P8106 HW3

Brian Jo Hsuan Lee

3/22/2022

Load packages

```
library(tidyverse)
library(AppliedPredictiveModeling)
library(caret)
```

Import and tidy data

```
data = read_csv("auto.csv") %>%
  mutate(
    year = factor(year),
    origin = factor(origin),
    mpg_cat = factor(mpg_cat)
  )
```

```
## Rows: 392 Columns: 8
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): mpg_cat
## dbl (7): cylinders, displacement, horsepower, weight, acceleration, year, or...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Partition the data for model training

```
set.seed(2022)

# partition data into training and testing sets into randomized 4:1 splits
train_index = createDataPartition(y = data$mpg_cat, p = 0.7, list = FALSE)
train_data = data[train_index, ]
test_data = data[-train_index, ]

# matrices of predictors
train_pred = model.matrix(mpg_cat ~ ., train_data)[ ,-1]
test_pred = model.matrix(mpg_cat ~ ., test_data)[ ,-1]

# vectors of response
train_resp = train_data$mpg_cat
test_resp = test_data$mpg_cat
```

Calculate descriptive statistics: quantile data for the continuous variables and count data for the categorical variables. Number of cylinders is arguably an ordinal categorical variable but is treated as a continuous variable here. Most cars have an American origin (category 1), and the number of high and low mileage car samples are the same.

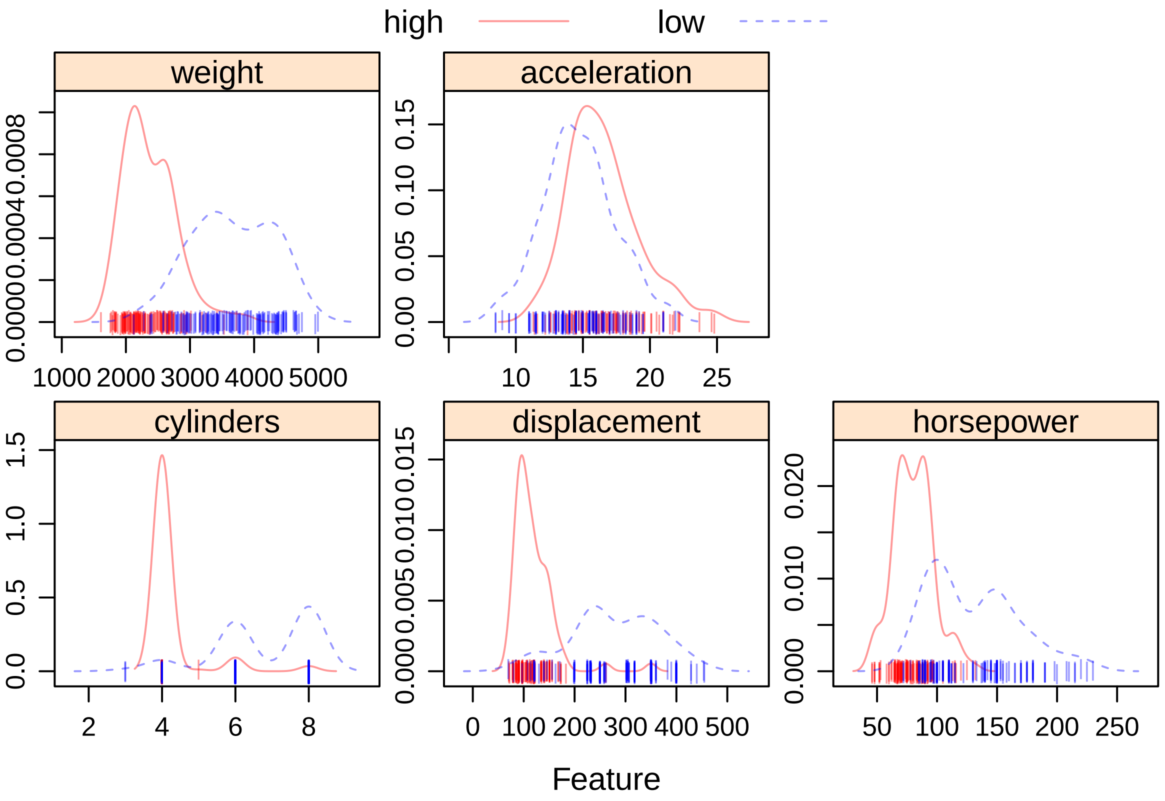
```
summary(train_data)
```

