131 HW1

2022-10-03

Question 1(From IBM Blog): Supervised learning is a machine learning approach defined by using labeled datasets. Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. Supervised machine learning relies on labeled input and output training data, whereas unsupervised learning processes use unlabelled or raw data.

Question 2 (From Lecture): Regression Model: The response Y is quantitative The result is numerical values (e.g., price, blood pressure) Classification Model: The response Y is qualitative The result is categorical values (e.g., survived/died, spam/not spam)

Question 3: Regression: Mean Squared Error (MSE). Root Mean Squared Error (RMSE). Classification: Accuracy. Confusion Matrix.

Question 4: A descriptive model describes a system or other entity and its relationship to its environment. It is generally used to help specify and/or understand what the system is, what it does, and how it does it. Inferential models are usually created not only for their predictions, but also to make inferences or judgments about some component of the model, such as a coefficient value or other parameter. Predictive models analyze past performance to assess how likely a customer is to exhibit a specific behavior in the future.

Question 5(From Lecture): Mechanistic： Assume a parametric form for f Won't match true unknown f Can add parameters = more flexibility

Empirically–driven: No assumptions about f Require a larger # of observations Much more flexible by default However, mechanistic and empirically–driven are all overfitting.

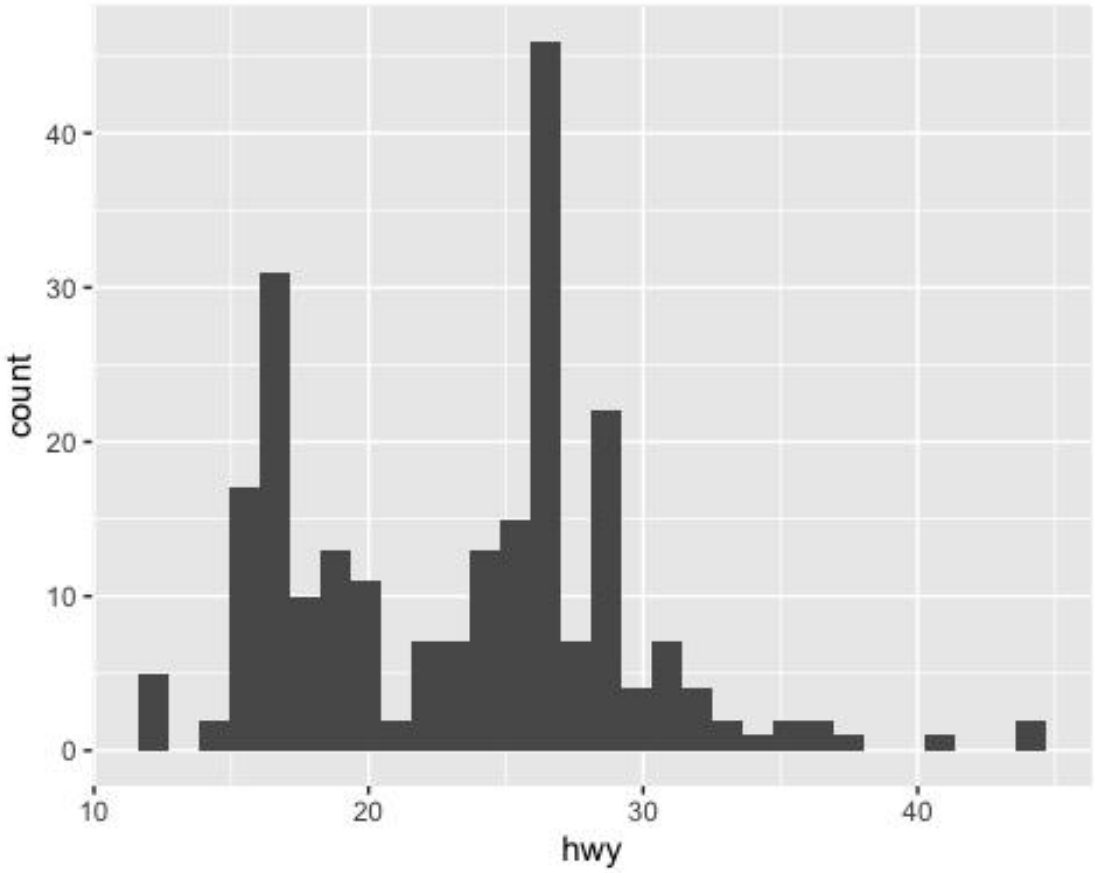
Mechanistic are easier to understand, because mechanistic models can provide insights and understanding into the mechanistic functions of treatments, and these are necessary to overcome the limitations of machine learning predictions.

