in the data set

First, we check whether customer's income is associated with annual income. As shown below, high income customers do not necessarily purchase expensive cars.

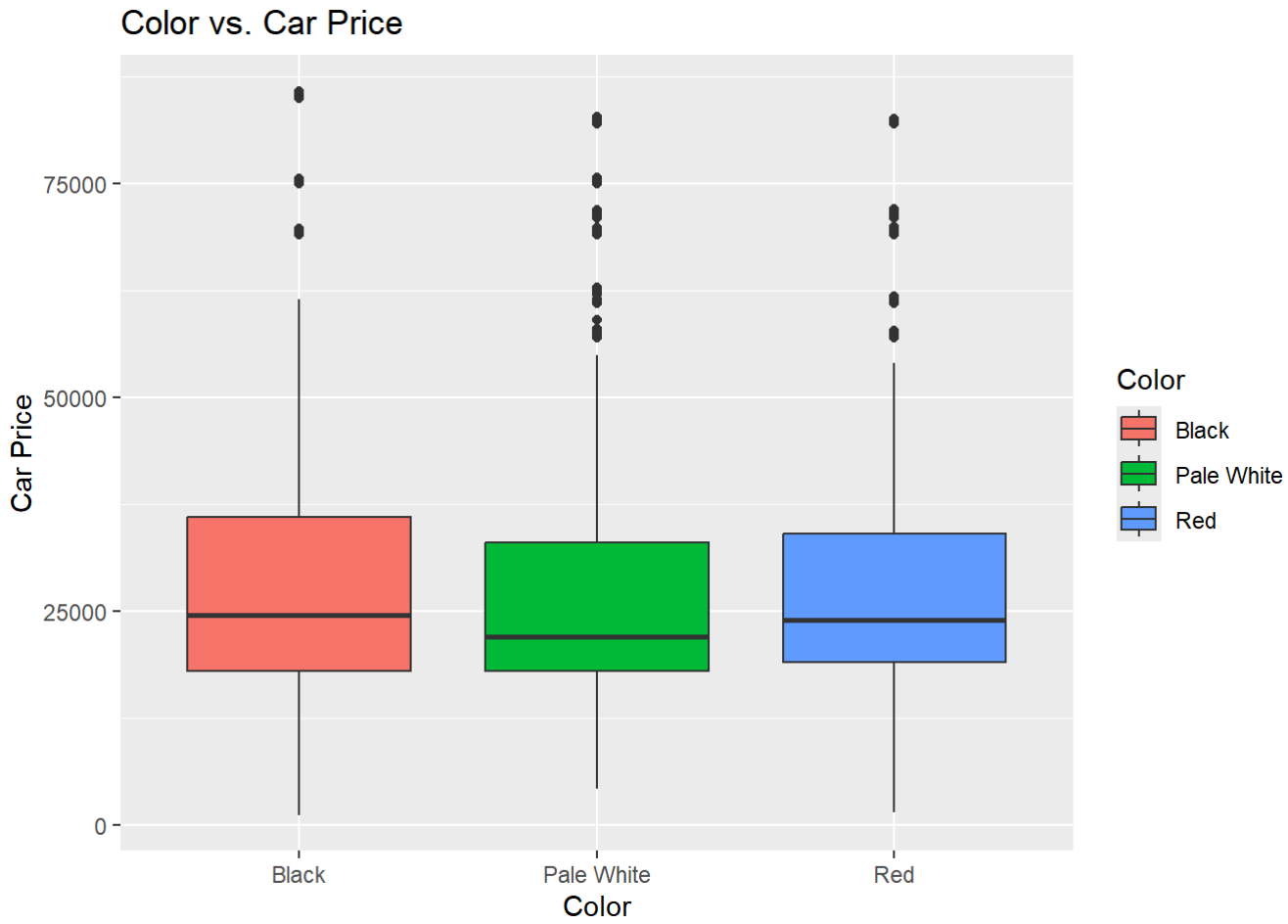



Next, we check the relationship between dealer region and customers' annual income. As we expected, Austin, the most common locations of dealers, has a slightly higher income distribution. However, the second common region, Janesville has a similar distribution to others.

