HW1_pstat131

Jiacong Wu

2022-09-29

Question 1

For supervised learning, there is a response variable, and we can know the accuracy of the model because we can get the data that how well the model fits the testing data.

However, for unsupervised learning, there is no response variable, the result would be a cluster or a pattern of the data.

The difference would be the supervised model has a predictor, but the unsupervised model does not.

Question 2

In the context of machine learning, the regression model will have quantitative response variable, but a classification model will have qualitative (categorical) response variable.

Question 3

commonly used metrics for regression ML problems: price(lecture slides), blood pressure(lecture slides), salary

commonly used metrics for classification ML problems: survived/died(lecture slides), spam/not spam(lecture slides), valid/invalid

Question 4

Descriptive models: Choose model to best visually emphasize a trend in data i.e. using a line or a scatterplot(from lecture slides)

Inferential models: Aim to test theories, causal claims, stating relationship between outcomes and preditors (from lecture slides)

Predictive models: Aim is to predict Y with minimum reducible error (from lecture slides)

Question 5

Mechanistic models are parametric models. Mechanistic assume a parametric form for f, and we can add parameters to get more flexibility (from lecture slides)

Empirically-driven models are non-parametric models which there is no assumption about the f, and it requires a large number of observations. (From lecture slides)

Difference: mechanistic requires a parametric form for f, while empirically-driven does not

Similarity: There is a f for both methods,$Y = f(X_1, ..., X_p) + \varepsilon$.

In my opinion, for people don't know anything about models, the empirically-driven is easier to understand because intuitively it makes sense that when we have a large number of observations, the data will have certain pattern which is the model we want.

For mechanistic, when we add more parameters, the model will have lower bias, and fitting the data better. However, we don't want there to be too many parameters, as that will overfit. In that case, the model will indeed fit the data pretty well, but it will behave too bad for new data that we need to predict.

For empirically-driven, a large number of observations is needed, and the more the data, the better the model fits the data, the lower the bias. However, if there is too many observations, overfitting will occur.

Question 6

(1) Predictive, as it wants to predict the likelihood of something

(2) inferential, as it wants the inference of how the data will behave if certain conditions are met

