

PSTAT131_HW1

Andrew Choi

2022-10-29

Machine Learning Main Ideas

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

Answer: Supervised learning is a subcategory of machine learning used to “accurately predict future response given predictors, understand how predictors affect response, find the”best” model for response given predictors,” and “assess the quality of our predictions and (or) estimation.” (from lecture)

Unsupervised learning is also a subcategory of machine learning used to “discover hidden patterns or data groupings without the need for human intervention.” It is useful when looking for similarities and differences in data. (from <https://www.ibm.com/cloud/learn/unsupervised-learning>)

The main difference between supervised and unsupervised learning is that supervised learning trains by using labeled data, which means that human intervention is required. Humans are needed to label data appropriately. This labeled data acts as an “answer key” that helps our supervised learning algorithm learn to predict/estimate more accurately. Unsupervised learning does not need labeled data, and instead works with large amounts of unlabeled data to train itself. (from <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>)

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

Answer: “The main difference between Regression and Classification algorithms is that Regression algorithms are used to predict the continuous values such as price, salary, age... and Classification algorithms are used to predict/classify discrete values such as Male or Female, True or False, Spam or Not Spam...” (from <https://www.javatpoint.com/regression-vs-classification-in-machine-learning>)

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

Answer: Two commonly used metrics for regression ML problems include Mean Squared Error (MSE) and R Squared. Two commonly used metrics for classification ML problems include Accuracy and Confusion Matrix. (from <https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/> and <https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/>)

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.

Answer: A descriptive model uses mainly unsupervised learning approaches “for outlining, classifying, and drawing out information for what had happened in the past.” (from <https://medium.com/appengine-ai/descriptive-analysis-machine-learning-268507a99e2>)

An inferential model runs “data points to a machine learning model to calculate an output such as a single numerical score.” (from <https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-syntax-inference-overview#:~:text=Anomaly%20Detection-,Introduction,machine%20learning%20model%20into%20product>)

A predictive model uses “machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data.” (from <https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml>)

Question 5: Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

Answer “A mechanistic model uses a theory to predict what will happen in the real world... Empirical modeling studies real-world events to develop a theory.” While they both have the same components, i.e. they make use of theories and real-world occurrences, the order in which these components are used are opposite. Mechanistic models use theories as inputs and real-world occurrences as outputs, whereas empirical models use real-world occurrences as inputs and theories as outputs. (from <https://smallbusiness.chron.com/mechanistic-model-12706.html>)

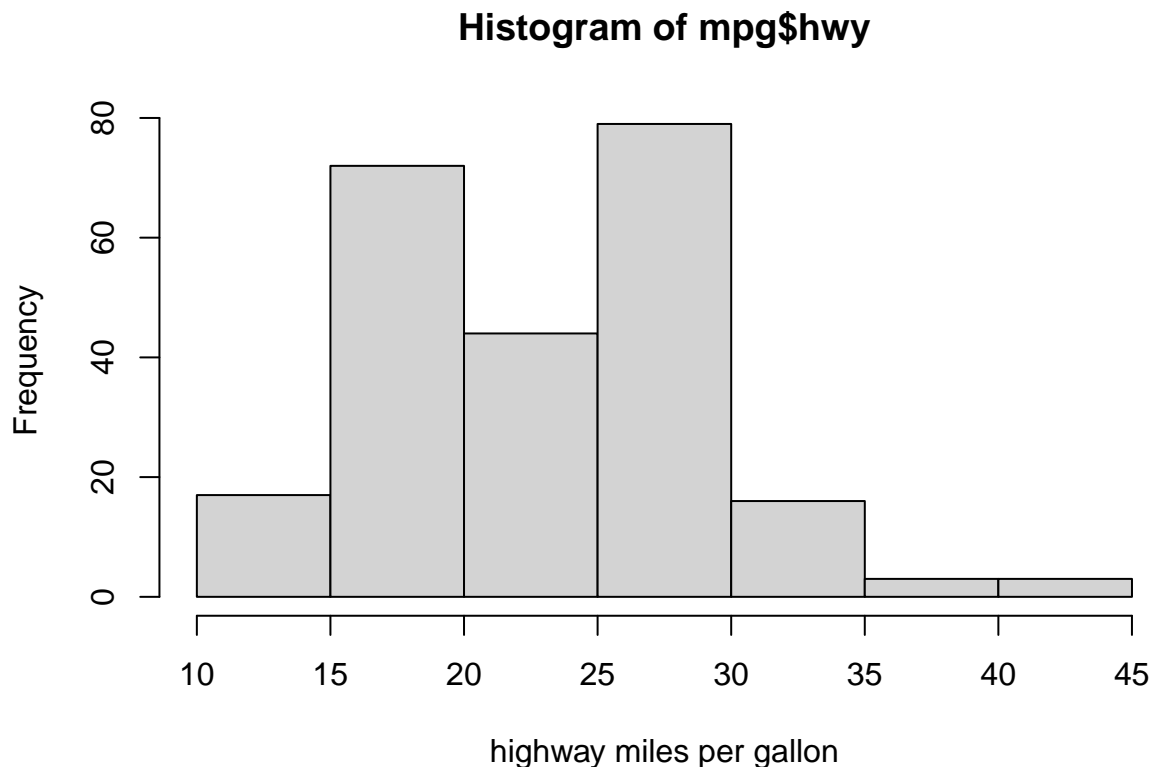
In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

