PSTAT131-HW1

2022-09-28

Q1.

Supervised learning is a machine learning approach in which we teach the machine through the use of labelled data sets. Unsupervised learning is when the machine is taught without the use of labelled data sets and guidance. The difference is that in supervised learning we are able to see the answer key, whereas in unsupervised learning we never see it.

Q2.

The regression model relates to quantitative or numerical data, whereas the classification model relates to qualitative or categorical data.

Q3.

In regression problems, two commonly used metrics are $\hat{f}(x_0) = \mathrm{E}[Y \mid X = x_0]$ and $\mathrm{Var}(\epsilon)$

In classification problems, two commonly used metrics are $\hat{y}_0 = \mathrm{argmax}_j\{\mathrm{Prob}(Y = j \mid X = x_0\}$ and $1 - \mathrm{E}[\max_j \mathrm{Prob}(Y = j \mid X)]$.

Q4.

Descriptive models: choosing a model that emphasizes a trend in the data.

Inferential models: seeing which data is significant and whether there is a relationship between the predictors.

Predictive models: predicting a response while trying to get the smallest error and focuses on getting a proper set of predictors.

Q5.

mechanistic: making predictions based on your understanding of the real world.

empirically-driven: making predictions by experimenting and observing the outcome when given certain circumstances.

They differ by mechanistic and empirically-driven base their assumptions are different material where mechanistic focuses on their belief of what would happen based on what they know, while empirically-driven focuses on their prediction based on what they know happened before given similar circumstances. They are the same because they are both based on previous understandings of the circumstance.

I think empirically-driven is easier to understand because it is solely based on prior experiments and is much more conclusive where mechanistic could be different depending on the person you're talking to.

Depending on question and which method is being done, it is possible that there would be a higher bias especially when given a mechanistic approach because it is of your own belief of the world which would in turn mean that there is some bias in play.

Q6.

Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?

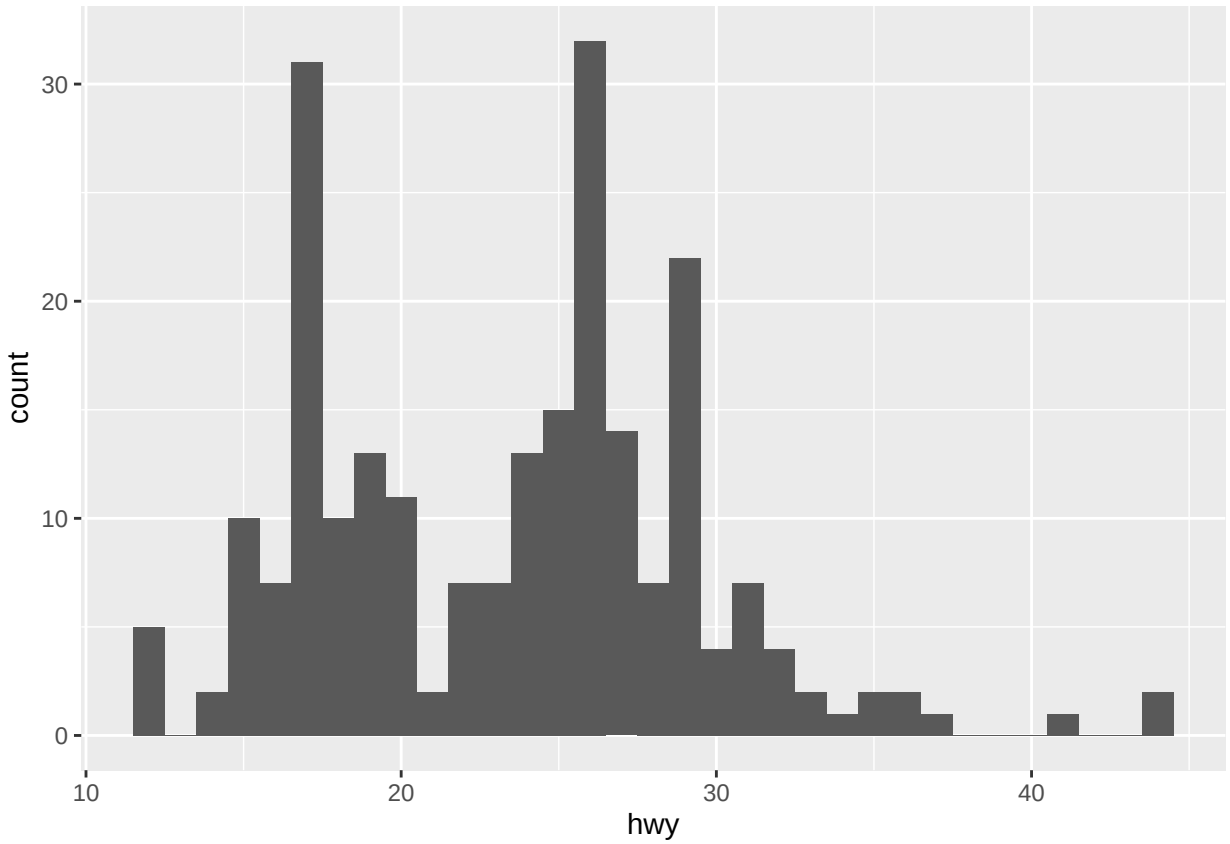
Predictive because the end goal is to predict the Y given certain predictors. In this case, they are trying to predict if the voter will vote for the candidate given their profile/data.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

