CS 360/530: Foundations of Machine Learning

Spring 2020

Clint P. George

January 8, 2020

Course overview

This course will give a broad introduction to foundations of machine learning and statistical learning.

Prerequisites. Familiarity with the following is required.

- Basic computer programming—we use R/Python.
- Probability theory (CS 215, MA 605)
- Multivariable calculus and linear algebra (MA 105, MA 106, EE 611)

If you do not have the necessary background, please meet me before registering this course.

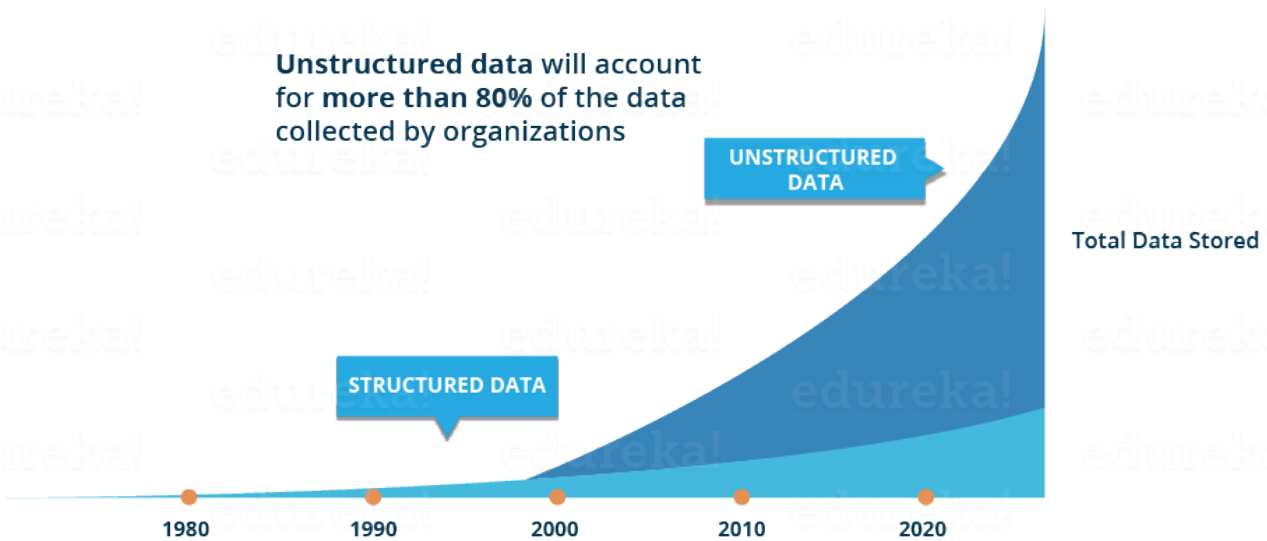
Course page and other resources

- `www.iitgoa.ac.in/~clint/courses/ml-spring-2020.html`
- We use Google Classroom
- Readings will be assigned for every lecture.

Data science

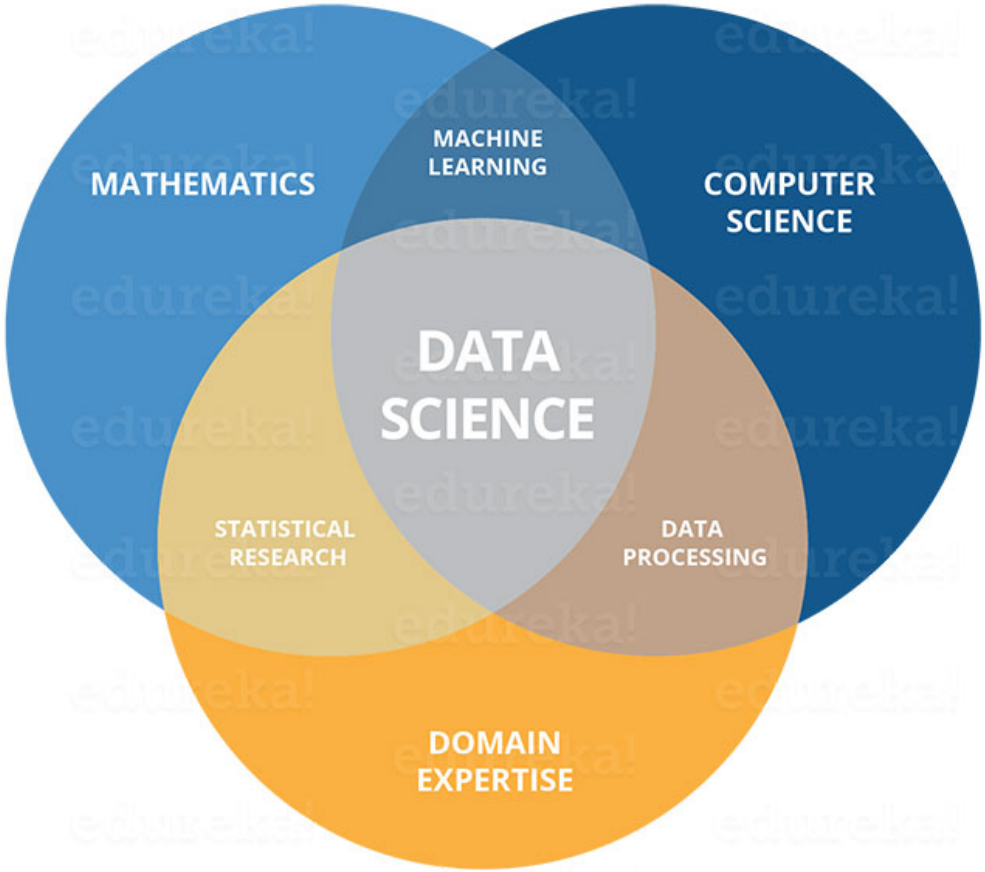
Data science—a field of study that aims to use a scientific approach to extract meaning and insights from data

Why do we need data science?—edureka.co



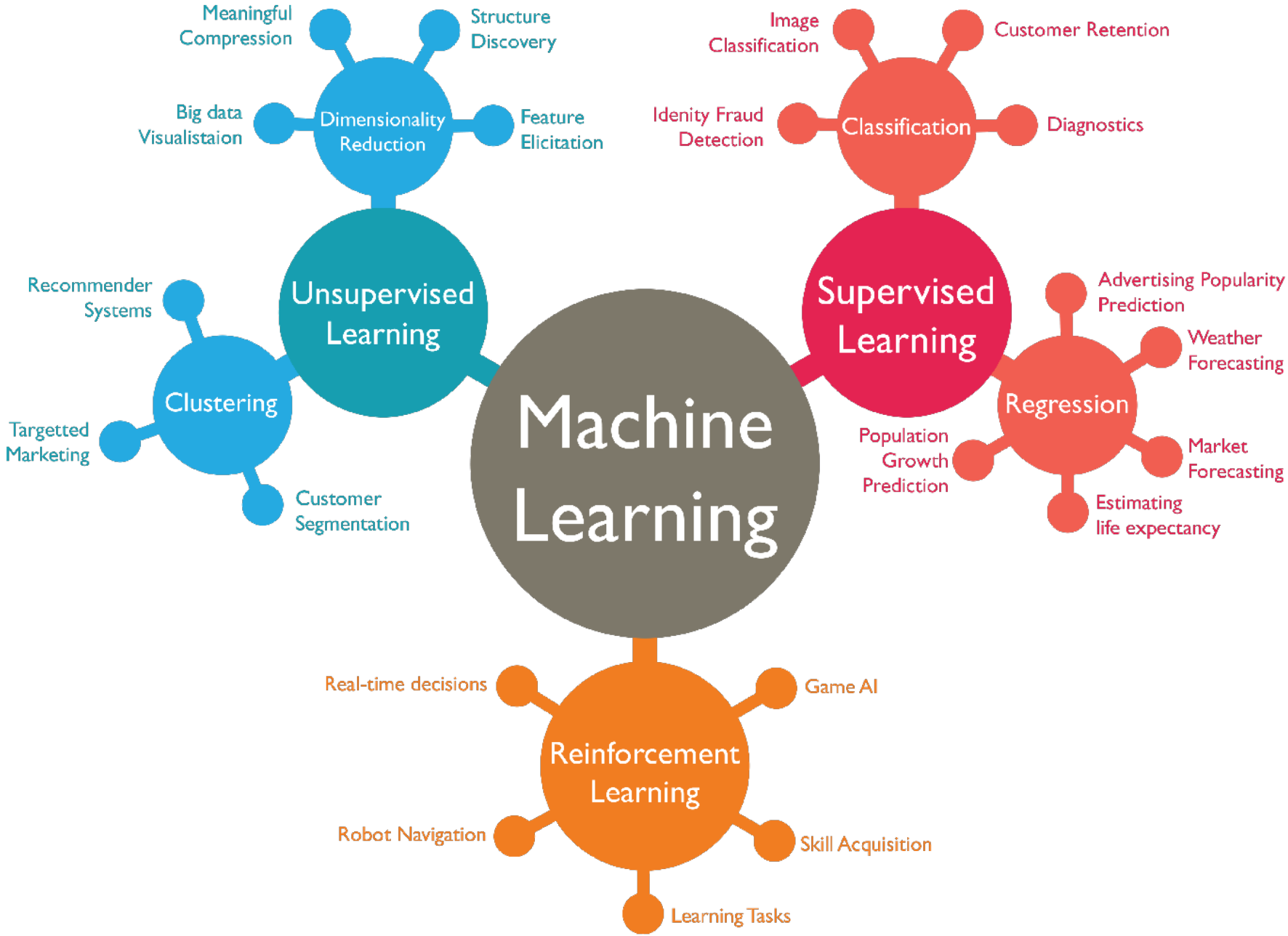
- sources: financial logs, text files, multimedia forms, sensors, and instruments
- Simple BI tools are incapable of processing the data

Skills required for data science



(edureka.co)

- Data science: focus is on the **data**
- Machine learning: focus is on **learning methods**


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Data science — outline

We learn how to load data into **R** (or Python), get it into the most useful structure, transform it, visualize it, and model it.

- Data visualization, manipulation, and exploratory study
- Expected value, variance, the Central Limit Theorem
- Hypothesis Testing—a method that is used in making statistical decisions using experimental data. E.g. A/B testing
- High-dimensional space and problems, Dimensionality reduction—e.g. Principal Component Analysis (PCA)

Machine learning — outline

Supervised learning

- Linear regression, logistic regression, Perceptron (review)
- Generative learning algorithms, Gaussian discriminant analysis
- Maximum likelihood estimation (MLE)
- Support Vector Machines (SVMs)

Machine learning in practice

- Bias–Variance tradeoff and error analysis
- Regularization and model selection
- Experimental evaluation of learning algorithms, cross-validation
- Learning Theory, Generalization errors + model selection, VC dimension — if time permits.

Machine learning — outline

Unsupervised learning

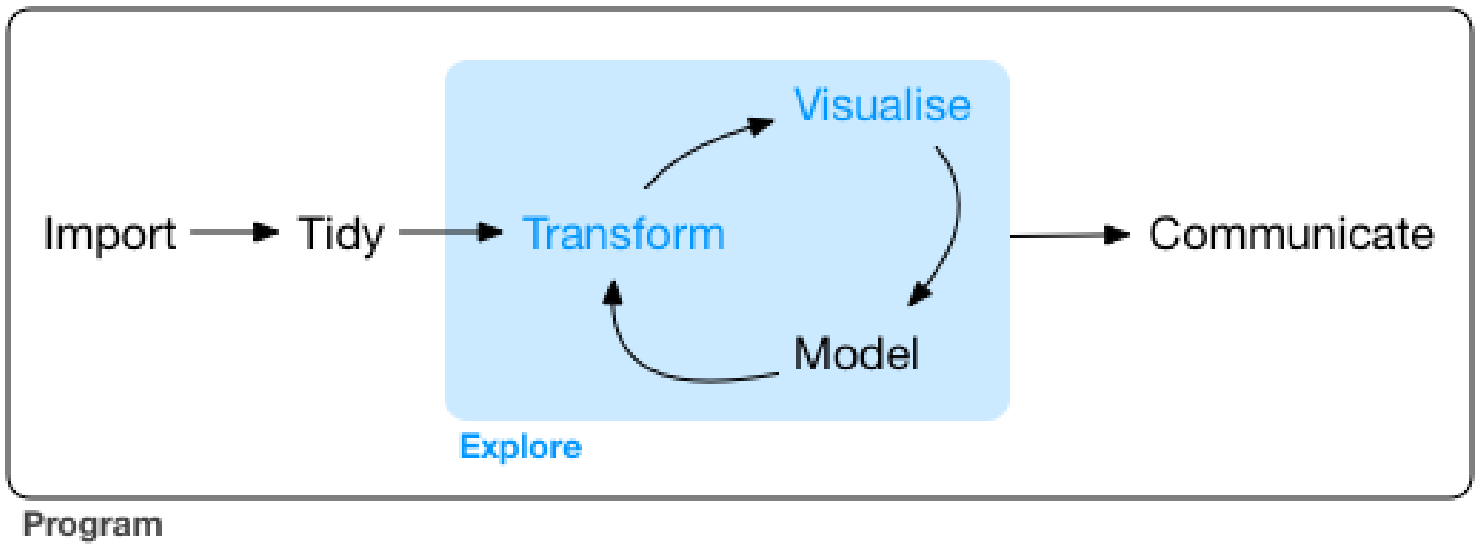
- Mixture models and mixture of Gaussians
- The expectation maximization (EM) algorithm
- Probabilistic topic models

Deep learning — outline

- Multilayer neural networks
- Back-propagation algorithm
- Auto-encoders

Data exploration cycle — some basics

Data exploration is a way to look at your data, rapidly generate hypotheses, quickly test them, then repeat again and again[1].



