MA575 Lab

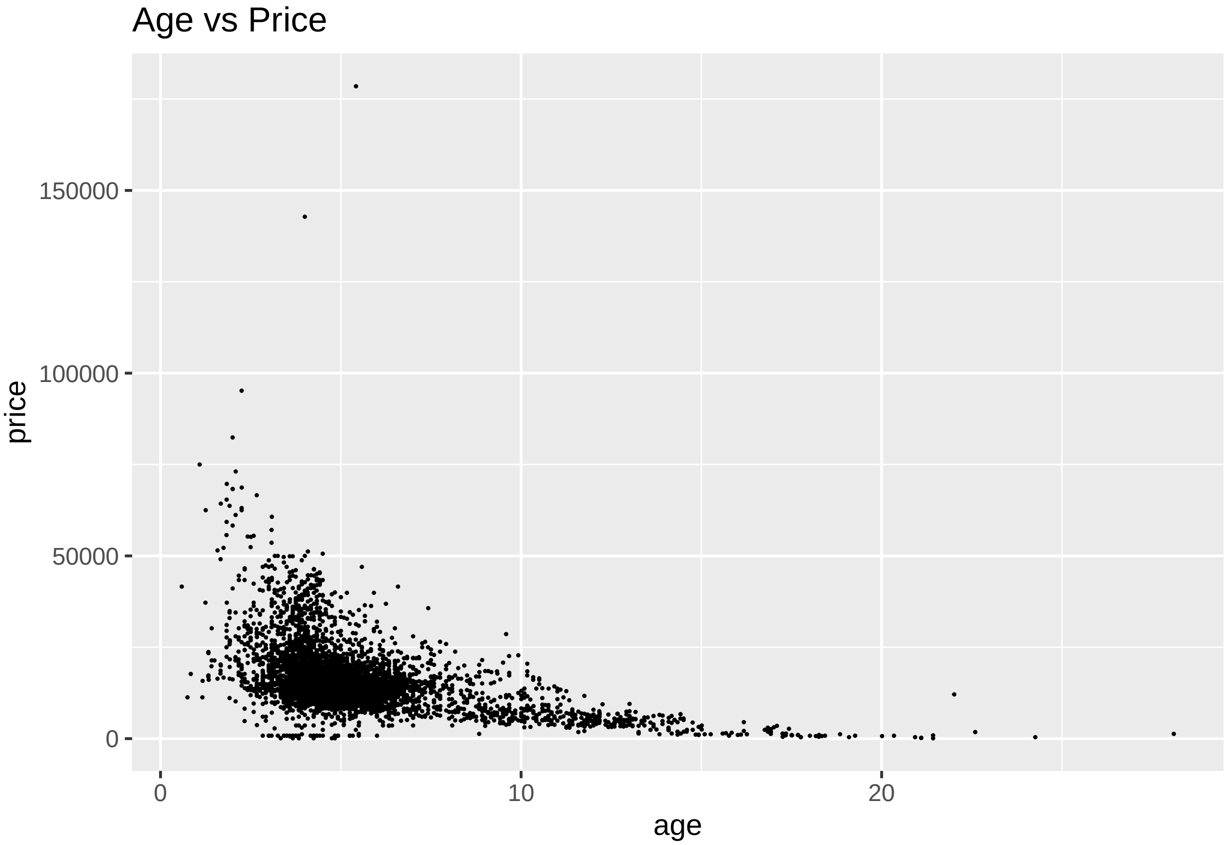
Henry Bayly

2023-02-12

We will build a linear model predicting the price of each BMW as a function of the age of the car. To do this we will first take the difference between the registration date and sale date of the car. After finding the difference in days, we divide by 365 to get the age in years (as a decimal). Below is a summary of the ages of our cars.

Measure	Value
Min.	0.5890411
1st Qu.	4.0794521
Median	4.8356164
Mean	5.4348618
3rd Qu.	5.8356164
Max.	28.1041096

Let's plot the data to understand the relationship.