Inferential because the end goal is to figure out whether the feature is significant to the problem. In this case, it would be whether personal contact is significant to voter support.

E1.

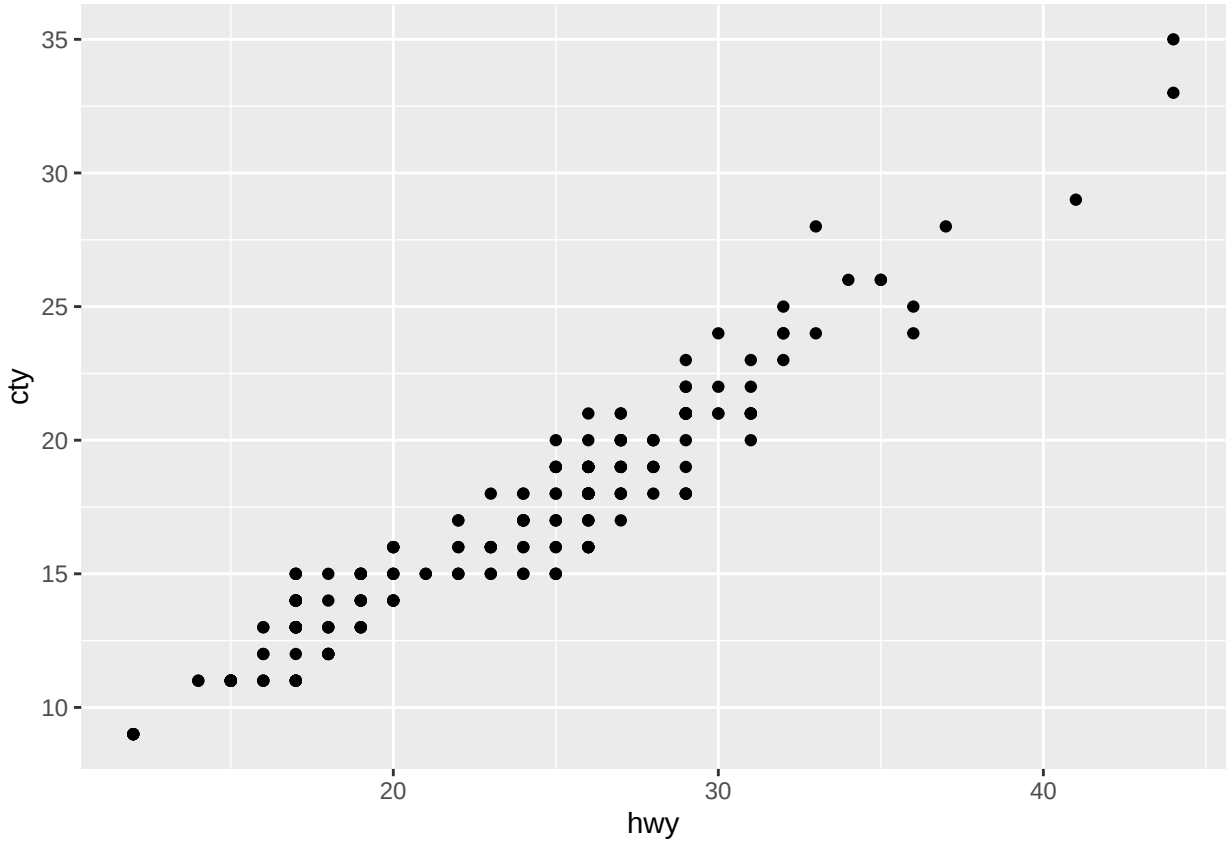
```
ggplot(data=mpg, aes(x=hwy)) + geom_histogram(binwidth = 1)
```