[1] Wickham

Prerequisites

- **R**—a language and environment for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques, and is highly extensible.
- Install RStudio—an IDE for R development.
- Install R package `tidyverse`

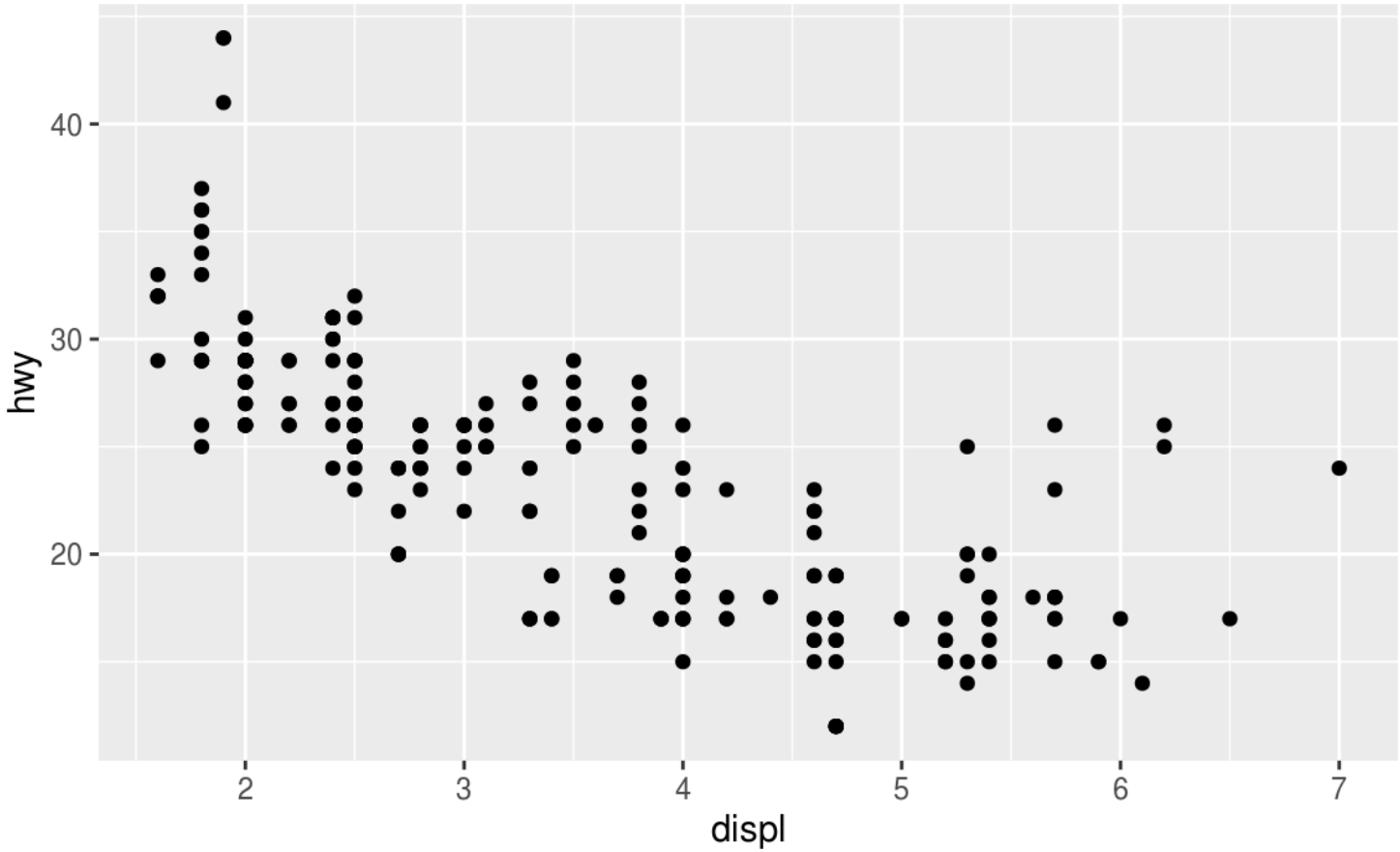
Reference: `https://r4ds.had.co.nz` (online book)
Reading: Sections 1–3

Example data and tools

- We consider the *data frame* `mpg` in package `tidyverse`. A data frame is a rectangular collection of variables (in the columns) and observations (in the rows).
- `mpg` contains observations collected by the US Environmental Protection Agency on 38 car models.
- We use `ggplot2` for visualization.

Scatterplots via `ggplot2`

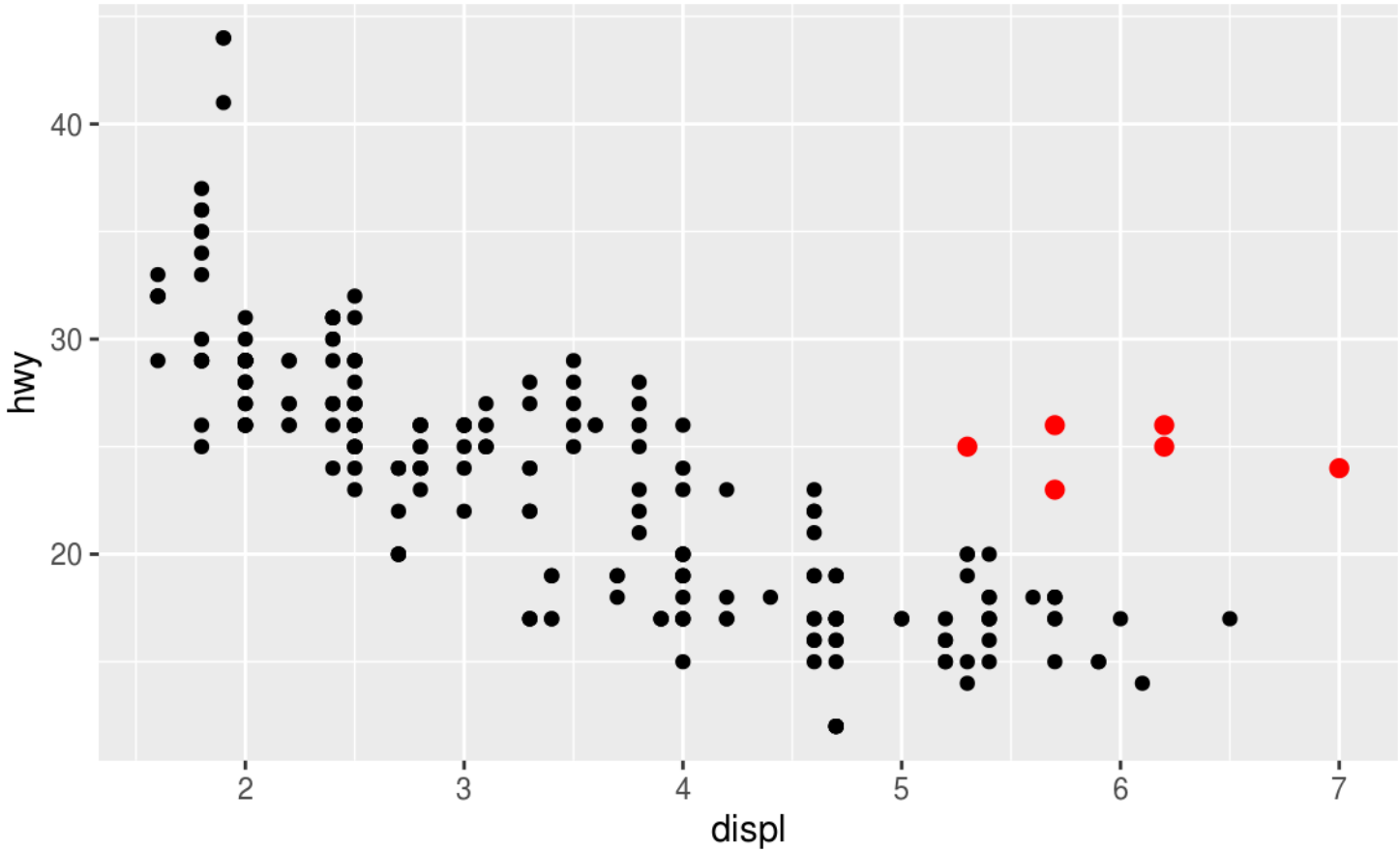
Scatterplots shows relationship between two variables or features.



The plot shows a negative relationship between engine size (displ) and fuel efficiency (highway miles per gallon)

Scatterplots — Changing aesthetics

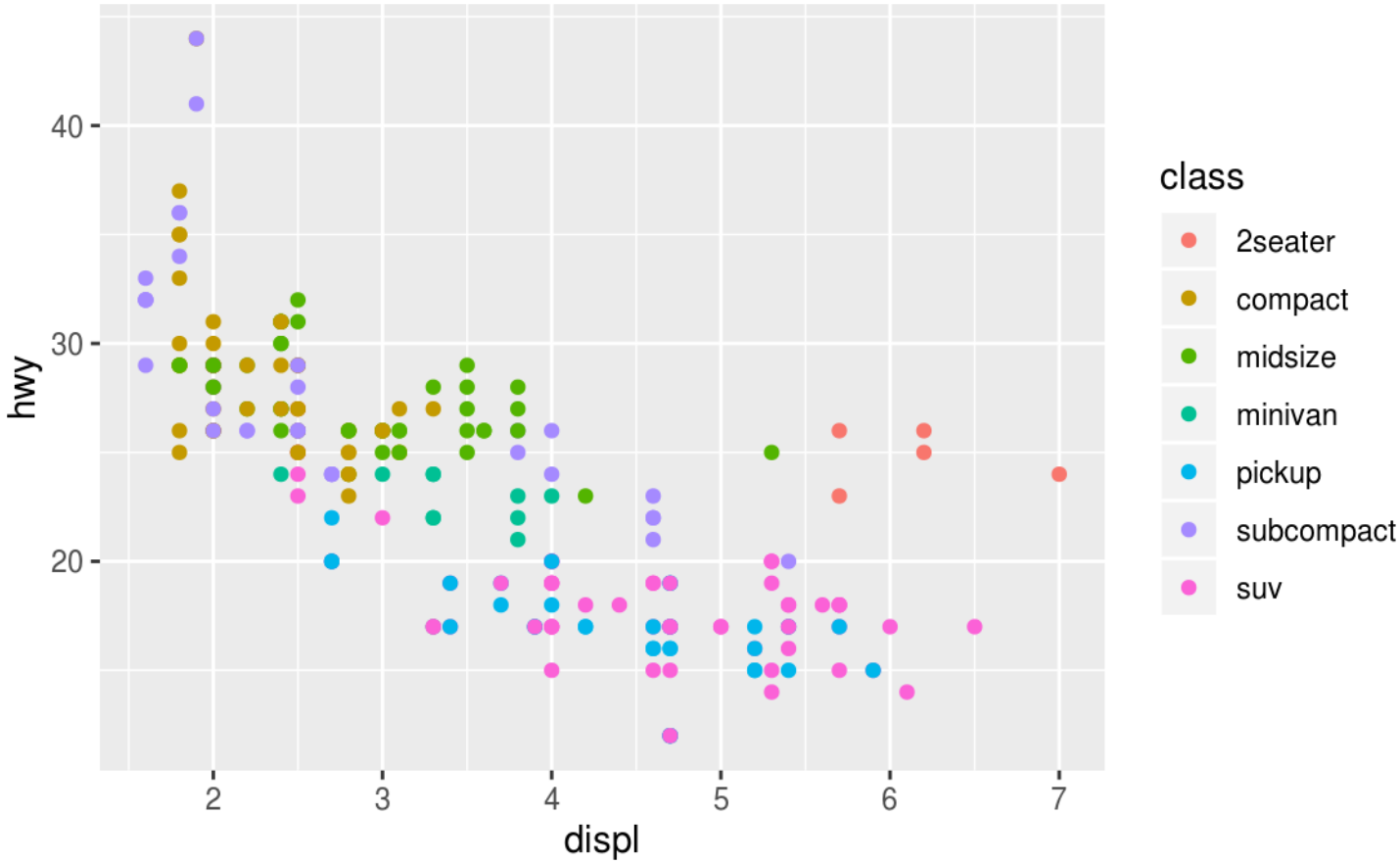
One group of points (highlighted in red) seems to fall outside of the linear trend, with higher mileage.



