Lab 2 model_Catherine

2024-02-25

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
bmw <- read.csv("/Users/rui/OneDrive/Documents/BU/MA575 Linear Models/Labs/Lab2/BMW Price Data/BMW-pric
# create summary
summary(bmw)
```

```
##   maker_key          model_key            mileage         engine_power
##  Length:4843        Length:4843        Min.   :    -64   Min.   :  0
##  Class :character   Class :character   1st Qu.: 102914   1st Qu.:100
##  Mode  :character   Mode  :character   Median : 141080   Median :120
##                                        Mean   : 140963   Mean   :129
##                                        3rd Qu.: 175196   3rd Qu.:135
##                                        Max.   :1000376   Max.   :423
##  registration_date      fuel           paint_color          car_type
##  Length:4843        Length:4843        Length:4843        Length:4843
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  feature_1       feature_2       feature_3       feature_4
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2181      FALSE:1004      FALSE:3865      FALSE:3881
##  TRUE :2662      TRUE :3839      TRUE :978       TRUE :962
##
##
##
##  feature_5       feature_6       feature_7       feature_8
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2613      FALSE:3674      FALSE:329       FALSE:2223
##  TRUE :2230      TRUE :1169      TRUE :4514      TRUE :2620
##
##
##
##      price          sold_at            obs_type
##  Min.   :   100   Length:4843        Length:4843
##  1st Qu.: 10800   Class :character   Class :character
##  Median : 14200   Mode  :character   Mode  :character
```

```
##  Mean   : 15828
##  3rd Qu.: 18600
##  Max.   :178500
```

```
sold_at_split <- strsplit(bmw$sold_at, "/")

registration_split <- strsplit(bmw$registration_date, "/")

# assign month only; all sold in 2018
bmw$month_sold <- sapply(sold_at_split, function(x) as.integer(x[1]))

bmw$year_sold <- sapply(sold_at_split, function(x) as.integer(x[3]))

bmw$month_registered <- sapply(registration_split, function(x) as.integer(x[1]))

bmw$year_registered <- sapply(registration_split, function(x) as.integer(x[3]))

price <- bmw$price # our y variable
engine_power <- bmw$engine_power # our x variable

