Hw1

Machine Learning Main Ideas

Question 1: Define supervised and unsupervised learning. What are the difference(s) between them?

Answer: In Supervised learning the machine using data which is already tagged with the correct answer. Unsupervised learning is a machine learning technique where you do not need to have answer for data. Instead, you need to allow the model to work on its own to discover information.

The main difference between supervised vs unsupervised learning is the need for labelled training data. Supervised machine learning relies on labelled input and output training data, whereas unsupervised learning processes unlabelled or raw data.

Question 2: Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

Answer: In Regression model Y is quantitative which means Y is a numerical value. In Classification model Y is qualitative which means Y is categorical values.

Question 3: Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Answer: For Regression model Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). For Classification model: Accuracy. Confusion Matrix.

Question 4: As discussed, statistical models can be used for different purposes. These purposes can generally be classified into the following three categories. Provide a brief description of each.

Fom Lecture day2 page 7 Descriptive models:Choose model to best visually emphasize a trend in data

Inferential models:used to compare the differences between the treatment groups it use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects.

Predictive models: designed to assess historical data, discover patterns, observe trends and use that information to draw up predictions about future trends

Question 5: Predictive models are frequently used in machine learning, and they can usually be described as either mechanistic or empirically-driven. Answer the following questions.

Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

Mechanistic model is parametric model it uses a theory to predict what will happen in the real world. You have to assign parameters to the model more parameters makes the model have more flexibility.

Empirical modeling is non-parametric model it uses real world observations to develop a theory.

Mechanistic assume a parametric f() while empirically driven does not have assumptions about f(). Both of them have overfitting problems.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice by default it have more flexible.

For me empirically-driven model is easier to understand because using real observations to estimate

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models. Bias and variance are inversely connected for example low bias will increase variance the model will fit with

the data set while increasing the chances of inaccurate predictions.when we have higher bias it will reduce the risk of inaccurate predictions, the model will not properly match the data set.

Question 6: A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:

Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? predictive, in example we are given datas and there is no visible true and false condition.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? inferential, in this example there is a change of condition which is personal contact with the candidate that separate the dataset into two groups one has personal contacted with the candidate and other group is not.

Exploratory Data Analysis

Exercise 1: We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

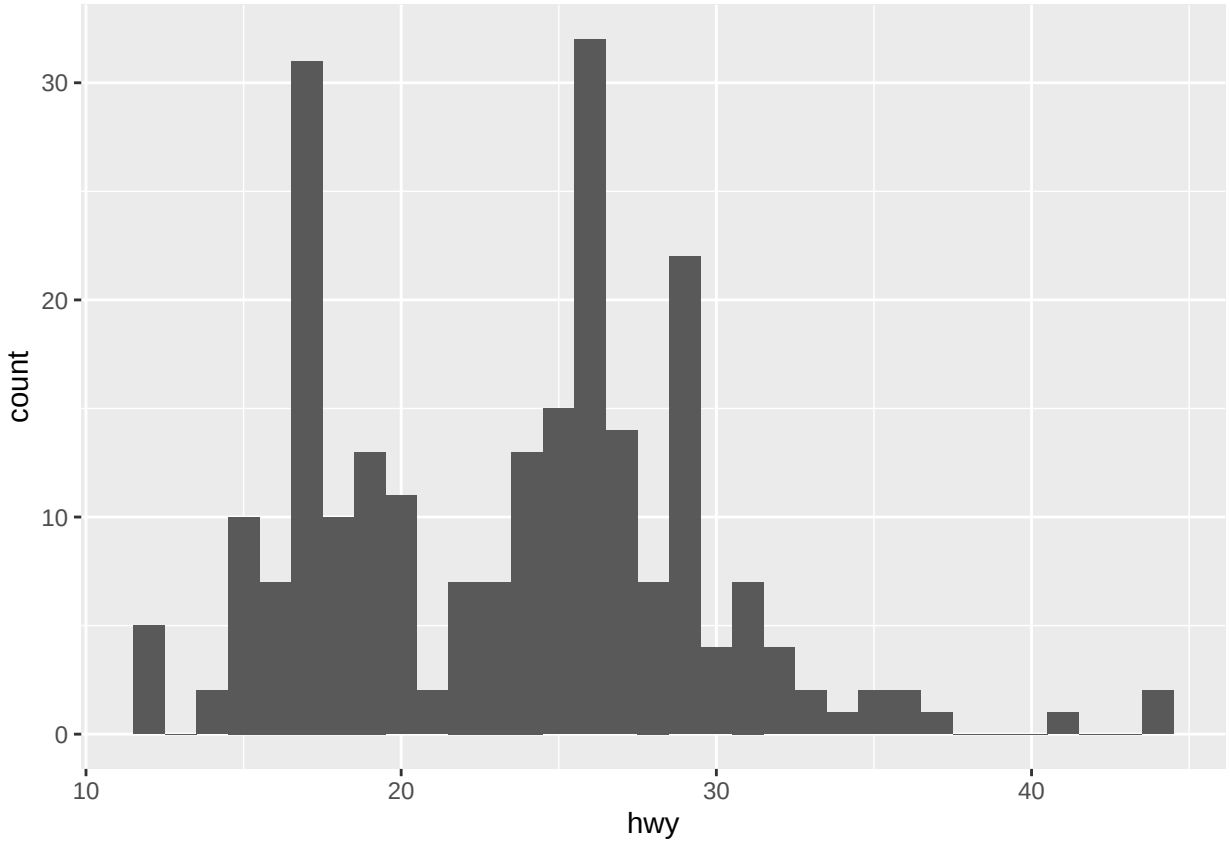
```
library(tinytex)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
data(mpg)
```