One can then hypothesize that the cars are hybrids.

Scatterplots — Aesthetic mappings

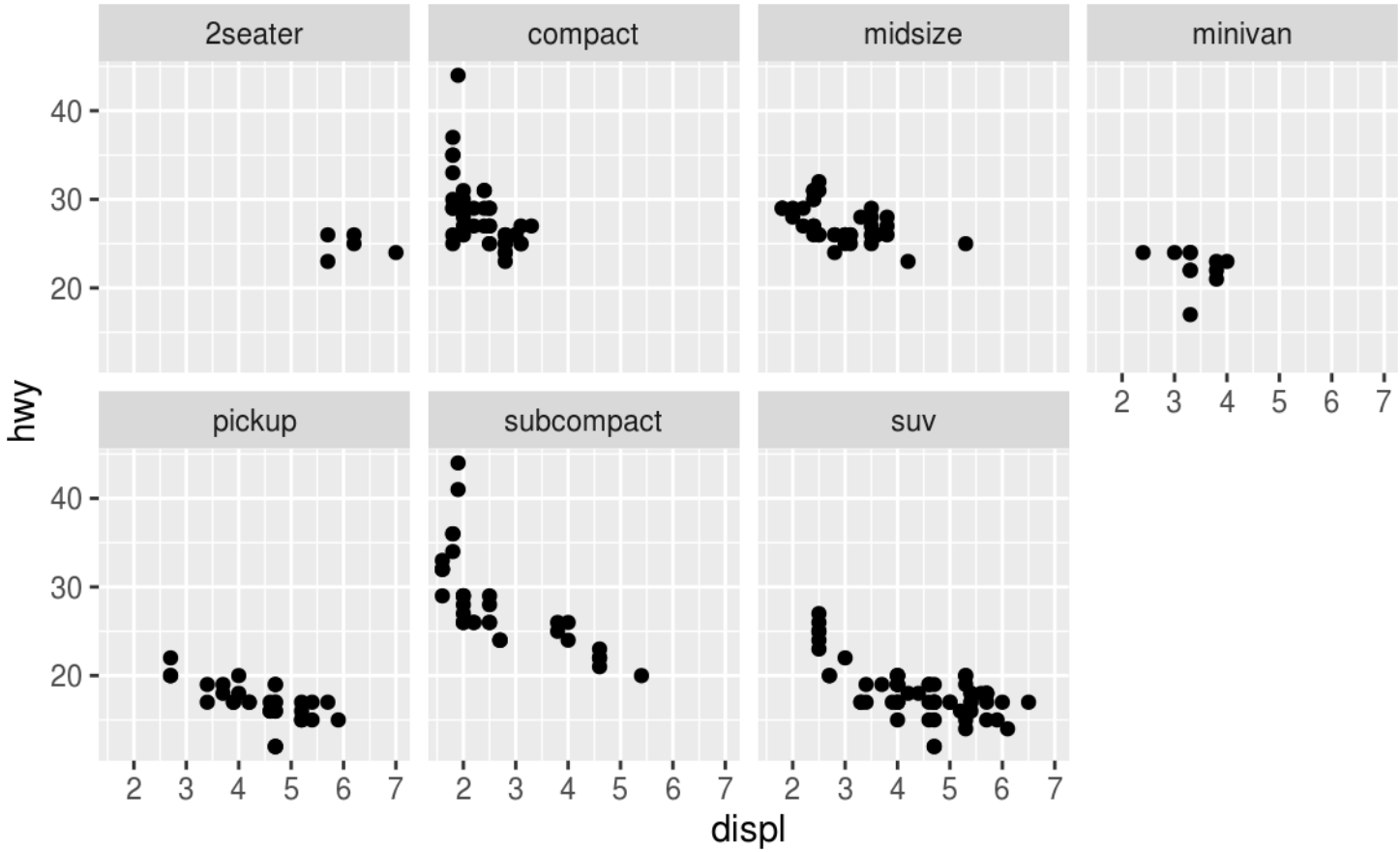
Convey information about your data by changing the aesthetics of individual variables in the data.



For example, map the colors of your points to each point's class.

Facets

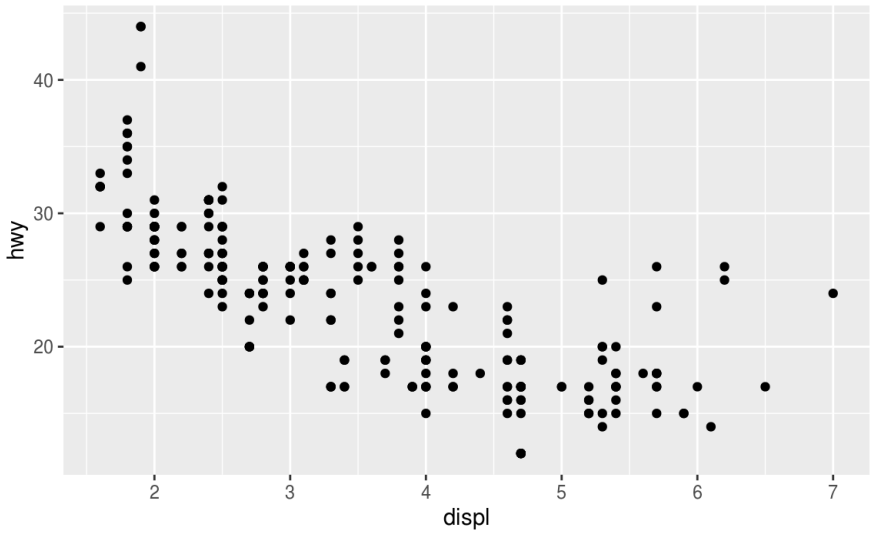
Another way to visualize more variables is to split your plot into facets—each subplot is a subset of the data.



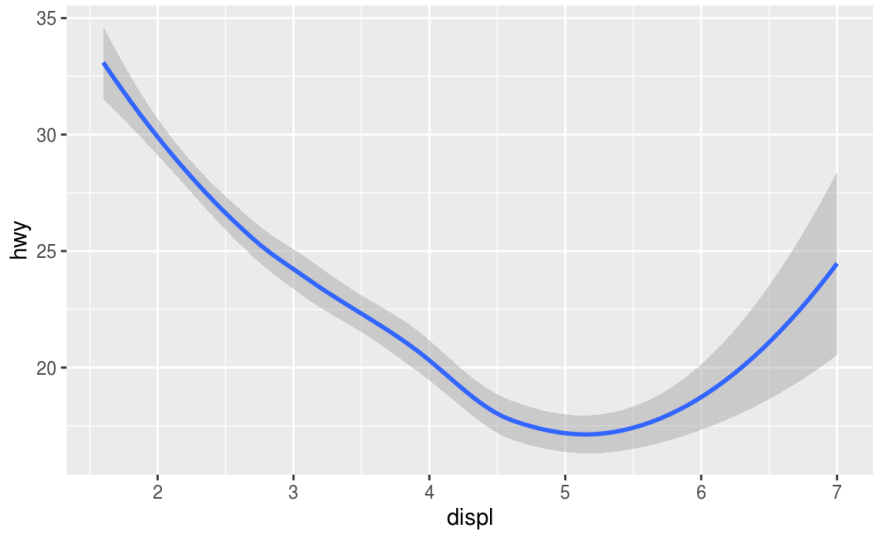
We can also have a combination of two variables.

Changing the geometric properties

Each plot uses a different visual object to represent the same data. This helps us to understand trends in the data.



(a)

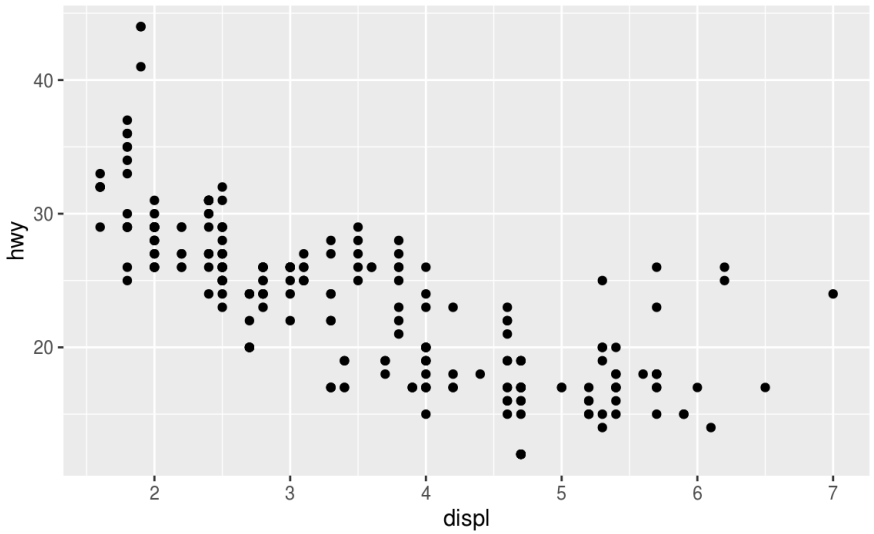


(b)

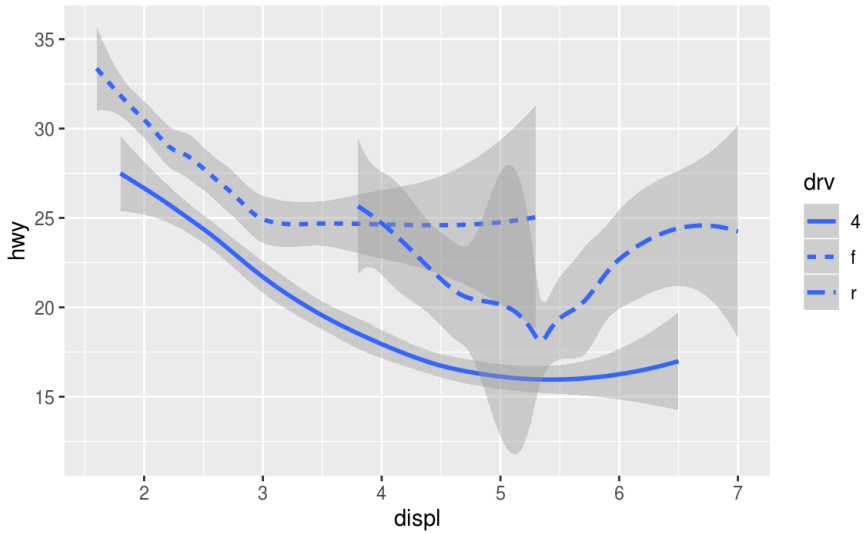
How do we get the line and confidence region?—we study that in this course.

Changing the geometric properties

Each plot uses a different visual object to represent the same data.



