home work 1 cristian razo 7009434

1

Supervised and unsupervised learning are two different type of tasks/models used for machine learning.

Supervised learning deals with features/inputs that have a predictor/output. The labels for the predictors are known and the Y variable is the supervisor.

Unsupervised learning is when we are given a set of inputs but the outputs / labels are unknown.These type of models learn without a supervisor and the true labels of our data is unknown.

2

When we work with machine learning model we can categorize our Y variable into two categories. When we are dealing with numerical value and it is continous Y is quantitative .When Y is quantitative a regression model is the most appropriate to use.When we are dealing with categorical values then Y is qualitative. The appropriate model when Y is qualitative is a CLassification model.

3

The most used scoring/accuracy metric used for regression model is MSE and RMSE,MAE and Cross-validation.The most used scoring/accuracy metric used for classification model is calculating the error rate, bayes classified, and precision recall.

4

The descriptive model is able to describe/represent a system of features and its relationship to the predictor variable. Model that visually emphasize a tend or seasonality happening in our data.

Inferential models use measurements from the sample of the data to compare the different groups and make generalizations about the larger population of subjects in our data.States and finds relationship between outcome and predictors.

predictive model are used to regress/predict future values/behavior based on analyzing past datas trend and behavior. Aims to find a combo of features that best fit our data. Predict y with minimum reducible error.

5

Mechanistic Parametric assumes a form of linear function with coefficients weights holding down the features column variables . It wont match true unknown f .We are able to add more parameters to add and allow more flexibility to our model.

Empirically - driven are non parametric so has no assumption about being a linear function. They require a large # of observation and are much more flexible by default.

Both of these approaches can have the issue of overfitting our data meaning it will do a better job at explaining the variance than explaining the relationship between the features used.

A mechanistic model is the easier model to understand . This is because by default the Empirically-driven model is much more flexible and we prefer a model with the less flexibility and a model that has better interpretability.

The bias variance trade off allows our mechanistic or empirically-driven model to be able to find the best simple model that will give us our lowest test mse. Being able to have a low variance and add bias to our model allows us to have a more simple model and allows the model to not be overfitted.

6

Given a voter’s profile/data, how likely is it that they will vote in favor of the candidate. This question seems that it can be answered with a predictive model approach. This is because we are given a select number of features and based on the features or profile given we want to predict what choice they will make.

How would a voter’s likelihood of support for the candidate change if they had personal contact with the candidate? This type of questions seems that it will be best to use an inferential model. This model is the best to use because We want to understand how the features affect the outcome and the relationship they share with each other.

Exercise 1

```
library(tsdl)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 4.1.3
## -- Attaching packages -------------------------------------- tidymodels 0.2.0 --
```

```
## v broom        0.7.12     v rsample      0.1.1
## v dials        0.1.0      v tune         0.2.0
## v infer        1.0.0      v workflows    0.2.6
## v modeldata    0.1.1      v workflowsets 0.2.1
## v parsnip      0.2.1      v yardstick    0.0.9
## v recipes      0.2.0
```

```
## Warning: package 'broom' was built under R version 4.1.3
```

```
## Warning: package 'dials' was built under R version 4.1.3
```

```
## Warning: package 'infer' was built under R version 4.1.3
```

```
## Warning: package 'modeldata' was built under R version 4.1.3
```

```
## Warning: package 'parsnip' was built under R version 4.1.3
```

```
## Warning: package 'recipes' was built under R version 4.1.3
```

```
## Warning: package 'rsample' was built under R version 4.1.3
```

```
## Warning: package 'tune' was built under R version 4.1.3
```

```
## Warning: package 'workflows' was built under R version 4.1.3
```

```
## Warning: package 'workflowsets' was built under R version 4.1.3
```

```
## Warning: package 'yardstick' was built under R version 4.1.3
```

```
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```r
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.1.3
```