Exercise 1

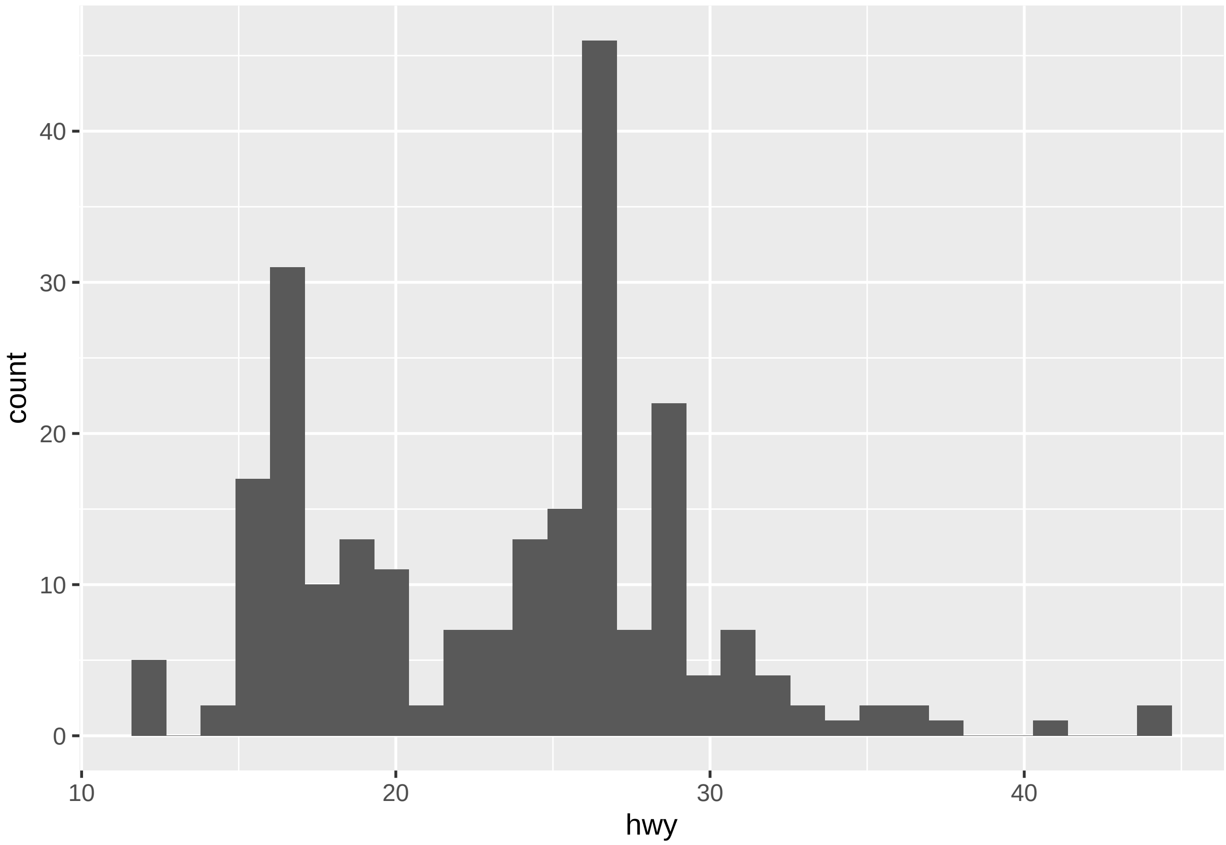
```
#install.packages("ggplot2")
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#mpg
```

```
ggplot(mpg,aes(hwy))+geom_histogram()
```

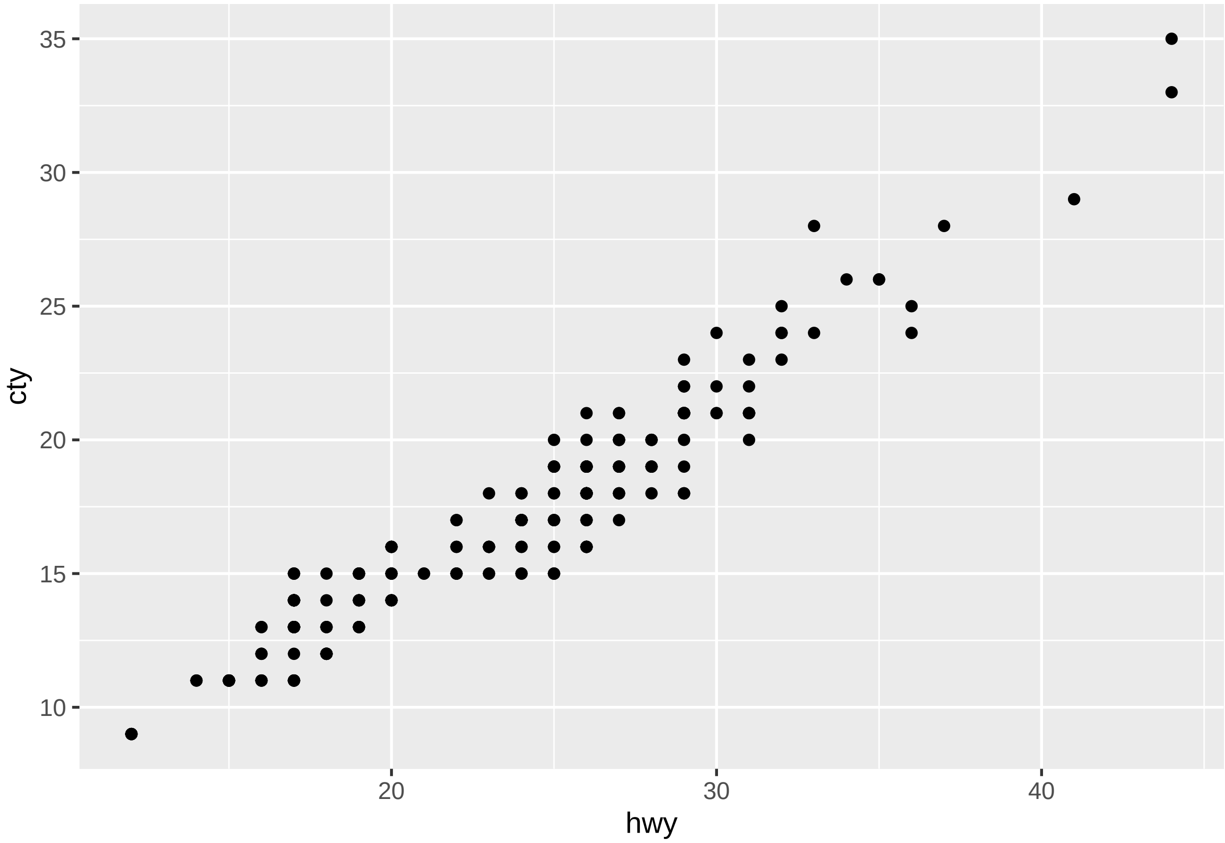
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It is more likely that the hwy mpg to be high in the center, and low for two tails. The distribution is left skewed.

