Chapter 02: Statistical Learning

Stan Piotrowski

September 24 2021

Contents

Notes

Prediction and inference

- If we have a quantitative response Y and a set of p predictors, we can write a general form of the relationship as $Y = f(X) + \epsilon$
- The error term (ϵ) is a random error term, also referred to as irreducible error, and has mean zero.
- In most situations, the true functional relationship between Y and a set of p predictors is not known, in which case we are estimating the function f in the prediction setting (i.e., where we don't know the output).
- In the prediction setting, we re-write the equation as $\hat{Y} = \hat{f}(X)$, and the process of estimating the functional relationship induces additional error, called the reducible error.
- In an inference setting, we aren't necessarily concerned with predicting a response, but understanding the association between Y and the set of p predictors.
- When we are interested in understanding the association between Y and p predictors, we are looking at identifying the exact functional form of the relationship.

Parametric vs. non-parametric methods

- The most important distinction between parametric and non-parametric methods is the former makes explicit assumptions about the shape of the function f, while the latter does not.
- An example of a parametric method is ordinary least squares regression, which assumes that the relationship between Y and a set of p predictors is linear and is solely explained by the set of p predictors.
- In contrast, an example of a non-parametric method is a thin-plate spline used for interpolation and smoothing, which instead instead of assuming a functional form for f, tries to develop an estimate that is as close as possible to the observed data.

Prediction accuracy and model interpretability trade-off

- When choosing a statistical learning method, there is a trade-off between flexibility and interpretability: flexible methods are generally better for prediction accuracy, but are less interpretable, while inflexible methods have relatively poor prediction accuracy but are relatively easier to interpret.
- For example, if we are interested in inference, inflexible models that reduce the problem to a simplified version of reality with just a few predictors is more desirable: the model performs relatively well on training data and easily generalizes to unseen test data.
- Further, inflexible methods allow us to generalize and understand the nature of the relationship (i.e., interpret the model) between the response and predictors.

- On the other hand, if we are interested in prediction, flexible models can capture further intricaces in the data, but suffer from overfitting (essentially capturing noise in the data rather than the true functional relationship).

Unsupervised vs. supervised learning methods

- In supervised settings, we are training a learning method using a set of predictors and known outcomes, then applying the model to unseen test data and essentially comparing how often the model identifies the true known response.
- In an unsupervised setting, we don't have outputs; rather, we are interested in discovering meaningful patterns just using the data.
- An example of supervised learning is the ordinary least squares regression; clustering using principal components analysis is an example of unsupervised learning.

Regression vs. classification problems

- In general, quantitative responses can be thought of as regression problems (although there are exceptions like logistic regression for categorical data), while qualitative responses can be thought of as classification problems.

Evaluating model accuracy

- The most important rule to remember is that there is no single "best" method for all data sets and careful exploration and interpretation is needed to select a model which balances the trade-off between flexibility and interpretability.
- The mean squared error (MSE) is a metric used in the regression setting to quantify how close the predicted response is from the observed response for a particular value of predictor.

- In essence, MSE is the mean of the squared differences between the predicted value and the observed response value.
- Interpreting the MSE is simple: the smaller the MSE, the closer the model predicted values are to the observed values.
- Importantly, simply because a given model out of a set has the lowest training MSE doesn't mean that it will have the lowest test MSE- this highlights the importance in balancing flexibility and the ability of the model to generalize to unseen test data.
- Degrees of freedom in statistical learning is the quantity that describes the flexibility of a curve to data: more flexible methods have higher degrees of freedom and generally fit the data more closely, but at the expense of potentially performing poorly on test data (generalization problem).
- When evaluating model flexibility and accuracy, as the flexibility of the method increases, the training MSE will consistently decrease, while the test MSE will exhibit a U-shaped curve: it will decrease up to a certain inflection point, beyond which it increases rapidly because of the generalization problem.
- We can decompose the expected test MSE using the following equation:

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

- The first term, $E(y_0 - \hat{f}(x_0))^2$, describes the expected test MSE, or the mean test MSE if we were to calculate the test MSE over successive training data sets.
- The second term, $Var(\hat{f}(x_0))$, describes the variance in the functional form of f: essentially, how much does f change over successive training data sets.
- The third term, $[Bias(\hat{f}(x_0))]^2$, is the absolute value of the squared bias, or the error introduced by approximating reality with a simplified model.
- Finally, the last term, $Var(\epsilon)$, is the variance in the irreducible error term over successive training data sets.
- In general, using increasingly flexible methods will lead to an increase in $Var(\hat{f}(x_0))$ and a decrease in $[Bias(\hat{f}(x_0))]^2$; the opposite is true for inflexible models.

