Homework01

Alex Lim

Machine Learning Main Ideas

1) Define supervised and unsupervised learning. What are the difference(s) between them? Supervised learning is when "... for each observation of the predictor measurement(s) x_i, i = 1,...,n there is an associated response measurement y_i"(Page 26). This method attempts to understand how the predictors affect the response, and involves us finding the most appropriate model given our predictors. However, for unsupervised learning, there is no associated response measurement y_i, making us unable to fit a linear regression model. As a result, supervised learning entails methods such as Linear regression, logistic regression, decision trees, while unsupervised learning entails PCA, k-means clustering, hierarchical clustering and Neural networks.

2) Explain the difference between a regression model and a classification model, specifically in the context of machine learning. In a regression model, the response variable Y is quantitative, such as weight, height, etc. However, in a classification model, the response is instead qualitative, or categorical, like eye color.

3) Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems. The two commonly used metrics for regression in ML problems are the test and training MSE's, while commonly used metrics for classification are the test and error rate.

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. Descriptive models: This model is used to emphasize a visual trend in a dataset

Inferential models:This model is used to find which features are significant, aims to test theories, and states the relationship between the outcome and predictors.

Predictive models:This model is used to predict the value of the response variable with minimal reducible error, wants to predict future behavior.

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 models involves making an assumption about the shape of f and then using data to fit or train the model. This helps simplify the problem because it is much easier to estimate a set of parameters. For these, you can add more parameters but too many can lead to over fitting. On the other hand, emprically-driven methods do not make assumptions about the shape of f. This way, they can potentially fit a wider range of possible shapes of f. However, these require far more observations as a result. Similarly, this predictive model is also prone to over fitting due to an excess of input.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice. Mechanistic models are easy to fit due to a small number of coefficients, and as a result, will likely have easier interpretations.In general, you don't have as many factors to consider.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models. Mechanistic models often have a high bias but a low variance, while empirically-driven often have a low bias but high variance. As a result,the bias-variance trade off should be considered when using either.

6) A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions: (Classify each question as either predictive or inferential. Explain your reasoning for each.)

Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? This is predictive because we are attempting to evaluate how likely a voter will vote for a candidate given a set profile/data.

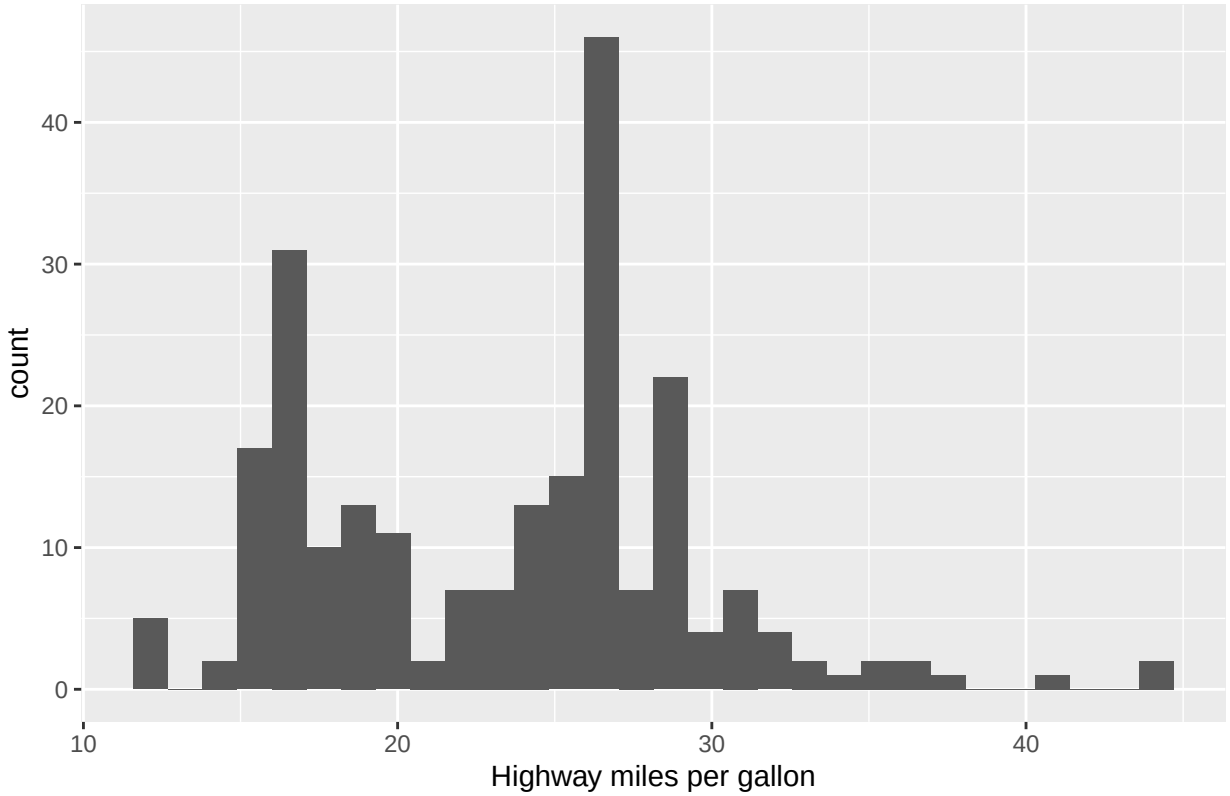
How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? However, this scenario involves more emphasis on the relationship between the response and all the predictor variables. We have to determine how a specific x changes the value of y.Therefore, it is inferential.