There are visible difference between highway mpg between different types of cars there are 2 category hwy about 15 hwy about 27.

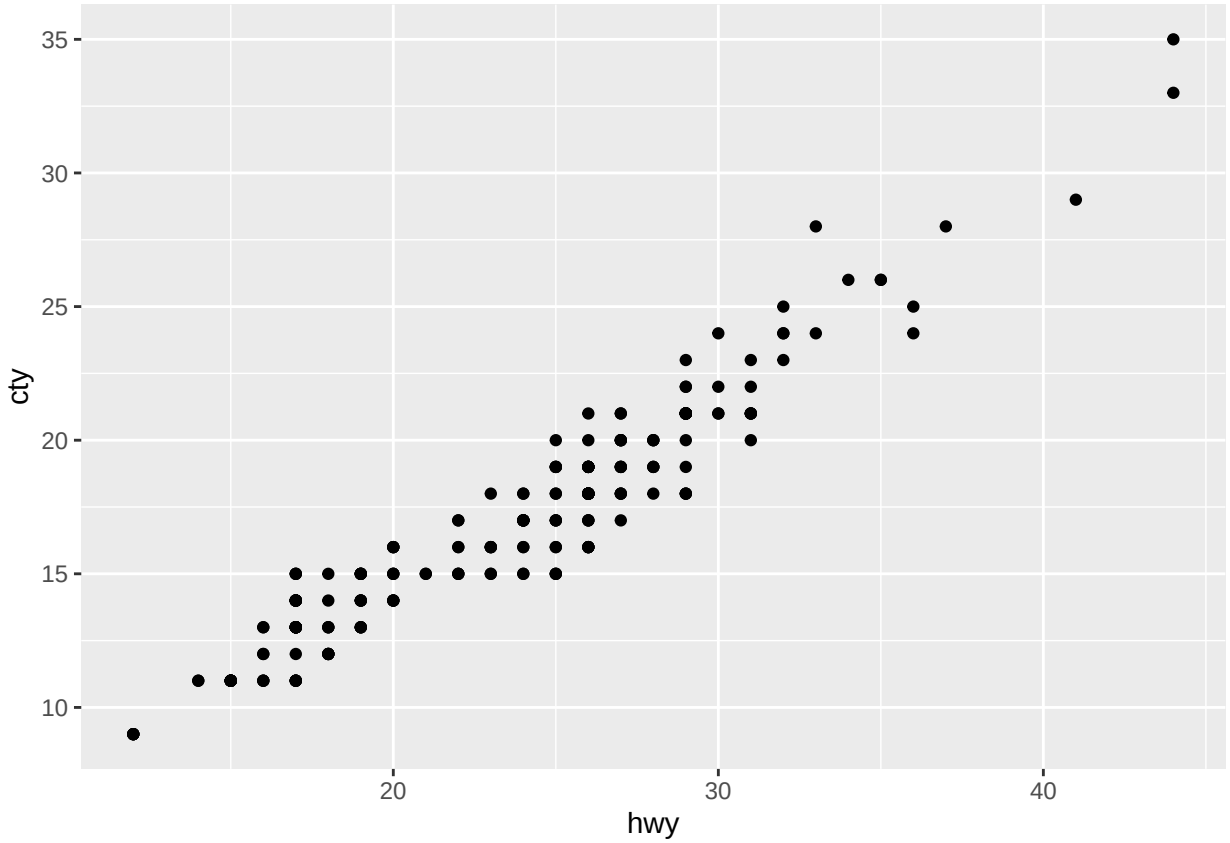
```
ggplot(mpg, aes(x=hwy)) + geom_histogram(binwidth=1)
```



