Cosumer Choice Model Regression Analysis

Nawal Zehra Hasan

Research Question

Is there a significant relationship between sales and price of vehicles? Does the relationship vary depending on other factors including EPA_class, footprint and miles per gallon?

Introduction

This project is a regression analysis on data about consumer vehicle choice model data, collected by Oak Ridge National Laboratory for US EPA. I am investigating the relationship between price and sales for the vehicles. To add more depth to my analysis, I have added a few control variables;MPG, EPA class and footprint. As an analyst I am given the task to understand if sales of a vehicle are impacted by the price and some other explanatory variables such as footprint, miles per gallon and EPA_class. This is crucial to understand consumer choice and can also be further used for making predictions about sales for future vehicles. We can also make generalizations about vehicle sales in USA in the year 2008 with findings from our data.

Motivation

When agencies such as the US Environmental Protection Agency (EPA) establish greenhouse gas emissions standards for vehicles, understanding sales due to changes in footprint, fuel economy and prices provides insight into regulatory impacts.

Data

The data set is part a trilogy of data sets for three different years; 2008, 2010 and 2016. For the purpose of my analysis, I have chosen the base data set which belongs to the year 2008, to which further details changes were added associated with 2010 vehicles(for predictive analysis). I cannot make statements how representative the sample is, as it depends on the method of data collection & this is a secondary data source. However, by looking at the data I can see that it includes several manufacturers and models as well as EPA_class which includes wagons, cars, trucks and other vehicle types. Hence, there is variety in the data set. The data set is available **here**.

Exploratory Data Analysis

Understanding the data

Below is the list of shortlisted variables and the reasons behind their selection

1. **Sales**: Sales of the vehicle according to the price for it and also different confounders that impact price. This is the dependent variable, against which all other variables will be regressed. What are the possible reasons for a vehicle's sales to be higher or lower?

2. **Price**: Gives information regarding the vehicles price by model type and other factors
3. **EPA_class**: This is the type of vehicle ranging from SUV's to compact vehicles so see sales difference by size. Does the size of a vehicle have an impact on the sales of a vehicle? Considering USA, where there is no public transport, do people prefer purchasing bigger vehicles?
4. **Footprint**: the amount of CO2 emitted by a vehicle annually, measured in tonnes, and its relationship with the sales of a car. Does higher carbon footprint hinder customers from purchasing a particular model?
5. **Miles_per_gallon**: Distance traveled by a vehicle per gallon of fuel can have a significant relationship with sales of a car. Does that stand correct for our data set?

Data Munging

I begin my analysis by selecting only the variables we have shortlisted above into a data-set for consistency. The trends that seem most interesting to us for further exploration in assessing what impacts sales of vehicles are explored in detail. My data has no missing values and has a total of 524 observations. *figure 1* gives an overview of the data set. We have no missing observations in our data and 25% of our variables are discrete while 75% are continuous. I have renamed a few variables and also rounded off sales, footprint and MPG to 2 decimals.I have also selected only those variables that are relevant for my study. EPA_class which is a categorical variable in our data set needs to be coded as dummy variables. After checking how vehicles are categorized by vehicle type I have coded these as dummy variables to be further used in regression models.

Descriptive Statistics

The summary statistics reveal an extreme value of price $1734000. I decided to keep this as an extreme x value to see why the price of this vehicle is so high. Although, it could be a measurement error and may attenuize our slope coefficient. The summary statistics also show that average price of a vehicle in my data set is $51,651. This can also be higher because of the extreme value present in our data.An average vehicle in our data set has a fuel economy on average of about 24.68 miles per gallon. We also have extreme sales values as shown in *table 1* with the minimum value as $1, which is quite unlikely. However, we must keep these as we are testing variability in y(sales). However, this tends to have an impact on our average sales as the table shows. The table also shows the average miles per gallon returned by vehicles in our data is approximately 25, with a standard deviation of 6.4 miles per gallon from the average.