I can see that the majority of the data has counts of around 15 or less, while 3 highways appear to be more frequent than the rest.

E2.

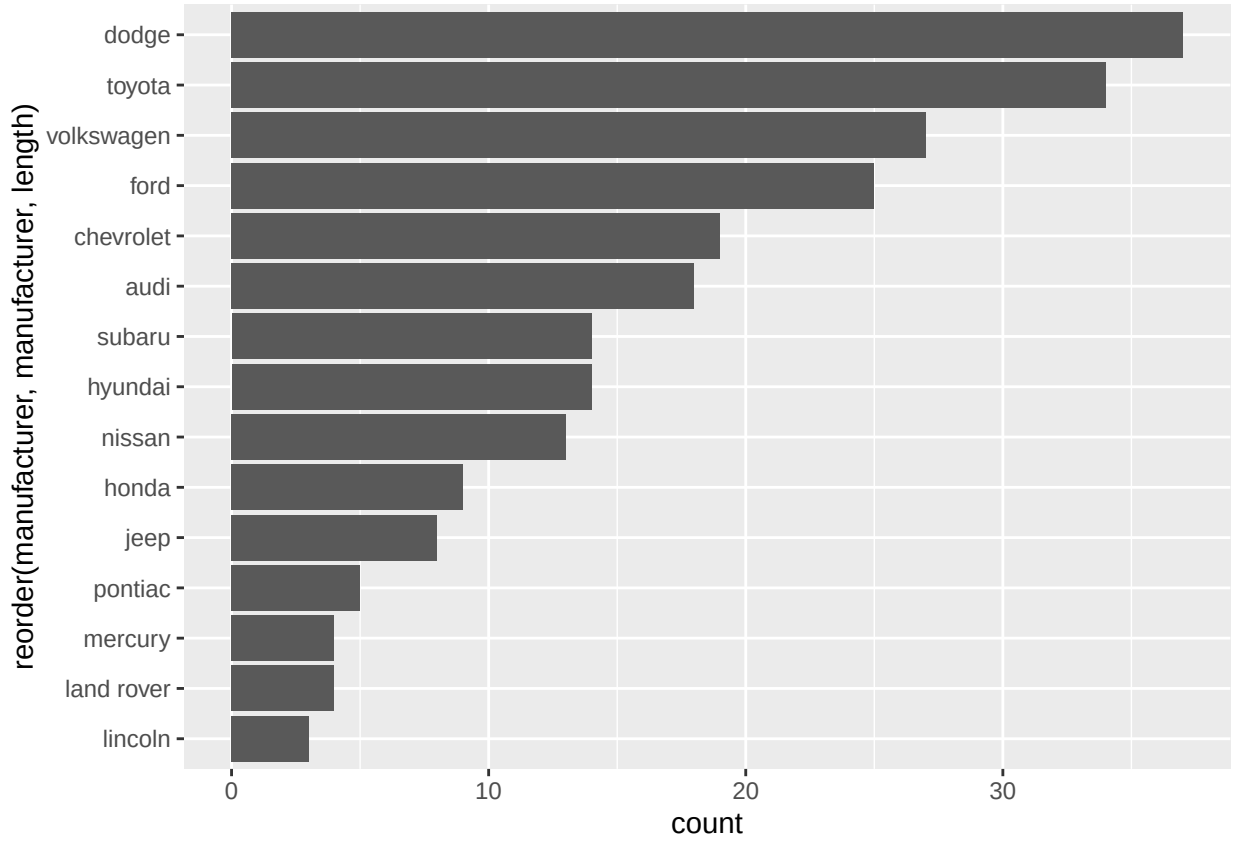
```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point()
```