Exercise 2: Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

In the graph hwy increase as cty increase therefore they are positive related. That means cars that have higher miles per gallon in city trend to have more miles per gallon on highway.

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



Exercise 3: Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least? Dodge produced the most cars and Lincoln produced the least.

```
library(tidyverse)
```

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

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

```
## 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
## v purrr   0.3.4
```

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

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

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

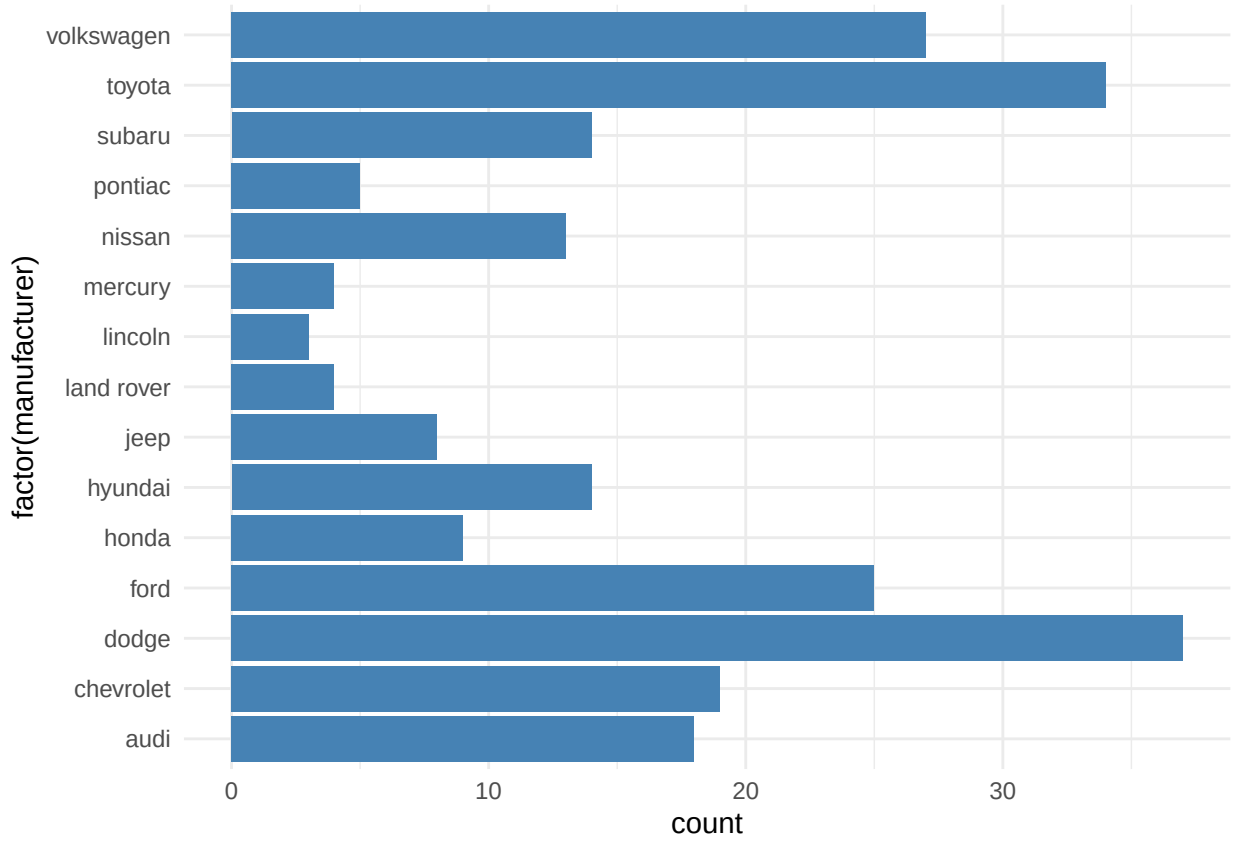
```
## Warning: package 'purrr' was built under R version 4.1.3
```