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
price	51 651.90	32 536.00	89 688.76	11 783.00	1 734 000.00	17 486.50	130 221.50
miles_per_gallon	24.68	23.78	6.43	12.00	65.78	17.06	36.69
sales	24 764.82	10 157.00	41 233.98	1.00	335 544.00	188.25	106 256.50
footprint	49.43	47.26	7.92	36.69	77.50	39.47	66.30

Variables Distribution

I checked the distribution of some variables including price. As price gave a right skewed distribution I chose to take the log of prices to give a close to normal distribution as depicted by *figure 4*. We can also observe that the distribution of sales variable is very close to normal if taken in log terms as shown in *figure 6*. Hence, we add another variable with log of sales. Both price and sales cannot be negative so transforming them was a possibility. Next, I used lowess as a method for nonparametric regression to check the association between price and sales of vehicles.

Multicollinearity

Before delving into the regressions, I used the correlation matrix to check whether independent variables in our analysis are correlated as shown in *figure 2*. This correlation is a problem as independent variables ought to be independent and it will impact our regression coefficient as it will show a biased average change in sales for a higher or lower price change. The heat map shows the relationship among all numerical variables of interest. ln(price) and ln(sales) are negatively correlated as we saw above with the lowess regression. ln(price) and footprint are negatively correlated, but it is not a strong correlation so I will keep both. Footprint and MPG are strongly negatively correlated. I have decided to drop miles per gallon as an explanatory variable as the footprint is an important variable in our data and analysis. This also makes intuitive sense, since vehicles with better mileage ought to have a lower carbon footprint.

Non-Parametric Regressions

With the help of a nonparametric regression such as lowess, I wanted to uncover the pattern of association between sales and the explanatory variables including price, footprint & EPA_class of the vehicle. We can see in *figure 7* that ln(sales) when regressed on ln(price) gives a negative slope in general. This was also visible in the correlation matrix. Using three types of transformations with log-log, level-log and log-level, we can observe that the most insightful of these is the log-log regression where we compare in relative terms as represented by *figures 7-9*. Regressing log sales on footprint we can see from *figure 10* that the slope is changing at approx 44 footprint so I chose to use spline with one knot show showing the change in slope.

Parametric Regressions

To model the regression I have chosen robust regression models as we can see from *table 1* that we have a few extra values in sales and price and since sales is our dependent variable we must be cautious in dropping these values as we ought to see the change in sales through our analysis and extreme values can be useful. Hence, lm robust seemed like a reasonable choice to model regressions.I used SUV's as reference category for *model 3*. This was a conscious choice as one way to choose a reference category is to see if it has more observations that others & in our case, SUV's had 160 observations so I chose that. Another important decision taken was to select a few types of vehicles as our data set has vehicles ranging from small to large for cars to trucks. Hence, for narrowing down my analysis, I selected all midsize vehicles.

Regression Models

$$ln(sales) := \beta_0 + \beta_1 ln(price)$$

1. **Model_1**: shows relative change in ln(sales) with a change in ln(price). I am analyzing whether sales and price have a significant relationship. In this case our null hypothesis will be that sales and price have no significant relationship i.e. our Beta coefficient is zero. While the alternate hypothesis is that sales and price have a significant relationship hence beta coefficient will not be zero.This is a simple linear regression model between one right hand side variable, with no controls. The intercept shows that when price is zero, sales is 27.47$ on average. This does not make sense as sales ought to be zero as price is zero. The slope coefficient shows that as price goes higher by 10%, sales go down by 18% on average. The confidence interval [-1.593, -1.941] can give us an understanding of the general population represented by our sample which is all the cars in USA sold in 2008. The sales of these cars will on average reduce between 16%-19% if prices go higher by 10%. The slope coefficient is also statistically significant at $p < 0.001$ so we can safely reject the null hypothesis in favor of our alternative and claim that price and sales of the vehicles have a significant relationship.

$$ln(sales) := \beta_0 + \beta_1 ln(price) + \beta_2(footprint < 44) + \beta_3(footprint >= 44)$$