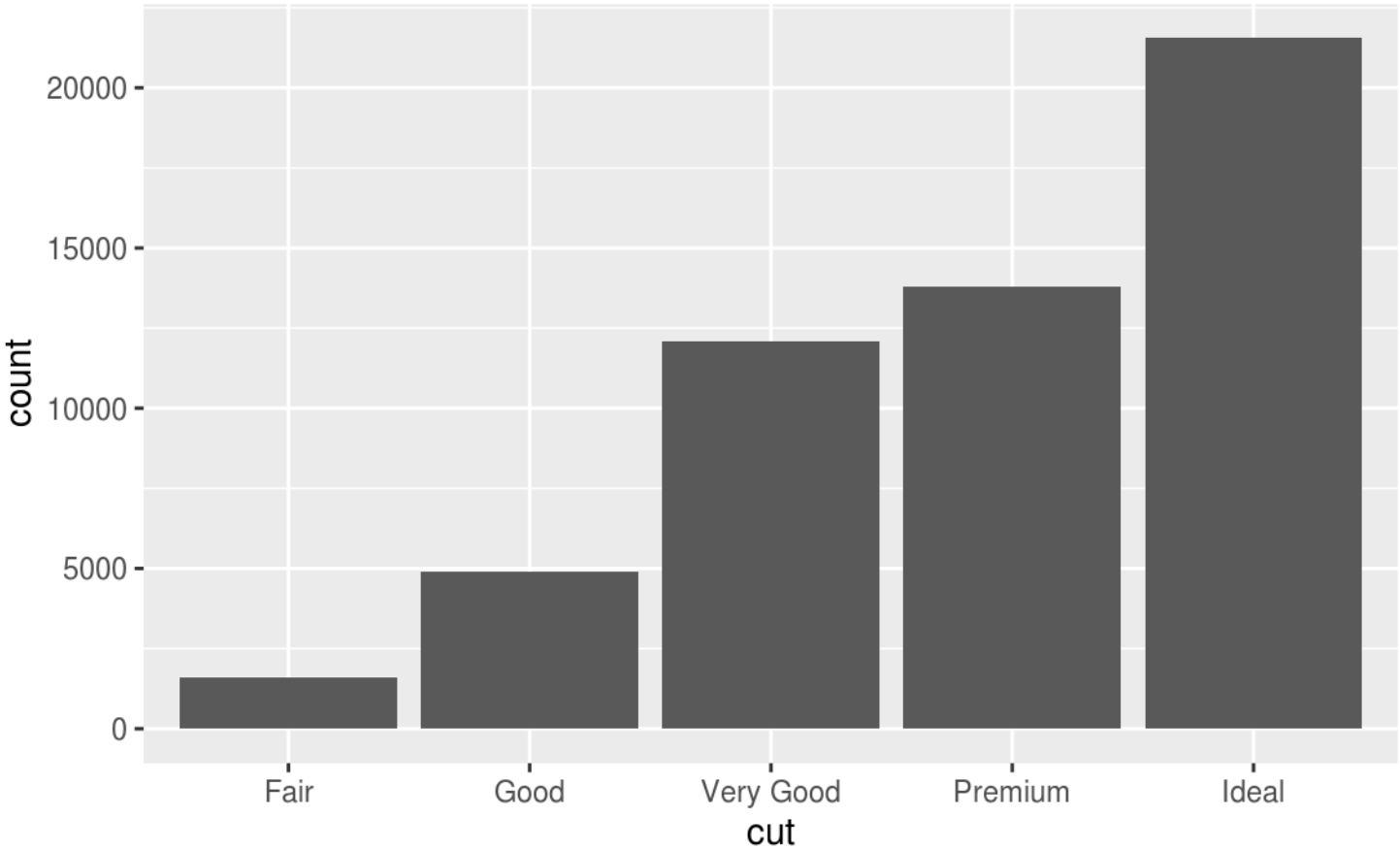
(a)



(b) linetype based on `drv`

Visualizing categorical data

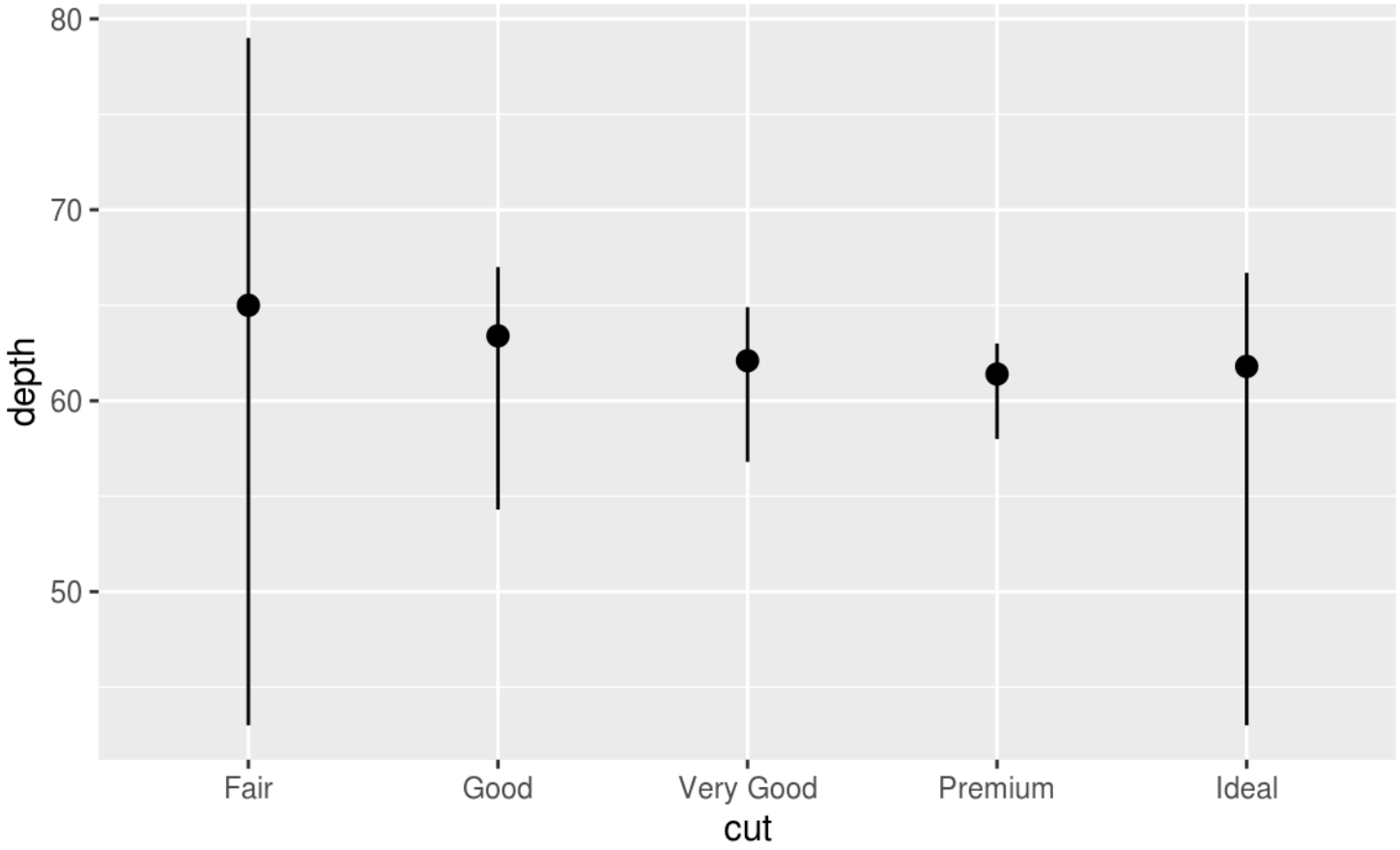
Visualizes the number of diamonds in the diamonds dataset—total 54,000 diamonds, including the price, carat, color, clarity, and cut of each diamond—grouped by cut.



More diamonds with high quality cuts than with low quality cuts.

Visualizing categorical data

Summarises depth values for each unique x value (e.g. cut) in the data, using a different stat object.



Data transformations

- Often you won't get the data in exactly the right form you need.
- To make the data easy to work, we
  - create some new variables or summaries
  - rename the variables or reorder the observations, etc.

Data transformations

- Often you won't get the data in exactly the right form you need.
- To make the data easy to work, we
  - create some new variables or summaries
  - rename the variables or reorder the observations, etc.

We illustrate data transformations using the data from the `nycflights13` package.

This data (frame) contains all 336,776 flights that departed from New York City in 2013.—from the US Bureau of Transportation Statistics.

Data manipulations using R — `dplyr`

- Subset observations by their values — `filter()`.
- Reorder the rows — `arrange()`.
- Pick variables by their names — `select()`.
- Create new variables with functions of existing variables — `mutate()`.
- Collapse many values down to a single summary — `summarise()`.

We can use the functions in conjunction with `group_by()`, which changes the scope of operation.

Data manipulations using R

All functions works similar

- The first argument is a data frame.
- The subsequent arguments describe what to do with the data frame, using the variable names
- The result is a new data frame.

`filter()`

`filter()` allows you to subset observations based on their values. The first argument is the name of the data frame. The second and subsequent arguments are the expressions that filter the data frame.

Examples:

- `filter(flights, month == 1, day == 1)`
- `filter(flights, month %in% c(11, 12))`

Includes rows where the condition is TRUE; excludes both FALSE and NA values.

arrange()

It takes a data frame and a set of column names to order by. If there is more than one column name, each additional column will be used to break ties in the values of preceding columns

Examples:

- `arrange(flights, year, month, day)`
- `arrange(flights, desc(dep_delay))`

NAs are sorted at the end.

`select()`

`select()` allows you to subset variables (features) using operations based on the names of the variables.

Examples:

- `select(flights, year, month, day)`
- `select(flights, year:day)`
- `select(flights, -(year:day))`

mutate()

Adds new columns that are functions of existing columns

```
mutate(flights_sml,
  gain = dep_delay - arr_delay,
  speed = distance / air_time * 60
)
```

summarise()

It collapses a data frame to a single row.

```
summarise(flights, delay = mean(dep_delay, na.rm = TRUE))
```

summarise() pairs well with group_by(), which helps us to apply summary operations on a grouped data.

```
by_day <- group_by(flights, year, month, day)
summarise(by_day, delay = mean(dep_delay, na.rm = TRUE))
```

Exploratory data analysis (EDA)

Exploration via visualizing and transforming data systematically:

1. Compose creative questions about your data
2. Search for answers by visualizing, transforming, and modeling your data
3. Based on the findings refine the questions and/or generate new questions.

Data cleaning is just one (important) application of EDA.

Exploratory data analysis (EDA)

A few questions that will always be useful for making discoveries within your data:

- What type of variation occurs within variables?
- What type of covariation occurs between variables?

Average Value (review)

When we have a large collection of numbers, as in a census, we are usually interested not in the individual numbers, but rather in certain descriptive quantities such as the average.

Average Value (review)

When we have a large collection of numbers, as in a census, we are usually interested not in the individual numbers, but rather in certain descriptive quantities such as the average.

Experiment: A die is rolled. If an odd number turns up, we win an amount equal to this number; if an even number turns up, we lose an amount equal to this number. Question: Is this a reasonable game to play?

Average Value (review)

When we have a large collection of numbers, as in a census, we are usually interested not in the individual numbers, but rather in certain descriptive quantities such as the average.

Experiment: A die is rolled. If an odd number turns up, we win an amount equal to this number; if an even number turns up, we lose an amount equal to this number. Question: Is this a reasonable game to play?—We will figure out via simulation.

Winning	n = 100		n = 10000	
	Frequency	Relative Frequency	Frequency	Relative Frequency
1	17	.17	1681	.1681
-2	17	.17	1678	.1678
3	16	.16	1626	.1626
-4	18	.18	1696	.1696
5	16	.16	1686	.1686
-6	16	.16	1633	.1633

Expected value (review)

Let X be a numerically-valued discrete random variable with sample space Ω and distribution function $m(x)$. The expected value $E(X)$ is defined by

$$E(X) = \sum_{x \in \Omega} x\, m(x)$$

provided the sum converges absolutely.

If the above sum does not converge absolutely, then no expected value exists.

Expected value—Example 1

Suppose we toss a fair coin until a head first comes up, and let X represent the number of tosses which were made.

Expected value—Example 1

Suppose we toss a fair coin until a head first comes up, and let X represent the number of tosses which were made. Then the possible values of X are $1, 2, \ldots$, and the distribution function of X

$$m(i) = \frac{1}{2^i}$$

Expected value—Example 1

Suppose we toss a fair coin until a head first comes up, and let X represent the number of tosses which were made. Then the possible values of X are $1, 2, \ldots$, and the distribution function of X

$$m(i) = \frac{1}{2^i}$$

We then have the expected value

$$E(X) = \sum_{i=1}^{\infty} i \frac{1}{2^i}$$

Expected value—Example 2

Suppose we toss a fair coin until a head first comes up, and if the number of tosses equals i, then we are paid 2^i dollars.

Expected value—Example 2

Suppose we toss a fair coin until a head first comes up, and if the number of tosses equals i, then we are paid 2^i dollars. Let X represents the payment, and the distribution function of X

$$m(i) = \frac{1}{2^i}$$

Expected value—Example 2

Suppose we toss a fair coin until a head first comes up, and if the number of tosses equals i, then we are paid 2^i dollars. Let X represents the payment, and the distribution function of X

$$m(i) = \frac{1}{2^i}$$

We write the expected value

$$E(X) = \sum_{i=1}^{\infty} 2^i \frac{1}{2^i}$$

Expected value—Example 2

Suppose we toss a fair coin until a head first comes up, and if the number of tosses equals i, then we are paid 2^i dollars. Let X represents the payment, and the distribution function of X

$$m(i) = \frac{1}{2^i}$$

We write the expected value

$$E(X) = \sum_{i=1}^{\infty} 2^i \frac{1}{2^i}$$

It is a divergent sum!