Exercise 2

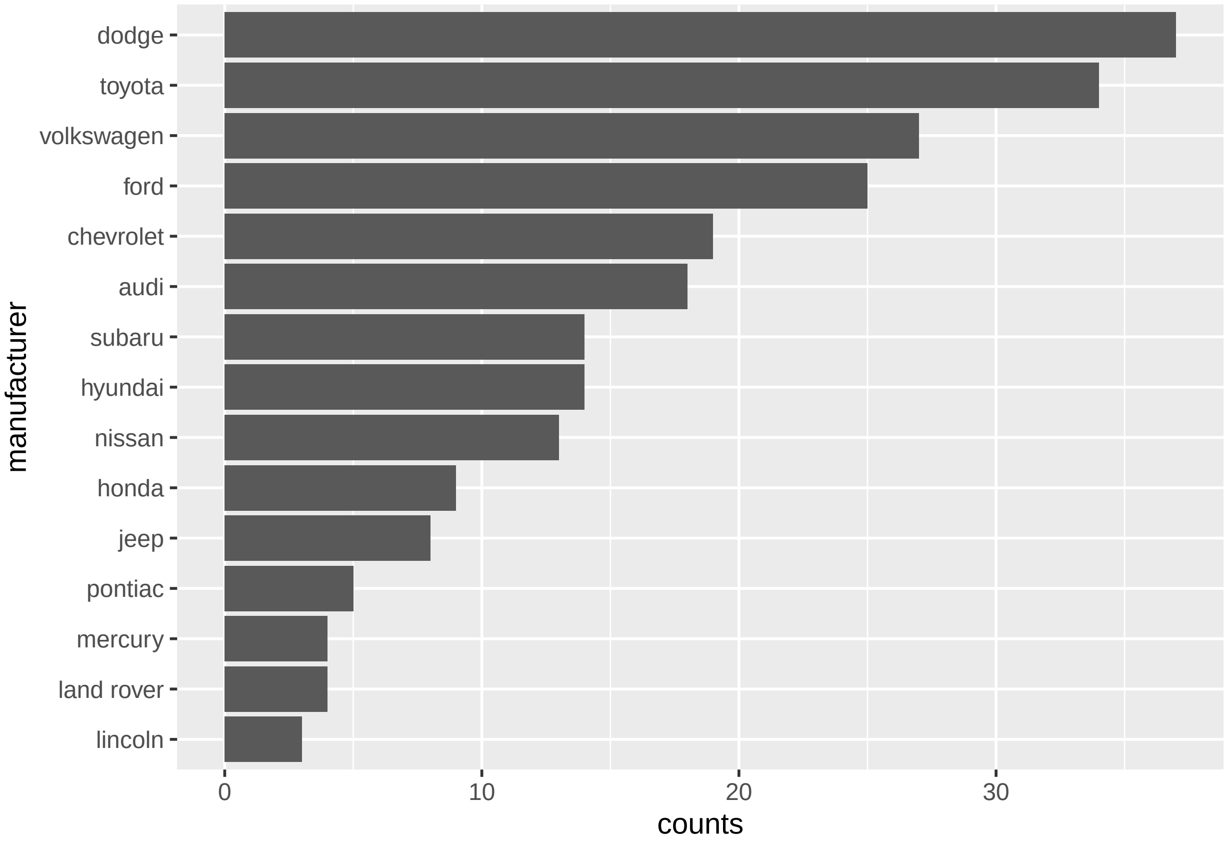
```
ggplot(mpg, aes(x = hwy,y = cty)) + geom_point()
```