length(price)
```

```
## [1] 4843
```

```
length(engine_power)
```

```
## [1] 4843
```

```
m2 <- lm(price~engine_power)
summary(m2)
```

```
##
## Call:
## lm(formula = price ~ engine_power)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30334  -3335    -36   2591 161764
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3661.247    352.266  -10.39   <2e-16 ***
## engine_power   151.094      2.614   57.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7093 on 4841 degrees of freedom
## Multiple R-squared:  0.4083, Adjusted R-squared:  0.4082
## F-statistic:  3341 on 1 and 4841 DF,  p-value: < 2.2e-16
```

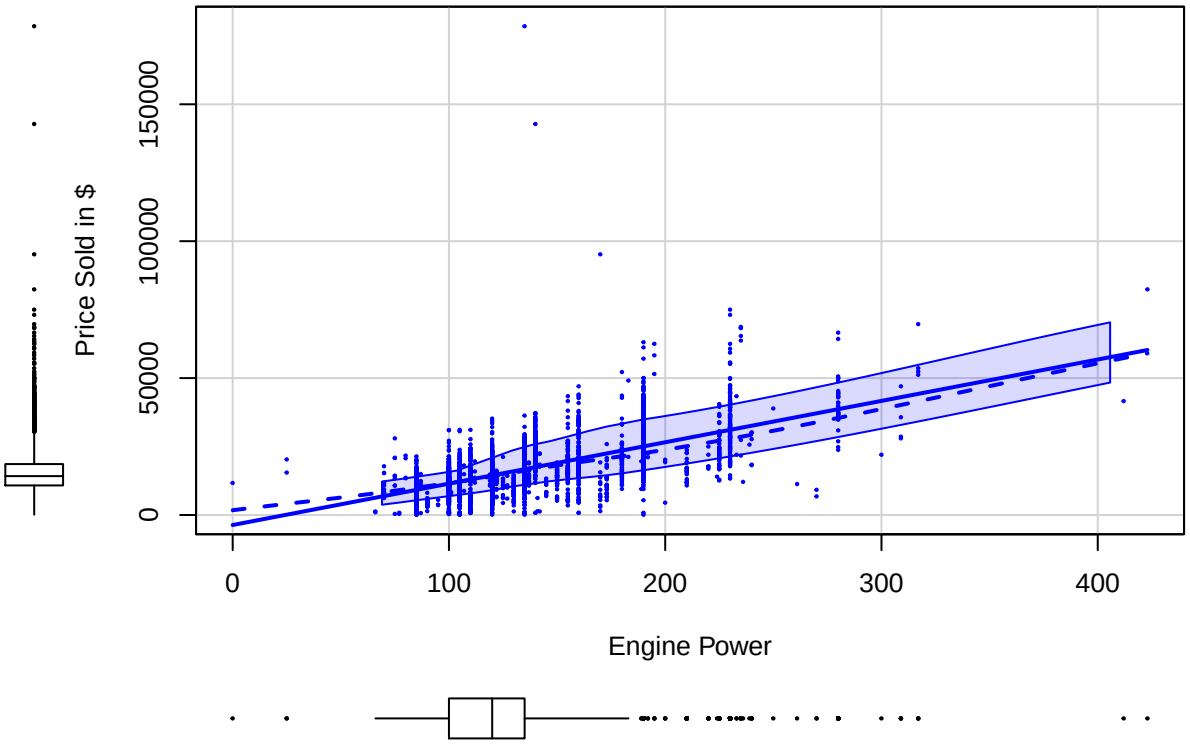
```r
# check the distribution of 2 variables
summary(engine_power) # mean > median, potentially right-skewed
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     100     120     129     135     423
```

```r
summary(price) # mean > median, potentially right-skewed
```

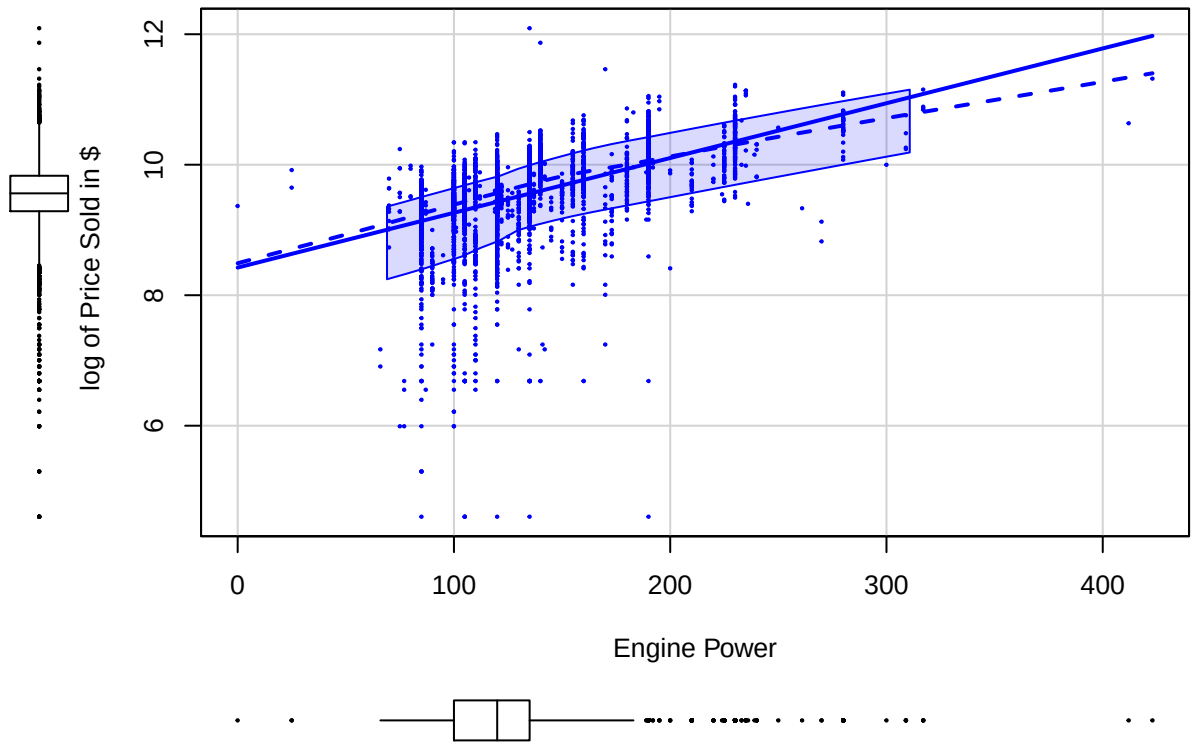
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     100   10800   14200   15828   18600  178500
```

```r
scatterplot(engine_power, price,
     ylab="Price Sold in $", xlab="Engine Power",
     pch=19, cex=0.2)
```




```r
# boxplot shows both variable is not normally distributed, scatterplot detects extreme outliers
```

```r
scatterplot(engine_power, log(price),
     xlab="Engine Power", ylab=" log of Price Sold in $",
     pch=19, cex=0.2)
```




```r
# make a new dataframe for cleaning outlier
model_data <- data.frame(engine_power = engine_power, price = price)
# Use leverage to check the outlier
lev <- hatvalues(m2)
model_data$filter1 <- lev <= (4/length(engine_power))
# use z-score for residuals to check the outliers
resid = residuals(m2)
z_resid = (resid - mean(resid))/sd(resid)
model_data$filter2 <- abs(z_resid)<=3
cleaned_data <- model_data[model_data$filter1 != FALSE & model_data$filter2 != FALSE, ]
cleaned_data <- cleaned_data[, -3:-4]
```