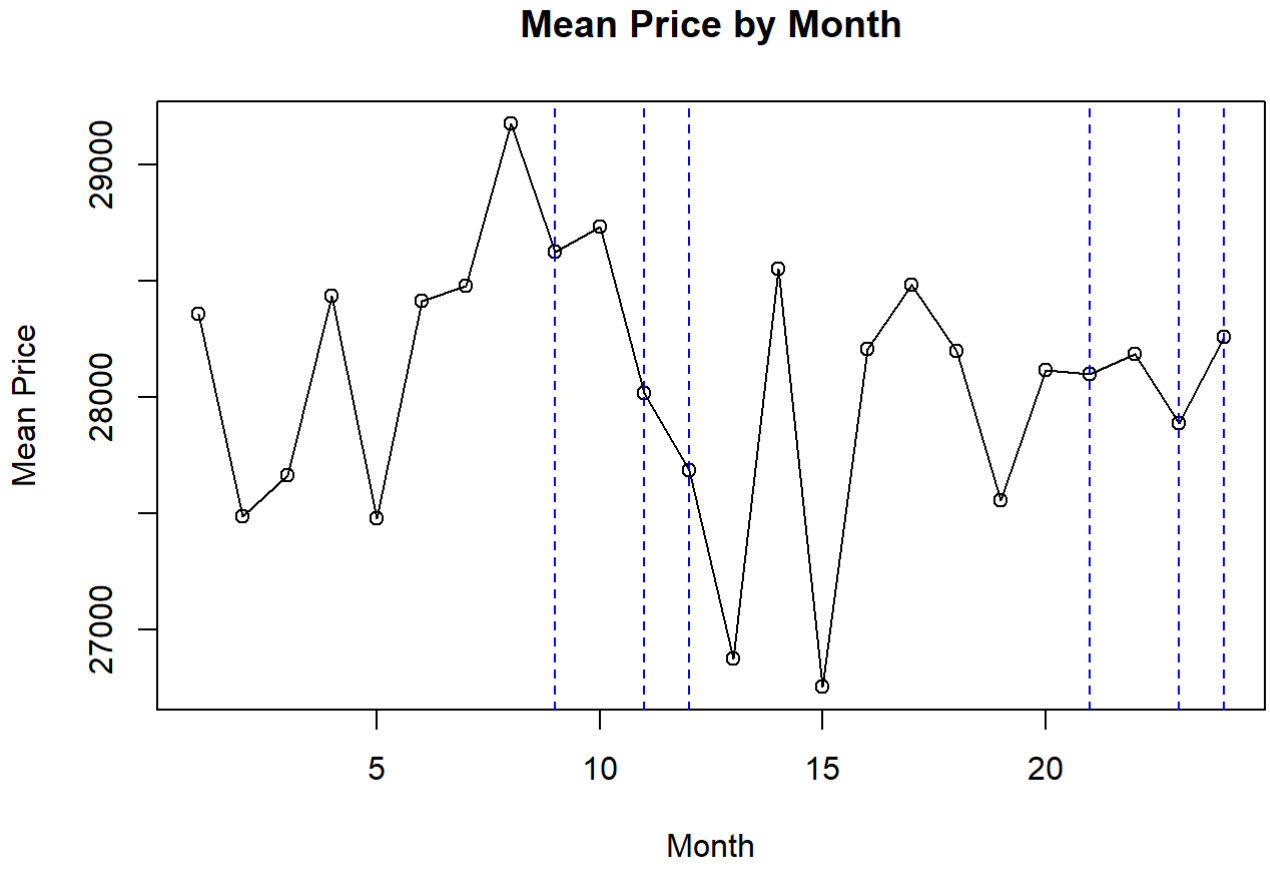


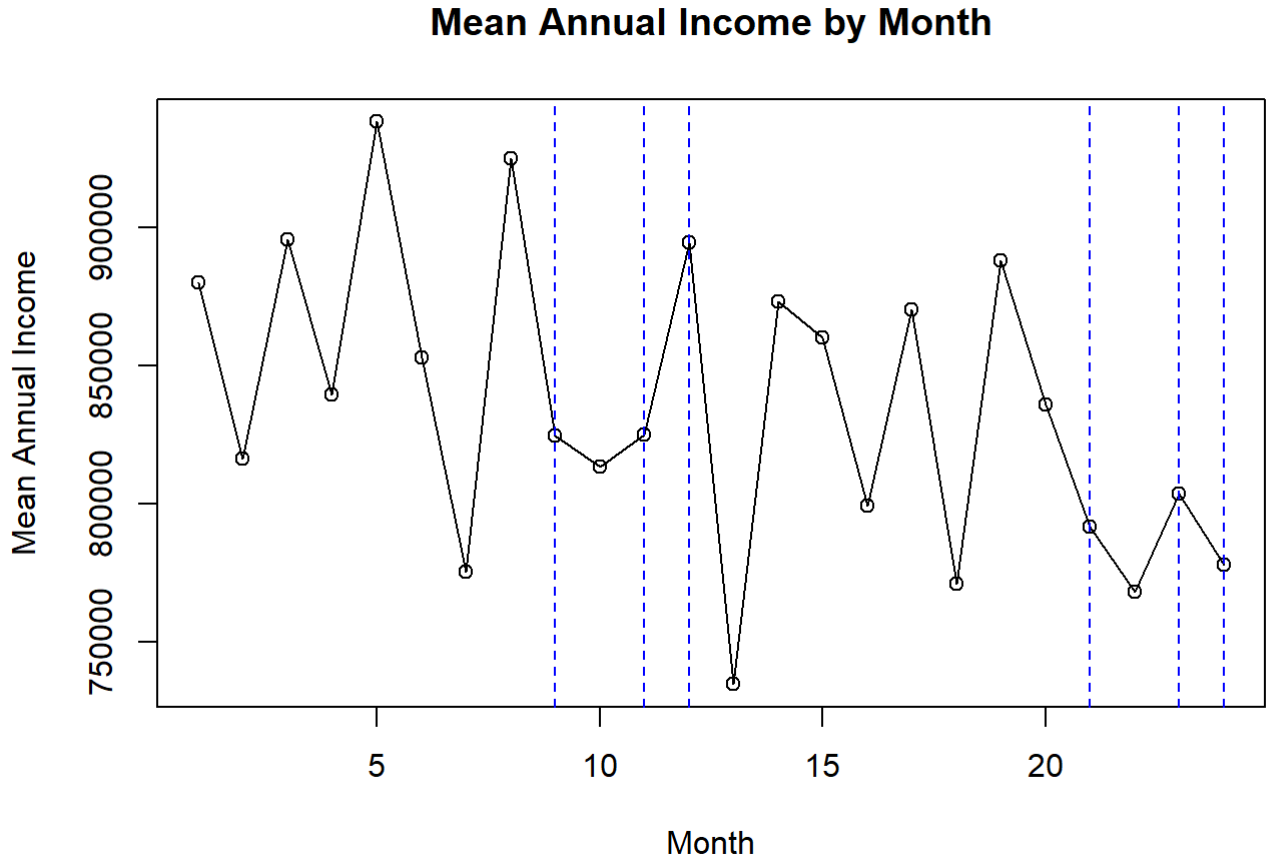




Furthermore, we check whether the well-selling color is cheaper than other colors. Overall, we can see Black is the most expensive and pale white is the least expensive color. Hence, this supports the assumption that customers prefer the color due to the lower price.




We also try to check whether car sales increases when the price becomes cheaper or customer's income becomes higher. Blue lines indicates months with higher car sales. As shown below, the car price does not seems to be associated with the months people purchased car the most. However, the income tends to be higher in summer, before the season of higher car sales. Hence, customer's income may be associated with car purchasing behavior.