Classification

- The Bayes classifier is considered the gold-standard and evaluates the conditional probability of a response belonging to a specific class given a value of the predictor variable: the assigned class is whichever has a probability > 0.5.
- However, using the Bayes classifier requires that we know the conditional distribution of Y given X, which is generally unknown.
- A close approximation is the K-nearest neighbors (KNN) classifier, which looks at each test observation in training data and assigns a conditional probability based on the fraction of K points that are closest to it; the assigned class is the one with the highest estimated conditional probability.
- In general, higher numbers of K using the KNN classifier decreases flexibility and the variance in functional form at the expense of increasing bias.

Exercises

Conceptual

1) This exercises poses a question regarding the performance of flexible statistical learning methods relative to inflexible methods.

    a) For each of these scenarios, we need to consider the variance of the function describing the relationship between the response and the predictor(s) between successive training sets and the bias, or the error introduced when models are used to simplify reality. Both of these metrics factor into the overall test mean squared error (MSE), or the mean squared difference between the predicted value or the value estimated by the function and the observed value (in the case of supervised statistical learning).

    b) In a case where there is a small sample size and a large number of predictors, an inflexible model would perform better than a flexible model. Although the bias would be larger using the inflexible model, the power of this approach is the ability to be able to identify the predictors that are associated with the response, assuming that only a small fraction are actually important.

    c) In a scenario where the relationship between the predictors and the response is highly non-linear, we would expect the flexible statistical learning method to be better than an inflexible method (generally). Inflexible learning methods, like ordinary least squares regression, for example, will likely have a low variance between successive training data sets, but have a large bias because we are attempting to explain a non-linear relationship with assumptions in a simplified linear reality.

    d) In contrast to c), in a scenario where the variance of the error terms is extremely high, an inflexible method may be preferred because fitting a flexible statistical model on successive training data sets that are substantially different could be driving the high variance.

2) The following questions ask to explain whether the scenario is a classification or regression problem, and whether the primary question of interest is inference or prediction. We are also asked to identify n, the number of samples, and p, the predictors.

    a) In this scenario, we are dealing with $n = 500$ samples (i.e., the 500 firms) and $p = 3$ predictors: profit, number of employees, and industry. In this example, the CEO salary is the response. This is a classic regression problem where the primary goal is inference, because we are interested in understanding the relationship between the set of predictors and CEO salary and the data generating mechanism. For example, we would expect profit to have a positive influence on CEO salary, but what about number of employees and industry? Further, we are not necessarily interested in predicting a CEO's salary based on historic data.

    b) This is a classic example of a classification problem where the primary goal is prediction using historical data We have $n = 20$ products and $p = 13$ predictors. Here, the response is whether the product was a success or failure, based on these predictors. I argue that the goal is prediction here because when we're launching a new product, we want to know how likely it is to succeed, plain

and simple, and the prompt does not suggest that we are interested in understanding *how* exactly each of these variables influence the response.

It seems plausible to pivot and frame this as an inference problem as well, as presumably you would want to know which variables were most influential in whether the product was a success or failure.

c) This is another example where the goal is prediction, but framed as a regression problem because we are interested in predicting a quantitative response. Since we have weekly data for all of 2012, we have $n = 52$ samples, with $p = 3$ predictors.

3) This question revisits the bias-variance decomposition and asks us to draw a sketch of a typical plot showing the bias, variance, training and test error, and Bayes or irreducible error curves. Instead, I'll explain how each curve would look, generally, and why they have the shapes they do.

On the x-axis we would display the flexibility of the model, going from less flexible (less) to more flexible (right). In general, less flexible models will have the following properties:

- **Bias:** less flexible methods suffer from the fact that they attempt to approximate complex reality using a simplified model, which inherently introduces error due to differences between the observed data and the responses predicted by the estimated function.
- **Variance:** less flexible methods generally have lower variance because the models may assume the relationship is linear (in the case of linear regression) and go through "most" of the data without trying to go through each point.
- **training error:** higher training error, again getting back to the fact that the model attempts to simplify the relationship, which introduces error if the relationship is not linear, for example.
- **test error:** lower training error, because the model is able to generalize well to unseen test data.
- **irreducible error:** this stays fixed, because it is not related to the functional form of the relationship betweeen the response and predictors.

On the other hand, more flexible models will generally show the opposite of each of the properties described above.

4) This question asks us to think of applications for statistical learning methods.

a) Classification applications:

i) In population genetics, we are often interested in exploring how genetic variation across the genome can be used for understanding population structure. A classification problem focused on inference would be building a statistical learning method which classifies samples into groups or "populations" based on the genetic data alone.

ii) Framing the above example slightly differently, we have some genotype data for fish spawning in different rivers and we know fish A only spawns in river A, and so on and so forth. We are then given several fish of unknown origin and asked to identify the population of origin using the genotype data alone. Here, we're presented with a classification problem where the primary goal is prediction to put the fish into the correct spawning river.

iii) In a different example, we might be interested in knowing which variables affect whether a person will default on a loan. Here, the question is one of inference, because we are interested in the functional form of the relationship between the binary outcome and the set of predictors.

b) Regression applications:

i) Let's say we are interested in understanding how a series of predictors influences the price of a home in a given city. Here, the predictors may be the lot size, school system rating in the area, square footage of the home, number of bathrooms and bedrooms, etc., and the response is the price of the home. This would be an inference question, as we're interested in the relationship between the predictors and the response.

