131-hw1

Isha Gokhale

2022-09-29

1. In unsupervised learning the outcome is not known, whereas in supervised learning the response/outcome is known.Additionally, supervised learning uses labelled input and output data. Typically, unsupervised learning uses clustering algorithms in order to analyze and detect unlabelled clusters.

2. Regression has quantitative outcomes, while classification has qualitative outcomes, such as categorical values.

3. Two commonly used metrics for regression ML problems include price and blood pressure. Two commonly used metrics for classification ML problems include categories such as: survived/died, spam/not spam (from lecture: course overview and introduction)

4. Descriptive models are usually used to best visually emphasize a trend in data. Predictive models are used to determine which combo of features fits best, and these models aim to predict Y with minimum reducible error. Lastly, they are not focused on hypothesis tests. Inferential models are used to determine which features are significant and they aim to test theories, causal claims, and state relationship between outcomes and predictor(s) (from lecture: course overview and introduction).

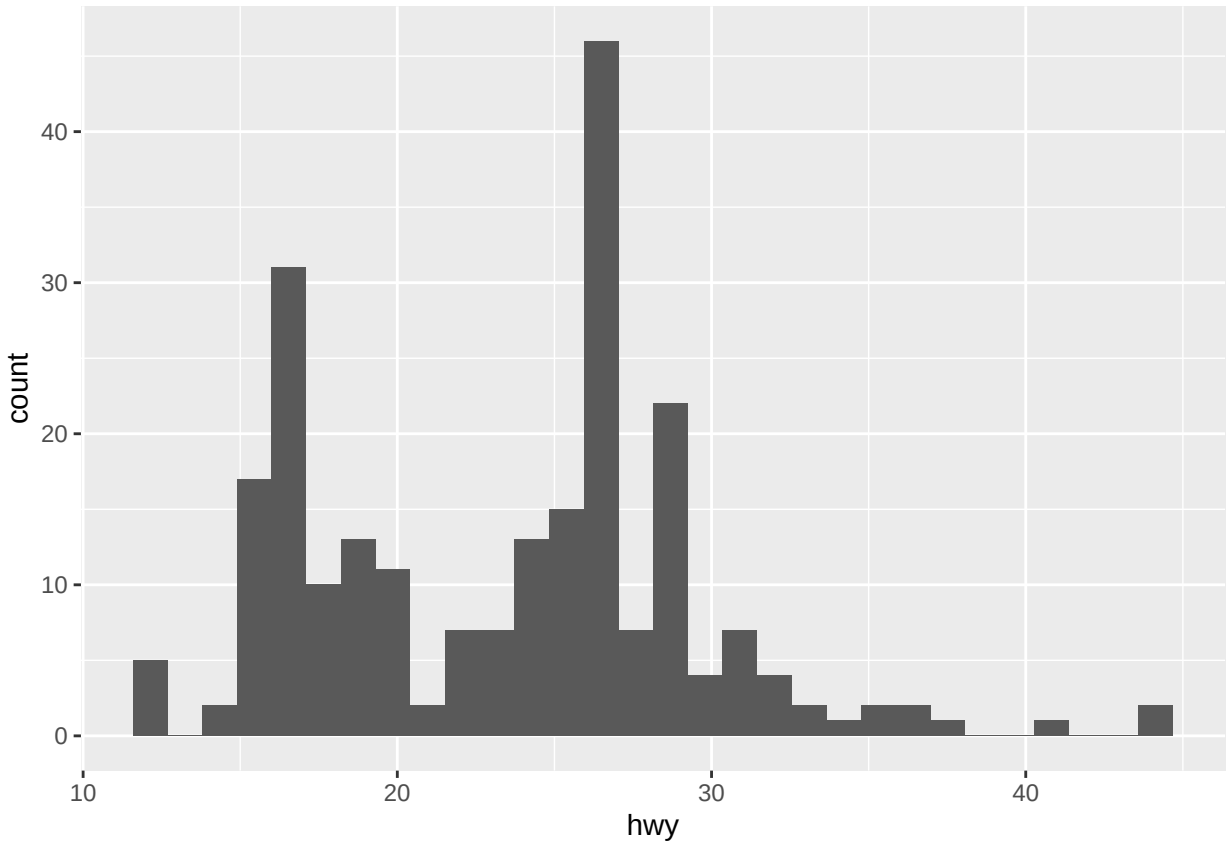
5. Mechanistic models use a theory in order to determine what the real world result will be. Empirical models, however, use real world events in order to create a theory. They differ in the sense that mechanistic models assume a parametric form for f, while empirically driven models do not make any assumptions about f. They are similar in the sense that they both fall risk to over fitting when too many parameters are used in the mechanistic model. In general, empirical models are easier to understand as we are given the facts and use them to develop a theory. Additionally, mechanistic information takes up a lot of computation or not always available, which is also why empirical models are easier to understand. Since empirical models are relatively simpler, they have high bias and low variance while a flexible model (such as mechanistic models) have low bias and high variance.