There is a tradeoff between a model's ability to minimize bias and variance which is referred to as the best solution for selecting a value of Regularization constant. Proper understanding of these errors would help to avoid the overfitting and underfitting of a data set while training the algorithm.

Question 6: Predictive, because this question is about the probability of something happening. Inferential, because the question asks about the relationship between two factors.

Exercise 1

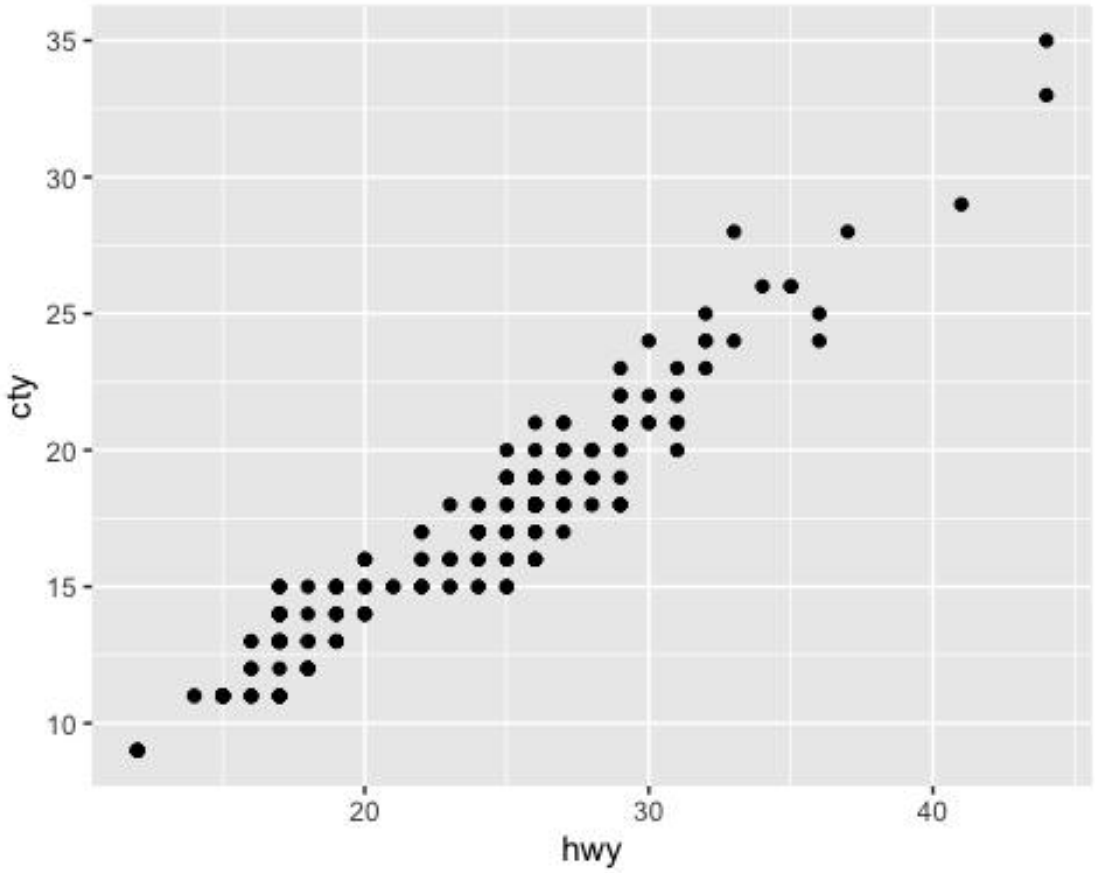
```
library(tidyverse)
ggplot(mpg, aes(x=hwy)) + geom_histogram(bins=30)
```



The distribution is multi-model and skewed to the right.

