1. **Introduction**

Drivetime is the second largest vehicle retailer in the US focused solely on used vehicles. Founded in 2002, the company had 76 dealers operating in 8 states. With the mission “To be the auto dealerships and finance company for people with less than perfect credit”, Drivetime typically sell about 4 thousand used vehicles and processes about 10 thousand credit applications per month.

Drivetime purchased most of its vehicles from the auctions and wholesalers. Vehicles include many makes and models of cars and trucks. After going through a process of checking and repair, the vehicles are sent to the dealership for sales. Normal dealer sales occur within 90 days of delivery to the dealership. If the vehicle does not sell within the 90 days, it’s called an overage vehicle, meaning that it has been on the lot too long to generate normal dealer profit.

Each overage vehicle has its sales price reduced in order to encourage the sales within 90–119 day period. Furthermore, if an overage vehicle fails to sell within 120 days, the vehicle is taken off the lot and sold at auction. These overage vehicles incurred a reduction in profit and even a loss to Drivetime. More specifically, while a non-overage vehicle is sold at a 20% markup price, an overage vehicle’s price is set at 10% markup and a vehicle is sold at auction at -20% markup.

This analysis is conducted to support purchasing managers at Drivetime for their selection process. The main objective is to increase the proportion of vehicles sold within 90 days as well as to reduce the risk of overage and auction vehicles, as well as maximize the profit for the company.

Our analysis includes 4 main parts:

1 — Data Characteristics and Exploration

2 — Model Selection, Interpretation & Prediction

3 — Recommendation & Conclusion

4 –--Appendix

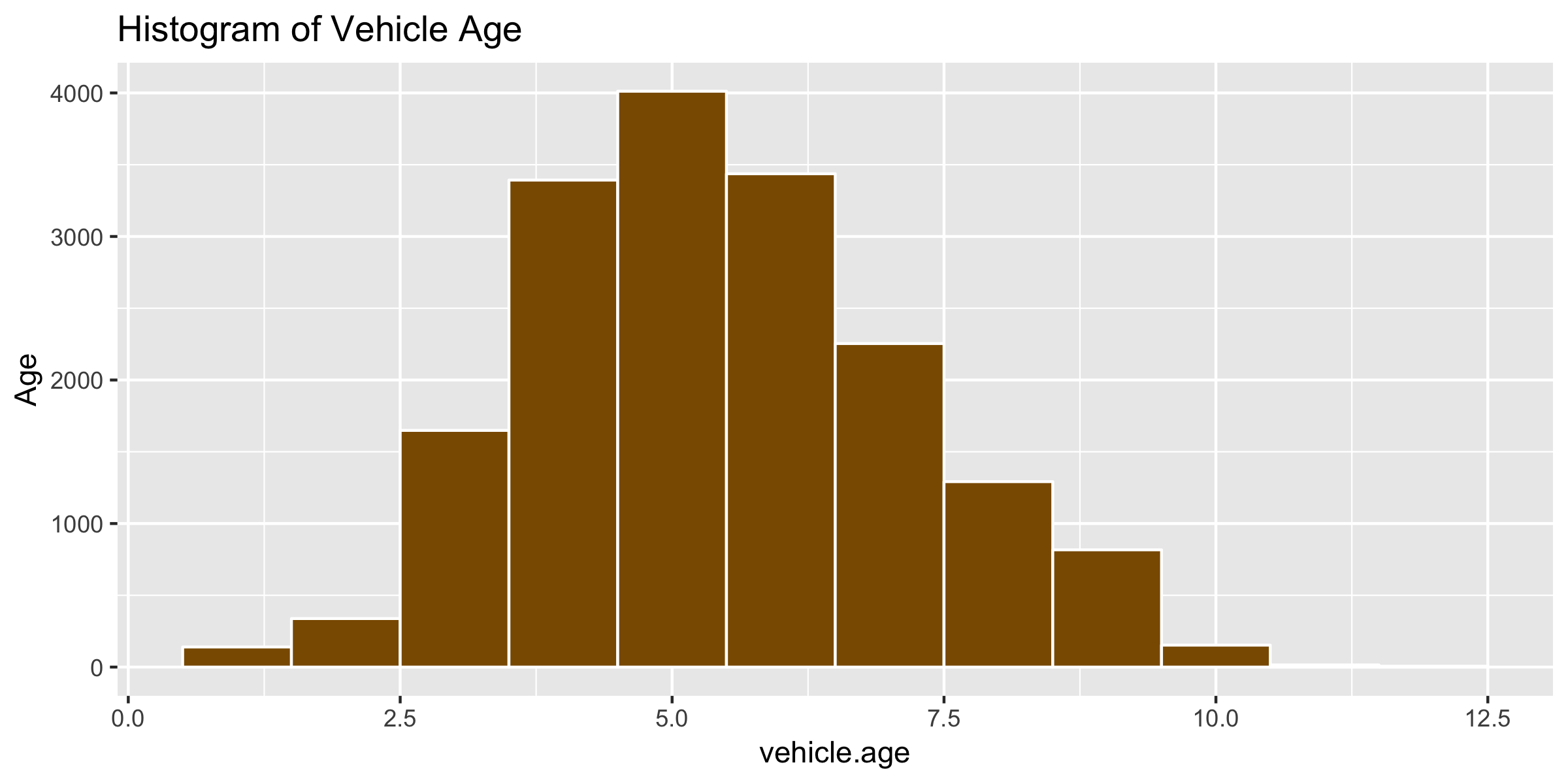
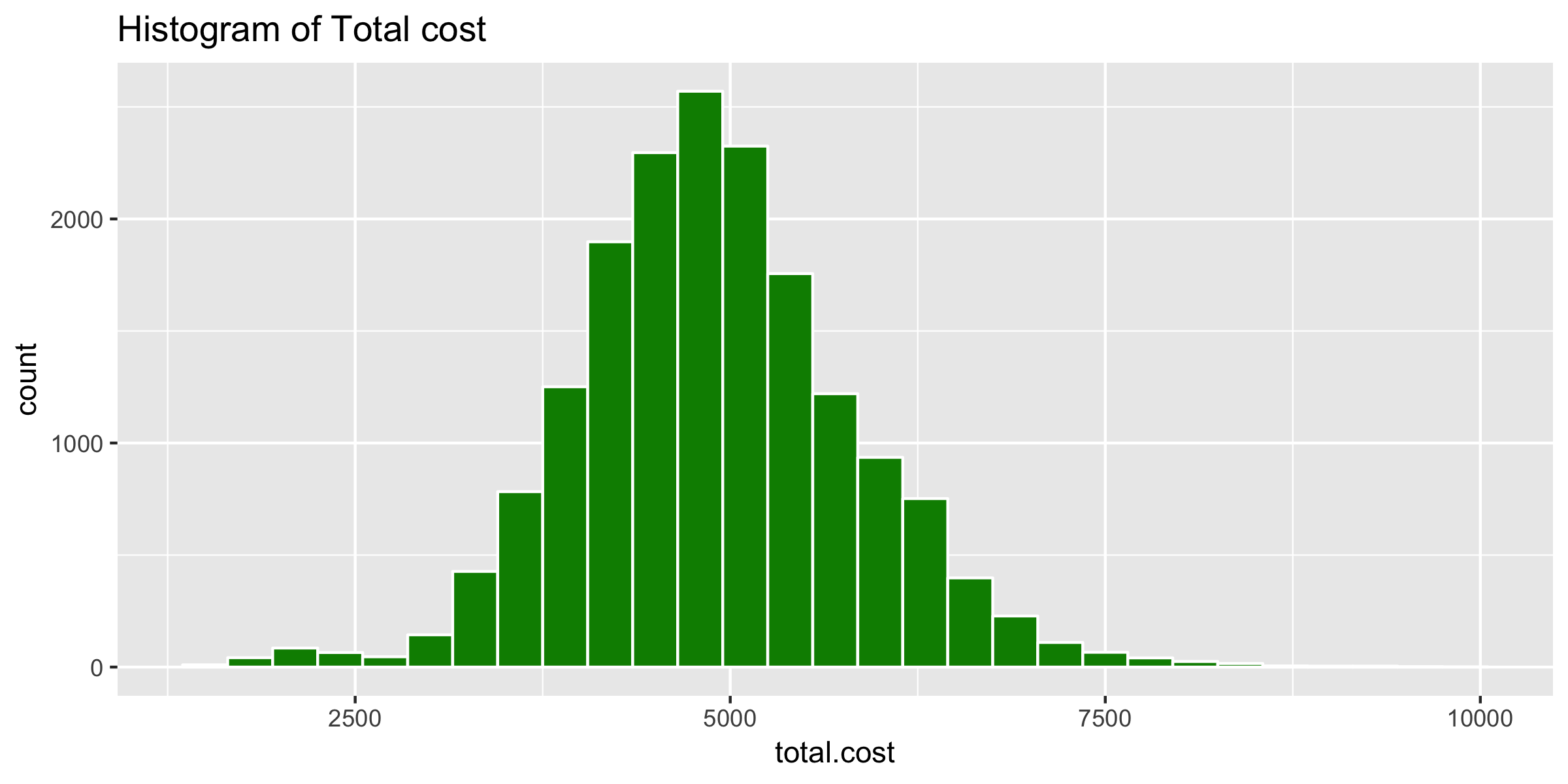
1. **Data Characteristics**