6. The first question is predictive as we are trying to predict Y: we are trying to forecast a person's vote. The second question, however, is inferential as we are aiming to test a theory involving how a voter's support changes based on personal contact with a candidate.

(1)

```
library(ggplot2)
#highway = mpg$hwy
#hist(highway)
ggplot(mpg, aes(x = hwy)) + geom_histogram()
```

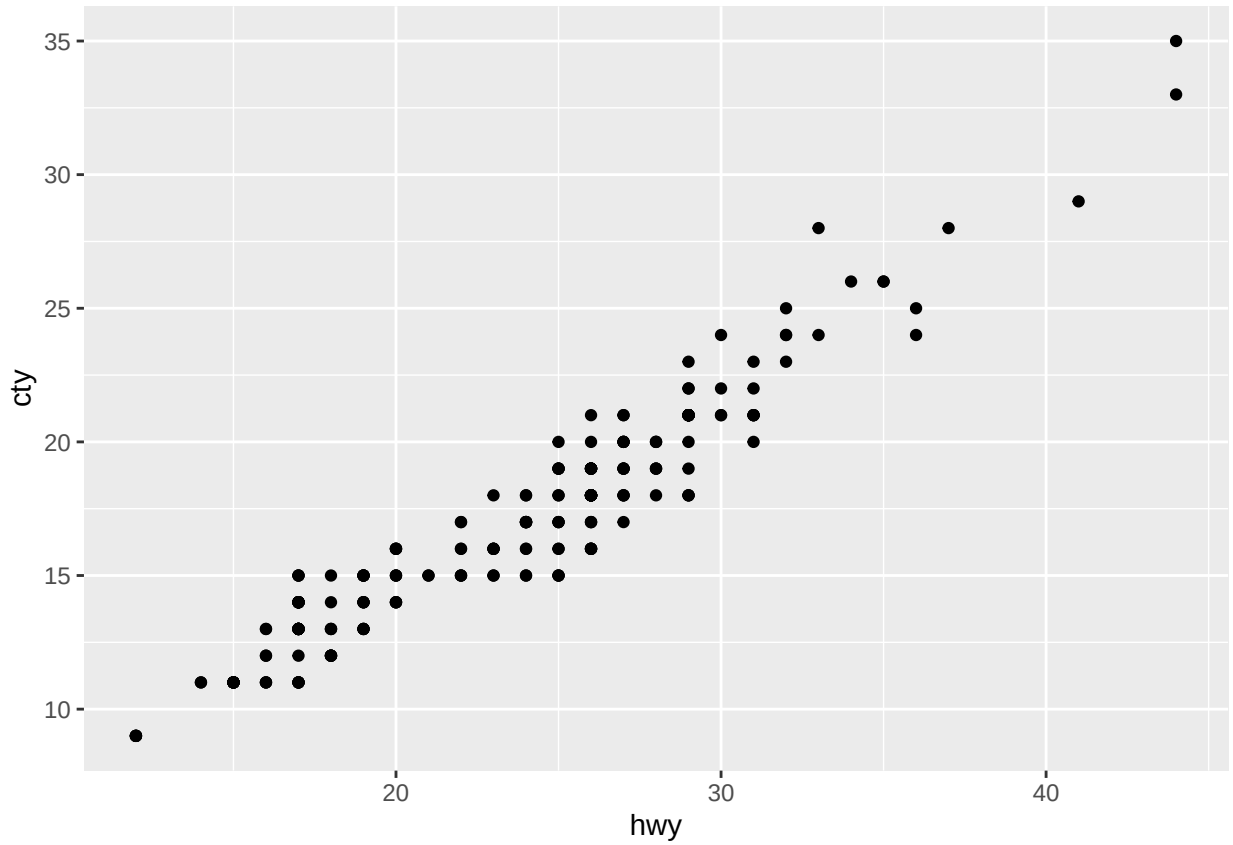
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see that the most common value of highway miles per gallon is around 26 highway mpg with a count of around 46.

