Report3

Lab2 Group A: Lee, Joshua; Liu, Kaiyi; Pulsone, Nathaniel; Wang, Mengyao; Xu, Zexian; Yang, Xiaojing

Loading the Library and DataSet

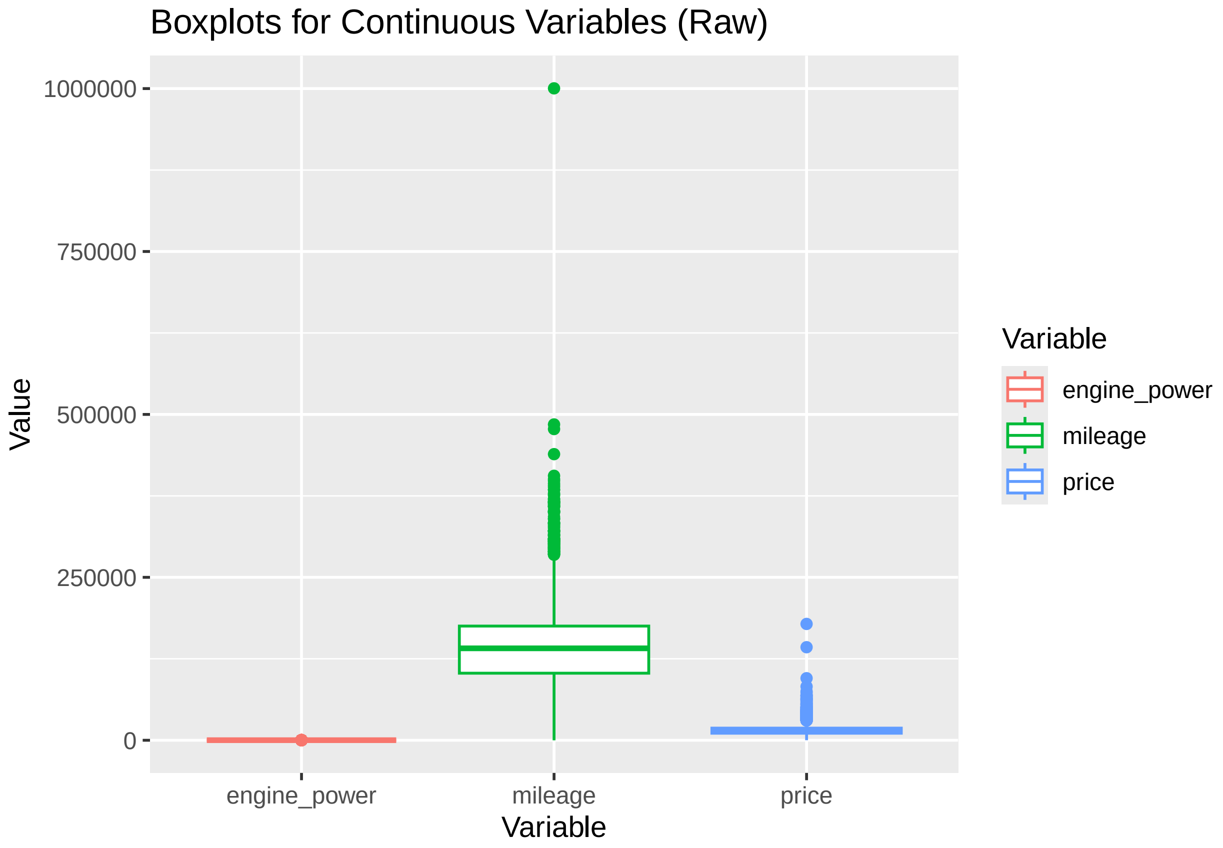
```r
mylibrary <- c("tidyverse", "cowplot", "GGally", "MASS")
invisible(lapply(mylibrary, library, character.only = TRUE))
```