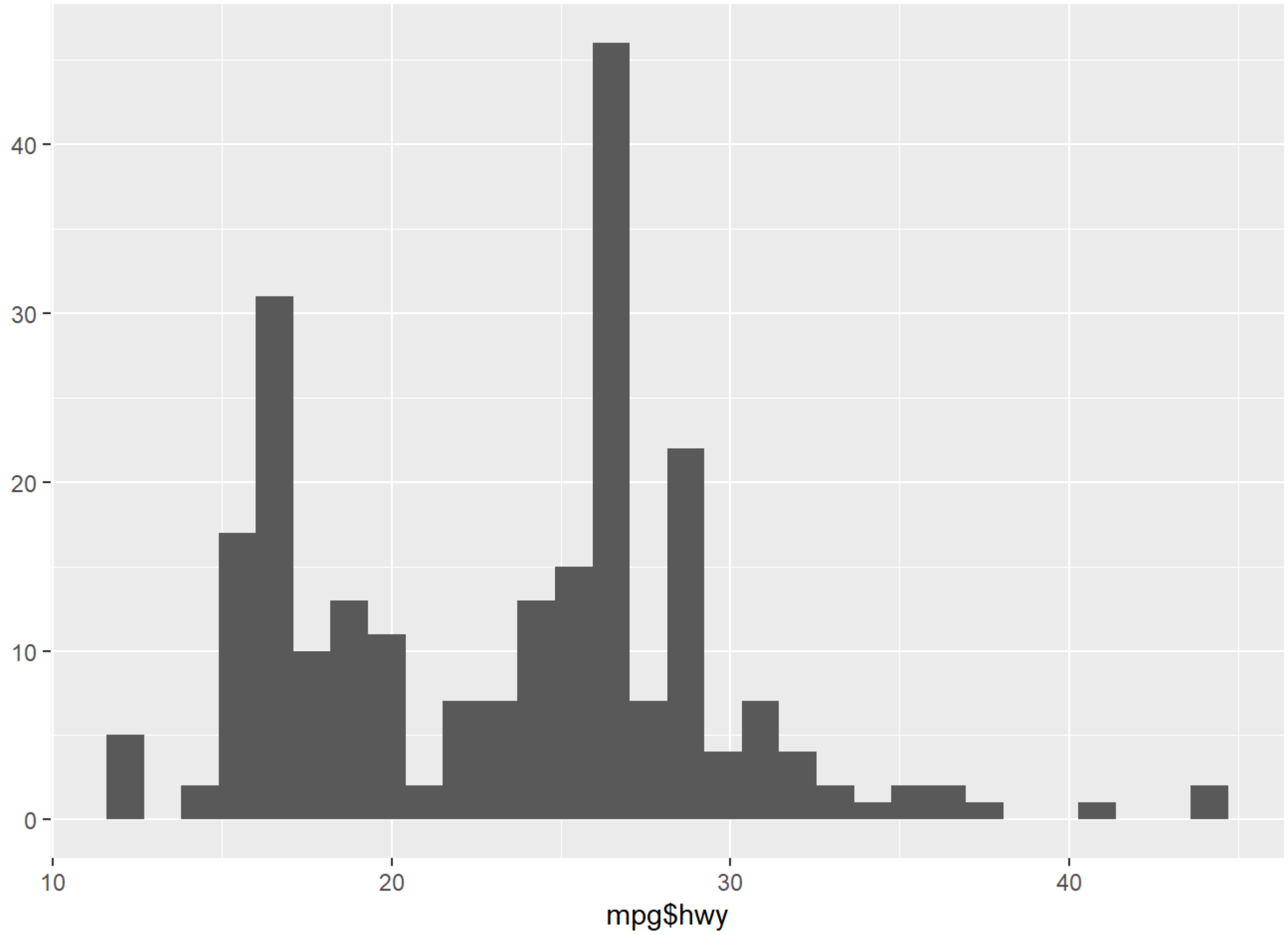
```
library(ggplot2)
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
library("dplyr")

qplot(mpg$hwy, geom="histogram")
```