Exercise 2

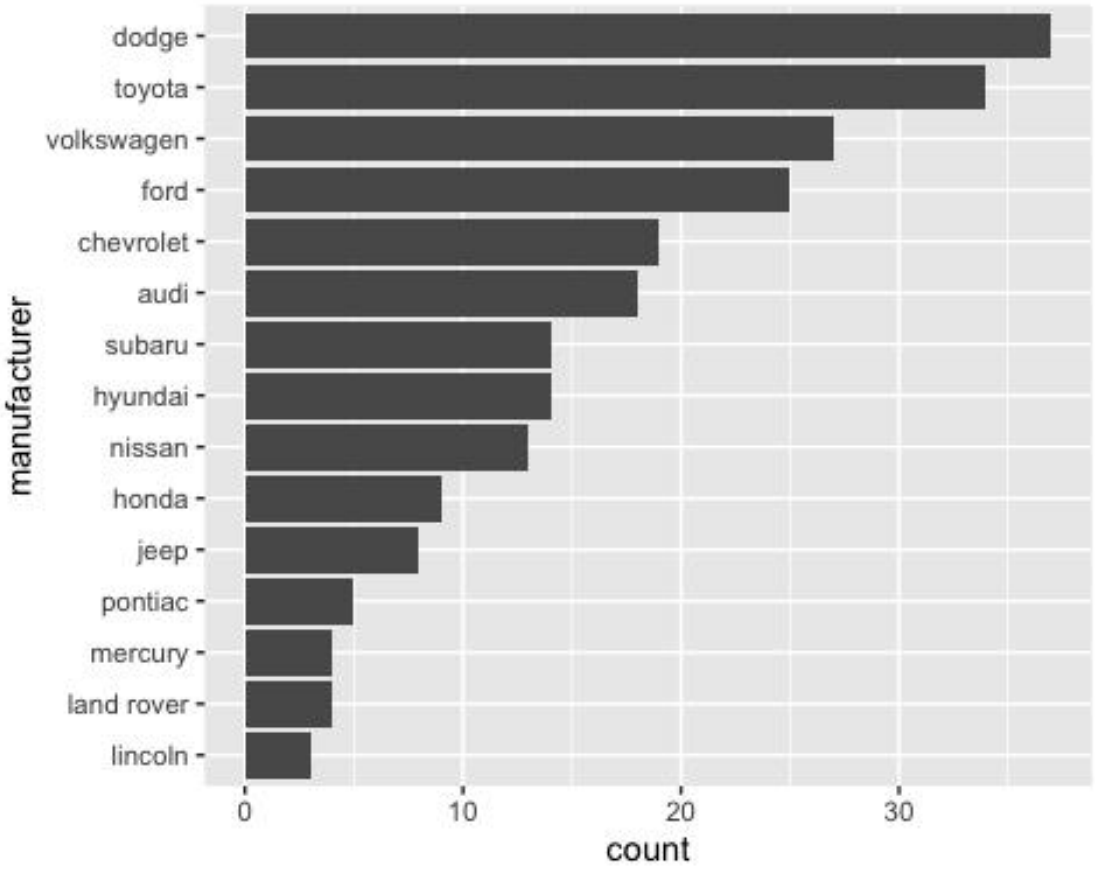
```
ggplot(mpg, aes(x=hwy, y=cty))+  geom_point()
```



There is a strongly positive relationship between cty and hwy, it means that when the hwy increases, the cty will also increase.

Exercise 3

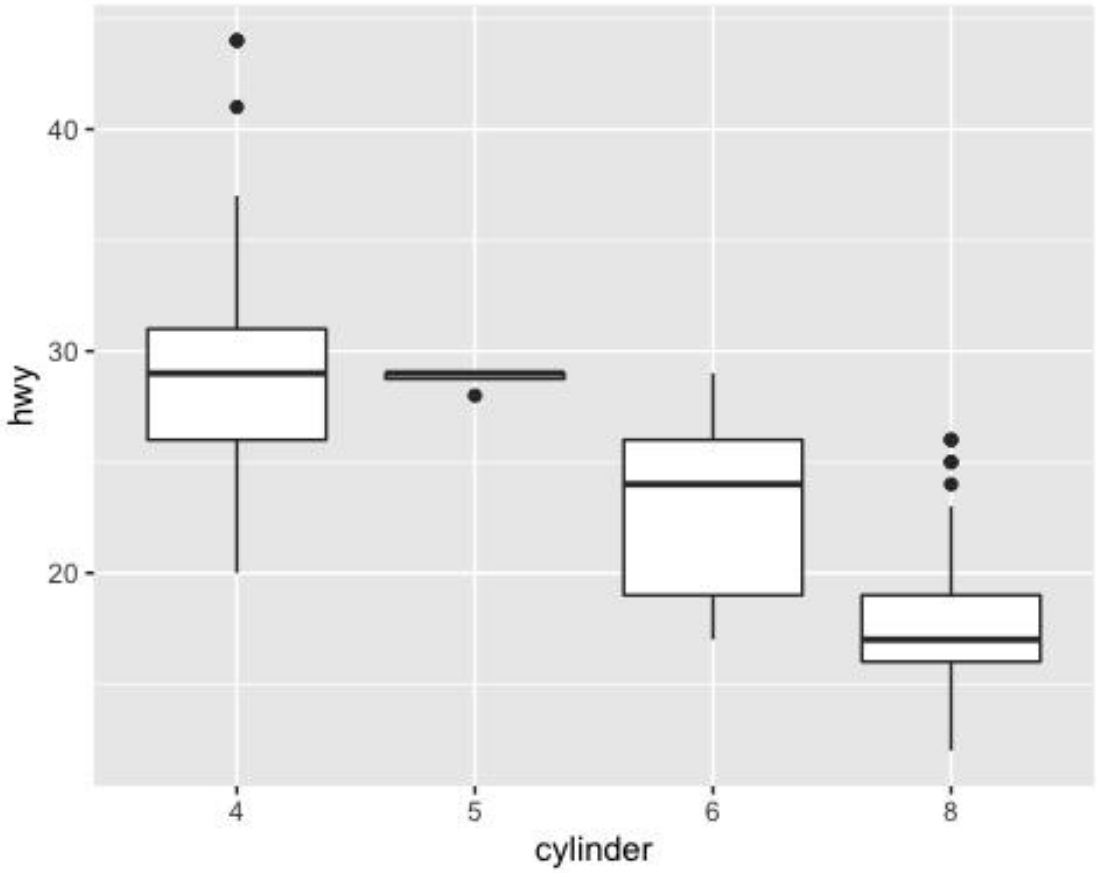
```
mpg %>% group_by(manufacturer)  %>% summarise(count=n()) %>%
  ggplot(aes(x=reorder(manufacturer, count), y=count)) + geom_col() +
  coord_flip() +
  labs(x="manufacturer")
```