Exploratory Data Analysis

```
ggplot(mpg, aes(x=hwy)) +
  geom_histogram() +
  labs(title="",x='Highway miles per gallon')
```

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

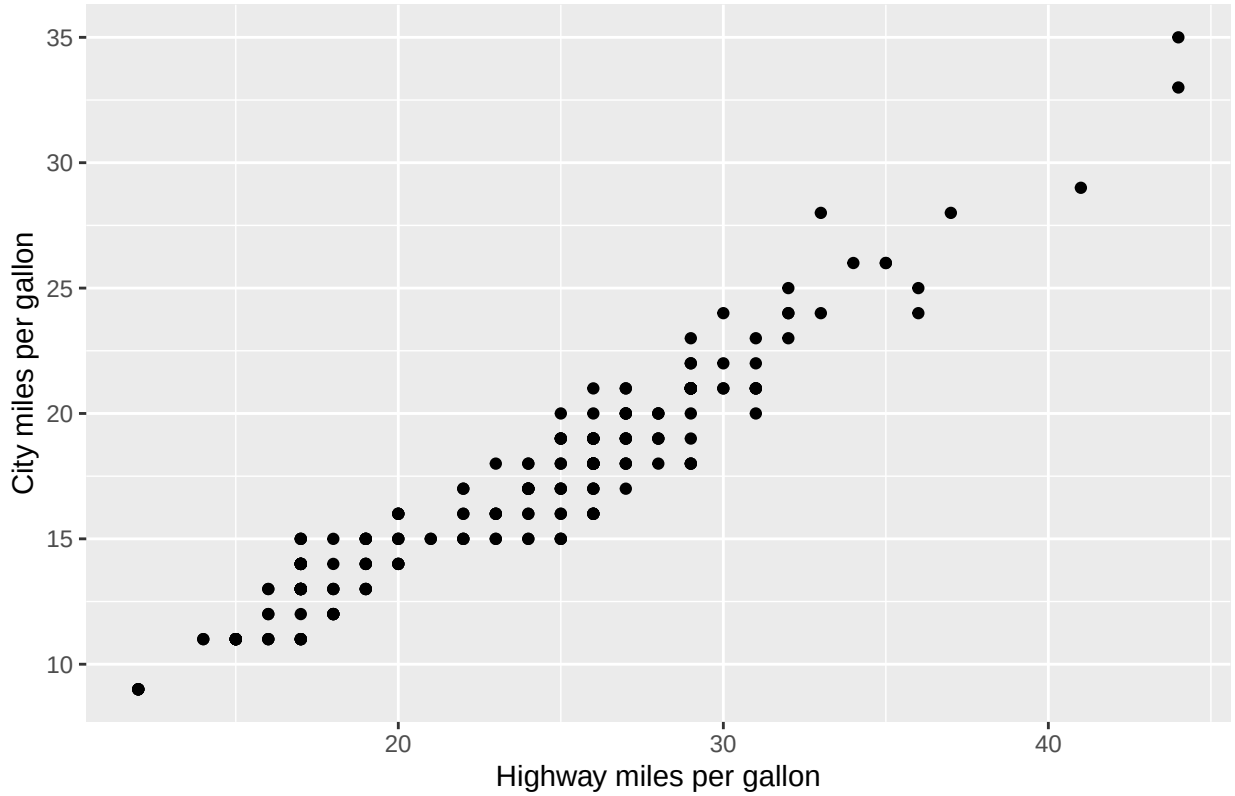
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This graph displays a bi modal distribution, in which it peaks around 17 and 26 miles per gallon. There are not a lot of values below 15, between 21 and 23, and above 29 miles per gallon on the highway.

```
ggplot(mpg, aes(x = hwy, y = cty)) +
  geom_point() +
  labs(title="",x='Highway miles per gallon', y="City miles per gallon")
```