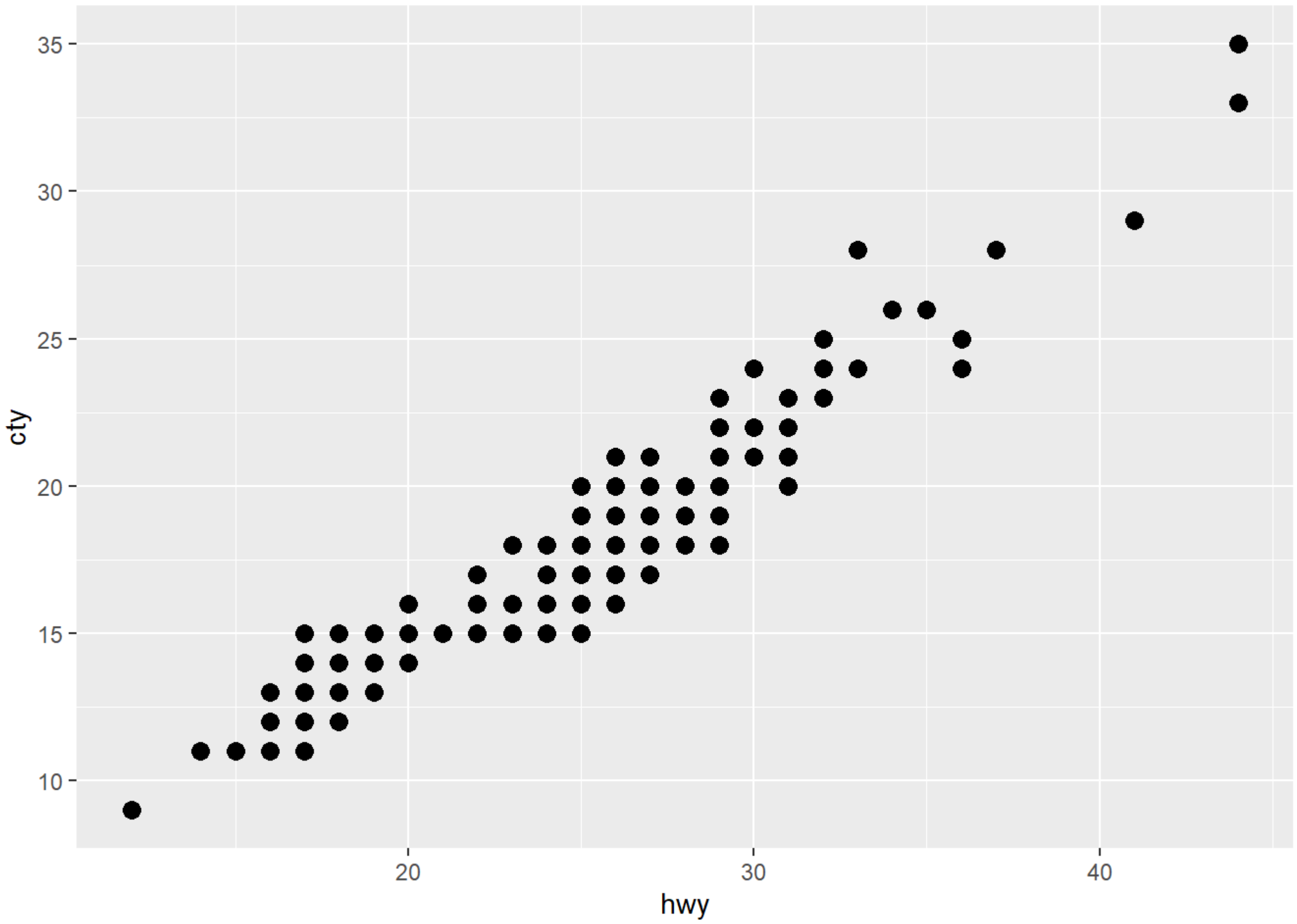
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



I see that this is left skewed type of data . The model doesnt follow a gaussian distribution and there is much variance between the vehicles MPG values.Most of our datas MPG lie in the interval of (20,30).