Obs.	X1	X2	X3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

ii) In another scenario, we may be interested in using regression to predict the highway fuel economy of a vehicle (the response) given a set of predictors like the model, the engine displacement, number of cylinders, transmission, and the fuel type. This is framed as a prediction question because we don't really care about the relationship; we're just interested in predicting a quantitative response correctly as often as possible.

iii) Framing the above question slightly different as an inference problem, I could could pose a question as an auto-maker: given all of these predictor variables, which subset has the most influence on the highway fuel economy. That could help direct effects towards improving those variables which have a positive influence, and trying to mitigate for the variables that have a negative influence.

c) Cluster analysis:

i) In a scenario where I had RNA-seq data generated from many different cell types in a piece of tissue, cluster analysis would facilitate the discovery of clusters using the data alone which may correspond to cell type.

ii) In another situation suppose I sampled fish in a lake and was interested in characterizing population structure using genetic data alone: this would be a great application for cluster analysis to reveal any interesting patterns in the data.

iii) Finally, in a different scenario (although not a very plausible one), say we collected data on vehicles including engine displacement, number of cylinders, transmission, drive train, and fuel economy. We could use cluster analysis to identify interesting groups within those data, which may correspond to vehicle class (e.g., compact, etc).

5) The advantages of using a very flexible model is that you don't have to explicitly assume the data follow a defined distribution (e.g., linear). This can be beneficial in scenarios where you aren't concerned with interpretability; rather, your sole focus may be prediction accuracy. On the other hand, very flexible models suffer from a general lack of interpretability, overfitting, and lack of generalizability, particularly as flexibility increas In some situations, a less flexible model may be preferred if the goal is inference and interpreting model coefficients with the added benefit of generally lower test mean squared error relative to more flexible models.

6) In essence, a parametric statistical learning method simplifies a problem by boiling down the shape of the function f down to just a few parameters and assumes that shape follows a known distribution. In a regression setting, parametric learning methods are advantageous in that increase interpretability and in general, simpler models are often preferred relative to overly complex models. Additionally, parametric methods can be used in situations with smaller sample sizes (generally far more samples are needed when using non-parametric methods). In contrast to parametric methods, non-parametric methods don't rely on explicit assumptions about the shape of the function f, so are much more flexible, but at the expense of interpretability: non-parametric methods aren't simplifying a functional form down to just a few parameters. Further, more flexible methods suffer from overfitting and high test error associated with poor performance on unseen test data.

7) The table below is re-recreated from the textbook:

a) First, we'll calculate the Euclidean distance between each observation as a test point with the values

Obs.	Y	Distance
5	Green	1.41
6	Red	1.73
2	Red	2.00
4	Green	2.24
1	Red	3.00
3	Red	3.16

$X1 = X2 = X3 = 0$. Note, the Euclidean distance formula is $\sqrt{(\sum((x_i - y_i)^2))}$. We'll sort the table below based on the calculated Euclidean distance.

b) For each test point, the KNN method looks at the K number of points closest to the test point using the Euclidean distance and estimates the probabilities of belonging to one of K classes. If we chose a highly flexible model of $K = 1$, we'll assign the test point to the class ("Red" or "Green") of the closest point (i.e., the point with the smallest Euclidean distance). In this case, the point with the smallest Euclidean distance 1.41 is "Green."

We can check this manual calculation using the `knn()` function from the `class` package. The function requires us to define a training set, a test set, the classes of the training set, and the number of neighbors to be considered in the KNN algorithm. We'll use the first table as the training data, the test point values for the single test case, and define the number of neighbors.

```
## [1] Green
## Levels: Green Red
```

c) If we modeled $K = 3$, looking now at the 3 closest points, we see that the most commonly occuring class (2/3) is "Red", so we'd assign the test point as "Red." We can use the same approach as we did above to check our manual calculations.

```
# Run KNN with K=3
knn(train = training_set,
    test = test_set,
    cl = classes$Y,
    k = 3)
```

```
## [1] Red
## Levels: Green Red
```

d) If the Bayes decision boundary in this particular problem is highly non-linear, the "best" value for K would be small. When the value for K is small, for each individiual test point we are only evaluating the closest neighbor to assign a class, which will allow the KNN decision boundary to be extremely flexible and essentially "custom catered" for each test point. When K is large, we are considering more neighbors for each individual test point to estimate a probability and assign a class, which generally forces the KNN decision boundary to follow a linear form.