Answer

In my opinion it’s easier to understand an empirically-driven model simply because it’s easier to generate a theory by studying real-world occurrences, than to predict a real-world occurrence using a theory.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

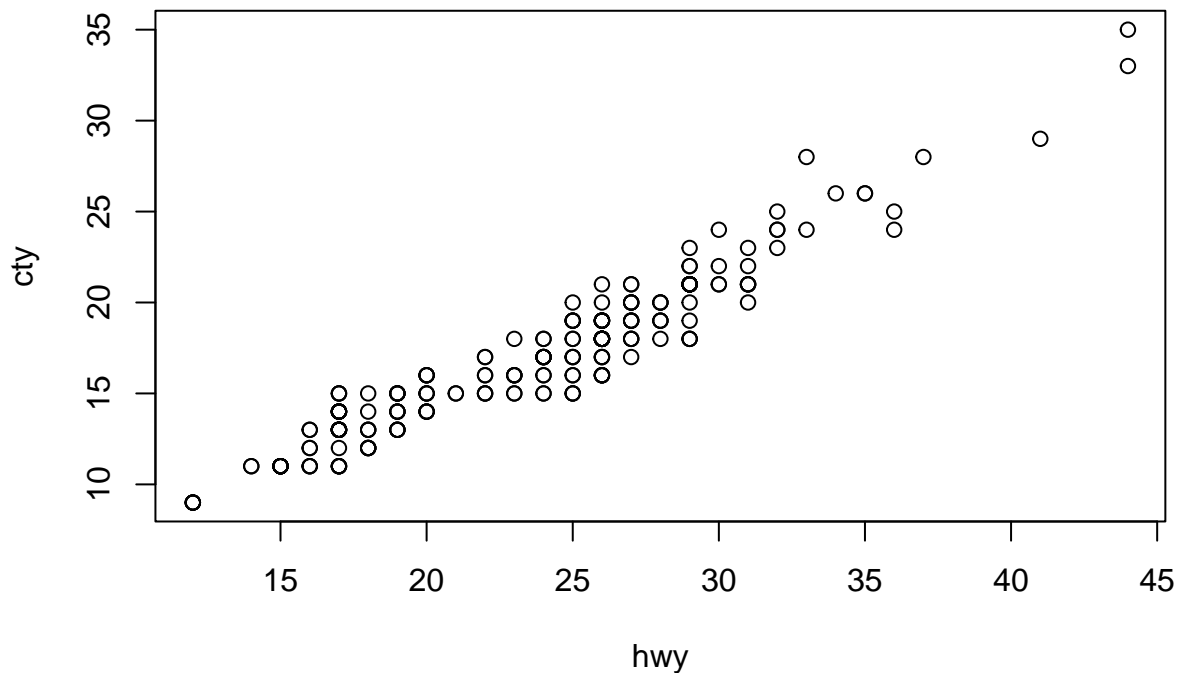
Answer



Cars that have between 15 to 20 highway miles and 25 to 30 highway miles appear the most frequently, while cars that have between 20 to 25 highway miles follow closely behind. Cars that don't have between 15 to 30 highway miles appear infrequently, with cars that have between 35 to 45 highway miles per gallon appearing the least frequently of all.

From this information, I've learned that cars made between 1999 and 2008 are likely to run between 15 to 30 highway miles per gallon of gasoline.