2. **Model_2**: This model incorporates piece wise linear spline with one knot on the footprint. Now, I have two right hand side variables. The average sales of a vehicle given the footprint is less than is approximately 24070, while that with a footprint higher than 44 is 24982 as shown in *figure 12*. The slope coefficient of price slightly changes with the addition of the new variable i.e. footprint. The slope coefficient for footprint < 44 explains that with price held constant, among vehicles with < 44 footprint, sales is 3.35% higher on average with 10% higher footprint. However, the slope coefficient for vehicles with > 44 tonnes footprint, we can see that ln(sales) reduce 0.24% as carbon footprint of car increases by 10 tonnes. This is also shown in *figure 10*. The slope coefficient of footprint < 44 is significant at $p < 0.001$, whereas the slope coefficient of footprint > 44 tonnes is significant at $p < 0.05$. Hence, we need less proof to support the alternative hypothesis that footprint & sales have a significant relationship.

$$ln(sales) := \beta_0 + \beta_1 ln(price) + \beta_2 footprint + \beta_3 cars + \beta_4 truck + \beta_5 wagon + \beta_6 vans$$

3. **Model_3**: For this model, I have added dummy variables for types of vehicles which is EPA_class and I have also used footprint in its original form i.e. without splines for the ease of interpretation. Keeping SUV as base for 2 reasons; a) SUV are usually high powered as compared to all other types of vehicles. b) The number of observations for SUV is highest in our data set as shown in *figure 11*. With all other factors held constant, as footprint increases my 1 tonne the sales of a vehicle is higher by 7.3% on average. The slope coefficient for price is quite small as compared to model 1 and model 2 so we can say that type of vehicle is a confounder. We can also see that the coefficient for cars is positive while for the rest of the types of vehicles it is negative. Holding the price and footprint variables constant, the sales of a vehicle that is a car is 47% higher than an SUV in our data set. This also make intuitive sense as cars are sold generally more for several reasons compared to SUV's.

$$ln(sales) := \beta_0 + \beta_1 ln(price) + \beta_2 footprint + \beta_3 cars + \beta_4 truck + \beta_5 wagon + \beta_6 vans + \beta_7 footprint * ln(price)$$

4. **Model_4**: For model 4, I have added an interaction term for ln(price) and footprint to see how they are associated with sales. The slope coefficient for the interaction term shows the difference between sales with 1 tonne more footprint is less by 2.3% on average, if the price is higher by 10%. With a standard error of 0.014, we can calculate the confidence interval at 95% is [-0.051,0.005]. This contains zero so we cannot reject the null that footprint & price are not associated with sales of vehicles. Thus, in the general pattern represented by the data which is all the cars in USA in 2008, it is possible that the association between price and sales is the same, whatever the amount of footprint is. This is also reaffirmed with the significance level of the coefficient as it is significant as $p < 0.1$ only.

External Validity

For testing external validity as part of robustness check, the sample collected for the years 2010 and 2016 can be used to run similar regression and test the results. This different setting of a different period of time can enable us to make generalizations whether our results stand correct for the population represented by the data using different data sets from 2010 and 2016.

Conclusion

In conclusion, we can say that price and sales are highly negatively correlated but the same cannot be said for control variables and sales of vehicles.My expectation was that higher footprint might result in lower sales and we did see that for model 2 which has splines for footprint. This was also confirmed by *figure 10* that sales tend to increase with footprint but beyond 44 they tend to stagnate and then also decrease at the end of the graph. Although, other factors such as mileage, colour of the vehicle, features of the vehicle may be perhaps better factors to incorporate but for my study I wanted to see if customers are conscious about purchasing the vehicle which have a higher footprin & how price and footprint interact to influence sales of the vehicle.