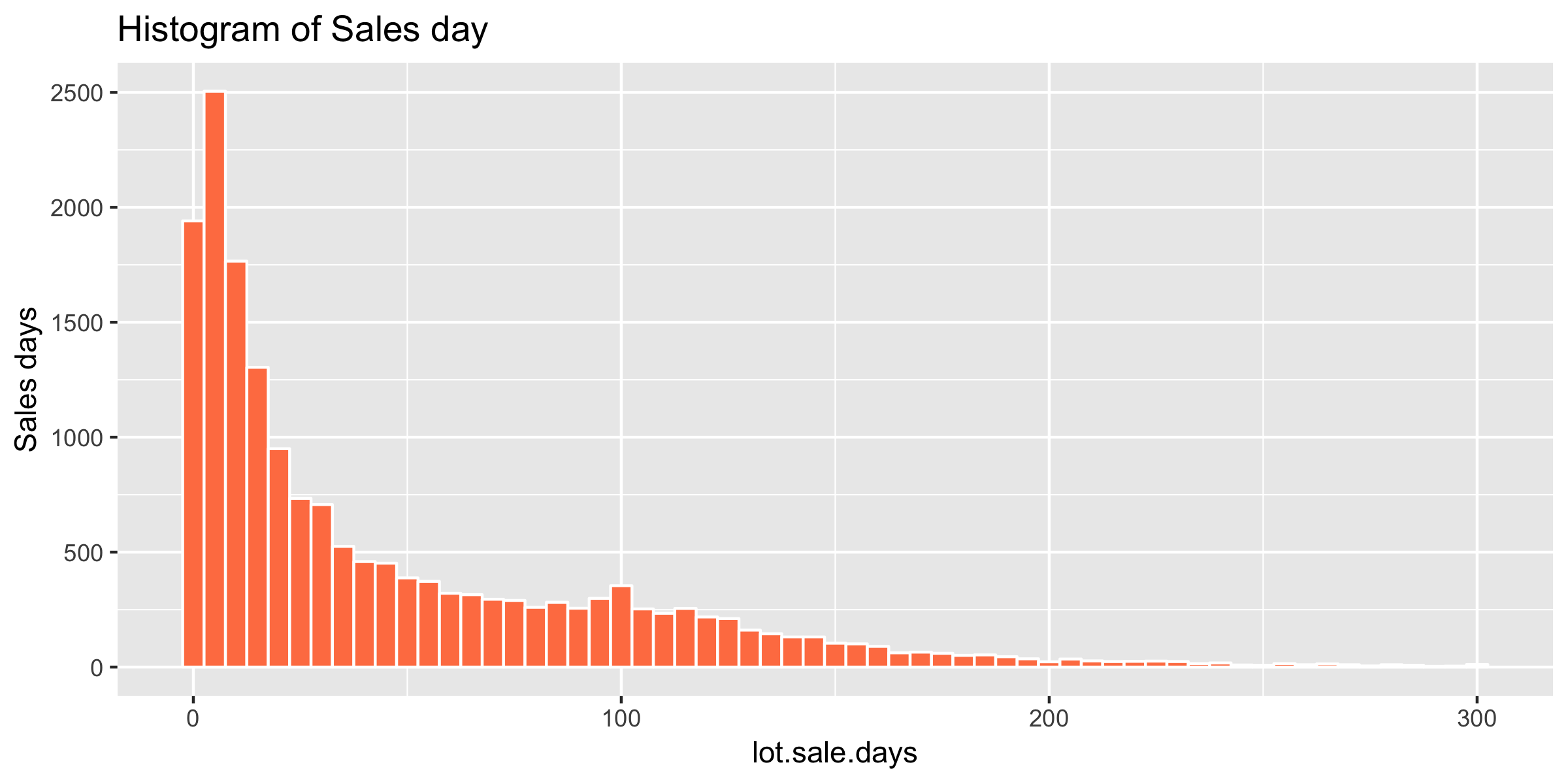
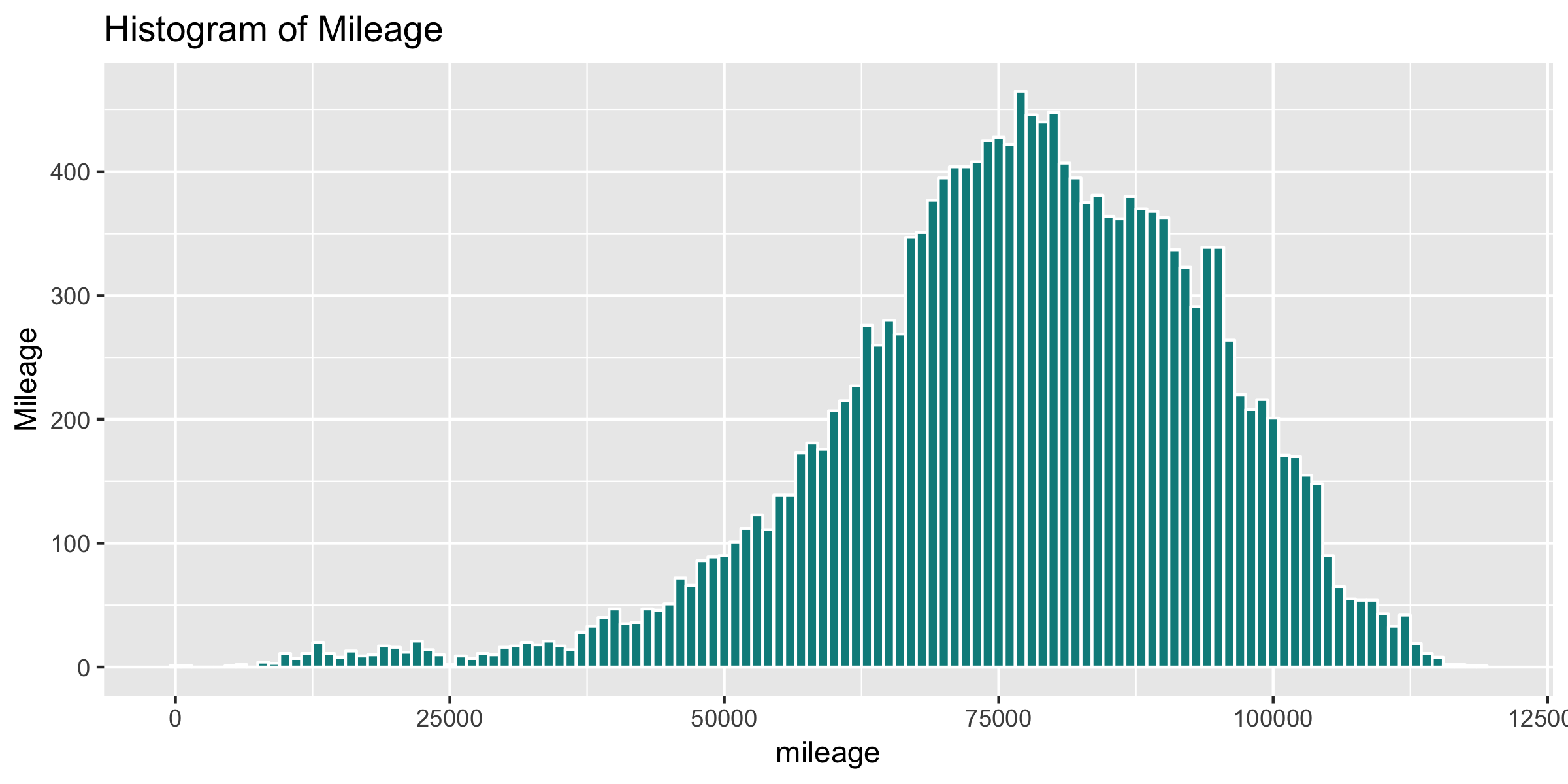
There are 17506 data points in this dataset, in which 8753 are marked as training data, 5376 are marked as test data and the rest is validate data. There is no missing data in the dataset.