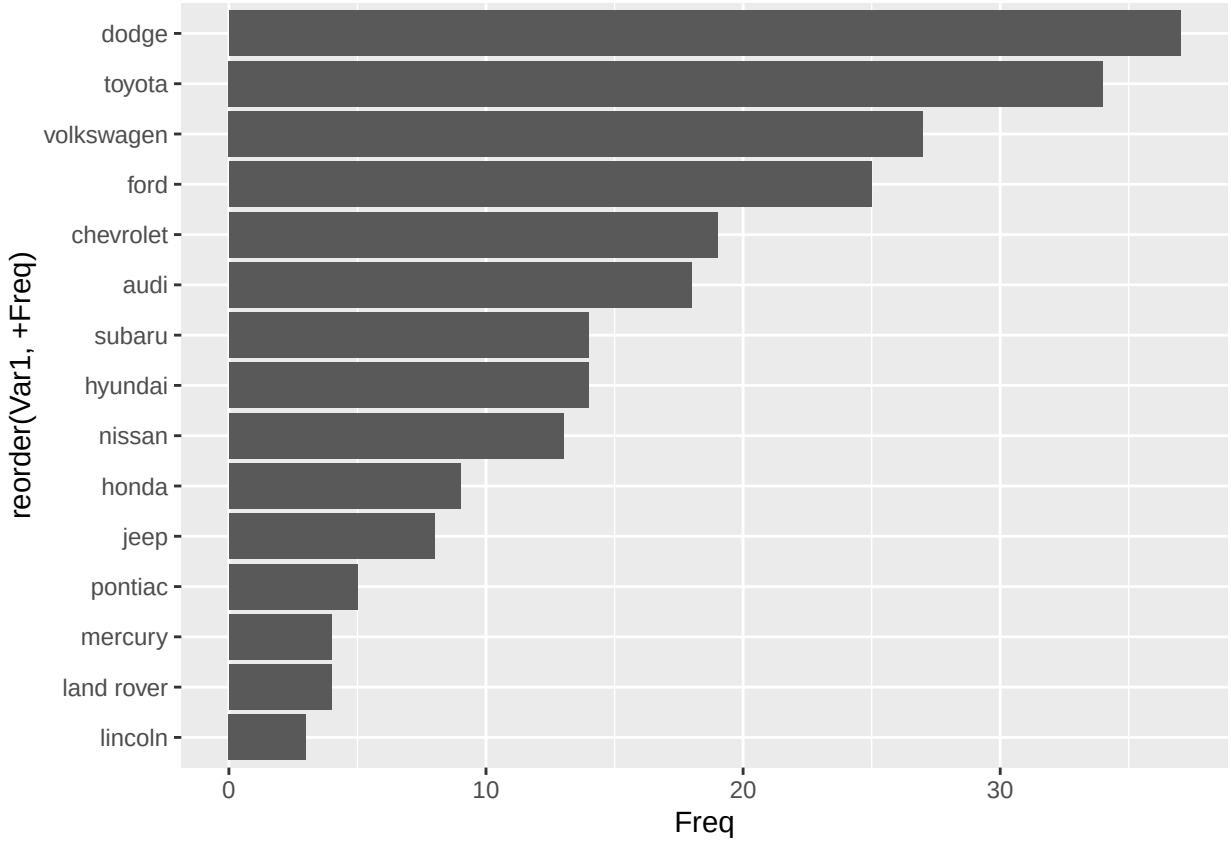
(2)

```
#attach(mpg)
#plot(hwy, cty, main = "Scatterplot of Highway vs Cty")
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point()
```



We can see that has highway mpg increases, so does cty, indicating there is a positive linear relationship between the two variables.
(3)