Applied

8) This question uses the `College` data set from the ISLR2 package.

a) This first questions asks us to load the `College` data, but it is already distributed in the ISLR2 package.

b) Add the rownames as a column. I'll also clean up the column names to make all of them lowercase and replace the period characters with underscores using the janitor package. Note how the data frame has changed from the original one distributed in the ISLR2 package.

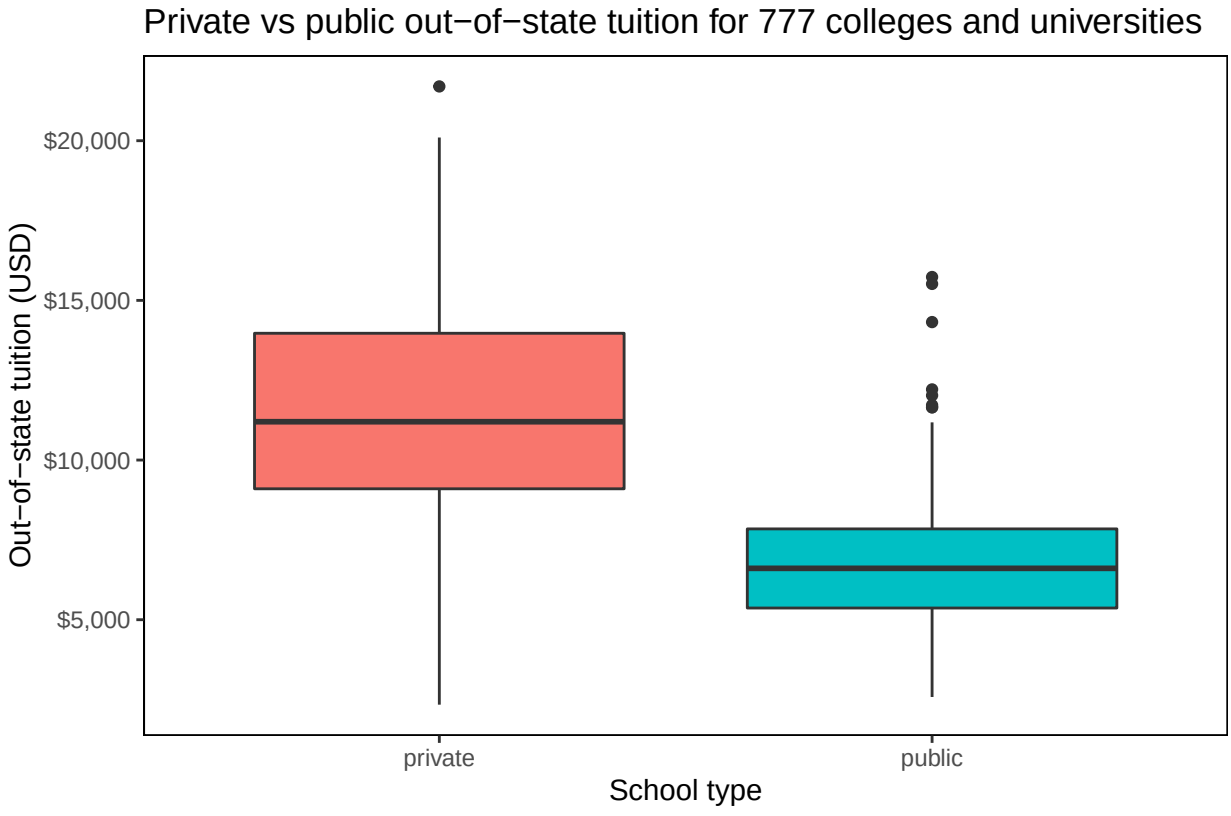
```
## Rows: 777
## Columns: 19
```

```
## $ college     <chr> "Abilene Christian University", "Adelphi University", "Adr~
## $ private     <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes~
## $ apps        <dbl> 1660, 2186, 1428, 417, 193, 587, 353, 1899, 1038, 582, 173~
## $ accept      <dbl> 1232, 1924, 1097, 349, 146, 479, 340, 1720, 839, 498, 1425~
## $ enroll      <dbl> 721, 512, 336, 137, 55, 158, 103, 489, 227, 172, 472, 484,~
## $ top10perc   <dbl> 23, 16, 22, 60, 16, 38, 17, 37, 30, 21, 37, 44, 38, 44, 23~
## $ top25perc   <dbl> 52, 29, 50, 89, 44, 62, 45, 68, 63, 44, 75, 77, 64, 73, 46~
## $ f_undergrad <dbl> 2885, 2683, 1036, 510, 249, 678, 416, 1594, 973, 799, 1830~
## $ p_undergrad <dbl> 537, 1227, 99, 63, 869, 41, 230, 32, 306, 78, 110, 44, 638~
## $ outstate    <dbl> 7440, 12280, 11250, 12960, 7560, 13500, 13290, 13868, 1559~
## $ room_board  <dbl> 3300, 6450, 3750, 5450, 4120, 3335, 5720, 4826, 4400, 3380~
## $ books       <dbl> 450, 750, 400, 450, 800, 500, 500, 450, 300, 660, 500, 400~
## $ personal    <dbl> 2200, 1500, 1165, 875, 1500, 675, 1500, 850, 500, 1800, 60~
## $ ph_d        <dbl> 70, 29, 53, 92, 76, 67, 90, 89, 79, 40, 82, 73, 60, 79, 36~
## $ terminal    <dbl> 78, 30, 66, 97, 72, 73, 93, 100, 84, 41, 88, 91, 84, 87, 6~
## $ s_f_ratio   <dbl> 18.1, 12.2, 12.9, 7.7, 11.9, 9.4, 11.5, 13.7, 11.3, 11.5, ~
## $ perc_alumni <dbl> 12, 16, 30, 37, 2, 11, 26, 37, 23, 15, 31, 41, 21, 32, 26,~
## $ expend      <dbl> 7041, 10527, 8735, 19016, 10922, 9727, 8861, 11487, 11644,~
## $ grad_rate   <dbl> 60, 56, 54, 59, 15, 55, 63, 73, 80, 52, 73, 76, 74, 68, 55~
```