The following table summarizes all the variables in the dataset:

|  |  |  |
| --- | --- | --- |
| **Variables** | **Type** | **Meaning** |
| Total cost | continuous | the cost ($) incurred to Drivetime to buy the vehicles |
| Mileage | continuous | the total miles that the vehicles have travelled at the time of purchase |
| Vehicle.age | ordinal | The age of the vehicle at the time it was purchased by Drivetime |
| Vehicle.age.group | categorical | which age group the vehicles belong to: one-three, four+, six+ and seven+ |
| color.set | categorical | the color of the vehicles |
| makex | categorical | manufacturer brands |
| state | categorical | where the vehicle is sold |
| make.model | categorical | the brands and the model of the vehicles |
| lot.sale.day | Continuous (independent variable) | Number of days the vehicles are in the lot before being sold |
| overage | Categorical  (class label) | “YES" means the lot.sale.days is higher than 90 days.  “NO" means the lot.sale.days is lower than 90 days. |

To better understand our dataset, we conduct an exploratory analysis of our individual variables and the relationship between them.

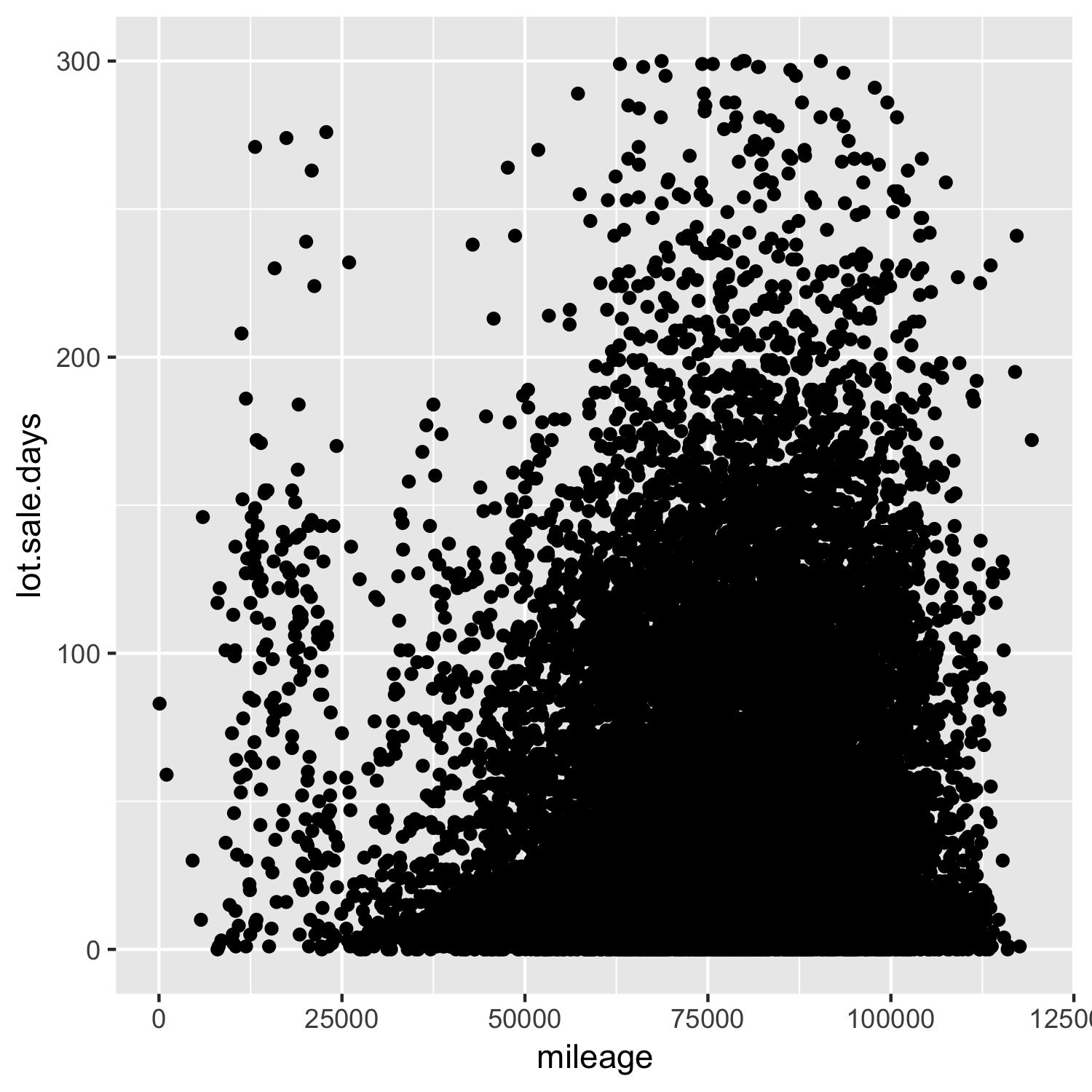
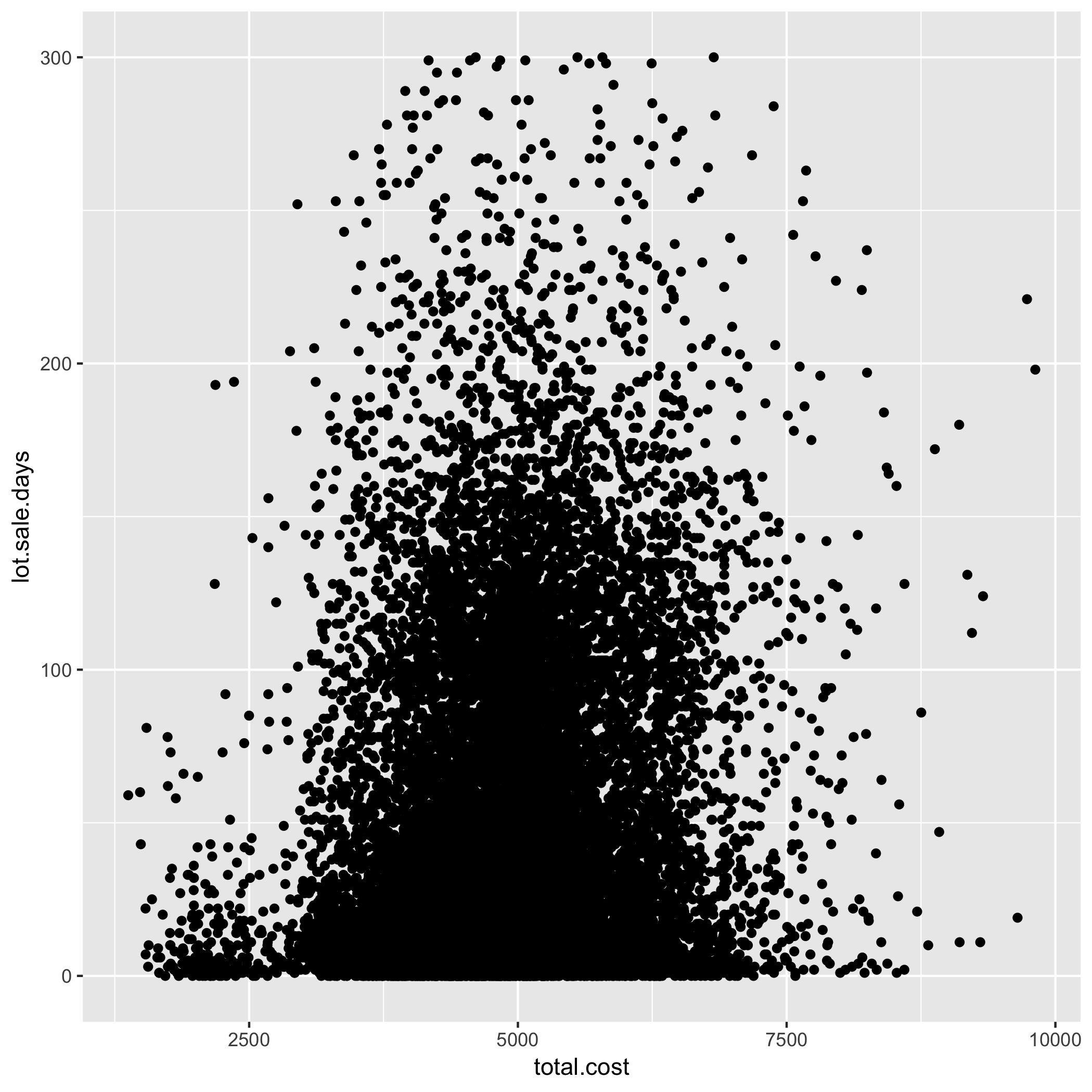