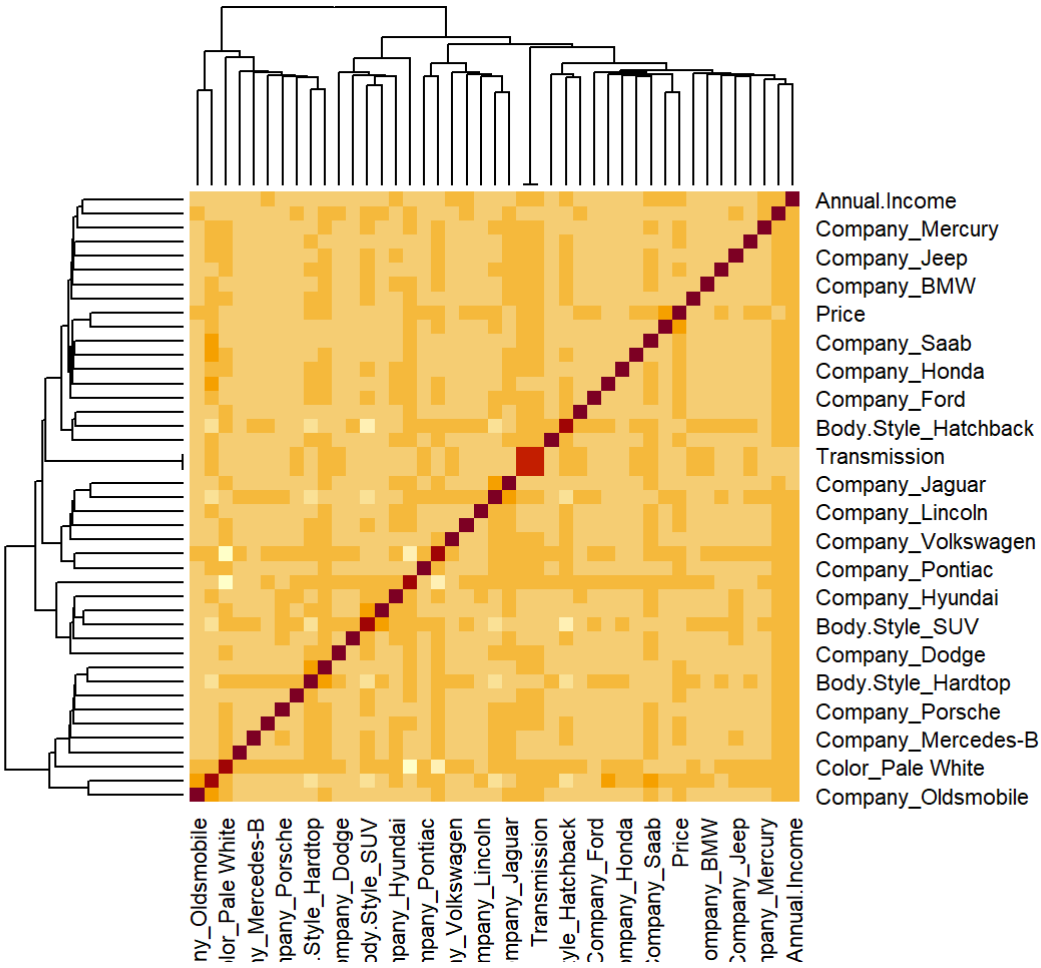

For correlation analysis, all variables used to check relationships were converted into binary. Between Price and other variables, we can clearly see Cadillac is positively associated with Price, and Hyndai is negatively associated with Price. Also, Engine type Double A is highly correlated with Auto transmission.

```
# Removing "Model" and "Dealer_Region"
df_cor <- cbind(df_car_binary[, 3:37], df_car_binary[, 192:199])
df_cor <- cbind('Price' = df_cor$Price, df_cor[, -5])
head(df_cor)
```