APPENDIX

Figure 1 - Data Overview

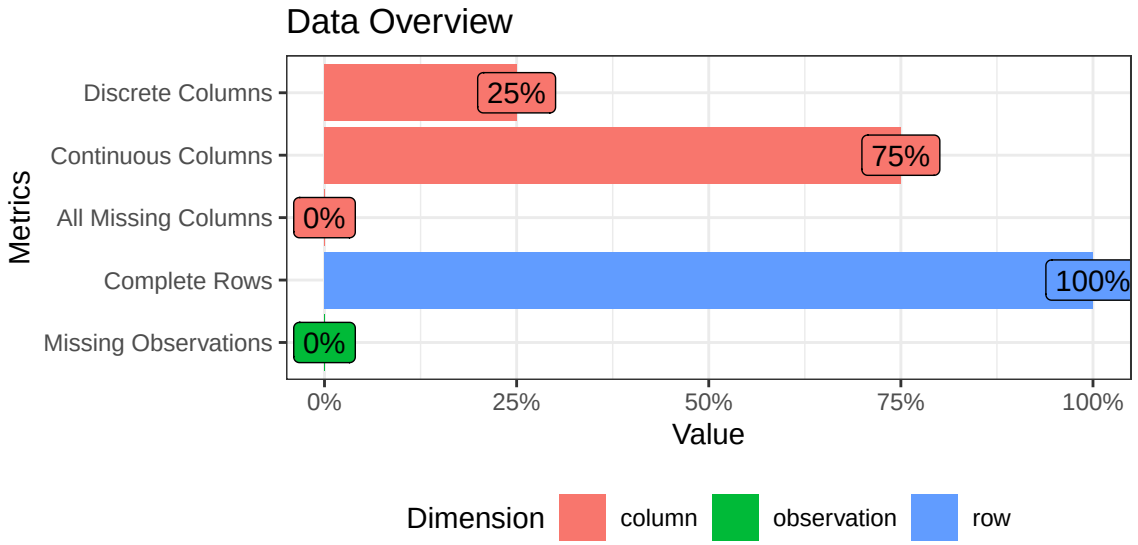

Figure 2- Correlation Heat Map

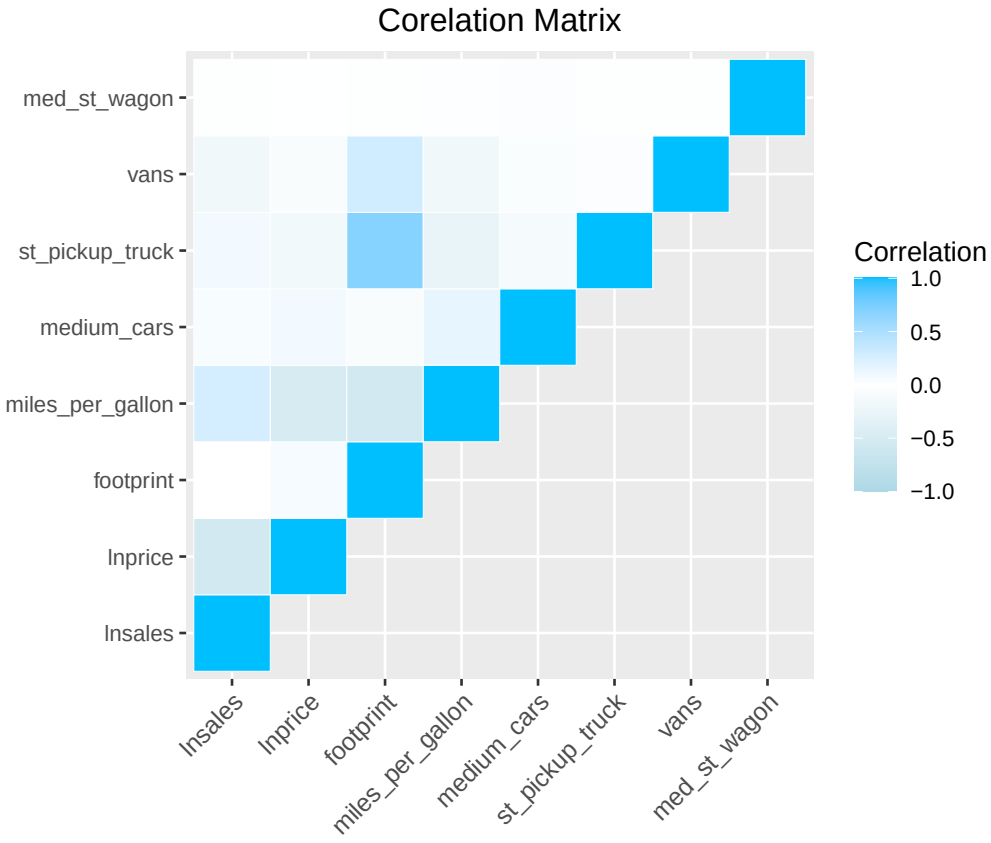

Figure 3 - Distribution of Price

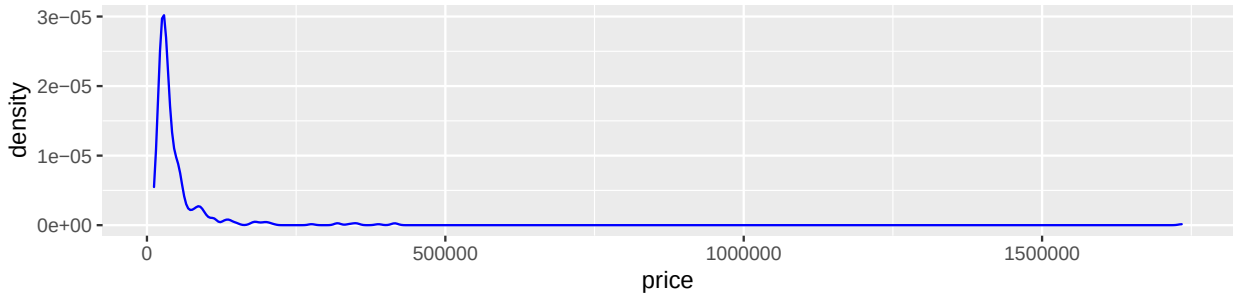