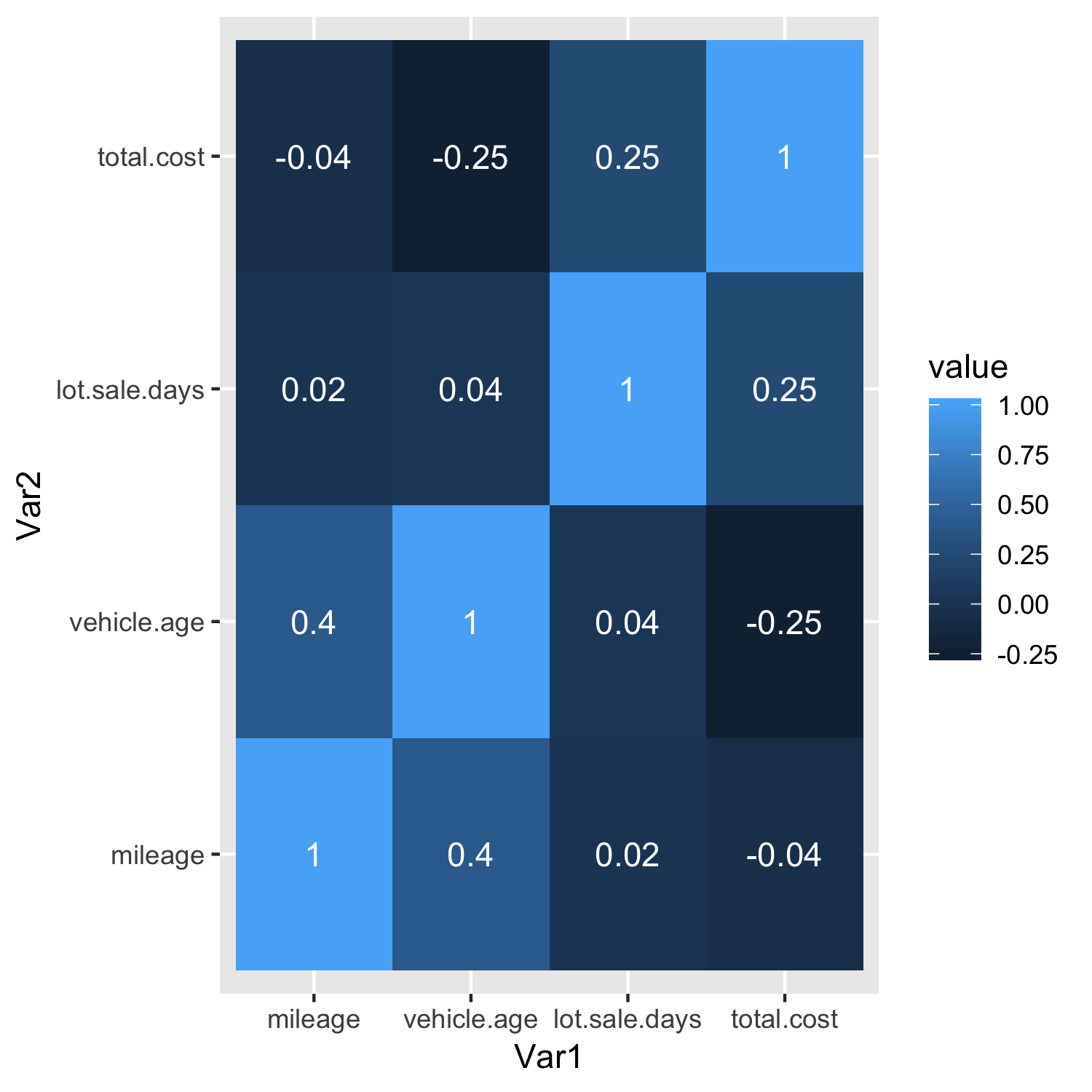


**Figure 1: Histogram of Cost, Vehicle Age, Mileage and Sales days**

The histograms of the continuous variables show that while the tocal cost, vehicle age and mileage follow a normal distribution, sales day data is left-skewed. The majority of vehicles at Drivetime have the cost of $4000–6000, with the age 3–7.5 years at the time of purchase, and their mileages range mainly from 50000–100000 miles. Moreover, 80% of the vehicles will be sold within 90 days.

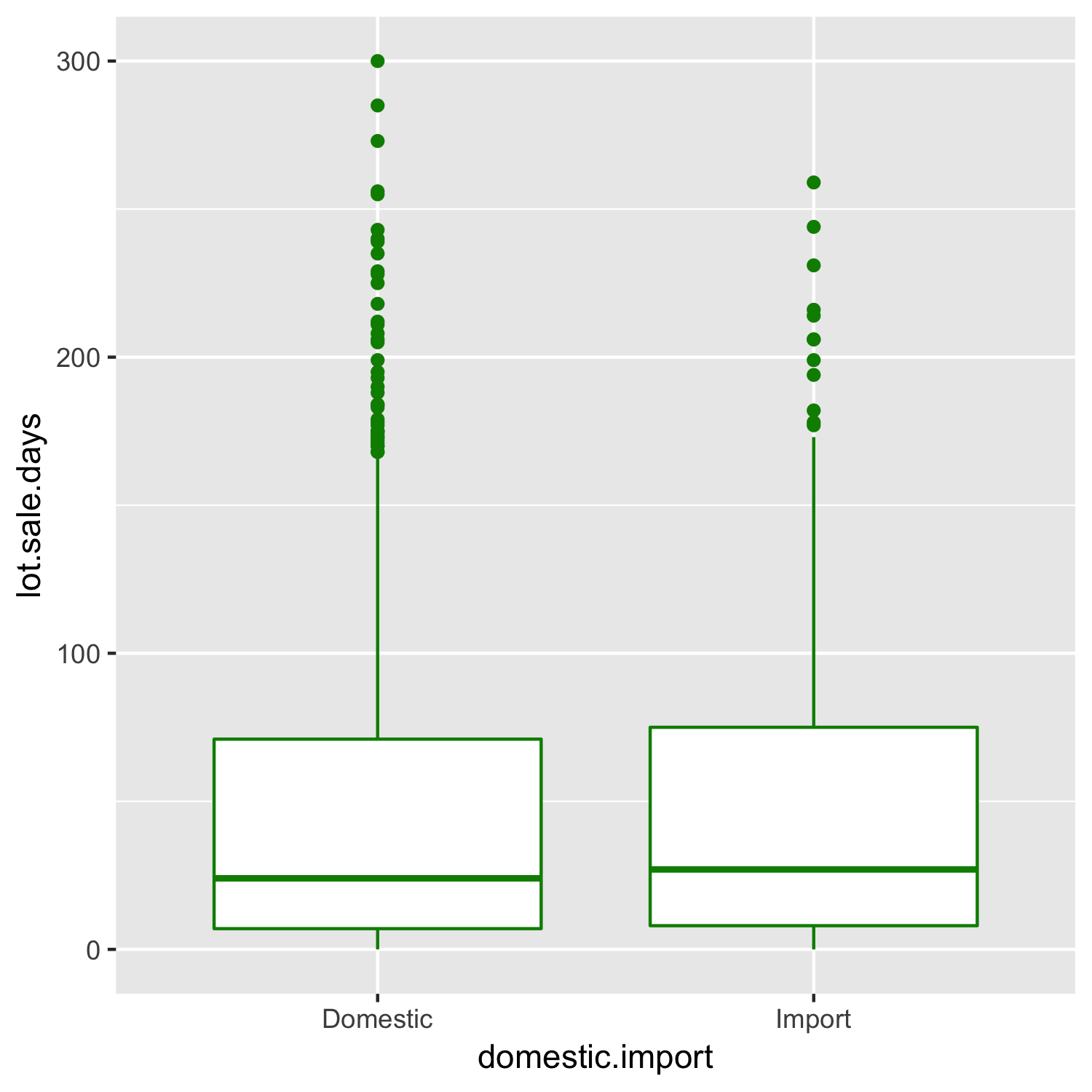