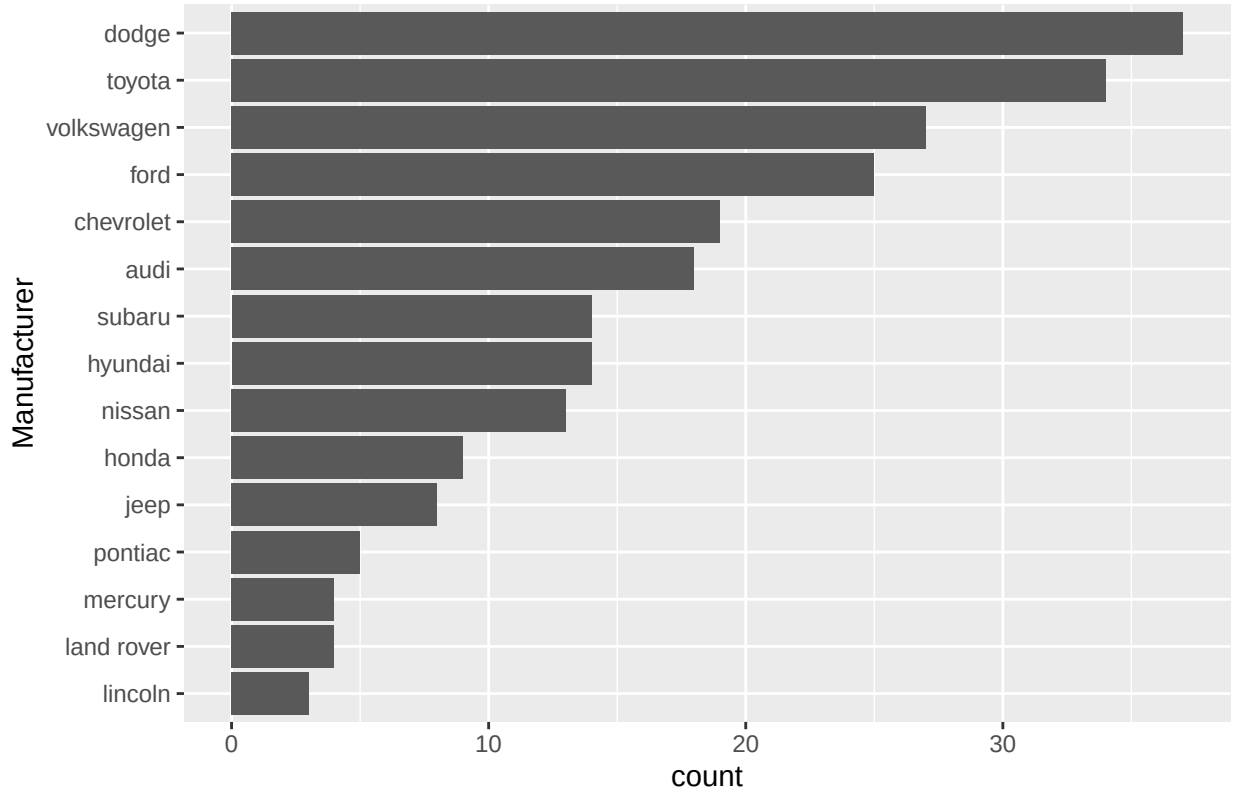
2) Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?**