c) i) I'm going to skip this step.

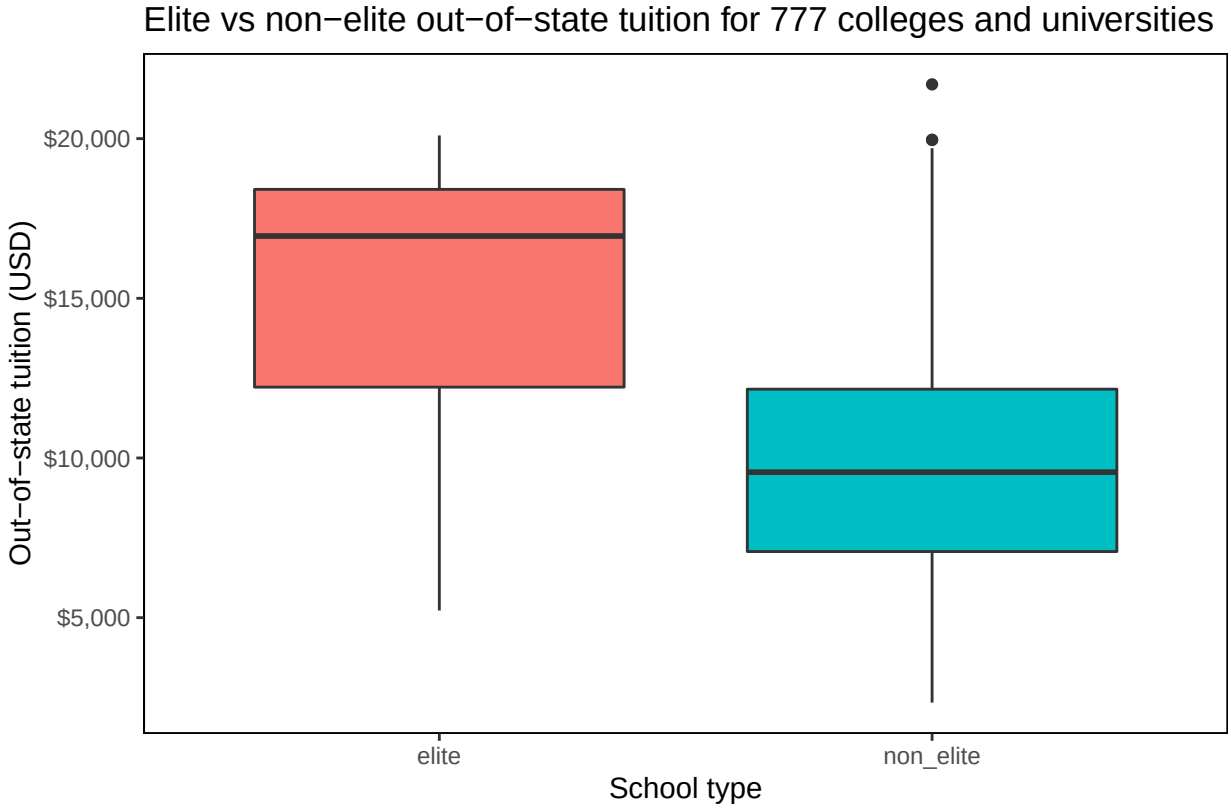
ii) I'm going to skip this question as well.

iii) Side-by-side boxplots of out-of-state tuition for private and public schools.