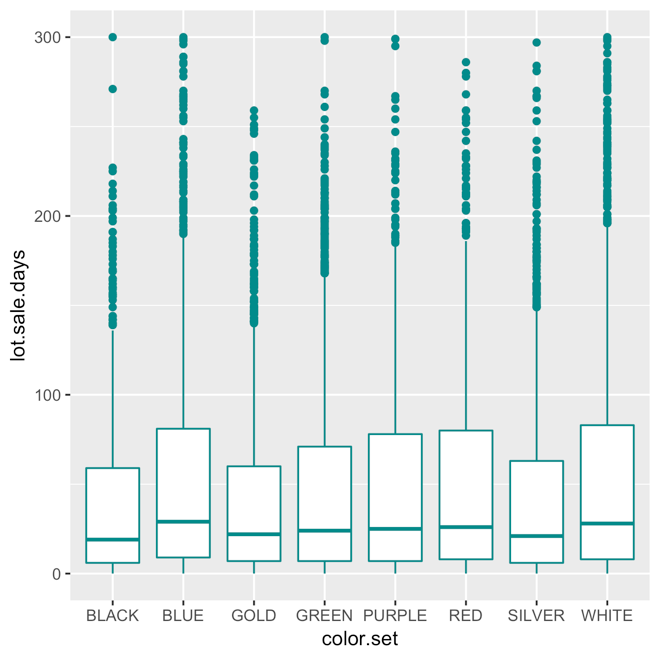
The below correlation heat map illustrates that sales days don’t have strong relationships with mileage and vehicle age and a slightly better positive correlation with the total cost. This means that generally, the vehicle tends to be sold in a longer period of time if its cost was higher than other vehicles. The scatterplot between sales days and other continuous variables confirms our observation that there is no strong and clear linear relationship between cost, mileage and sales days.