Figure 4 - Distribution of Price with ln(price)

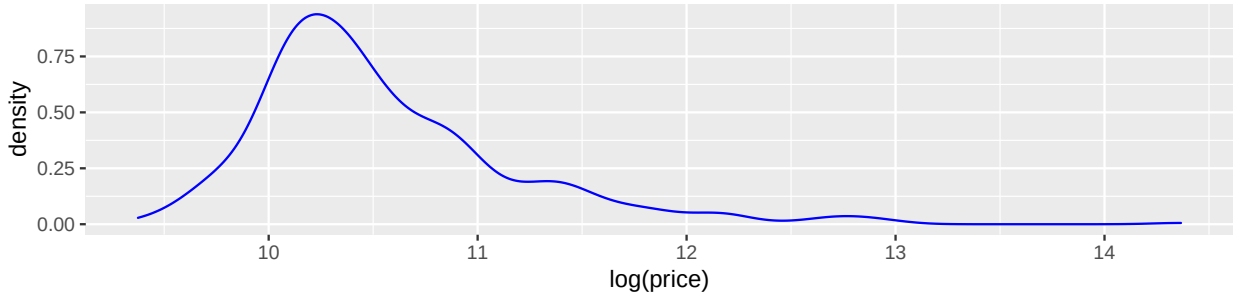

Figure 5 - Distribution of Sales

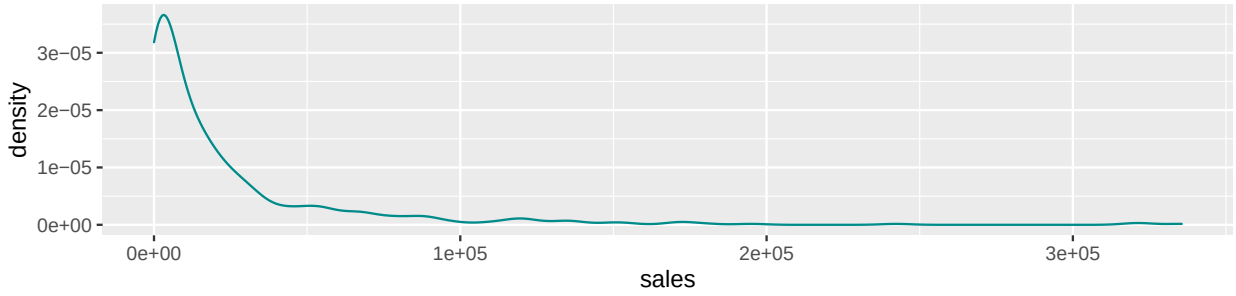

Figure 6 - Distribution of Sales with ln(sales)