Dodge produced the most cars, lincoln produced the least cars.

Exercise 4

```
ggplot(data=mpg, aes(x=factor(cyl), y=hwy)) + geom_boxplot() +
  labs(x="cylinder")
```



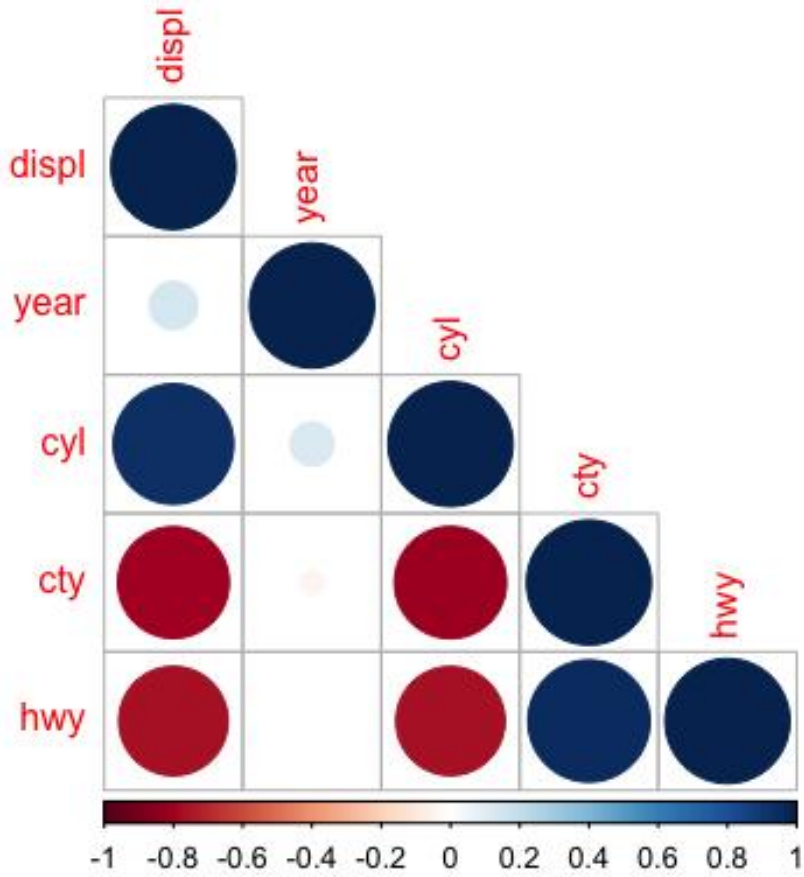
We can see a pattern in the plot, as the number of cylinders increase, the hwy decreases.

Exercise 5

```
library(corrplot)

## corrplot 0.92 loaded

corrplot(cor(mpg[, sapply(mpg, is.numeric)]), type="lower")
```



The displ is positively associated withe year and cyl, but negatively associated with cty and hwy. The year is negatively associated with cty but positively associated with hwy and cyl. The cyl is negatively associated with cy and hwy, the cty is positively associated with hwy, it makes sense to me, they are not surprise to me.