Expected value—Example 2

Suppose we toss a fair coin until a head first comes up, and if the number of tosses equals i, then we are paid 2^i dollars. Let X represents the payment, and the distribution function of X

$$m(i) = \frac{1}{2^i}$$

We write the expected value

$$E(X) = \sum_{i=1}^{\infty} 2^i \frac{1}{2^i}$$

It is a divergent sum!—This experiment is called St. Petersburg Paradox

Expected value—Some useful concepts

Function of a Random Variable: If X is a discrete random variable with sample space Ω and distribution function $m(x)$, and if $\phi : \Omega \to R$ is a function, then

$$E(\phi(X)) = \sum_{x \in \Omega} \phi(x) m(x),$$

provided the series converges.

Expected value—Some useful concepts

Function of a Random Variable: If X is a discrete random variable with sample space Ω and distribution function $m(x)$, and if $\phi : \Omega \to R$ is a function, then

$$E(\phi(X)) = \sum_{x \in \Omega} \phi(x) m(x),$$

provided the series converges.

Sum of Two Random Variables: Let X and Y be random variables with finite expected values. Then

$$E(X + Y) = E(X) + E(Y)$$

Expected value—Some useful concepts

Function of a Random Variable: If X is a discrete random variable with sample space Ω and distribution function $m(x)$, and if $\phi : \Omega \to R$ is a function, then

$$E(\phi(X)) = \sum_{x \in \Omega} \phi(x) m(x),$$

provided the series converges.

Sum of Two Random Variables: Let X and Y be random variables with finite expected values. Then

$$E(X + Y) = E(X) + E(Y)$$

and for a constant c, we have

$$E(cX) = cE(X)$$

Expected value—Some useful concepts

Function of a Random Variable: If X is a discrete random variable with sample space Ω and distribution function $m(x)$, and if $\phi : \Omega \to R$ is a function, then

$$E(\phi(X)) = \sum_{x \in \Omega} \phi(x) m(x),$$

provided the series converges.

Sum of Two Random Variables: Let X and Y be random variables with finite expected values. Then

$$E(X + Y) = E(X) + E(Y)$$

and for a constant c, we have

$$E(cX) = cE(X)$$

Homework 4.1: Prove that they are true!

Expected value—Some useful concepts

Let S_n be the number of successes in n Bernoulli trials with probability p for success on each trial. Then the expected number of successes is np.

Homework 4.2: Prove this!

Variance of Discrete Random Variables

Using the expected value as a prediction for the outcome of an experiment is not effective when the observations deviate too much from the expected value.

Variance of Discrete Random Variables

Using the expected value as a prediction for the outcome of an experiment is not effective when the observations deviate too much from the expected value.—We thus need a measure of this deviation.

Variance of Discrete Random Variables

Using the expected value as a prediction for the outcome of an experiment is not effective when the observations deviate too much from the expected value.—We thus need a measure of this deviation.

Variance: Let X be a numerically valued random variable with expected value $\mu = E(X)$. Then the variance of X is given by:

$$V(X) = E((X - \mu)^2)$$

Variance—Some useful properties

For a constant c and a random variable X, we have

$$V(cX) = c^2 V(X)$$

and

$$V(X + c) = V(X)$$

Variance—Some useful properties

For a constant c and a random variable X, we have

$$V(cX) = c^2 V(X)$$

and

$$V(X + c) = V(X)$$

Let X and Y be two *independent* random variables.

$$V(X + Y) = V(X) + V(Y)$$

Variance—Some useful properties

For a constant c and a random variable X, we have

$$V(cX) = c^2 V(X)$$

and

$$V(X + c) = V(X)$$

Let X and Y be two *independent* random variables.

$$V(X + Y) = V(X) + V(Y)$$

Expected value is a linear function, but variance is not.

Visualizing distributions

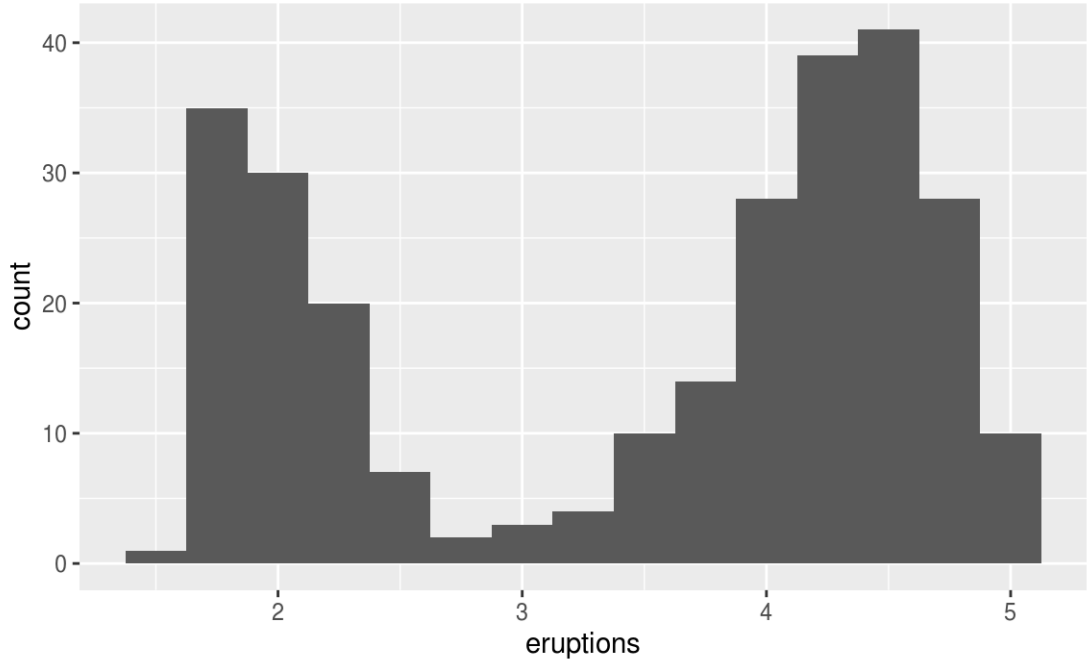
- Categorical variables — represented as counts, e.g., text data
- Continuous variables — can take any of an infinite set of ordered values, can be represented by histograms

`ggplot` tools available: `geom_bar()`, `geom_histogram()`, `geom_freqpoly()`, etc.

Inference: In both bar charts and histograms, tall bars show common values of a variable, and shorter bars show less-common values.

Visualizing distributions

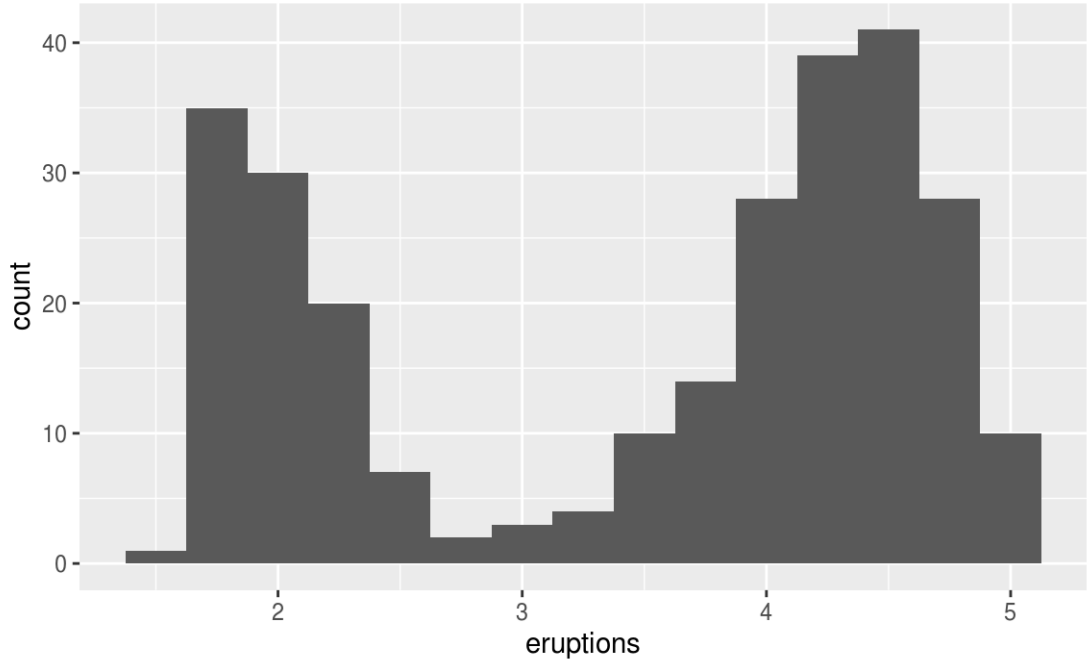
Clusters of similar values suggest that subgroups exist in a data.



Shows the length (in minutes) of 272 eruptions of the Old Faithful Geyser in the Yellowstone National Park.

Visualizing distributions

Clusters of similar values suggest that subgroups exist in a data.



Shows the length (in minutes) of 272 eruptions of the Old Faithful Geyser in the Yellowstone National Park.

Some modeling questions: How are the observations

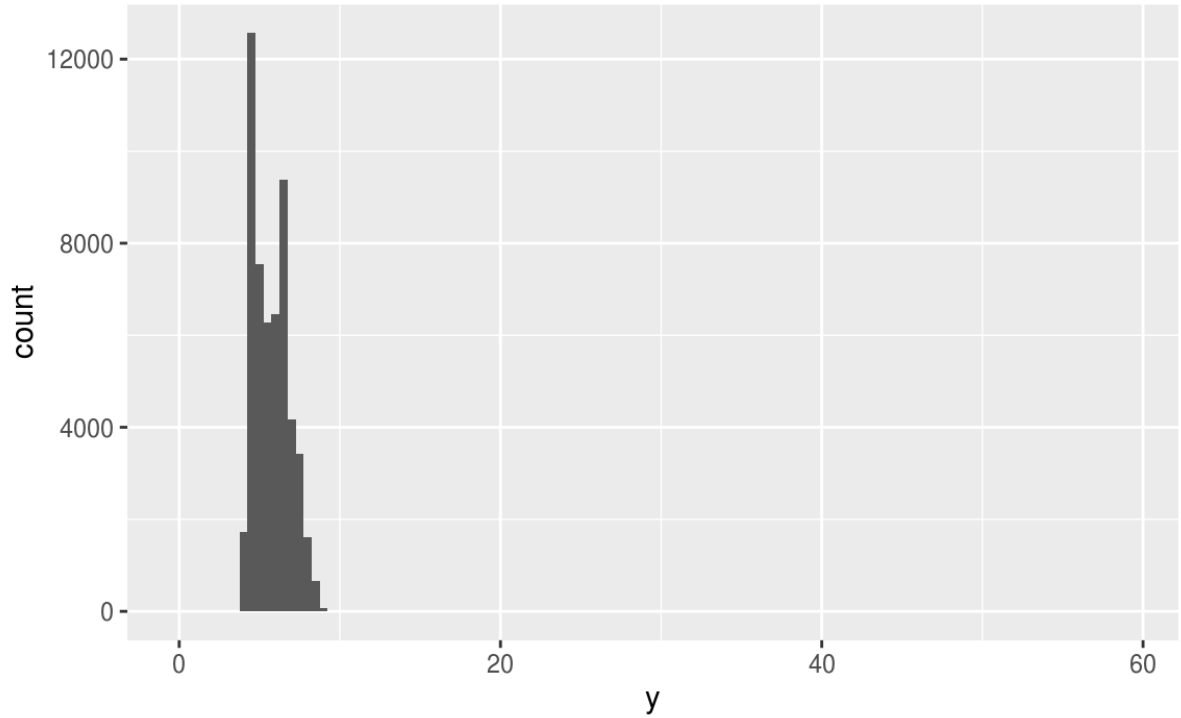
- within each cluster similar to each other?
- in separate clusters different from each other?
- describe clusters?

Detecting outliers

Outliers are observations that are unusual; data points that don't seem to fit the pattern. Sometimes outliers are data entry errors; other times outliers suggest important new science.