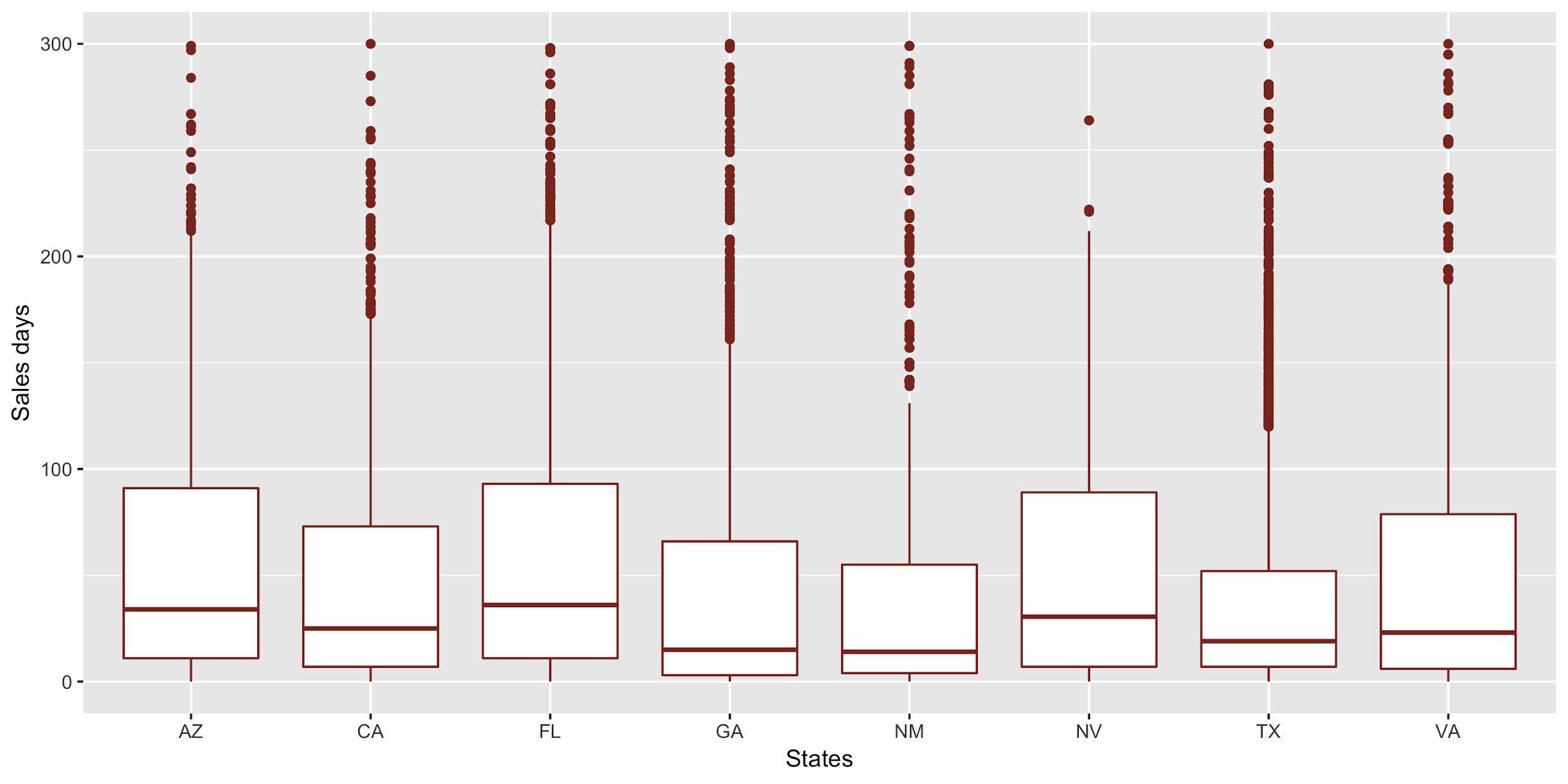
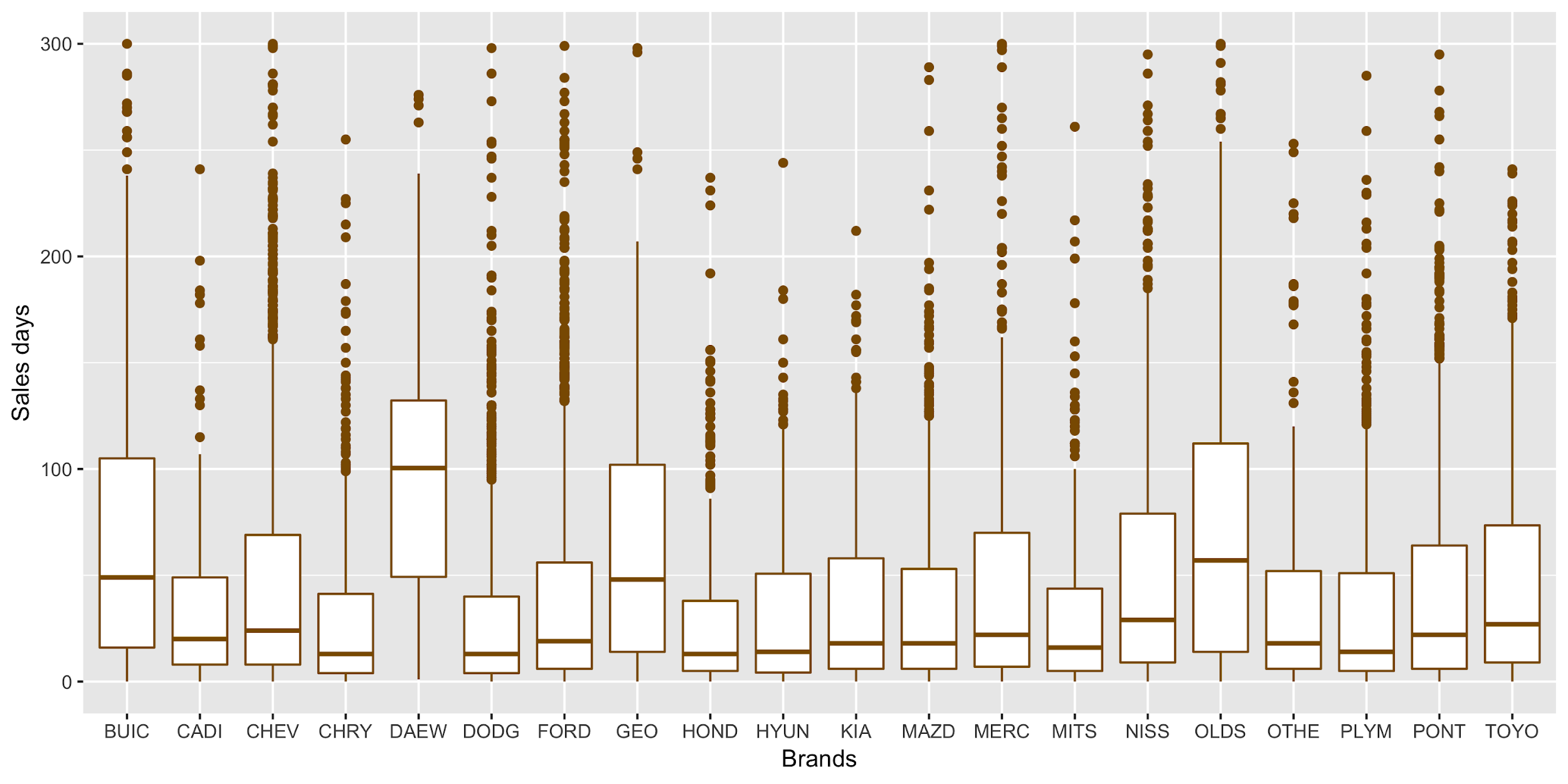
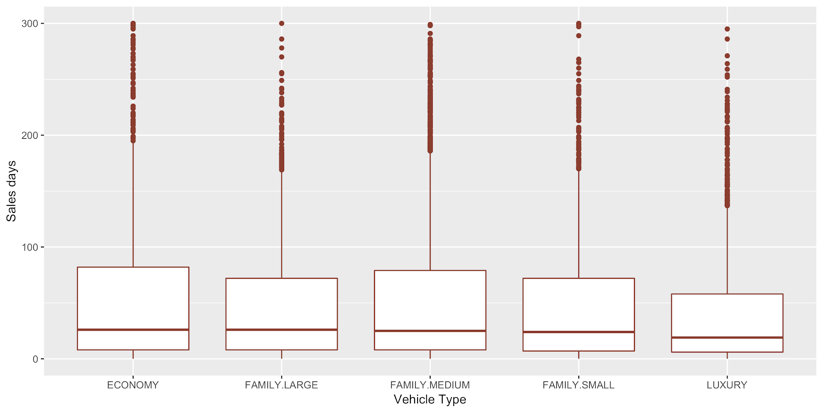