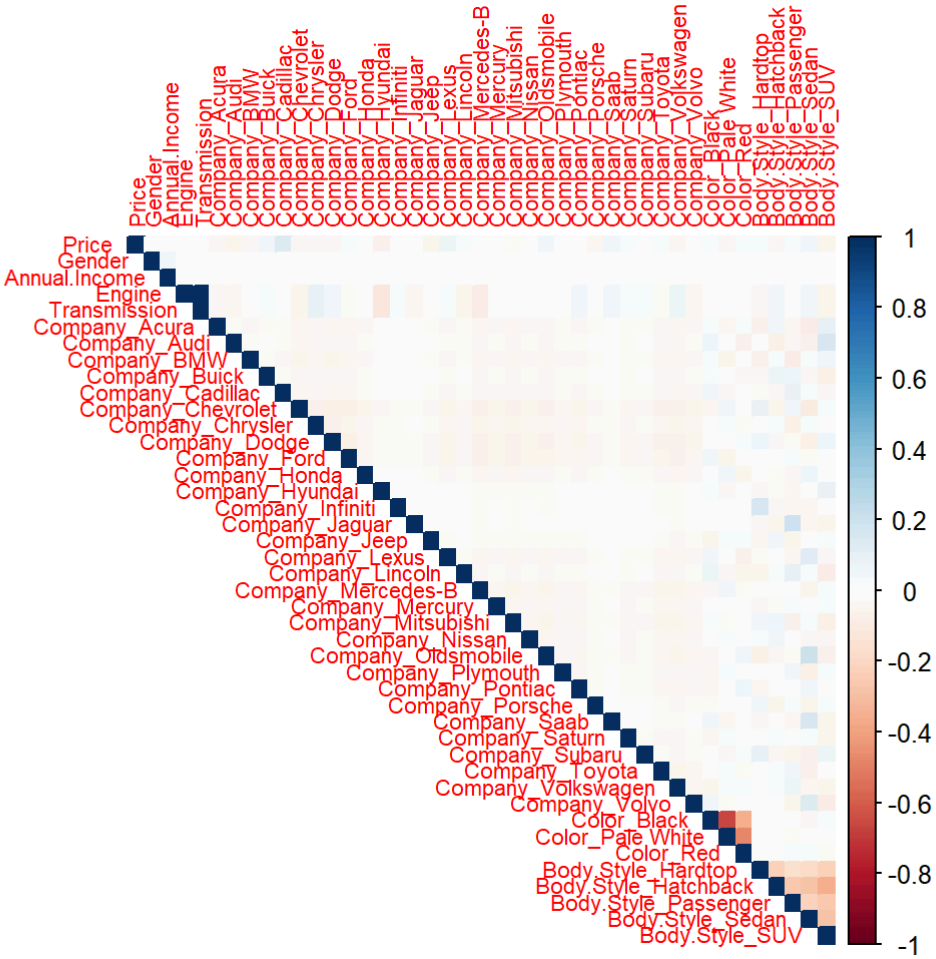
```
##   Price Gender Annual.Income Engine Transmission Company_Acura Company_Audi
## 1 26000      1         13500      1            1             0            0
## 2 19000      1       1480000      1            1             0            0
## 3 31500      1       1035000      0            0             0            0
## 4 14000      1         13500      0            0             0            0
## 5 24500      1       1465000      1            1             1            0
## 6 12000      1        850000      0            0             0            0
##   Company_BMW Company_Buick Company_Cadillac Company_Chevrolet Company_Chrysler
## 1           0             0                0                 0                0
## 2           0             0                0                 0                0
## 3           0             0                1                 0                0
## 4           0             0                0                 0                0
## 5           0             0                0                 0                0
## 6           0             0                0                 0                0
##   Company_Dodge Company_Ford Company_Honda Company_Hyundai Company_Infiniti
## 1             0            1             0                0                0
## 2             1            0             0                0                0
## 3             0            0             0                0                0
## 4             0            0             0                0                0
## 5             0            0             0                0                0
## 6             0            0             0                0                0
##   Company_Jaguar Company_Jeep Company_Lexus Company_Lincoln Company_Mercedes-B
## 1              0            0             0                0                 0
## 2              0            0             0                0                 0
## 3              0            0             0                0                 0
## 4              0            0             0                0                 0
## 5              0            0             0                0                 0
## 6              0            0             0                0                 0
##   Company_Mercury Company_Mitsubishi Company_Nissan Company_Oldsmobile
## 1               0                  0              0                  0
## 2               0                  0              0                  0
## 3               0                  0              0                  0
## 4               0                  0              0                  0
## 5               0                  0              0                  0
## 6               0                  1              0                  0
##   Company_Plymouth Company_Pontiac Company_Porsche Company_Saab Company_Saturn
## 1                0               0               0            0              0
## 2                0               0               0            0              0
## 3                0               0               0            0              0
## 4                0               0               0            0              0
## 5                0               0               0            0              0
## 6                0               0               0            0              0
##   Company_Subaru Company_Toyota Company_Volkswagen Company_Volvo Color_Black
## 1              0              0                 0             0           1
## 2              0              0                 0             0           1
## 3              0              0                 0             0           0
## 4              0              1                 0             0           0
## 5              0              0                 0             0           0
## 6              0              0                 0             0           0
##   Color_Pale White Color_Red Body.Style_Hardtop Body.Style_Hatchback
## 1                0         0                  0                   0
## 2                0         0                  0                   0
## 3                0         1                  0                   0
## 4                1         0                  0                   0
## 5                0         1                  0                   1
```

```
## 6                 1         0                   0                   1
##   Body.Style_Passenger Body.Style_Sedan Body.Style_SUV
## 1                    0                0              1
## 2                    0                0              1
## 3                    1                0              0
## 4                    0                0              1
## 5                    0                0              0
## 6                    0                0              0
```

```
correlation_matrix <- cor(df_cor)
heatmap(correlation_matrix)
```




```
corrplot(correlation_matrix, method = "color", type = "upper", tl.cex = 0.7)
```