```
##    cylinders      displacement     horsepower        weight      acceleration
##  Min.   :3.000   Min.   : 70.0   Min.   : 46.0   Min.   :1613   Min.   : 8.50
##  1st Qu.:4.000   1st Qu.:107.0   1st Qu.: 78.0   1st Qu.:2279   1st Qu.:13.88
##  Median :4.000   Median :151.0   Median : 95.0   Median :2866   Median :15.50
##  Mean   :5.504   Mean   :198.1   Mean   :105.6   Mean   :3018   Mean   :15.61
##  3rd Qu.:8.000   3rd Qu.:272.0   3rd Qu.:130.0   3rd Qu.:3666   3rd Qu.:17.23
##  Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :4997   Max.   :24.80
##
##       year     origin  mpg_cat
##  73     : 29   1:175   high:138
##  75     : 27   2: 48   low :138
##  78     : 25   3: 53
##  79     : 24
##  70     : 22
##  81     : 21
##  (Other):128
```

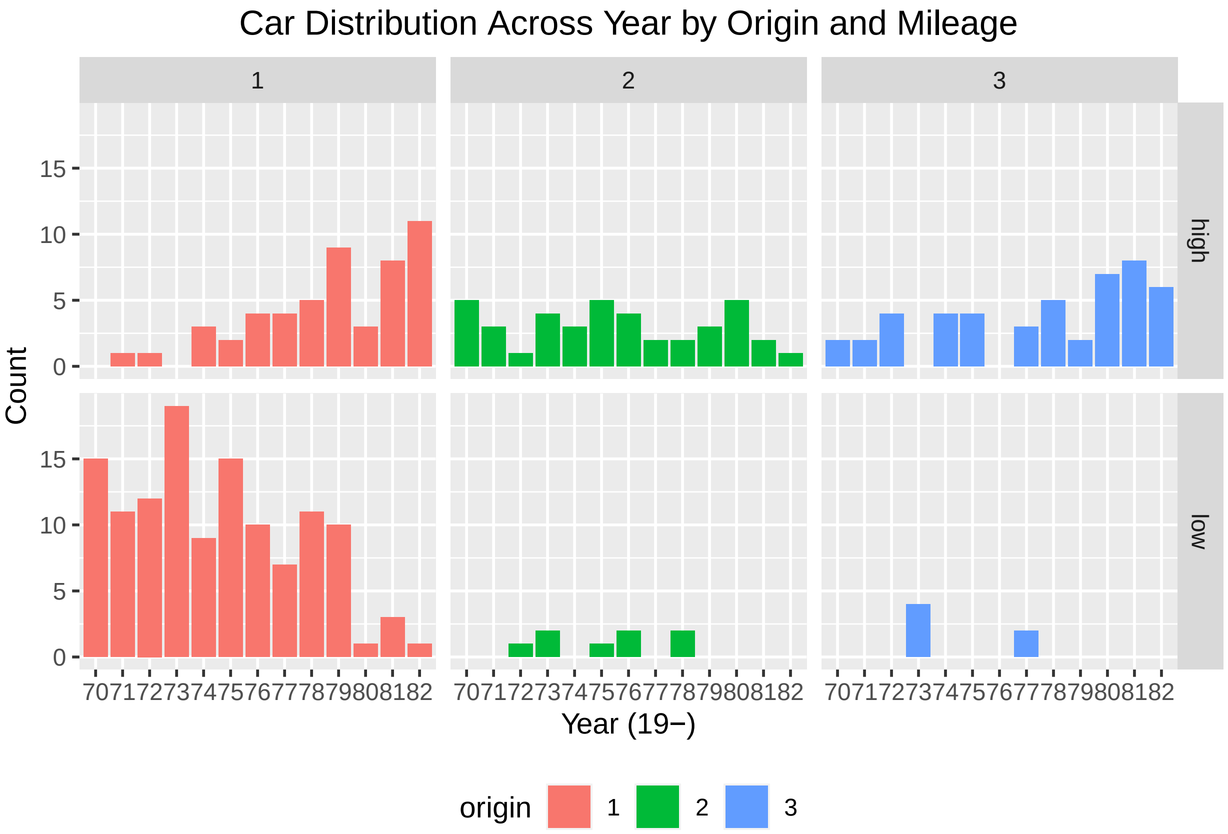
Visualize data distribution. In general, cars with high mileage have lower weights, cylinder count, engine displacement in inches, and horsepower. Note the unequal distribution of car count when conditioned on their origin and mileage.

```
trellis.par.set(transparentTheme(trans = .4))

featurePlot(train_pred[, 1:5], train_resp,
            scales = list(x = list(relation = "free"),
                          y = list(relation = "free")),
            plot = "density", pch = "|",
            auto.key = list(columns = 2))
```



```
train_data %>%
  count(year, origin, mpg_cat) %>%
  ggplot(aes(x = year, y = n, fill = origin)) +
  geom_col() +
  facet_grid(cols = vars(origin), rows = vars(mpg_cat)) +
  labs(
    title = "Car Distribution Across Year by Origin and Mileage",
    x = "Year (19-)",
    y = "Count"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "bottom"
  )
```


Car Distribution Across Year by Origin and Mileage
1
2
3
high
low
Count
15
10
5
0
15
10
5
0
707172737475767778798081 82
707172737475767778798081 82
707172737475767778798081 82
Year (19−)
origin
1
2
3