Exercise 2

```
ggplot(mpg,aes(x=hwy,y=cty)) + geom_point(size=3)
```



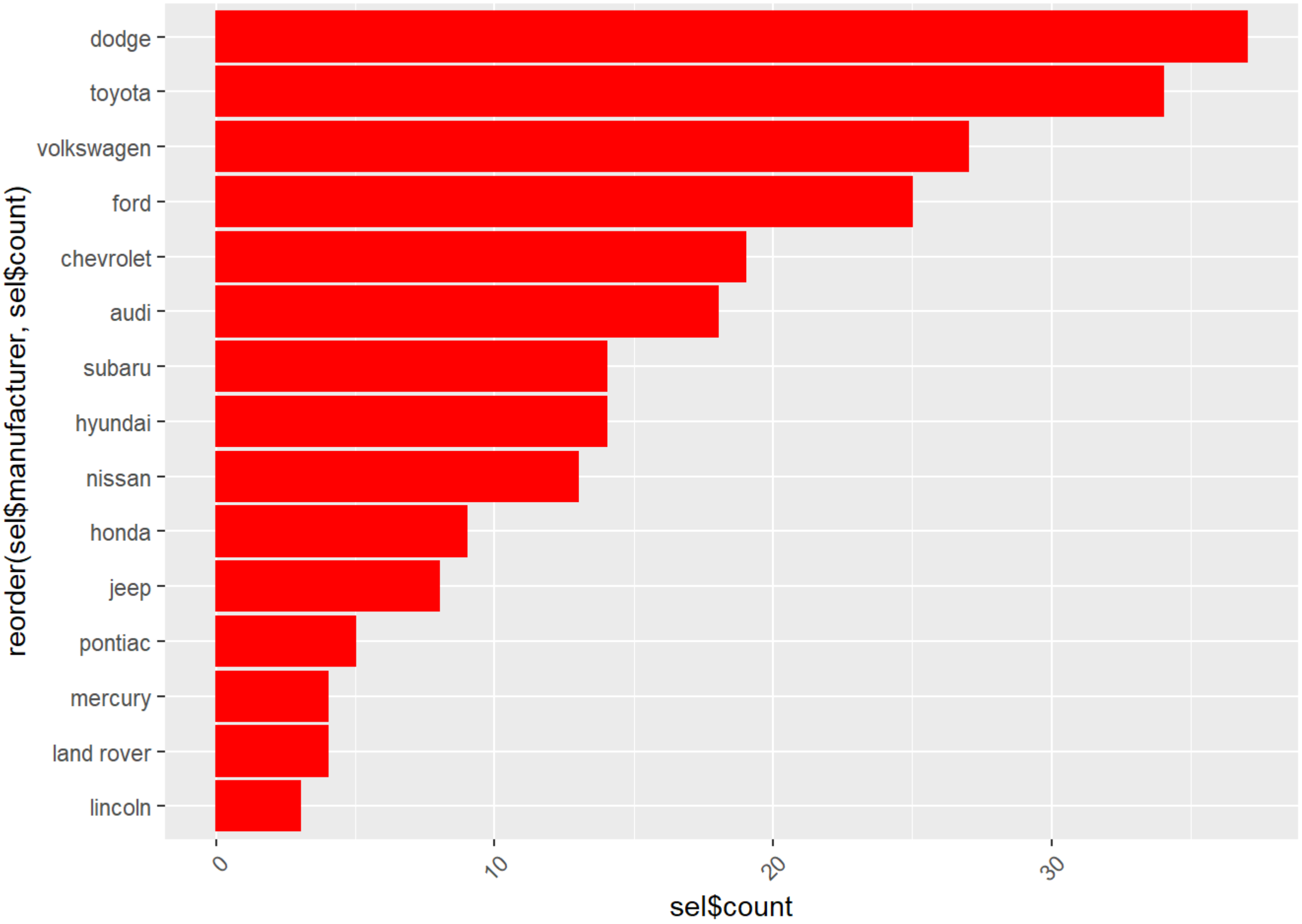
The is a postive correlation because if we fit a line it would show a positive linear relationship. This suggest that the MPG in the city rises with the MPG used in the high ways.