**Figure 2, 3, 4: Scatterplot between Cost and Sales days, between Mileage and Sales days and the correlation heat map between continuous variables**

Next, we look at the categorical variables to determine whether there are any significant factors that influence sales days.





**Figure 5,6,7,8: Comparison of the Domestic, Vehicle type, brands and States**

There seems to be no significant difference between domestic and import vehicles in term of how fast they are sold. Regarding vehicle type, economy vehicles seem to be sold slightly slower than family-type vehicles and luxury sedans. On average, it takes 51 days to sell an economy vehicle, 39 days to sell a luxury vehicle and 47 days to sell family-type vehicles.

Manufacturer brands seem to make a big difference on sales days. Some fast-selling used vehicles tend to be Honda, Chrysler, Dodge, Hyundai, Plymouth, etc. while Daewoo, Oldsmobile, Buick and GEO vehicles stay at the lot for a longer time before being sold. These vehicles tend to hold their values after several years of use. This time to sales depends on a number of factor such as a discount, value depreciation after 1–3 years, availability and popularity of the brands, etc.

Some states such as California, Georgia, Texas have significantly lower sales days than other states such as Arizona, Florida, Nevada. On average, vehicles in Arizona and Florida are sold within 53 and 57 days respectively; vehicles in Texas and Georgia are sold within 39 and 42 days respectively. This difference might due to the difference in state regulation regarding all types of fee such as dealer documentation fee, registration fee, local state tax which can make affect the demands for used vehicles in these state. Moreover, population and unemployment situation of the states might be a significant factor as well. For example, Florida and Arizona are well-known for retirement destination in the US, therefore the demand for used vehicles in these 2 states are usually lower than that in other states.

Another important variable in the dataset is make.model, indicating the brand and the model of the vehicles. There are 110 models observed in our data. Some models have a very high frequency such as Chevrolet Lumina, Ford Taurus, Nissan Sentra (more than 900 times) while some models appear rarely such as Acura Vigor, Buick Lasabre or Chevrolet celebrity (1 time). For the purpose of accurate generalization and prediction when constructing a prediction model on the data, we will remove the observations whose model frequency is less than 4. The dataset ready for modeling contain 17469 data points.

**3. Model Selection and Prediction**

We will use train data to build the model, test data to evaluate how fit the model is, and validate data to predict profit and see the impact of the model on the profit.

We will build 2 models using the regression model to see what is the difference between them:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Independent Variables | | | | | | | |
| Model 1 | Mileage | Total cost | Vehicle Age | Color set | States | Brand | Vehicle type | |
| Model 2 | Mileage | Total cost | Vehicle Age | Color set | States | Brand/ model | |  |

As mentioned above, the brand/model has more than 100 values and it pretty much indicates the brand and vehicle type information as well. It also incorporates other information such as fuel-efficiency, Electric or Hybrid Vehicles, etc. Therefore, it might be redundant to include all brand/model, brand and vehicle type in one single regression model.

The results of the 2 models show us the insights that can guide Drivetime’s vehicle selection process:

* An additional 10000 mile of the mileage increases the sales days by 2.53 days on average
* An additional $100 in the total cost increases the sales days by 2 days on average
* An additional year of the vehicle age increase the sales days by 9.83 days on average
* Color: In compared with Black, the color Blue, Green, Purple, Red and White increase the sales days by 9.29, 6.47, 8.72, 11.5 and 8.7 days respectively. There is no significant difference between Gold and Black, between Silver and Black in term of sales time.
* Brand: Comparing with Buick, the other brands tend to have shorter sales days, except Daewoo and Oldsmobile. Some brands have many shortest sales days are Honda (59 days shorter), Toyota (36 days shorter), Cadillac (57 days shorter), Chrysler (34 days shorter), Mazda (30 days shorter).
* State: Arizona is the state that has the highest sales days. Other states CA, VA, NV have 9–11 shorter in sales days. GA, FL, TX have 2–4 shorter sales days than AZ.
* Vehicle type: Economy cars take the longest time to sell. Luxury and Family Large and Family Medium take the shortest time to sell (40 days shorter than economy)
* Some fastest-selling models are Infiniti J30, Infiniti Q45, Mazda Millenia, Honda Accord, Cadillac Serville, Toyota Camry, Nissan Maxima, etc. Slowest-selling models are Plymouth Acclaim, Oldsmobile Cierra, Geo Metro, Hyundai Accent, Dodge Spirit…etc.

The following table compares the performance of 2 models:

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Model 1** | **Model 2** |
| R square | 20.2% | 32% |
| Residual mean error | 47.21 | 43.41 |
| Non-overage accuracy | 82% | 84% |
| Effect on Profit | +$10 720 | +$26 406 |

Model 2 seems to be better in predicting the sales days. While model 1 explains 20.2% of the variations in sales days, model 2 can explain 32% of the variations in sales days, leaving 68% unexplained.