It looks like a linear relationship between the two, this means that when a car have a high highway mpg, than it is likely to have a high city mpg.

Exercise 3

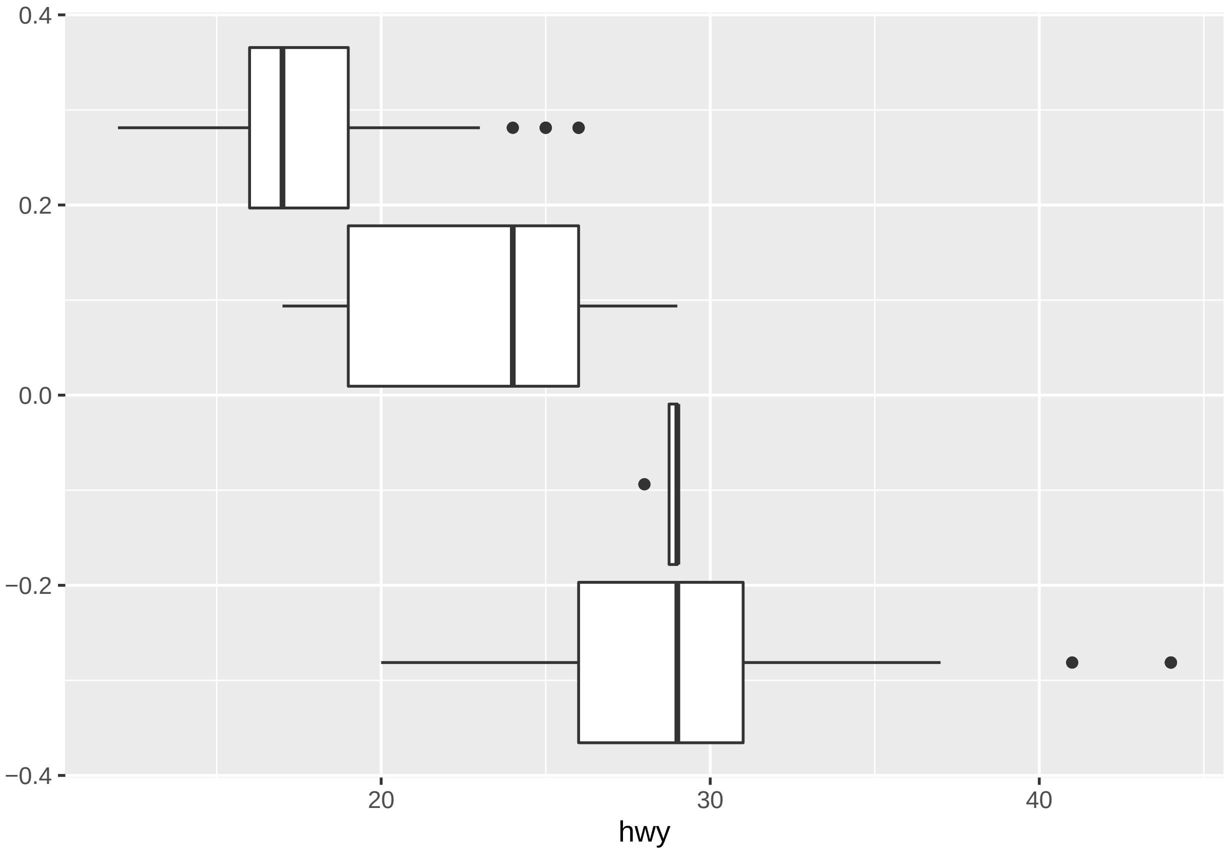
```
m = mpg %>%
  group_by(manufacturer) %>%
  summarise(counts = n()) %>%
  arrange(counts) %>%
  mutate(manufacturer = factor(manufacturer,manufacturer))%>%
  ggplot(aes(x = counts, y = manufacturer)) + geom_bar(stat = "identity")
m
```