Exercise 3

```
sel=mpg %>%
  group_by(manufacturer) %>%
  summarise(count = n())  %>%
  arrange(count)

x <- ggplot(sel, aes(x =sel$count
,y = reorder(sel$manufacturer,sel$count)))

x <- x + geom_bar(stat="identity", color='red',fill='red')

x <- x + theme(axis.text.x=element_text(angle=45, hjust=0.9))
x
```

```
## Warning: Use of `sel$count` is discouraged. Use `count` instead.
## Warning: Use of `sel$manufacturer` is discouraged. Use `manufacturer` instead.
```

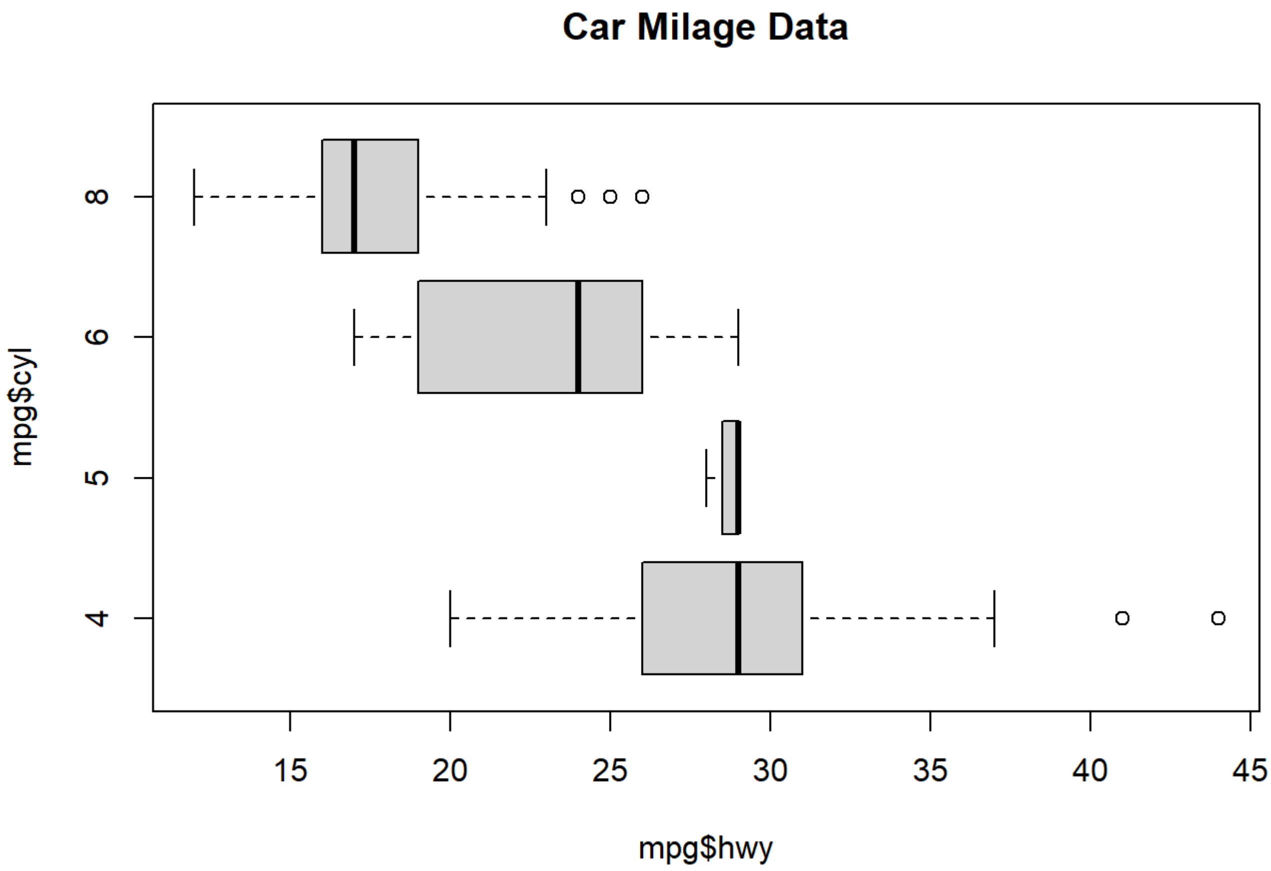
```
## Warning: Use of `sel$count` is discouraged. Use `count` instead.
```