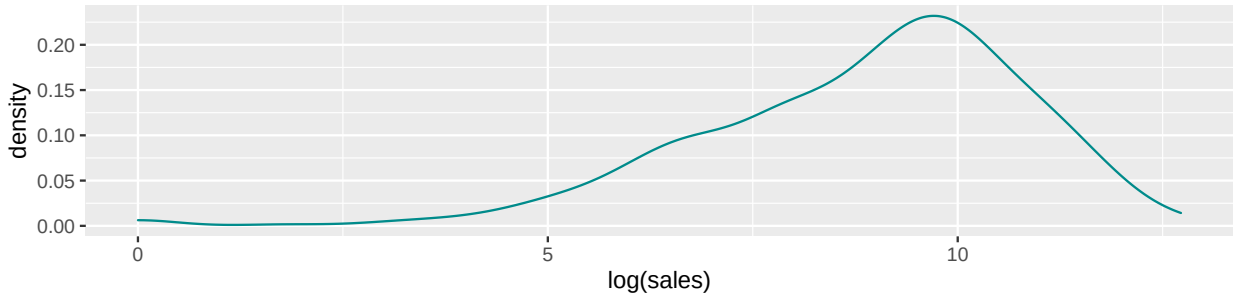

Figure 7 - Lowess Regression with scatter plot - ln(price) & ln(sales)

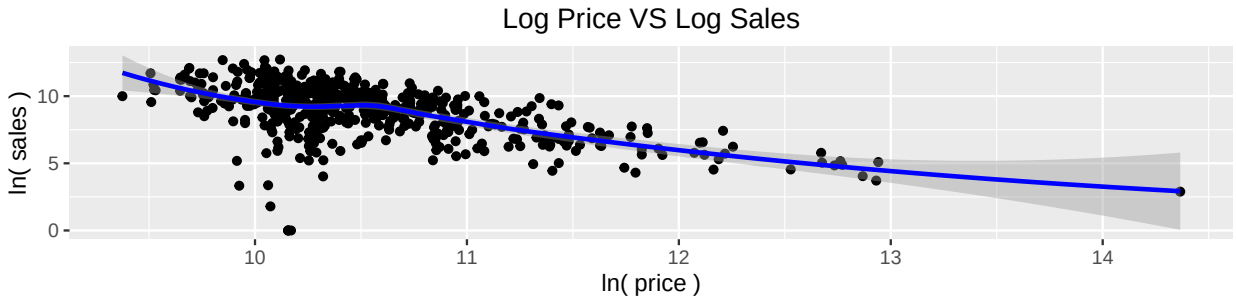

Figure 8 - Lowess Regression with scatter plot - price & ln(sales)