I notice that the points form almost a straight line and if there were a line through the center of the points, the majority of them would fall close to that line. There appears to be a positive linear relationship between highway and city. The mpg in the city and the mpg on highways are correlated in some way.

E3.

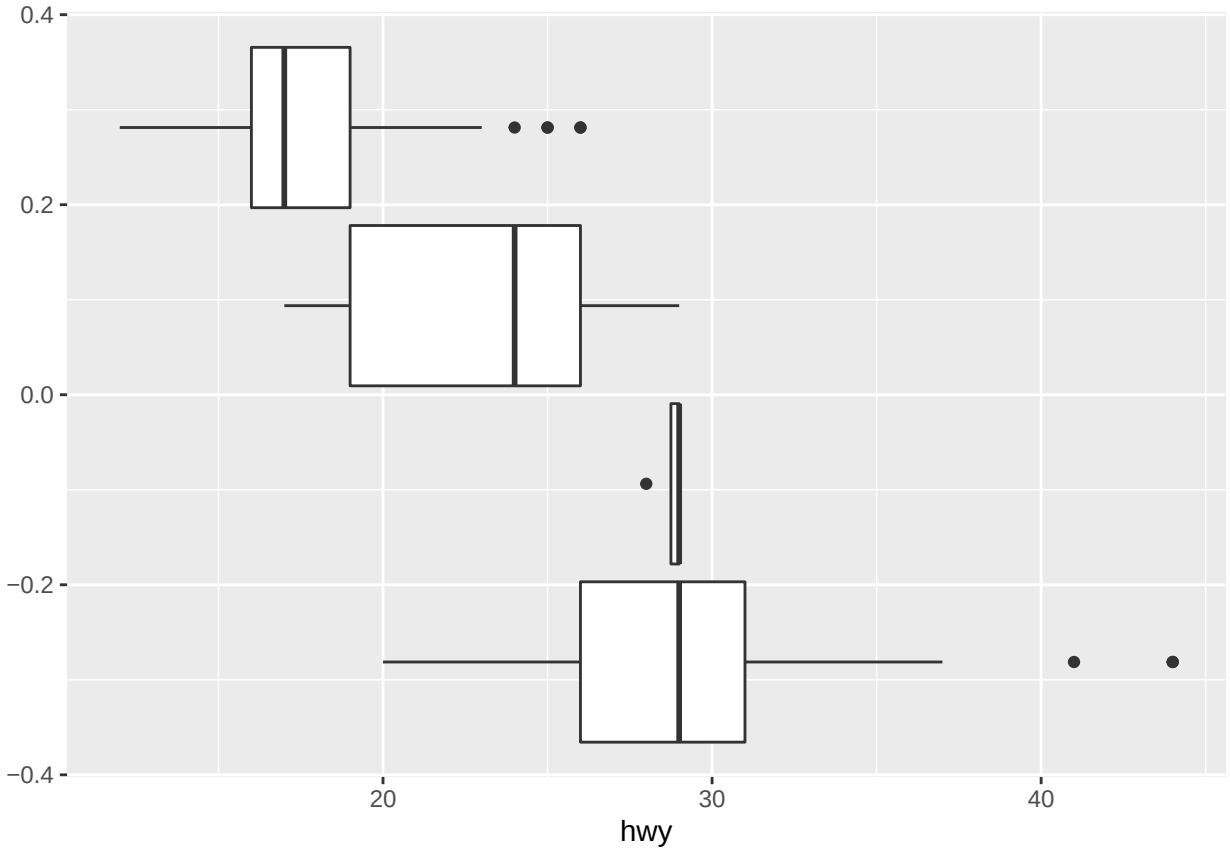
```
ggplot(data=mpg, aes(x=reorder(manufacturer, manufacturer, length))) + geom_bar() + coord_flip()
```