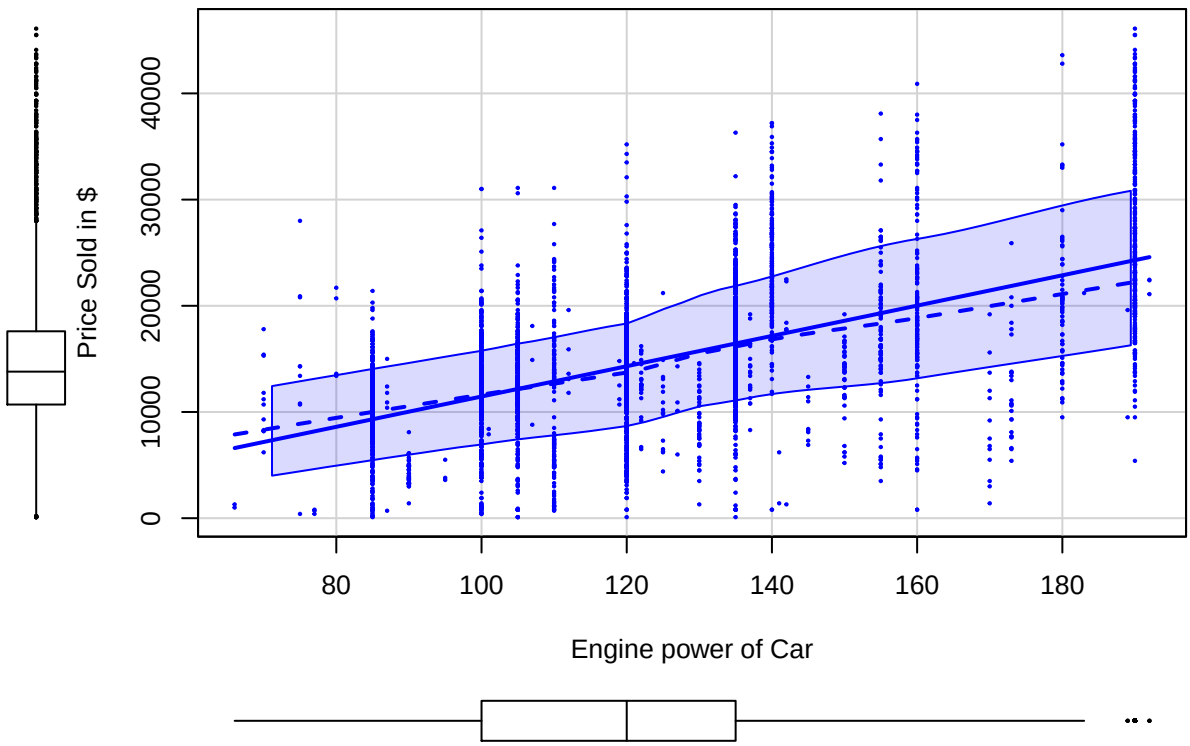
```r
summary(cleaned_data$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     100   10700   13800   14666   17600   46100
```