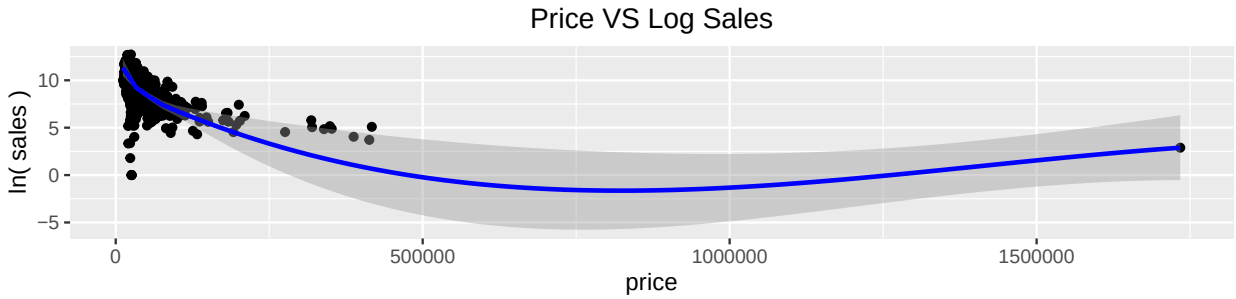

Figure 9 - Lowess Regression with scatter plot - ln(price) & sales

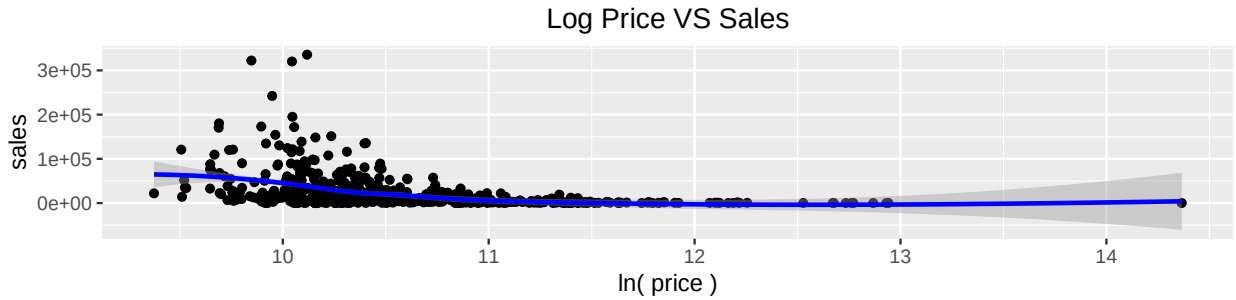

Figure 10 - Lowess Regression with scatter plot - footprint & lnsales

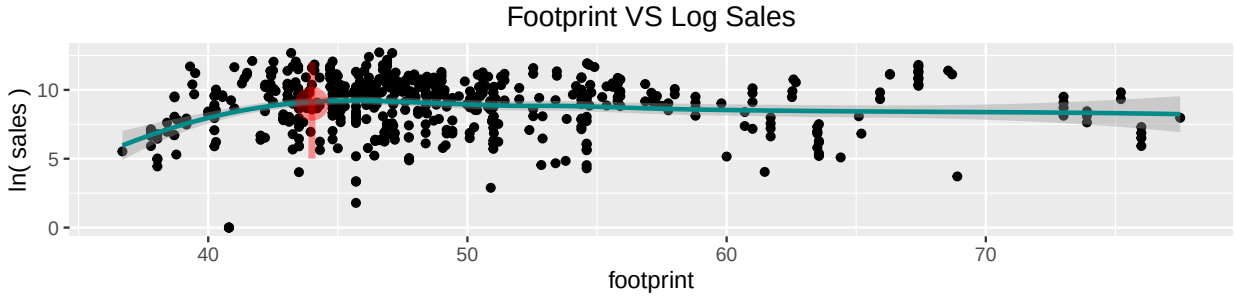

Figure 11 - EPA_class - Types of Vehicles

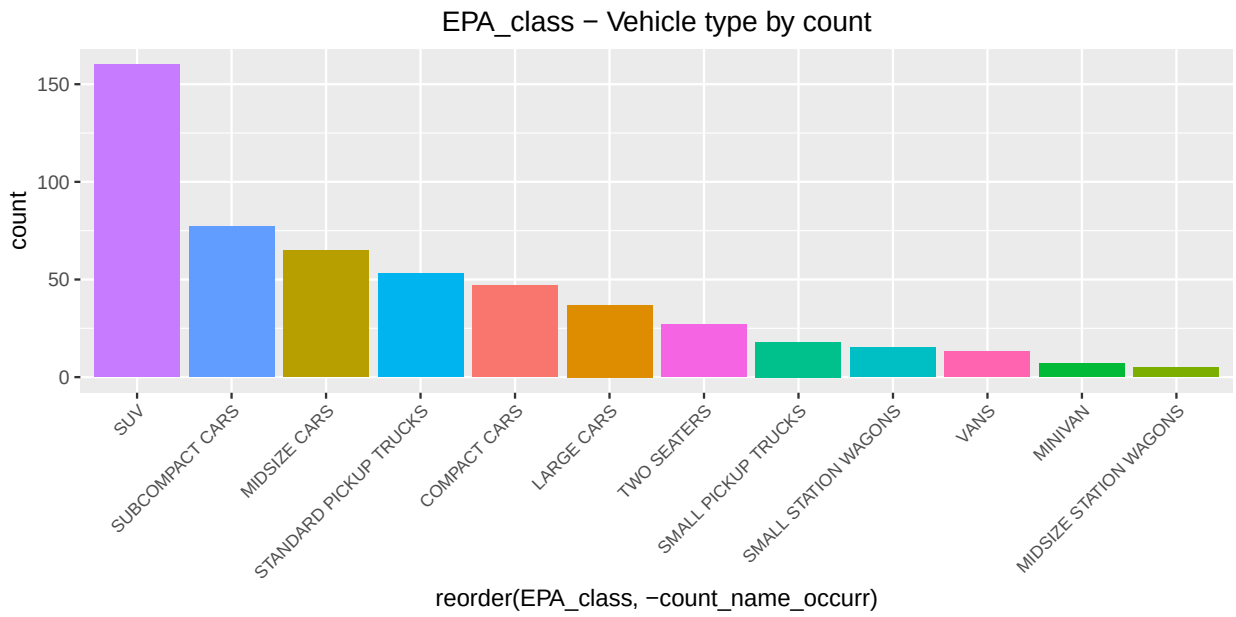

Figure 12 - Average Sales by footprint size

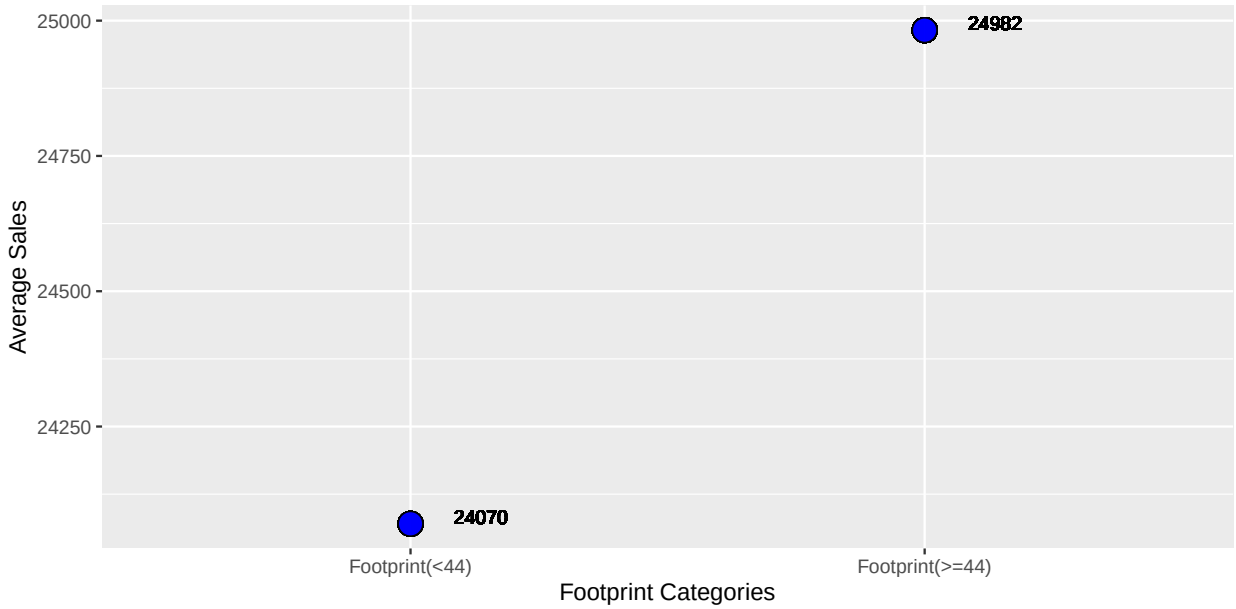

Table 2: Regression table of Consumer Choice Model Vehicles

	Model_1	Model_2	Model_3	Model_4
Intercept	27.470 ****	12.883 ****	26.960 ****	14.821 *
	(0.957)	(2.789)	(1.170)	(7.642)
ln(price)	-1.767 ****	-1.750 ****	-2.042 ****	-0.909
	(0.087)	(0.087)	(0.095)	(0.680)
Footprint<44		0.335 ****		
		(0.057)		
Footprint>=44		-0.024 **		
		(0.011)		
Footprint			0.073 ****	0.325 **
			(0.014)	(0.157)
Cars			0.472 ***	0.494 ***
			(0.171)	(0.171)
Trucks			-1.463 ****	-1.681 ****
			(0.299)	(0.347)
Vans			-4.248 ****	-4.461 ****
			(0.329)	(0.384)
Wagon			-0.610 **	-0.618 **
			(0.245)	(0.253)
Footprint:ln(price)				-0.023 *
				(0.014)
N	524	524	524	524
R2	0.304	0.354	0.398	0.400

***** $p < 0.001$; *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.*