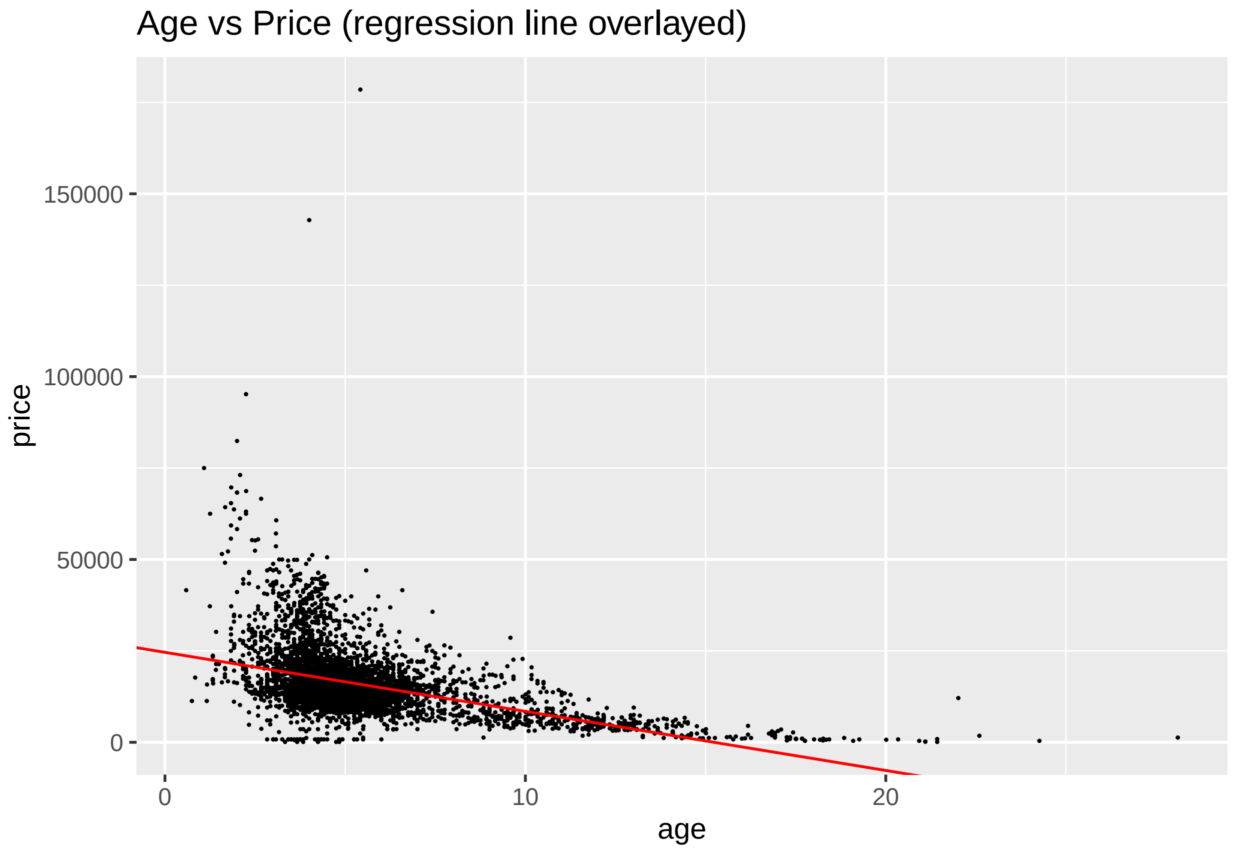
It is difficult to say whether or not there is a significant linear trend in the data. However, generally as the age of the car increases the plot shows that the price of the car decreases. There seem to be a couple outliers that are around 5 years old but sold for an unusual amount of money. Further inspection will be done later to see if these points should be removed.

Now we will create a linear model of the form Price = $\beta_0 + age * \beta_1$

```
##
## Call:
## lm(formula = price ~ age, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19227  -4761  -1644   2408 162651
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24606.82     280.23   87.81   <2e-16 ***
## age         -1615.26      46.71  -34.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8258 on 4841 degrees of freedom
## Multiple R-squared:  0.1981, Adjusted R-squared:  0.1979
## F-statistic:  1196 on 1 and 4841 DF,  p-value: < 2.2e-16
```

Our model shows that age is a significant linear predictor of age (p $< 2.2e^{-16}$, F = 1196, df = (1,4841)). Ourregression coefficient for age is -1615.23 (SE = 46.71,t=-34.58,p$<2e^{-16}$). This means that for every 1 year increase in the age of the car we would predict the price of the car to drop by about 1,600 dollars. Our model had an adjusted R^2 of 0.1979. This means that, when accounting for the number of predictors, age captures about 19.79% of the variability in price.

Let's overlay the regression line onto the scatterplot to see how it fits the data.



While the model seems to capture the general trend of the data, we can see a good amount of variability on both sides of the line. We also see that the variability on each side of the regression line decreases as the age increases. This indicates that there could be some aspect of the relationship missing from our model. Since this is only a univariate analysis, multiple regression models will be built later to investigate this.