Looking at the scatter, you can see that highway and city miles per gallon are positively correlated. This means that a car's mileage will perform similarly in highways and cities relative to other cars.

```
ggplot(mpg, aes(x=reorder(manufacturer,manufacturer,function(x)+length(x)))) +
       geom_bar() +
       coord_flip() +
        labs(title="", x="Manufacturer")
```

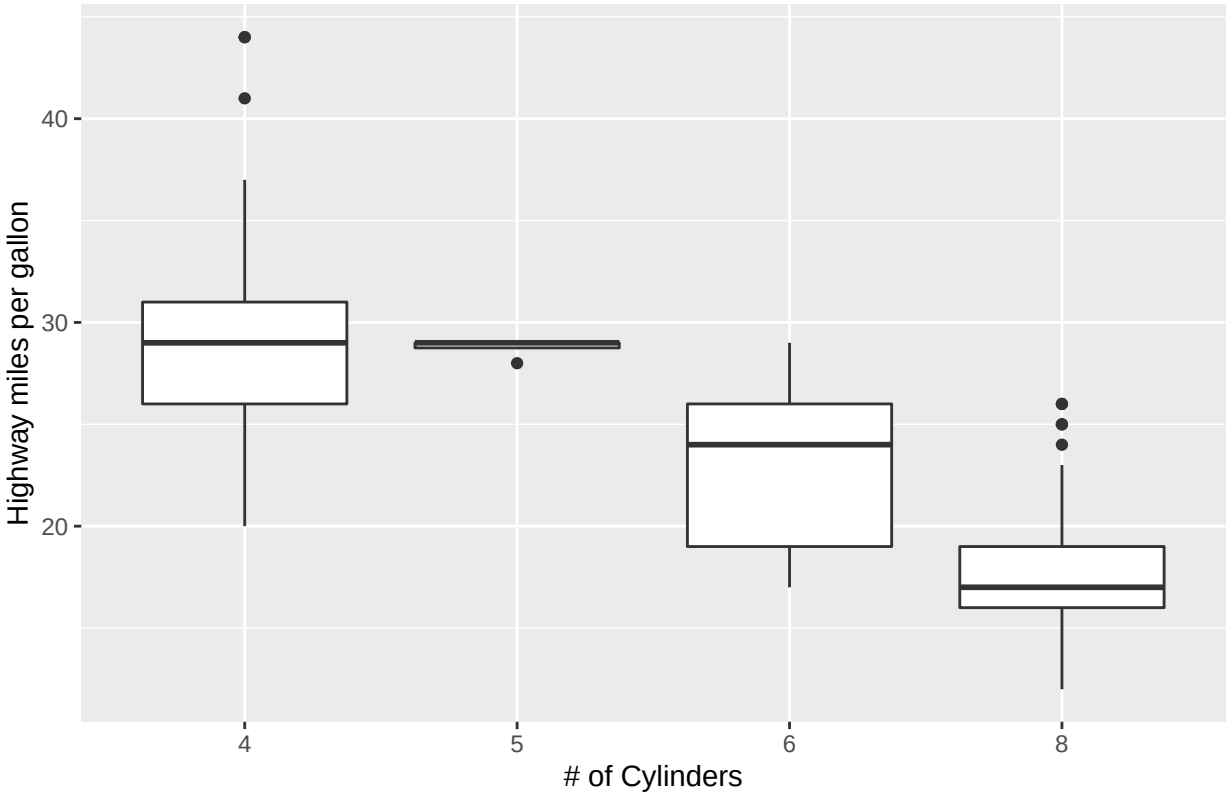
3) Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?