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

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

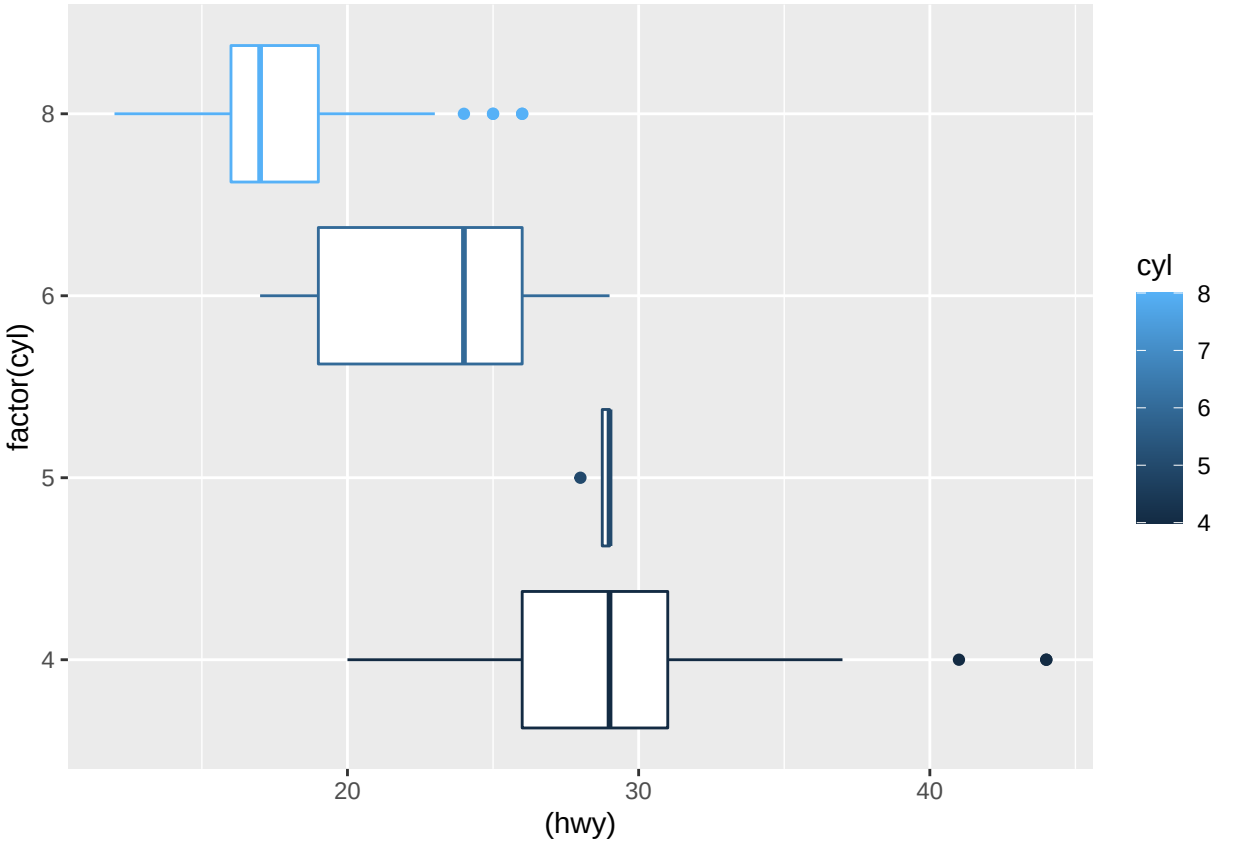
```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ggplot(mpg, aes(x=factor(manufacturer)))+
  geom_bar(stat="count", width=0.9, fill="steelblue")+
  theme_minimal()+ coord_flip()
```