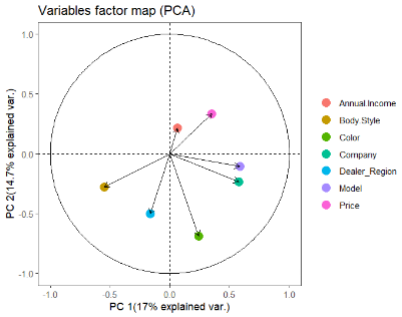

```
#jpeg(file = "correlation_matrix.jpeg", width = 200, height = 200, units = "mm", res =300)
#corrplot(correlation_matrix, method = "color", type = "upper", tl.cex = 0.7)

#dev.off()
```

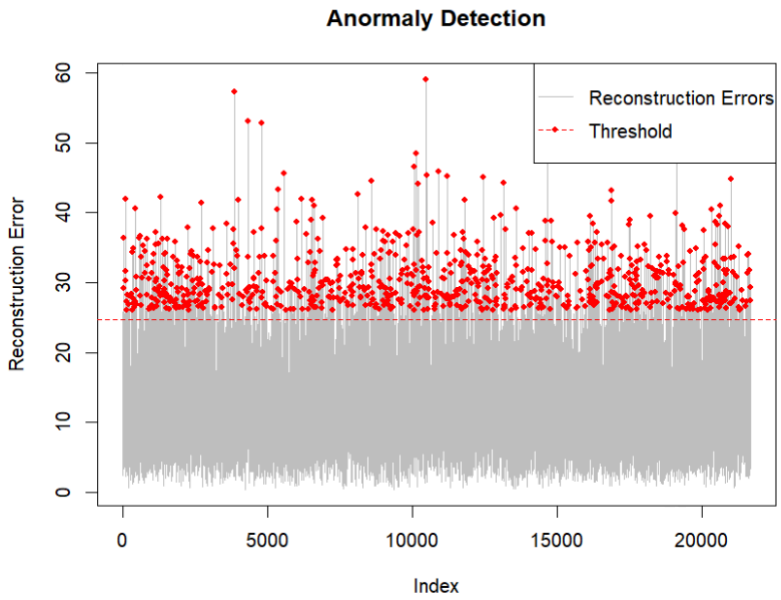
# 2-4. Robust PCA analysis

For PCA analysis, the data was modified to have a correct type and class, and then outlying observations were removed based on IQR.