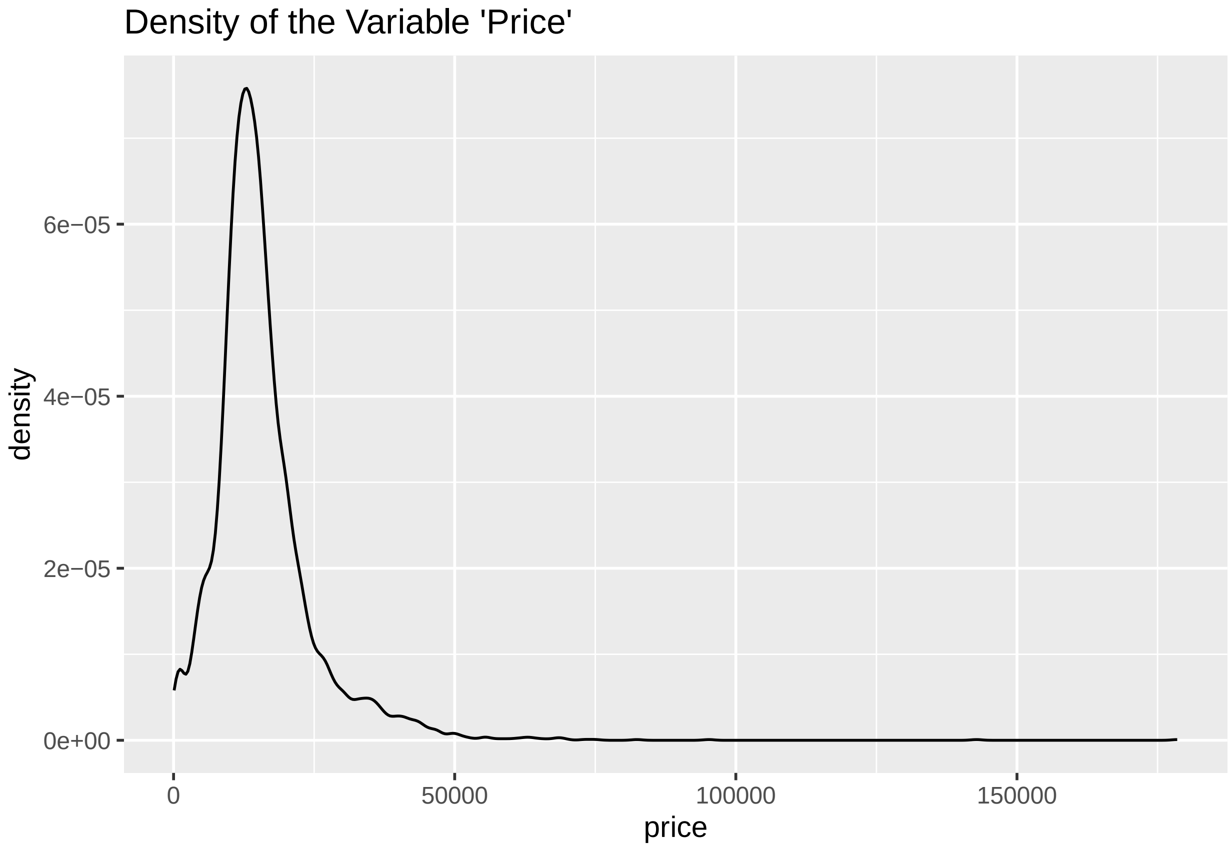
```
## Warning: package 'lubridate' was built under R version 4.4.2

## Warning: package 'cowplot' was built under R version 4.4.2

## Warning: package 'GGally' was built under R version 4.4.2
```