Dodge produced the most cars, Lincoln produced the least.

```
ggplot(data=mpg, aes(x = as.factor(cyl), y = hwy)) +
      geom_boxplot()+
      labs(title="",y = "Highway miles per gallon", x="# of Cylinders")
```

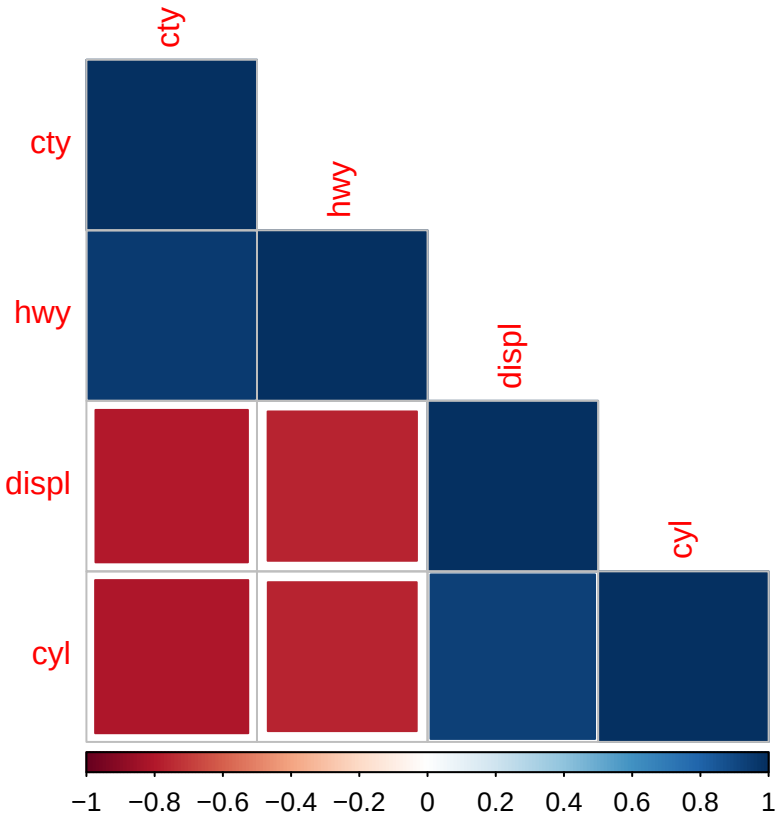
4) Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?



Observing the boxplot, you can see that cars with less cylinders tend to have higher highway miles per gallon.

```
mpg2 <- mpg[,sapply(mpg,is.numeric)]
mpg3 <- subset(mpg2, select = -c(year))
M = cor(mpg3)
corrplot(M, method = 'square', order = 'FPC', type = 'lower')
```

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



that surprise you?

From the map, we can see that highway/city and cylinder/displacement are both positively correlated. On the other hand, displacement/city, cylinders/city, highway/displacement, and cylinder/highways are all negatively correlated. Intuitively, the relationship between city and highway miles per gallon makes sense. Additionally, the relationship between displacement and cylinders makes sense since displacement is the total volume of all the cylinders in an engine, meaning more cylinders likely means more displacement. Adding onto this, given our analysis in question 4, less cylinders means more highway miles per gallon. It would follow that this would be similar for city miles per gallon since they're similar. Finally, since displacement is highly positively correlated with cylinders,so it would make sense that displacement also has a negative correlation with highway and city miles per hour. Therefore, most of what is shown in the heat map makes sense.