We can see that dodge produced the most cars while lincoln produced the least amount of vehicles represented in our data set.

Excercise 4

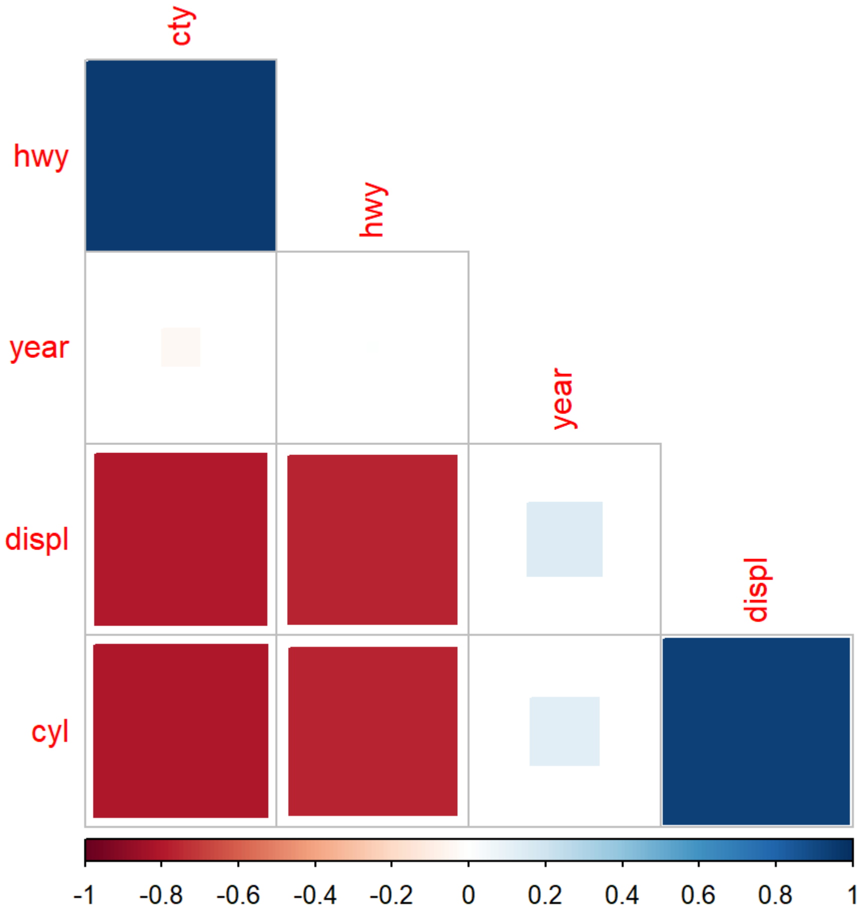
```
boxplot(mpg$hwy~mpg$cyl,data=mpg, main="Car Milage Data",horizontal = T)
```



We can see that cars with high cylinders have less MPG while low cylinders have higher MPG. Cylinders and MPG have a negative pearson R correlation relationship .

Excercise 5

```
mp<- select_if(mpg, is.numeric)

corrplot(cor(mp), method = 'square', order = 'FPC', type = 'lower', diag = FALSE)
```



We can see a strong positive relationship with the relations of cty ~ hwy and cyl ~ displ column variables. We that there is a negative relationship between cyl ~ hwy,cyl ~ cty , and also these columns displ ~ hwy and displ ~ cty based on the heat map correlation matrix we made. Yes these relationship makes sense because the higher the cylider is the more gas the car will use and will take in less miles per gallon.