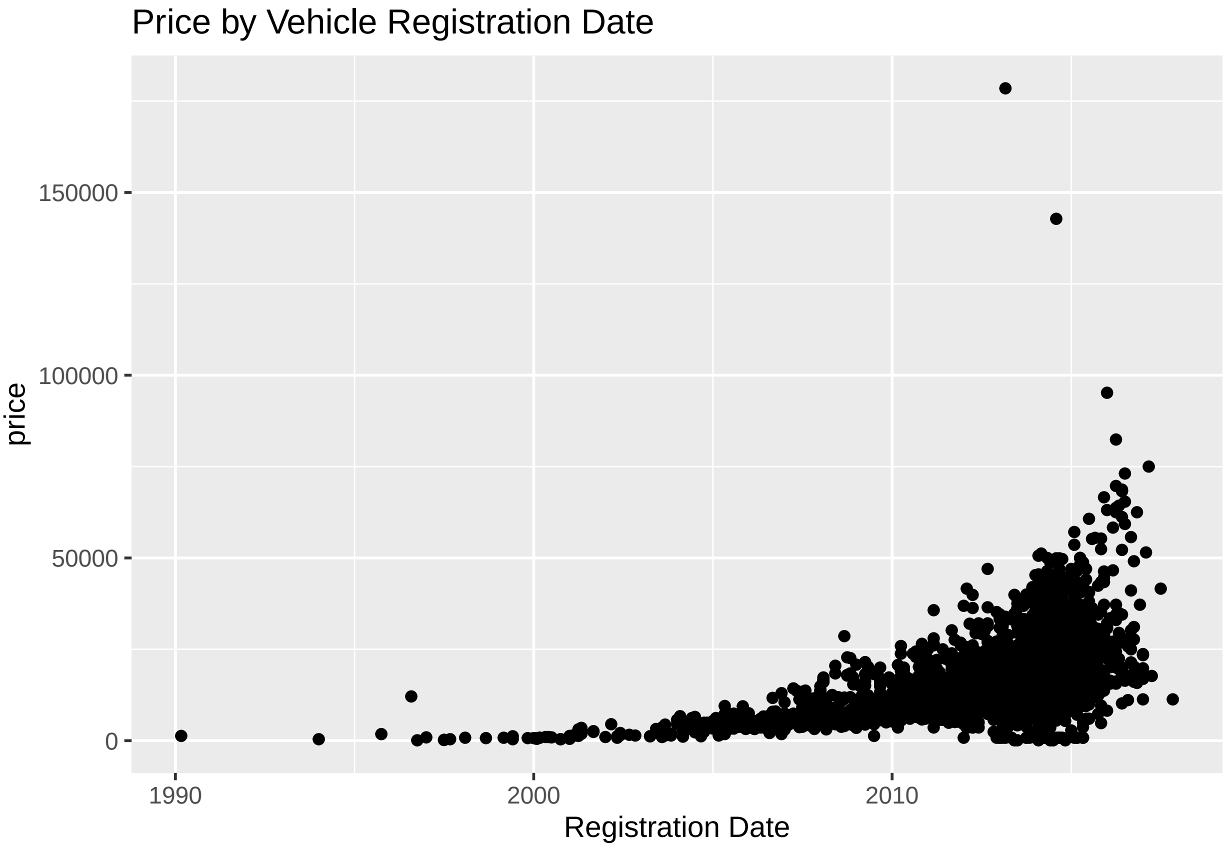
```r
### load the data
dat_bmw <- read.csv("BMWpricing_updated.csv")
```