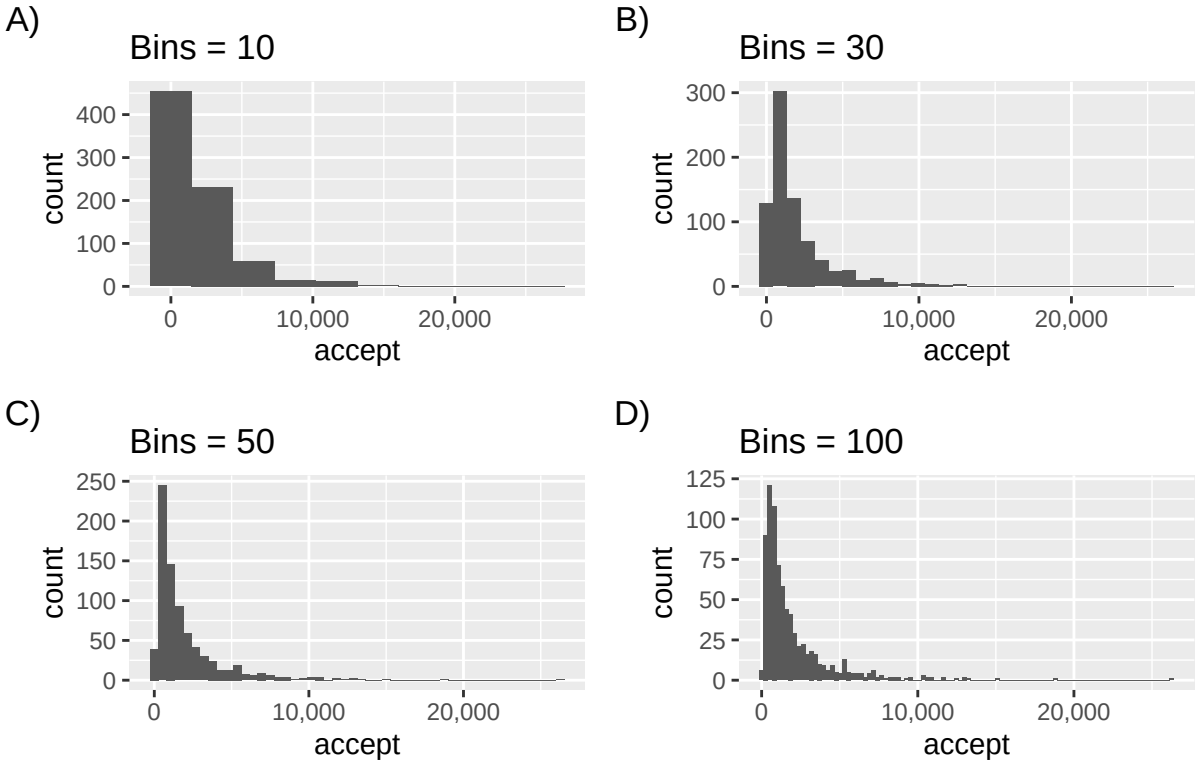
iv) Here we'll create a new variable, `elite`, labeling each school where the proportion of their incoming students coming from the top 10% of their high school classes exceeds 50% as `elite`. The remaining schools will be labeled as `non-elite`. Then, we'll create a side-by-side box plot of the `elite` vs

`non-elite` out-of-state tuition. There are 78 elite schools and 699 non-elite schools based on the criterion defined above.