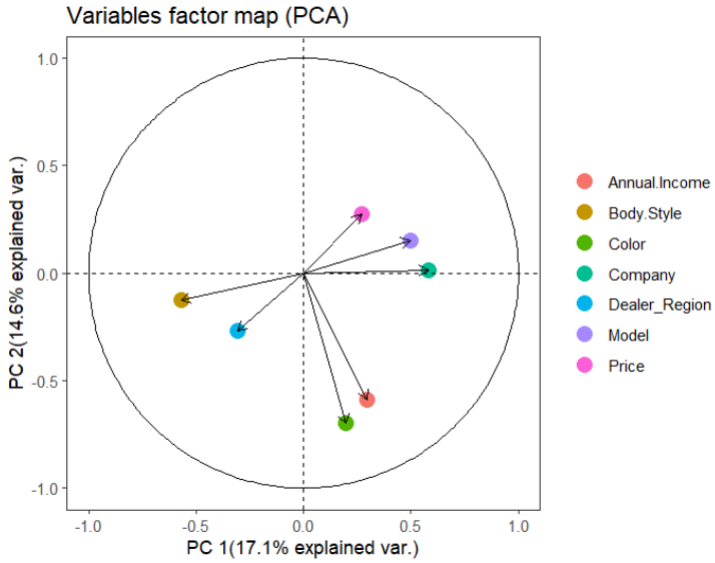
Based on the robust PCA result, we can see Price and Model are highly correlated, and Company and Dealer_Region are also grouped in the same group, suggesting that they have similarity. In addition, abnormalities were detected based on this result, with a threshold of 3* sigma. After removing these observations, we tried fitting the robust PCA on the cleaned data.

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Explained variance | 1.192 | 1.027 | 1.008 | 0.989 | 0.964 | 0.922 | 0.898 |
| Standard deviations | 1.092 | 1.013 | 1.004 | 0.995 | 0.982 | 0.960 | 0.948 |
| Proportion of variance | 0.170 | 0.147 | 0.144 | 0.141 | 0.138 | 0.132 | 0.128 |
| Cumulative proportion | 0.170 | 0.317 | 0.461 | 0.602 | 0.740 | 0.872 | 1.000 |






Here, the result of robust PCA on cleaned data is shown. Although variations explained by the first and second principle components are not large, variables were clustered clearer than the uncleaned data. Now, Price, Model, and Company were in the same group, suggesting their positive correlations. This result indicates that the car price is likely to get influenced by car models and company, which is reasonable and consistent to the previous explanatory analysis result. Additionally, the result shows that annual income and color of the car is also associated, and dealer region and body styles also showed they are related to each other.




## Summary

To sum up, we introduced the trends in the car sales data that the majority (78%) of customers are male, and the car price tends to be influenced by car model and brands (company). The popular brands are Chevrolet, Dodge, and Ford, while popular models are Diamante (Mitsubishi), Prizm (Chevrolet), Silhouette (Oldsmobile), which is different from the brands ranking. Therefore, some specific models seem to attract customer's interests. In addition, SUV and Hatchback are two top selling style, with popular color, pale white. This preference can be explained by price of car by colors and the purpose of car purchasing as customers may prefer to have durable, long-lasting car, rather than appearance. The car sales were made mostly in Austin; however, there was not clear association between the region and other variables. Furthermore, associations between variables were also explained by the robust PCA result, showing that similar tendency with the findings previously explained, the car price seems to be determined by car model and brands rather than customer's income or regions. This result is interesting and useful for machine learning applications to predict the car sales price. Working on larger data with more variables associated with customers and car price may help assuming reasons and causes of these results we have seen in this project.