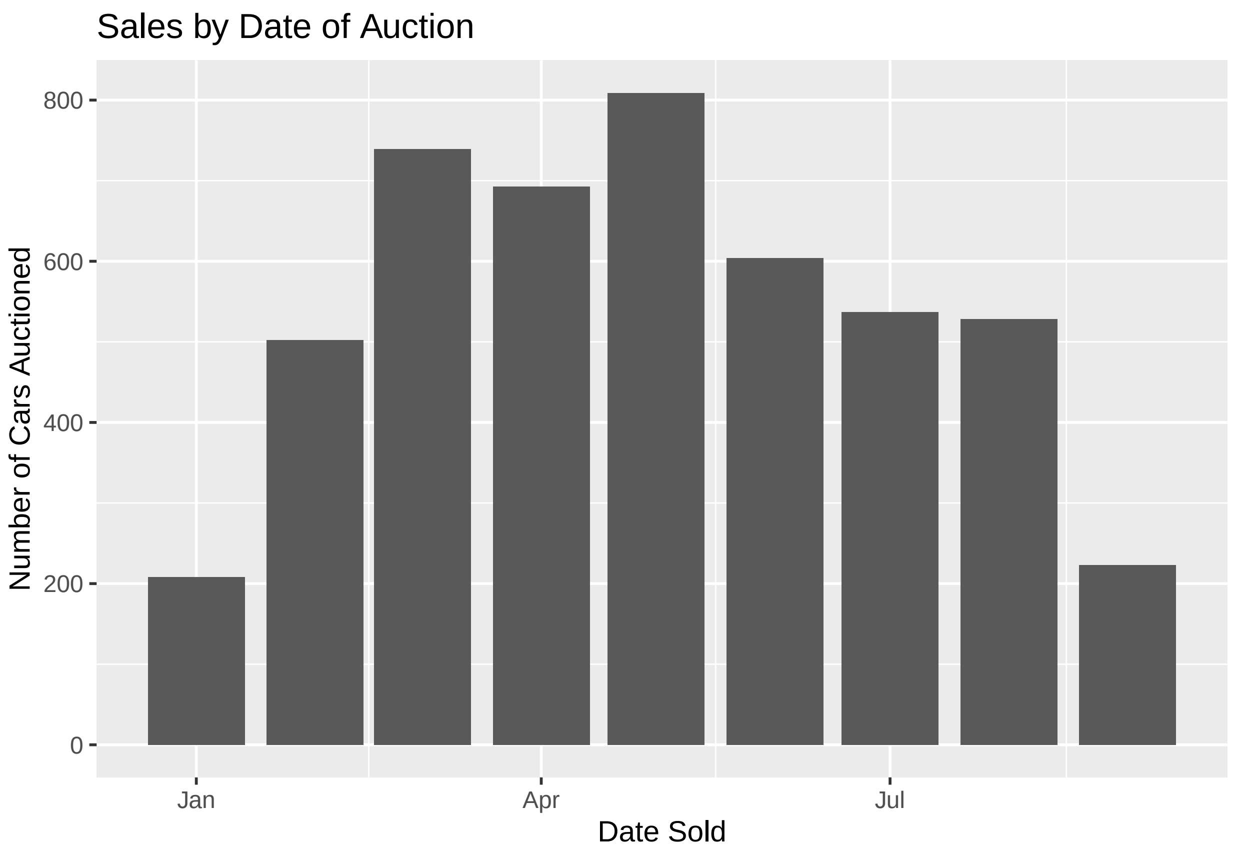
Data Overview

Sourced from Kaggle, the dataset provides information on roughly 5000 used BMW cars sold in a business-to-business auction. Notable variables included are the price, car model, color, mileage, engine power, and various categorical descriptors.




Density of the Variable 'Price'
density
6e−05
4e−05
2e−05
0e+00
0
50000
100000
150000
price


Price by Vehicle Registration Date
price
150000
100000
50000
0
1990
2000
2010
Registration Date



From the graphs above, along with analysis that can be found in the Appendix section, we can see that the cars from the auction were all registered between March 1990 and November 2017, and the auction took place from January to September 2018. The cars spanned 75 different BMW models, 10 different colors, and 4 different fuel types.

The pricing of the cars is most concentrated around 15000, with a median price of 14200. The distribution of price is skewed heavily to the right. Because price is our main response variable, we will see that this causes the residuals of the constructed models to be right-skewed as well. The skewness of the variables can be somewhat remedied with a log-transformation on price.

Although it is hard to identify the feature variables in the dataset, we can see That most of the vehicles posses features 2 and 7, while most do not have features 3, 4 and 6. Each of the features when present tend to increase the price by $5000 to $10000, except for feature 7, which actually decreases the price by about $185