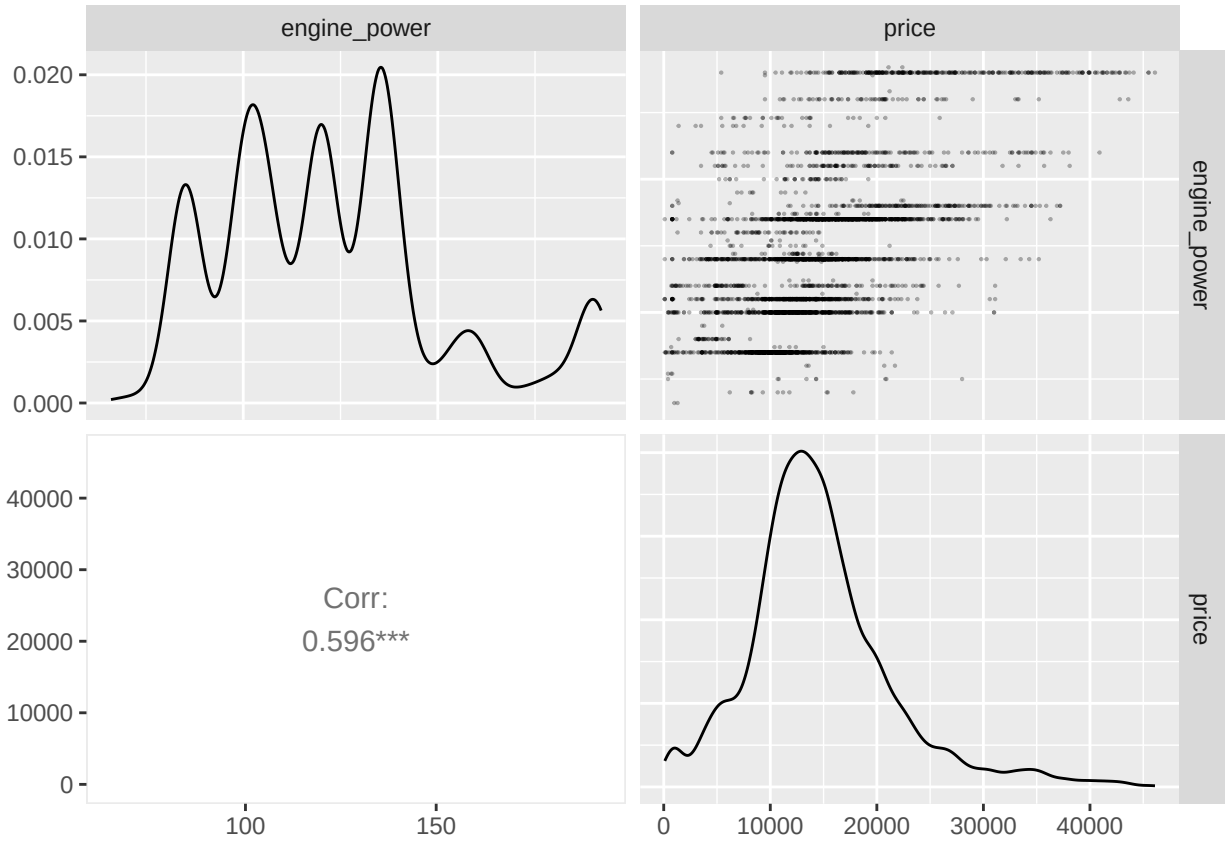
```r
summary(cleaned_data$engine_power)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    66.0   100.0   120.0   122.5   135.0   192.0
```

```r
# plot the cleaned data (without outlieers)
scatterplot(cleaned_data$engine_power, cleaned_data$price,
     ylab="Price Sold in $", xlab="Engine power of Car",
     pch=19, cex=0.2)
```