Non-overage accuracy is one of the important factors that illustrate the improvement of using statistical modeling in guiding Drivetime’s selection process. As mentioned in the Data Characteristic section, about 80% of the vehicles purchased by Drivetime are sold within 90 days without the presence of any statistical models. After we applied the model 1 or model 2 on test data, we suggest Drivetime to purchase only the vehicles that are predicted to be non-overage and avoid the vehicles that are predicted to be overage or auction. 82% and 84% of these predicted non-overage vehicles are actually sold within 90 days. The accuracy is improved by 2% and 4% by using model 1 and model 2 respectively.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Without Model** | | **Model 1** | **Model 2** |
| Number of predicted non-overage vehicles | | 4364 | 4168 | 4018 |
| Number of vehicles are actually sold under 90 days (non-overage) | | 3510 | 3421 | 3377 |
| % non-overage vehicles | | 80% | 82% | 84% |

Next, we use the validate data to examine the effect of using the statistical model on Drivetime’s profit. We know that a non-overage vehicle is sold at a 20% markup price, an overage vehicle’s price is set at 10% markup and a vehicle is sold at auction at -20% markup. Therefore, in comparison with selecting vehicles without the guidance of the statistical model, Model 1 is expected to increase the profit by $10,720 and Model 2 is expected to increase the profit by $26,406, equivalent to 1.06% of the profit.

1. **Recommendation & Conclusion**

Used vehicle market involves many factors when it comes to predicting the fast-selling vehicles that maintain profit and reduce inventory cost for the retailers. Colors, vehicle type, Brand, model, mileage, cost, and vehicle ages are some of the significant factors we had a chance to analyze that affect the time a used vehicle stays in the lot. In general, the customer’s preference is varied and complex. However, there are some characteristics that are highly desirable among used vehicles that Drivetime can prioritize to purchase:

* Customer seems to prefer a vehicle that is Black, Silver, Gold or White
* Highly demanded vehicle brands are Honda, Dodge, Chrysler, Toyota, Cadillac, Mazda. Drivetime should avoid Daewoo, Geo, Oldsmobile or Buick since these brands are quite unpopular and take a very long time to sell
* Luxury and Family-type vehicles are sold faster than Economy vehicles.
* Location affects sales days significantly. When purchasing and selling used vehicles in the low-selling states such as Florida and Arizona, Drivetime should focus on other highly desirable characteristics that colors, brands, low mileage and age.
* Drivetime should prefer to purchase vehicles that have small mileage, low cost and small age when it comes to choosing between 2 equivalent vehicles in term of other characteristics.
* Have a list of 20 fastest-selling vehicles and purchase them more.

Drivetime should adopt Model 2 to guide their selection process from now on. This model can improve Drivetime profit by 1.06%.

Future research is recommended to explore other factors that influence the sales period of a used vehicle. For example, the level of fuel-efficiency, whether the vehicle is electric or hybrid, level of discount from the original price, etc. Incorporating these factors in the analysis can improve the accuracy to choose non-overage vehicles and have a positive impact on profit.

**Appendix**

A - Data Exploration R code

## Histogram of total cost

>data %>%

>ggplot(aes(total.cost)) +

>geom\_histogram(binwidth = 300, fill = "green4", color = "white") +

>ggtitle('Histogram of Total cost')

>ggsave('histogram1.png', width = 8, height = 4)

#Histogram of lot.sale.days

>data %>%

>ggplot(aes(lot.sale.days)) +

>geom\_histogram(binwidth = 5, fill = "coral", color = "white") +

>ggtitle('Histogram of Sales day') +

>labs(y = "Sales days")

>ggsave('histogram2.png', width = 8, height = 4)

#correlation heatmap

>cormat <- round(cor(data[ ,c('mileage', 'vehicle.age', 'lot.sale.days', 'total.cost')]),2)

>melted\_format <- melt(cormat)

>ggplot(data = melted\_format, aes(x=Var1, y=Var2, fill=value )) +

>geom\_tile() + geom\_text(aes(label = value), color = 'white')

>ggsave('Correlation\_heatmap.png', width = 5, height = 5)

#Color set boxplot

>data %>% ggplot(aes(color.set, lot.sale.days)) +

>geom\_boxplot(color = "cyan4")

>ggsave('colorset.png', width = 5, height =5)

B - Model Building & Prediction

>model <- lm(lot.sale.days ~ mileage + total.cost + vehicle.age +

color.set + makex + state + vehicle.type, data = data.new[1:8738,])

>model2 <- lm(lot.sale.days ~ mileage + total.cost + color.set + make.model + vehicle.age + state, data = data.new[1:8738,])

#Histogram of the residuals

>ggplot(data = train, aes(model2$residuals)) +

geom\_histogram(binwidth = 10, color = "white", fill = "cyan4") +

labs(title = "Histogram of residuals 2.png")

test\_pred <- data.frame(predict = predict(model, data.new[8739:13102,]))

test['Pred\_sales\_type\_1'] <- as.factor(ifelse(test\_pred$predict > 119, "Auction",

ifelse(test\_pred$predict < 90, "Non-ovarage", "Overage")))

test['Act\_sales\_type'] <- as.factor(ifelse(test$lot.sale.days > 119, "Auction",

ifelse(test$lot.sale.days < 90, "Non-ovarage", "Overage")))

test\_pred2 <- data.frame(predict = predict(model2,data.new[8739:13102,]))

test['Pred\_sales\_type2'] <- as.factor(ifelse(test\_pred2$predict > 119, "Auction",

ifelse(test\_pred2$predict < 90, "Non-ovarage", "Overage")))

confusionMatrix(test$Act\_sales\_type, test$Pred\_sales\_type\_1)

confusionMatrix(test$Act\_sales\_type, test$Pred\_sales\_type2)

validate\_pred <- data.frame(predict = predict(model, data.new[13103: 17469,]))

validate['Pred\_sales\_type\_1'] <- as.factor(ifelse(validate\_pred$predict > 119, "Auction",

ifelse(validate\_pred$predict < 90, "Non-ovarage", "Overage")))

validate['Act\_sales\_type'] <- as.factor(ifelse(validate$lot.sale.days > 119, "Auction",

ifelse(validate$lot.sale.days < 90, "Non-ovarage", "Overage")))

validate\_pred2 <- data.frame(predict = predict(model2, data.new[13103: 17469,]))

validate['Pred\_sales\_type\_2'] <- as.factor(ifelse(validate\_pred2$predict > 119, "Auction",

ifelse(validate\_pred$predict < 90, "Non-ovarage", "Overage")))

validate$profit\_no\_model <- ifelse(validate$Act\_sales\_type == "Non-ovarage", validate$total.cost\*0.2, ifelse(validate$Act\_sales\_type == "overage", validate$total.cost\*0.1, -validate$total.cost\*0.2))

chosen\_vehicles <- validate %>% filter(Pred\_sales\_type\_2 == "Non-ovarage")

chosen\_vehicles$profit <- ifelse(chosen\_vehicles$Act\_sales\_type == "Non-ovarage", chosen\_vehicles$total.cost\*0.2, ifelse(chosen\_vehicles$Act\_sales\_type == "overage", chosen\_vehicles$total.cost\*0.1, -chosen\_vehicles$total.cost\*0.2))

(sum(chosen\_vehicles$profit) - sum(validate$profit\_no\_model))/sum(validate$profit\_no\_model)