Dodge produced the most cars and Lincoln produced the least cars.

E4.

```
ggplot(data=mpg, aes(x=hwy , group=cyl)) + geom_boxplot()
```

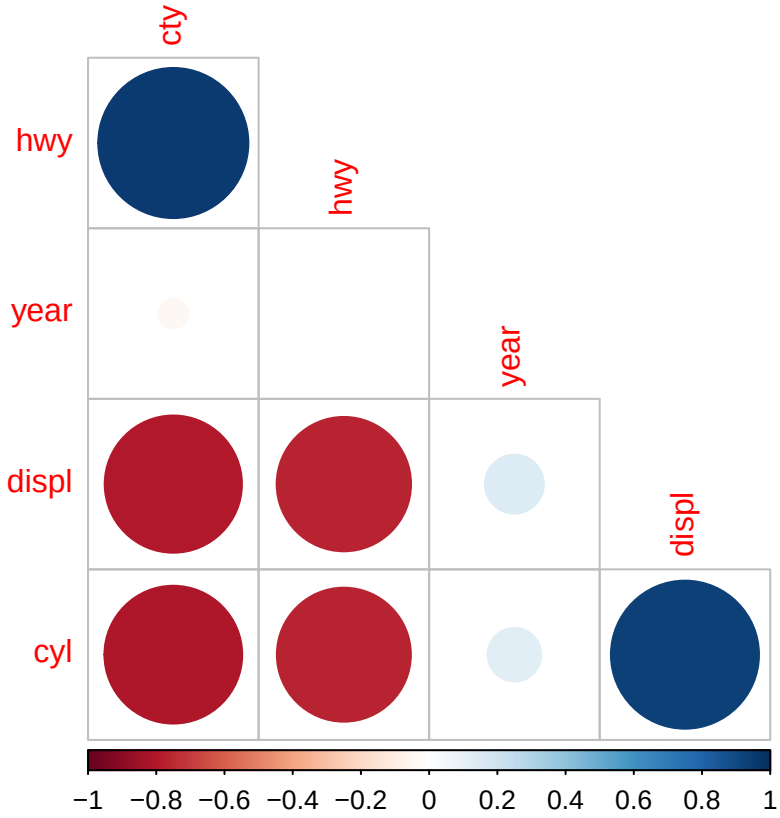


There does not seem to be any clear patterns except that the means for the 3rd and the 4th box are around the same value.

E5.

```
mpg_new <- mpg[, unlist(lapply(mpg, is.numeric))]
mpg_corr <- cor(mpg_new)

corrplot(mpg_corr, order = 'FPC', type = 'lower', diag = FALSE)
```



(cty + hwy), and (displ + cyl) are postively correlated. (year + displ), and (year + cyl) are very slightly positively correlated, as well. (displ + cty), (displ + hwy), and (cyl + cty), (cyl + hwy) are negatively correlated. (year + cty) are very slightly negatively correlated, as well. For the most part, I think the relationships do make sense. However, it is a little odd how the cty and displ is negatively correlated with cty and hwy.