Detecting outliers

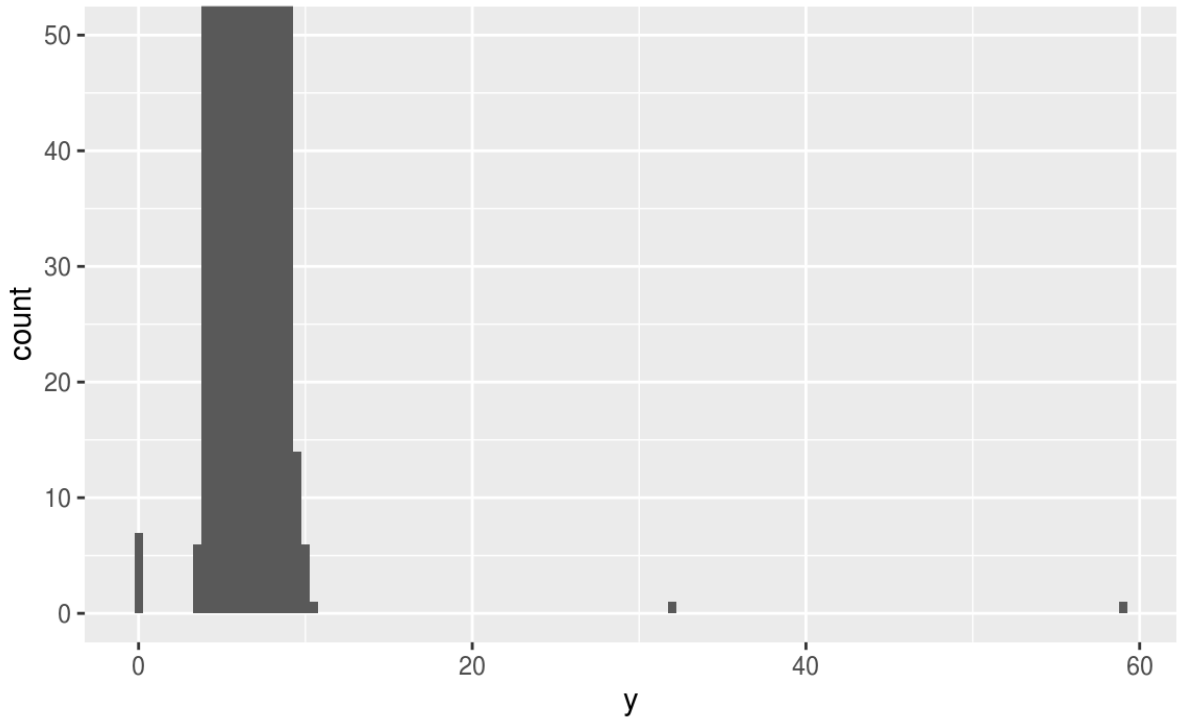
Outliers are observations that are unusual; data points that don't seem to fit the pattern. Sometimes outliers are data entry errors; other times outliers suggest important new science.



The distribution of the y variable from the diamonds dataset. The only evidence of outliers is the unusually wide limits on the x-axis.

Detecting outliers

Outliers are observations that are unusual; data points that don't seem to fit the pattern. Sometimes outliers are data entry errors; other times outliers suggest important new science.

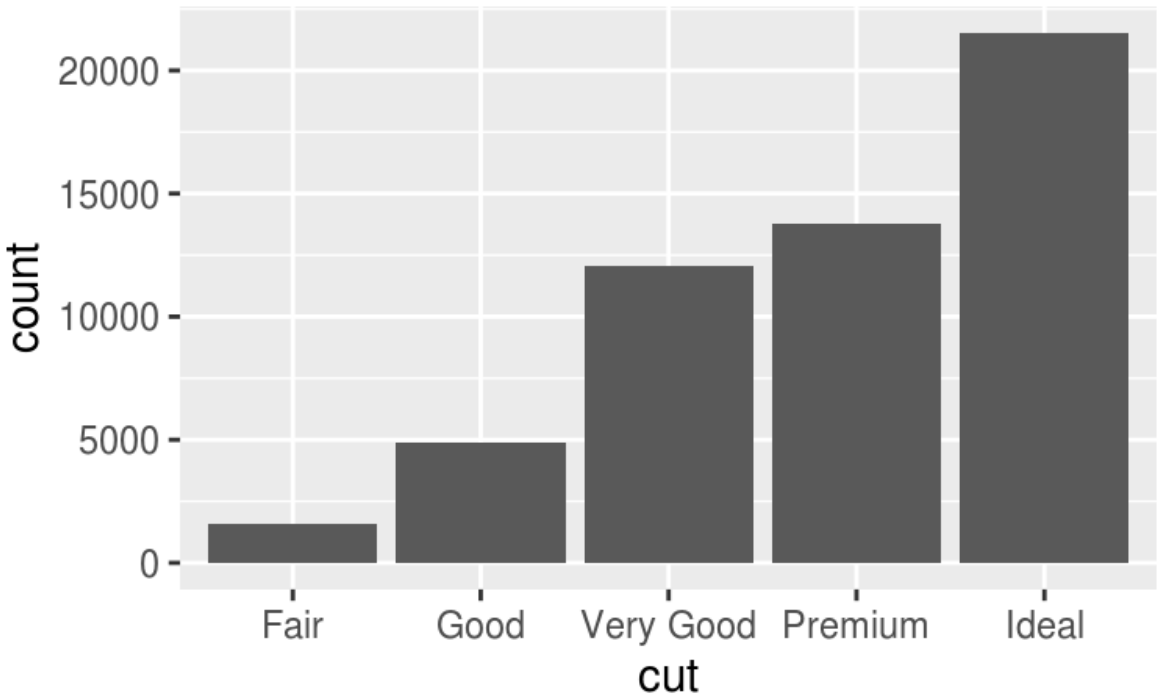


There are so many observations in the common bins that the rare bins are so short that you can't see them.—Zoom that portion!

Detecting covariation

Covariation is the tendency for the values of two or more variables to vary together in a related way.

A categorical and continuous variable—e.g. how the price of a diamond varies with its quality:

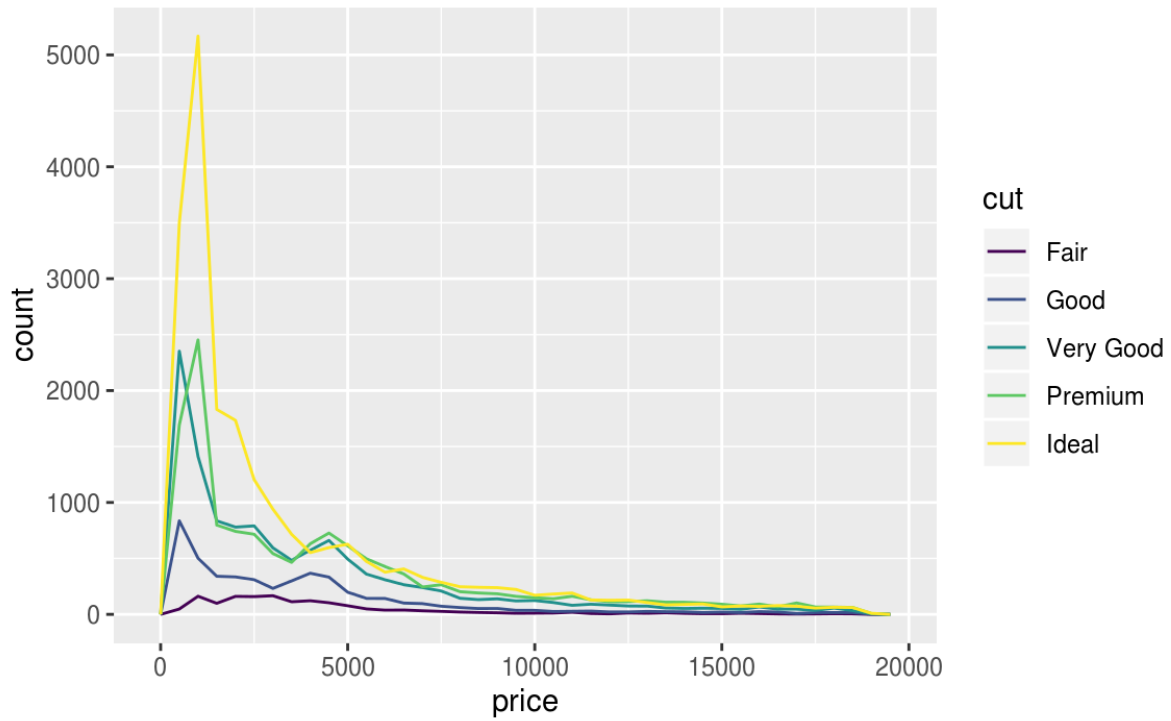


It's hard to see the difference in distribution because the overall counts differ so much.

Detecting covariation

Covariation is the tendency for the values of two or more variables to vary together in a related way.

A categorical and continuous variable—e.g. how the price of a diamond varies with its quality:

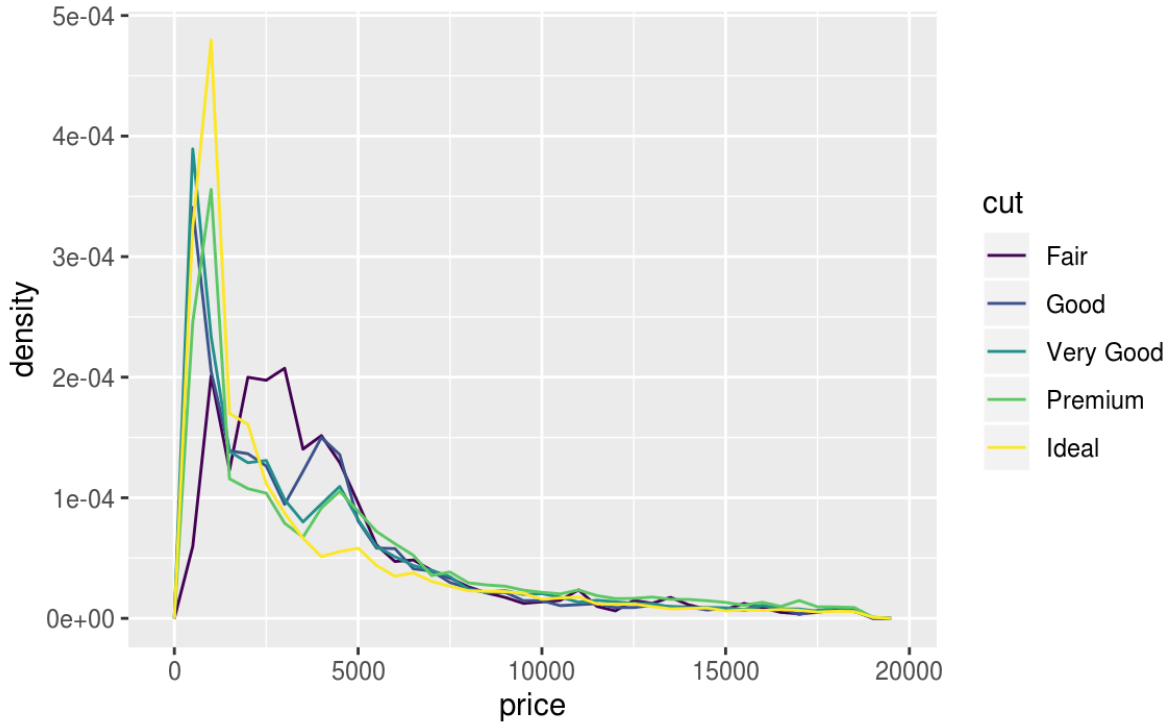


The distribution of a continuous variable broken down by a categorical variable.—not very useful due to the counts on y-axis

Detecting covariation

Covariation is the tendency for the values of two or more variables to vary together in a related way.