Lastly, there are a few unusual values and outliers to consider in the dataset. There are two cars in the dataset that were sold for more than $100,000 which is unusual enough to provide a high leverage and skew our constructed models. For this reason, we will remove that observation from the dataset. In addition, there is a single car with over 1 million miles on it, which we speculate was an error in data entry, so we will also remove this from the dataset. Finally, there are a number of observations with a negative mileage, or an engine power of zero. These values are also either errors in data entry, or indicative of a special case, such as scrapped or salvaged car. For this reason, we will not remove the observations, but instead set the negative and zero values to NA.

```r
dat_bmw_clean <- dat_bmw|>
  filter(mileage < 500000)|>
  filter(price < 100000)|>
  mutate(mileage = ifelse(mileage < 0, NA, mileage))|>
  mutate(engine_power = ifelse(engine_power <= 0, NA, engine_power))

head(dat_bmw_clean)
```

```
##   maker_key model_key mileage engine_power registration_date   fuel paint_color
## 1       BMW       118  140411          100          2/1/2012 diesel       black
## 2       BMW        M4   13929          317          4/1/2016 petrol        grey
## 3       BMW       320  183297          120          4/1/2012 diesel       white
## 4       BMW       420  128035          135          7/1/2014 diesel         red
## 5       BMW       425   97097          160         12/1/2014 diesel      silver
## 6       BMW       335  152352          225          5/1/2011 petrol       black
##      car_type feature_1 feature_2 feature_3 feature_4 feature_5 feature_6
## 1 convertible      TRUE      TRUE     FALSE     FALSE      TRUE      TRUE
## 2 convertible      TRUE      TRUE     FALSE     FALSE     FALSE      TRUE
## 3 convertible     FALSE     FALSE     FALSE     FALSE      TRUE     FALSE
## 4 convertible      TRUE      TRUE     FALSE     FALSE      TRUE      TRUE
## 5 convertible      TRUE      TRUE     FALSE     FALSE     FALSE      TRUE
## 6 convertible      TRUE      TRUE     FALSE     FALSE      TRUE      TRUE
##   feature_7 feature_8 price  sold_at obs_type
## 1      TRUE     FALSE 11300 1/1/2018 Training
## 2      TRUE      TRUE 69700 2/1/2018 Training
## 3      TRUE     FALSE 10200 2/1/2018 Training
## 4      TRUE      TRUE 25100 2/1/2018 Training
## 5      TRUE      TRUE 33400 4/1/2018 Training
## 6      TRUE      TRUE 17100 2/1/2018 Training
```

Appendix

The following data chunk was used to create summary statistics for the data overview section

```
length(unique(dat_bmw$model_key))
```

```
## [1] 75
```

```
unique(dat_bmw$paint_color)
```

```
##  [1] "black"  "grey"   "white"  "red"    "silver" "blue"   "orange" "beige" 
##  [9] "brown"  "green"
```

```
unique(dat_bmw$fuel)
```

```
## [1] "diesel"        "petrol"        "hybrid_petrol" "electro"
```

```
range(as.Date(dat_bmw$registration_date, format = "%m/%d/%Y"), na.rm=T)
```

```
## [1] "1990-03-01" "2017-11-01"
```

```r
range(as.Date(dat_bmw$sold_at, format = "%m/%d/%Y"), na.rm=T)
```

```
## [1] "2018-01-01" "2018-09-01"
```

```r
min(dat_bmw$mileage)
```

```
## [1] -64
```

```r
min(dat_bmw$engine_power)
```

```
## [1] 0
```

```r
min(dat_bmw$price)
```

```
## [1] 100
```

```r
median(dat_bmw$price)
```

```
## [1] 14200
```

```r
range(dat_bmw$engine_power)
```

```
## [1]   0 423
```

```r
mean(dat_bmw$feature_1 == T)
```

```
## [1] 0.5496593
```

```r
mean(dat_bmw$feature_2 == T)
```

```
## [1] 0.7926905
```

```r
mean(dat_bmw$feature_3 == T)
```

```
## [1] 0.2019409
```

```r
mean(dat_bmw$feature_4 == T)
```

```
## [1] 0.1986372
```

```r
mean(dat_bmw$feature_5 == T)
```

```
## [1] 0.4604584
```

```r
mean(dat_bmw$feature_6 == T)
```

```
## [1] 0.2413793
```

```r
mean(dat_bmw$feature_7 == T)
```

```
## [1] 0.9320669
```

```r
mean(dat_bmw$feature_8 == T)
```

```
## [1] 0.540987
```

```r
coef(lm(dat_bmw$price ~ dat_bmw$feature_7))
```

```
##           (Intercept) dat_bmw$feature_7TRUE 
##            16010.3343             -195.5359
```

```r
max(dat_bmw$mileage)
```

```
## [1] 1000376
```