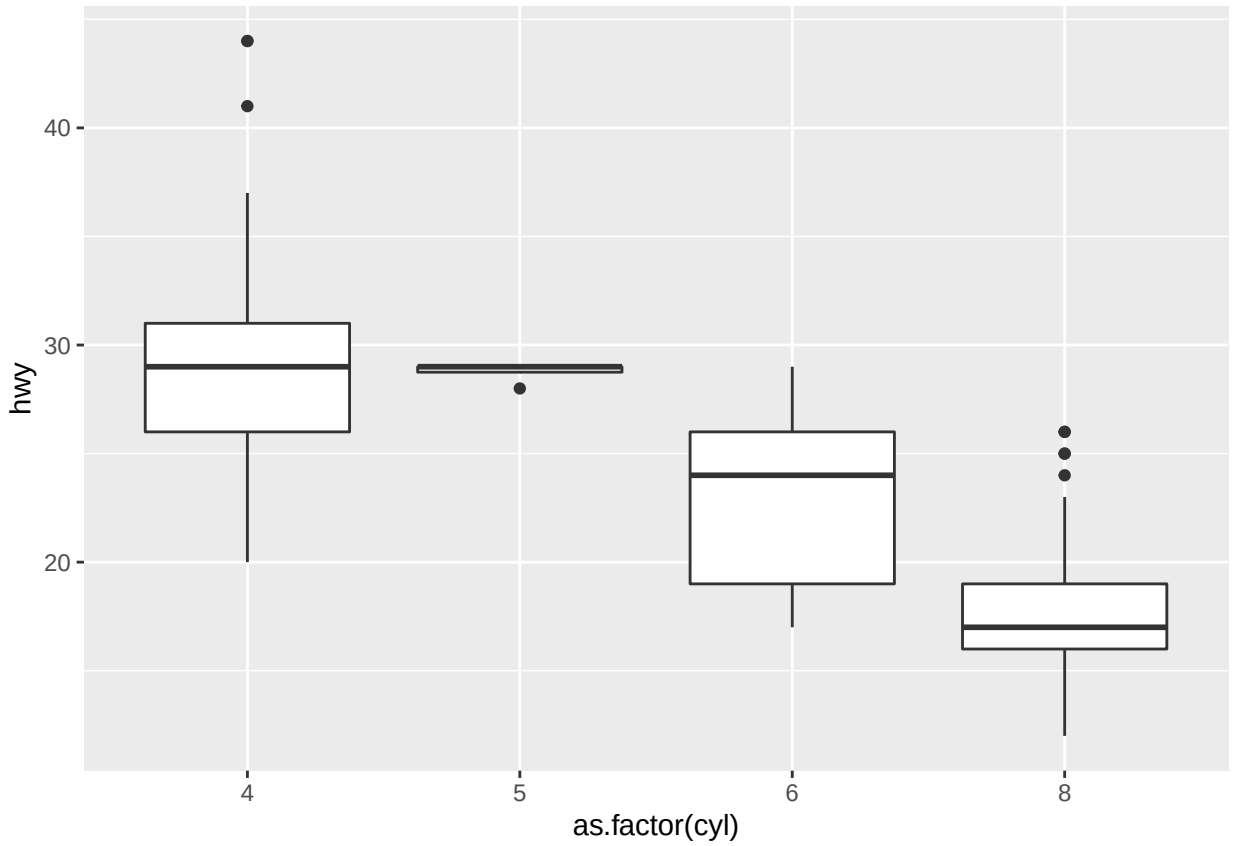
```
df = as.data.frame(table(mpg$manufacturer))

p = ggplot(data=df, aes(x=reorder(Var1, +Freq), y = Freq)) + geom_bar(stat="identity")

p + coord_flip()
```



Dodge produced the most cars while Lincoln produced the least.

(4)

```
b <- ggplot(mpg, aes(x=as.factor(cyl), y=hwy)) +
  geom_boxplot()
b
```

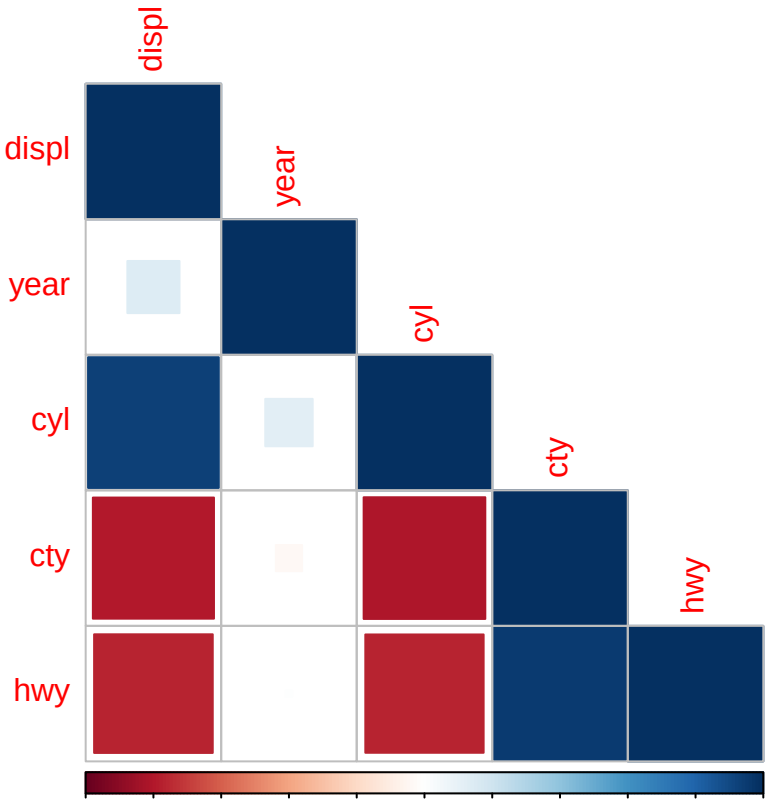


(5)

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpg2 = mpg[,-c(1,2,6,7,10,11)]

M<-cor(mpg2)

corrplot(M, type="lower", method = 'square')
```



Cyl and displ have a strong positive correlation, whereas hwy and displ have a strong negative correlation. These relationships make sense to me as displ is engine displacement while cyl is number of cylinders, so as the engine is displaced more, the number of cylinders increase.