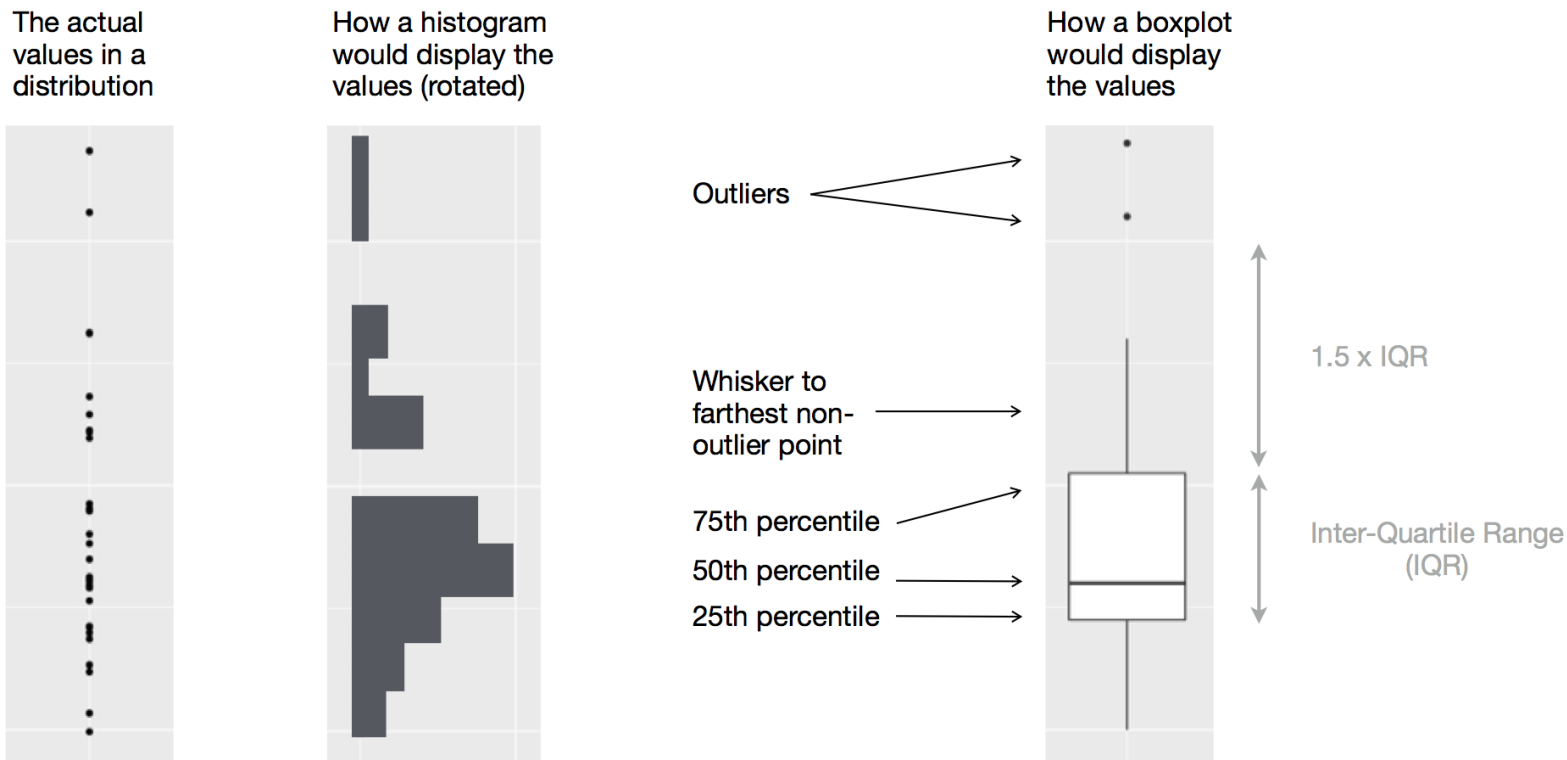
A categorical and continuous variable—e.g. how the price of a diamond varies with its quality:



We display density on y-axis, which is the count standardised.

Detecting covariation

The boxplot is an alternative to display the distribution of a continuous variable broken down by a categorical variable.

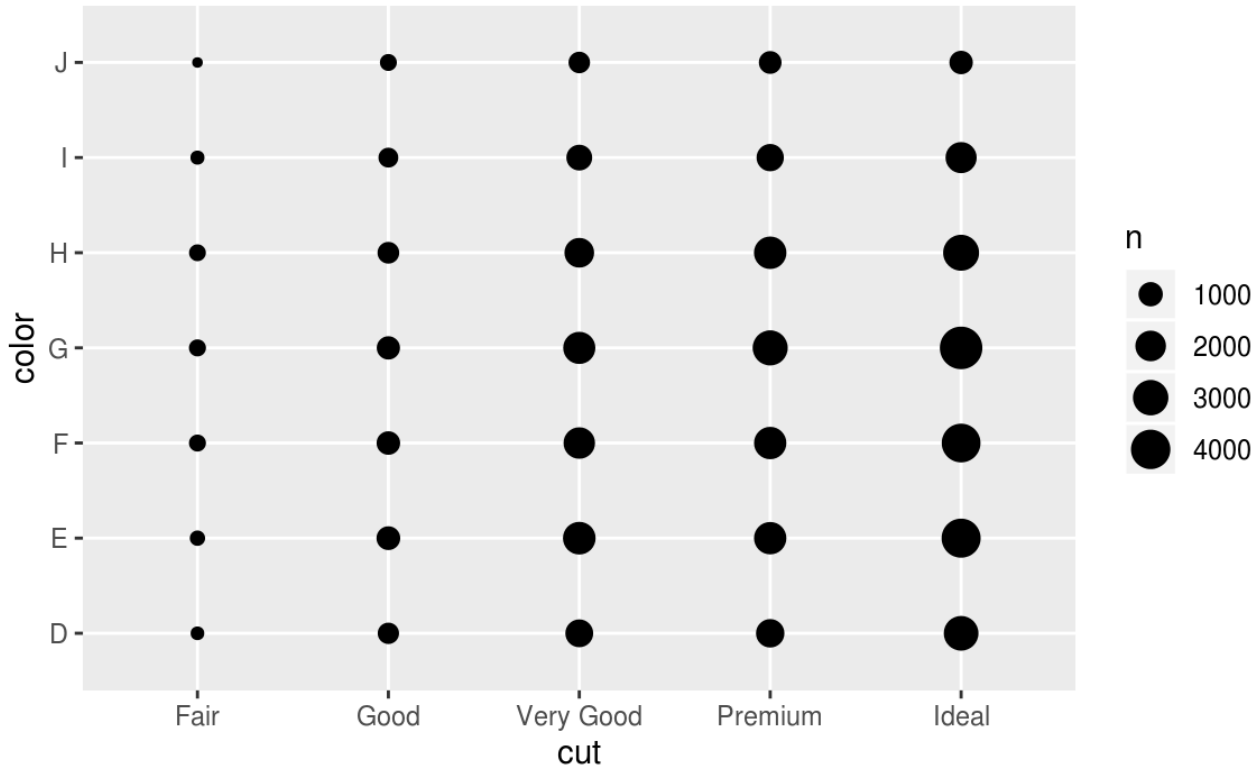


Inference: whether or not the distribution is symmetric about the median or skewed to one side.

Detecting covariation

Covariation is the tendency for the values of two or more variables to vary together in a related way.

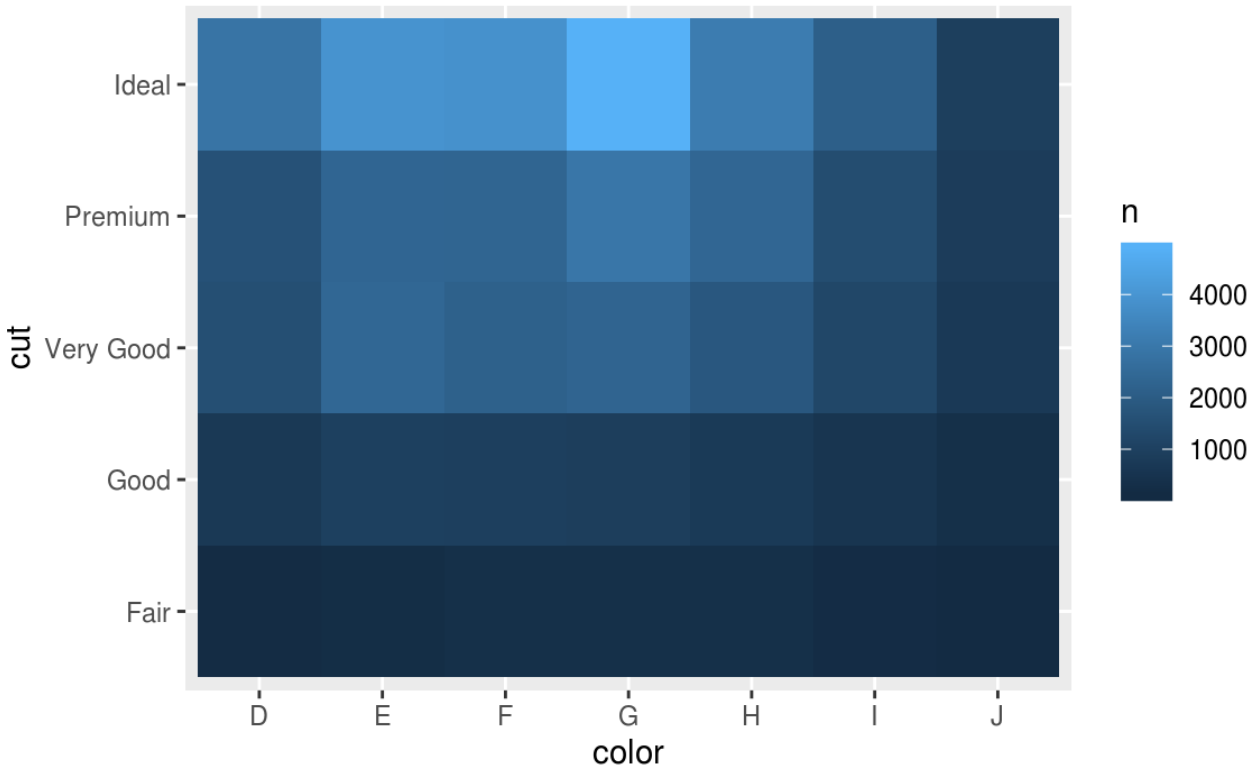
Two categorical variables: count the number of observations for each combination



Detecting covariation

Covariation is the tendency for the values of two or more variables to vary together in a related way.

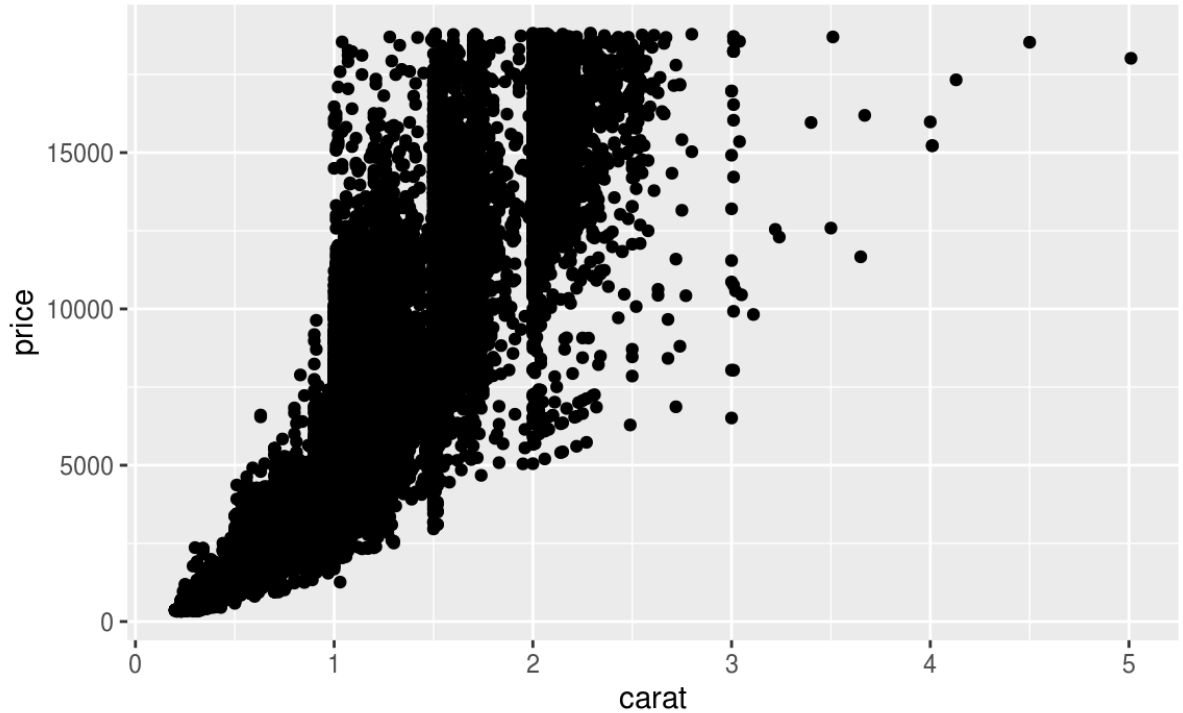
Two categorical variables: count the number of observations for each combination



First count then visualize using tile.