Exercise 4: Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
ggplot(mpg, aes(x=(hwy),y=factor(cyl),color=cyl)) +
  geom_boxplot()
```



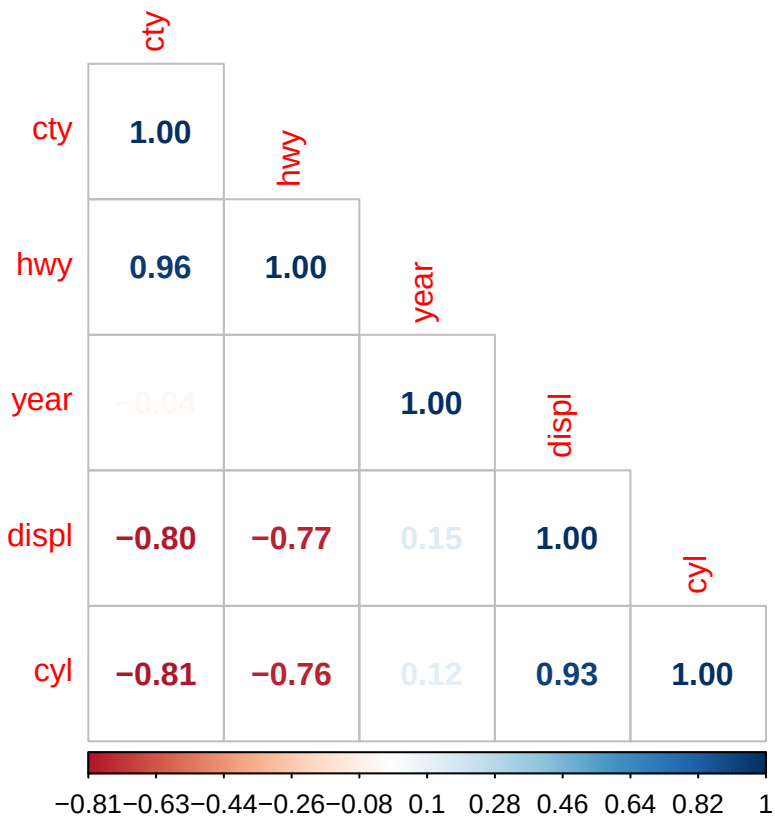
Exercise 5: Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. (Hint: You can find information on the package here.)

```
library(corrplot)
```

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

```
## corrplot 0.92 loaded
```

```
e <- cor(mpg[,c(3,4,5,8,9)])
corrplot(e, method = 'number', order = 'FPC', type = 'lower',is.corr=FALSE)
```



Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

(hwy and cty), (cyl and displ) are positively correlated. (displ and cty/hwy), (cyl and cty/hwy) are negatively correlated. year is not correlated with any other values. those relationships makes sense to me they do not surprise me.