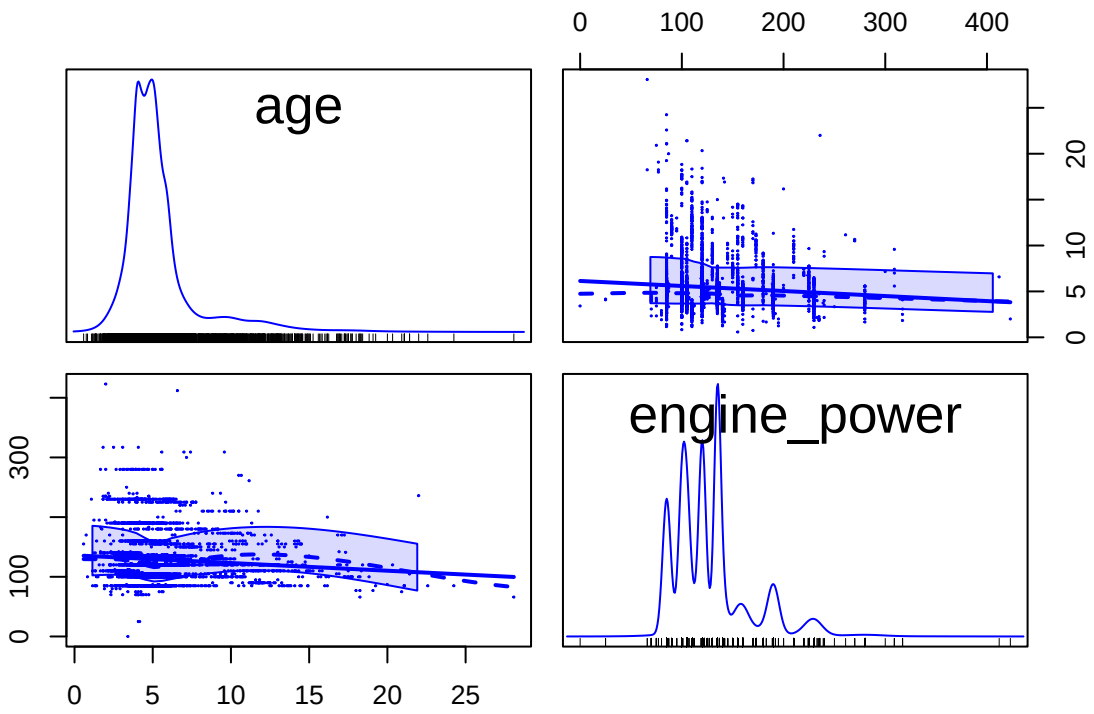



```r
# use the correlation matrix plot to check linear association and distribution
ggpairs(cleaned_data,
        upper=list(continuous=wrap("points", alpha=0.3, size=0.1)),
        lower=list(continuous=wrap('cor', size=4)))
```




```
# age normally distributed
# price right skewed, need log transformation
# some negative linear association as r = -0.357
```

```
# check correlation between age and engine power - almost no correlation
age <- bmw$year_sold-bmw$year_registered + (1/12)*(bmw$month_sold - bmw$month_registered)
scatterplotMatrix(~ age + engine_power,
                  pch=19, cex=0.1)
```


age
engine_power
0
100
200
300
400
0
5
10
20
0
100
300
0
5
10
15
20
25