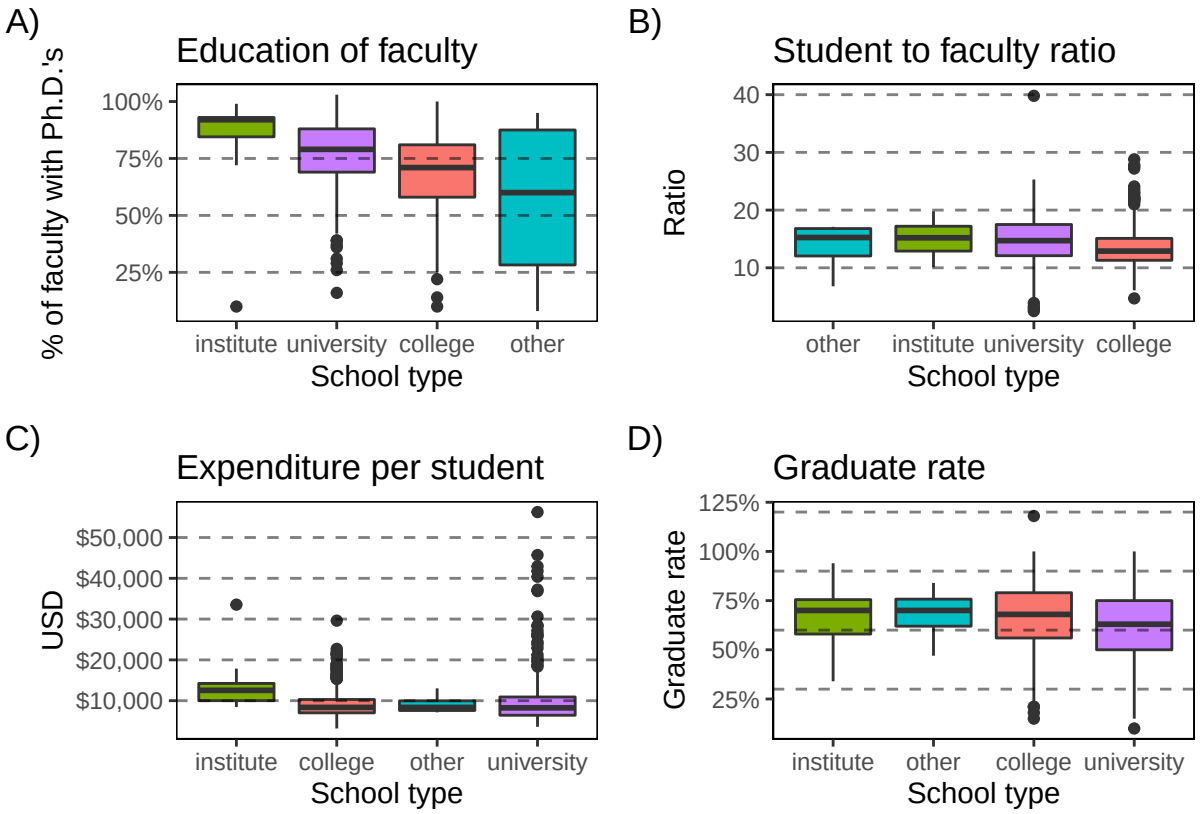


v) Here, I'll produce histograms to visualize the distribution of the number of students accepted but use different binning schemes for each plot.

Distribution of the number of accepted students



iv) Here we're asked to keep exploring the data and report a brief summary of what we find. I was interested in exploring the distribution of different quantitative variables like the percentage of faculty with Ph.D.'s, the student-to-faculty ratio, the instructional expenditure per student, and the graduation rate in colleges vs universities. Note, there was some additional data wrangling required as some schools were labeled as abbreviations (e.g., SUNY for State University of New York), or weren't labeled by type (e.g., Rutgers, which is the State University of New Jersey). I further categorized schools as "Institutes" (e.g., Massachussets Institute of Technology) and schools that could not be categorized as either colleges, universities, or institutes were classified as "Other."



From these exploratory plots, it was interesting to note that the institutes have the highest median percentage of faculty with Ph.D.'s, and while the distribution was similar for universities and colleges, there was quite a bit of wide distribution with the "other" category. This category only includes four schools: the Center for Creative Studies (a private art school in Detroit, MI), Milwaukee School of Engineering, The Citadel (also known as The Military College of South Carolina), and Virginia Tech. It seems plausible that most of the faculty would have Ph.D.'s and each of these schools except the Center for Creative Studies. Indeed, only 8% of faculty at the latter school have Ph.D.'s.

When we look at the student to faculty ratio, it's quite interesting that across the board, the median value is around the same, but there are some obvious outliers in many different colleges and one university. Incredibly, one school, the University of Charleston, has a reported student to faculty ratio of just 2.5. On the other end of the spectrum, Indiana Wesleyan University boasts a reported student to faculty ratio of almost 40.

Comparing the expenditure per student across school types, we can see that in general, institutes spend the highest median amount per student. According to the reported data, Johns Hopkins University spends the most on each student, a whopping $56,233! This doesn't seem suprising, given that Johns Hopkins also has a large medical school, which may be disproportionately influencing these data.

Finally, when we compare graduate rates, a very similar story: most have very similar median graduation rates, although there is one school, Cazenovia College, that reported a 118% graduation rate. This seems like a recording mistake, although perhaps there is more to the story that we don't know from the description of the data alone. Interestingly at the other end of the spectrum, Texas Southern University only reported a 10% graduation rate.