Detecting covariation

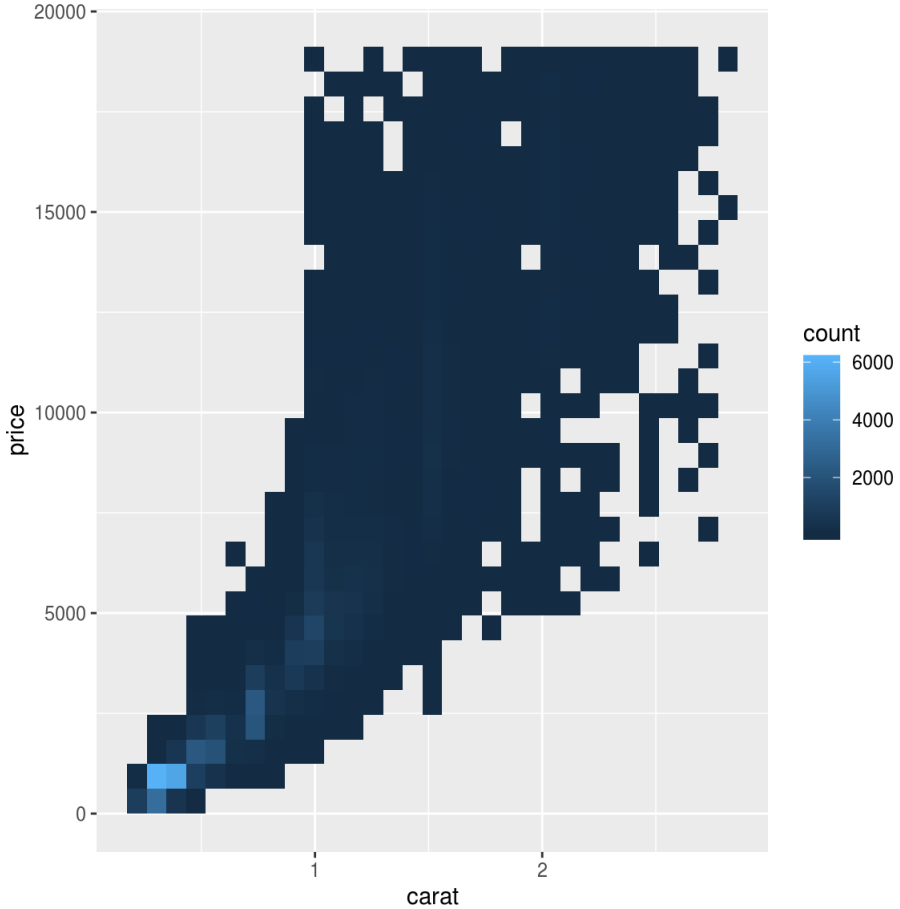
Two continuous variables: scatterplots



Scatterplots become less useful with more data points, because points begin to overplot, and pile up into areas of uniform black.

Detecting covariation

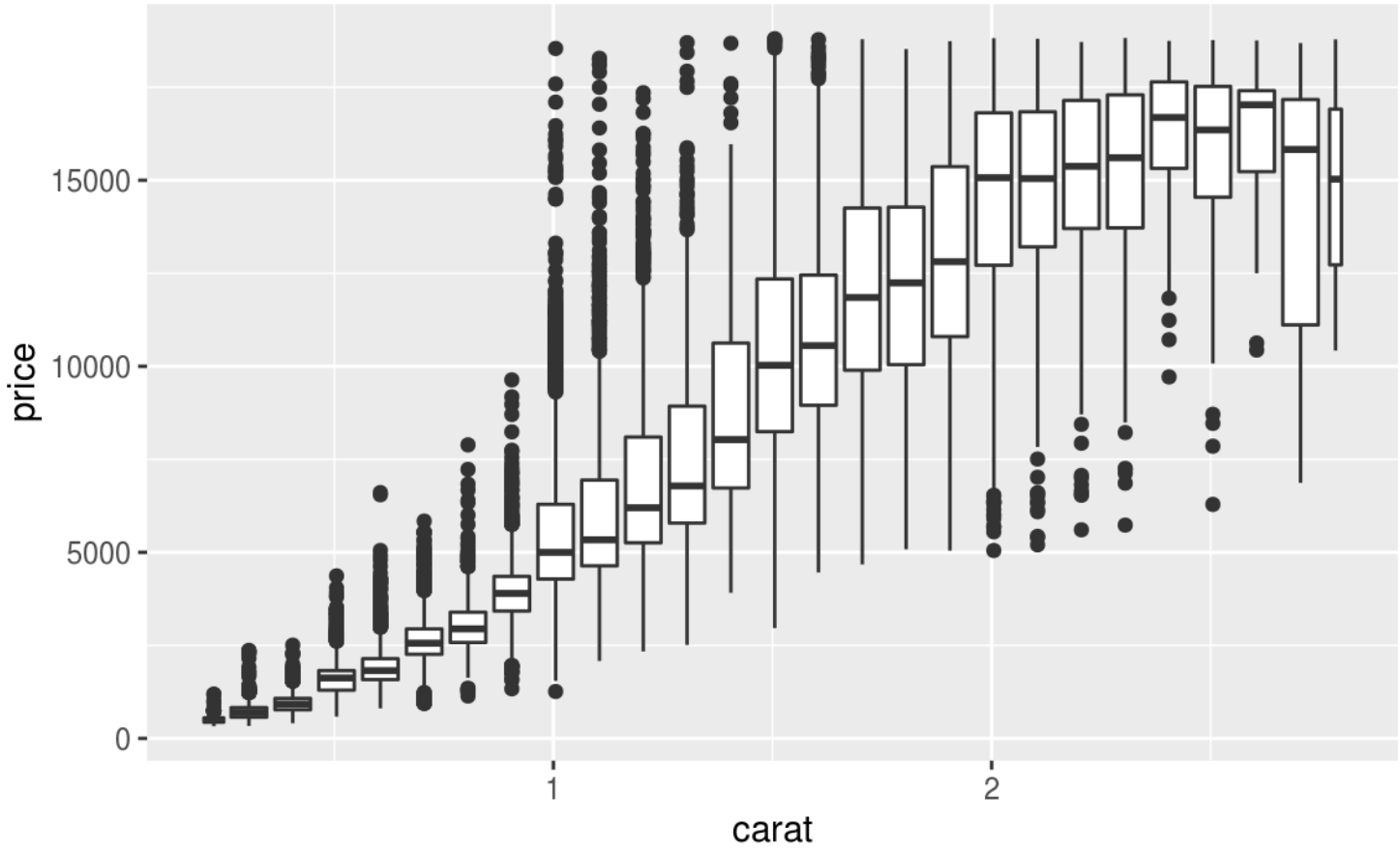
Two continuous variables: alternative



Divide the coordinate plane into 2d bins and then use a fill color to display how many points fall into each bin

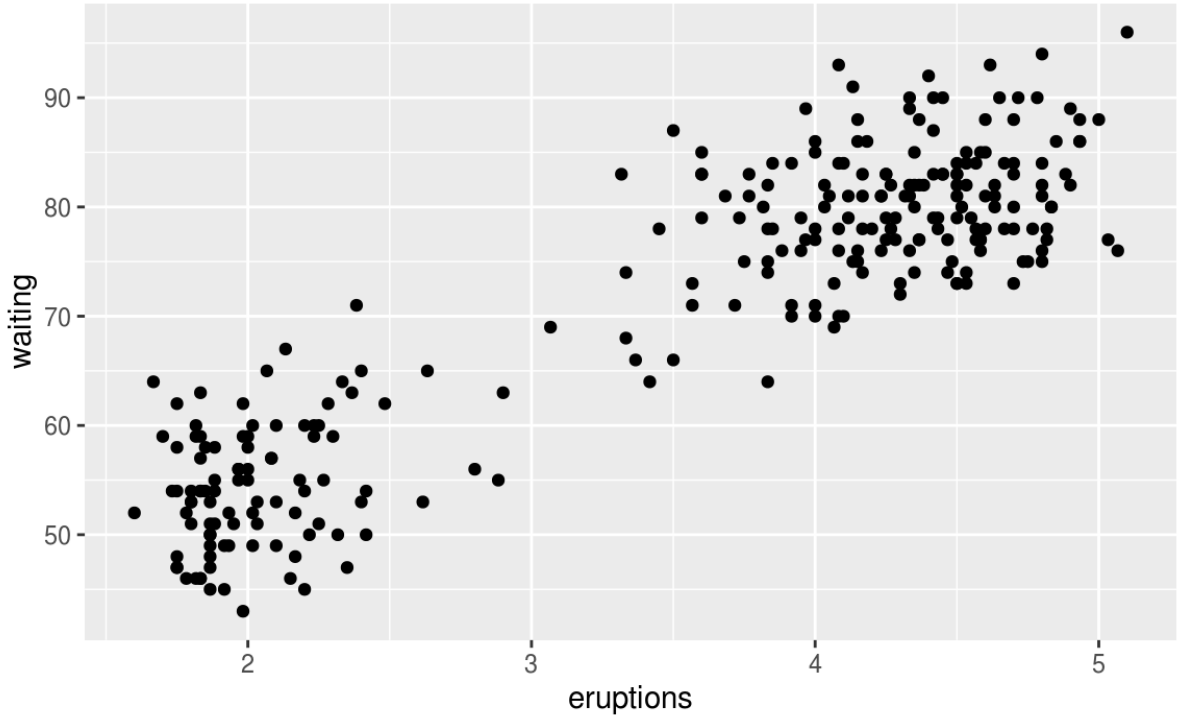
Detecting covariation

Two continuous variables: another alternative



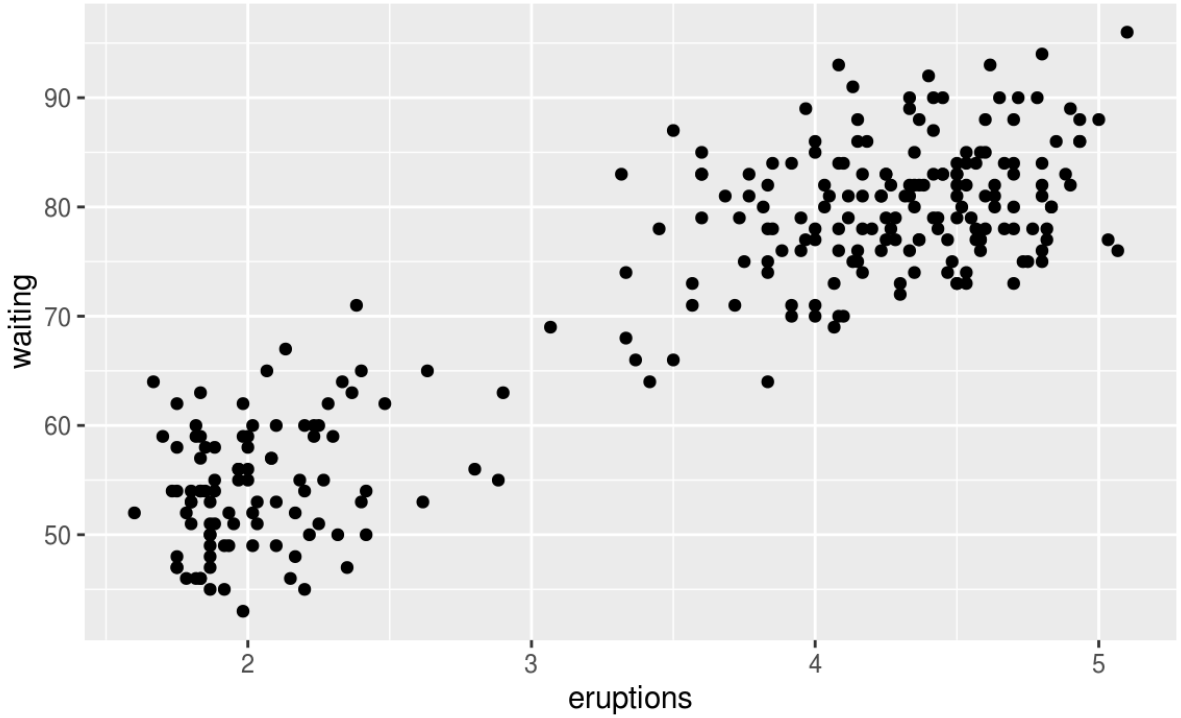
Patterns and Models

If a systematic relationship exists between two variables it will appear as a pattern in the data



Patterns and Models

If a systematic relationship exists between two variables it will appear as a pattern in the data



A scatterplot of Old Faithful eruption lengths versus the wait time between eruptions shows a pattern:

- longer wait times are associated with longer eruptions
- two visible clusters exist

Patterns and Models

Models are tools to extract patterns out of data.

For example, in diamonds data, it hard to understand the relationship between cut and price, because cut and carat, and carat and price are tightly related. One may use a model to remove the very strong relationship between price and carat so we can explore the subtleties that remain.

Quiz 1 — 10min — 10marks

Let X and Y be random variables with finite expected values. Then prove that

$$E(X + Y) = E(X) + E(Y).$$

Let S_n be the number of successes in n *Bernoulli* trials with probability p for success on each trial. Then prove that the expected number of successes is np.