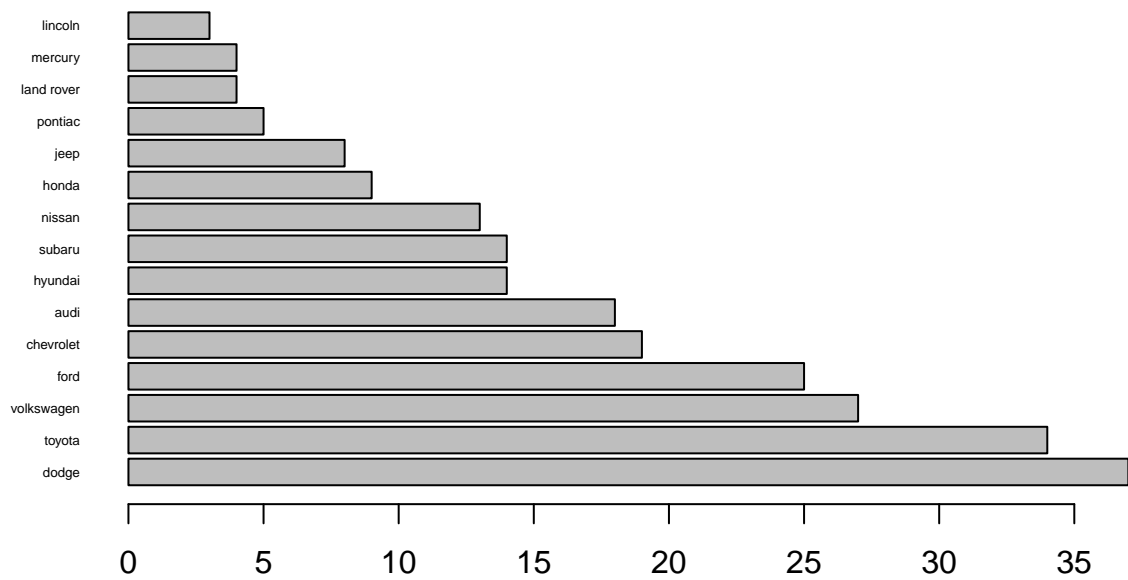
Exercise 2:



There appears to be a strong positive correlation between the variables `hwy` and `cty`. This tells us that as the number of highway miles per gallon of gasoline a car has rises, the number of city miles per gallon of gasoline tends to rise as well. This does not mean that these variables are dependent, they are simply related.

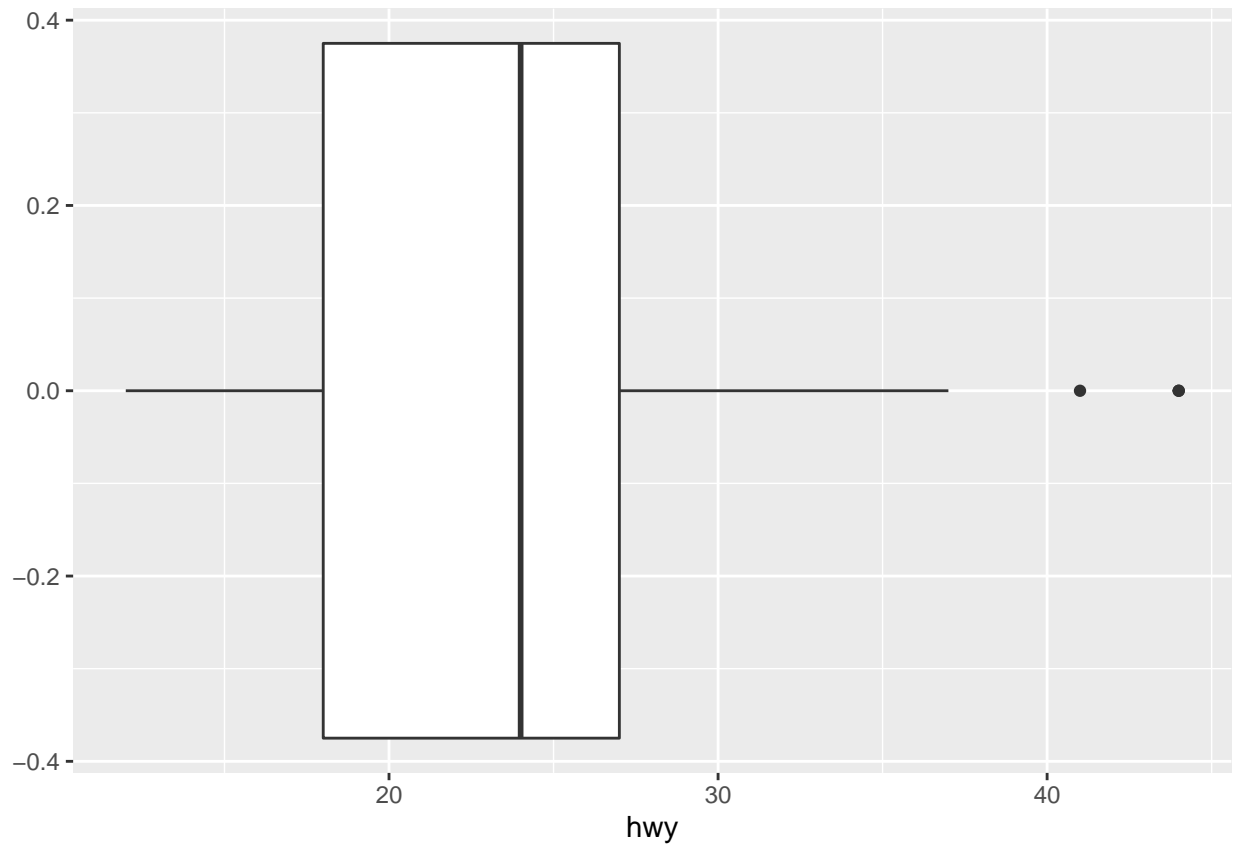
Exercise 3:

The following website was used to understand how to order my barplot from highest to lowest values:
<https://statisticsglobe.com/table>



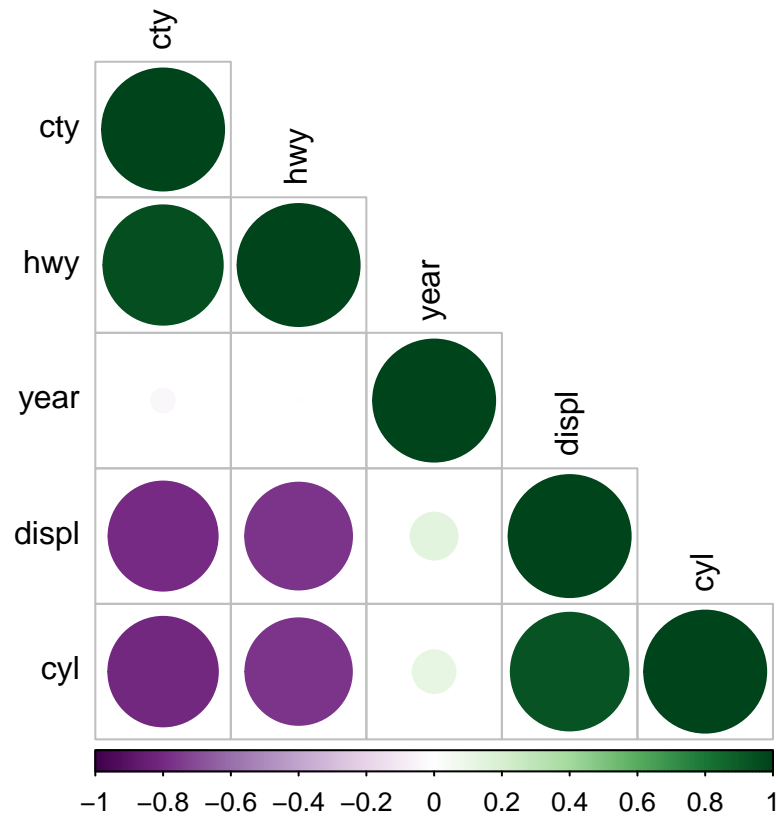
Lincoln produced the least amount of cars, while Dodge produced the most.

Exercise 4:



Exercise 5:

```
##      table_trans
## 1:           5
## 2:           2
## 3:          83
## 4:          39
## 5:           6
## 6:           3
## 7:           3
## 8:          16
## 9:          58
## 10:         19
```



Things I need help with

question 5.3

which one has higher variance and which one has

How do I submit the homework the way Coburn wants? (how to upload knitted html file to github?)

Should I include nonnumeric variables in my correlation matrix for #5?