9) The next set of questions will use the `Auto` data set from the ISLR2 package. Note, I've removed the records with missing values before proceeding with the analyses.

```
## Rows: 392
## Columns: 9
## $ mpg          <dbl> 18, 15, 18, 16, 17, 15, 14, 14, 14, 15, 15, 14, 15, 14, 2~
```

variable	min	max
acceleration	8	24.8
cylinders	3	8.0
displacement	68	455.0
horsepower	46	230.0
mpg	9	46.6
weight	1613	5140.0

variable	mean	sd
acceleration	16	3
cylinders	5	2
displacement	194	105
horsepower	104	38
mpg	23	8
weight	2978	849

```
## $ cylinders    <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 4, 6, 6, 6, 4, ~
## $ displacement <dbl> 307, 350, 318, 304, 302, 429, 454, 440, 455, 390, 383, 34~
## $ horsepower   <int> 130, 165, 150, 150, 140, 198, 220, 215, 225, 190, 170, 16~
## $ weight       <int> 3504, 3693, 3436, 3433, 3449, 4341, 4354, 4312, 4425, 385~
## $ acceleration <dbl> 12.0, 11.5, 11.0, 12.0, 10.5, 10.0, 9.0, 8.5, 10.0, 8.5, ~
## $ year         <int> 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 7~
## $ origin       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, ~
## $ name         <fct> chevrolet chevelle malibu, buick skylark 320, plymouth sa~
```

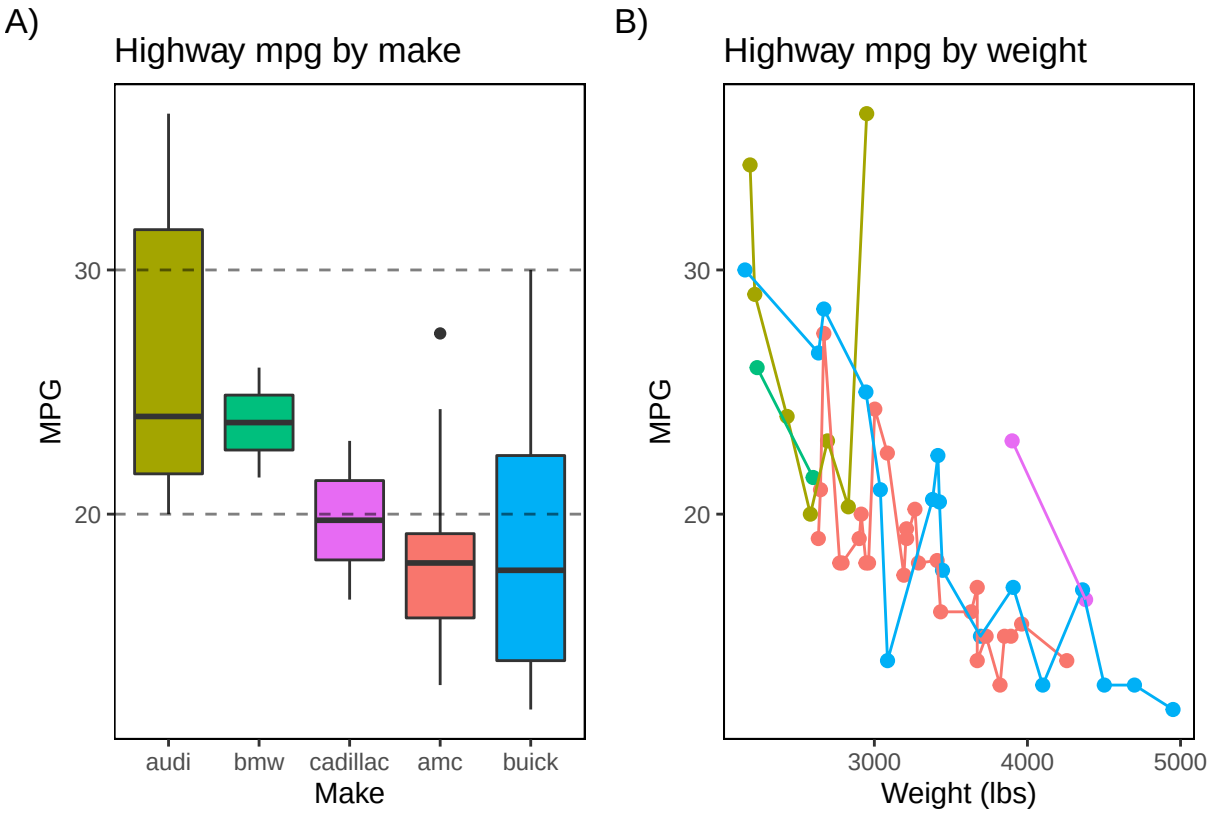
```
a) Using the `glimpse()` function, we can see that the following variables are quantitative: mpg, cylin
```

```
b) The following table displays the range of each quantitative predictor using a custom function which
```

c) We can use a similar approah as we did above to calculate the mean and standard deviation of each quantitative variable.

d) I'm going to skip this question.

e) This question asks us to investigate the predictors graphically using different methods.



f) Based on the plots above and the other variables in the data set, it is plausible that many variables would be useful in predicting highway fuel economy. From plot A) above, it is clear that not all car models are alike in the distribution of highway fuel economy, and this categorical variable, along with other quantitative variables like weight, displacement, number of cylinders in the engine, etc., should be used in building a statistical model.