From the data, I see that the Dodge produce the most cars, and Lincoln produce the least cars

Exercise 4

```
mpg %>%
  mutate(cyl = factor(cyl))%>%
  ggplot(aes(group = cyl,x = hwy)) + geom_boxplot()
```



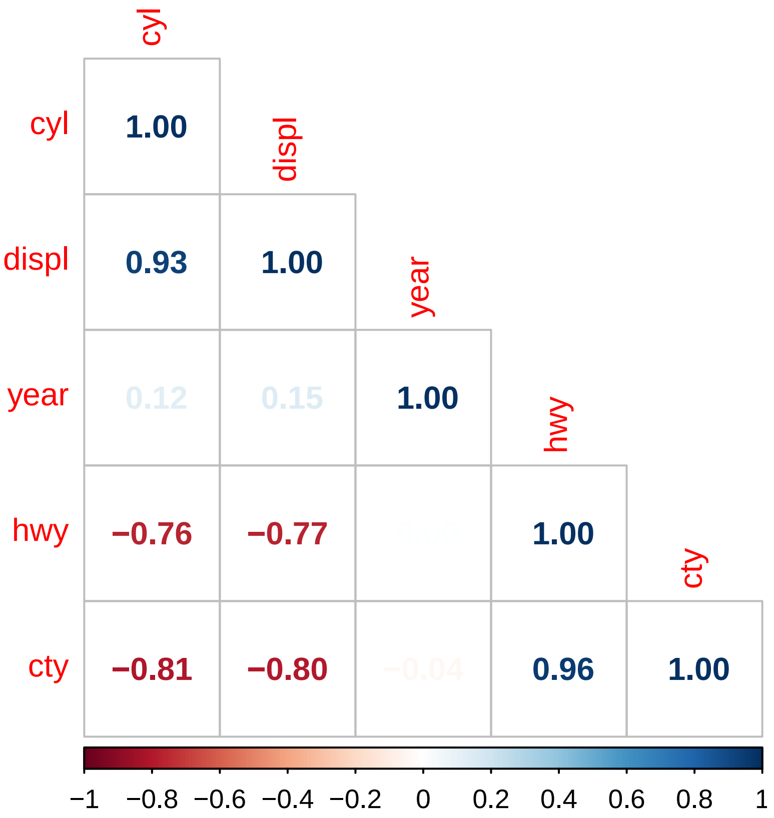
Yes, I can see a pattern. The more cylinders a car has, the lower the highway mpg would be.

Question 5

```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpg %>%
  select_if(is.numeric) %>%
  cor()%>%
  corrplot(method = 'number', order = 'FPC', type = 'lower')
```



From the graph, the displ(engine displacement) and cyl(number of cylinders) are positively correlated. Also, highway and city mpg are correlated. Displ and cyl are negatively correlated to city and highway mpgs.

Nothing surprising here for me. It makes sense that hwy and city are both mpgs, so they are highly correlated. The more cylinders a car have, the less miles a car can run per gallon, so the negative correlation is also intuitive.