Data Characteristics

Laura E. Barnes
and
Julianne Quinn

Week (08/30-09/3)

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

LSM 9/ 27

Baee, Barnes & Quinn

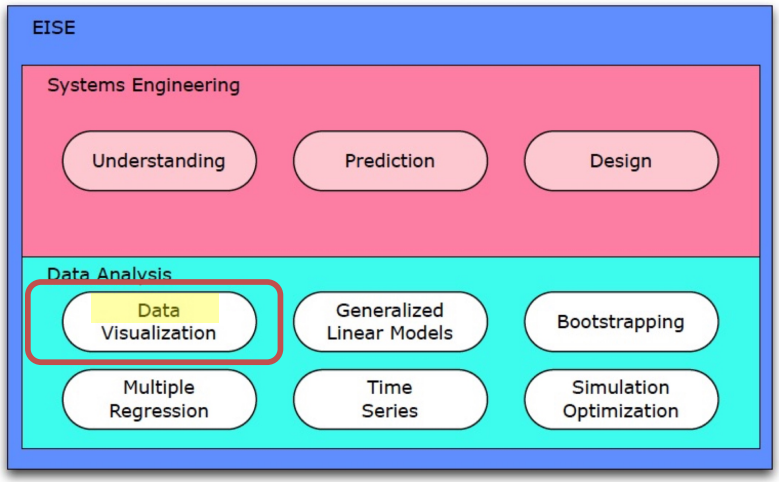
We want to use evidence-informed systems engineering for good.

- Modeling can be a powerful tool for understanding and improving systems. But if unchecked, models can become “Weapons of Math Destruction” (O’neil, 2016).
- The goal of this course is to learn how to ask the right questions, use appropriate metrics, and constantly re-evaluate our models to solve real-world problems.

* O’neil, Cathy. **Weapons of math destruction: How big data increases inequality and threatens democracy**, Broadway Books, 2016.

LSM 26/ 27

Baee, Barnes & Quinn



Agenda

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Objectives
- Data
- Data collection methods
- Data sampling methods
- Data types
- Temporal data
- Summary
- Others (Jefferson Literary and Debating Society)

Objectives

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Understand the important characteristics of data
- Understand the different data collection and sampling methods
- Understand potential issues in the data collection and sampling process
- Understand the different data types and numerical summarization methods
- Understand the different types of temporal data

Data

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Statistics is a tool for turning data into information
- **Goal:** make inferences about a population based on a data sample
  - If the sample is not representative of the whole population, we cannot make inference about the population from that sample.

Data

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Where does data come from?
- How is it gathered?
- How do we ensure it is accurate?
- Is it representative of the population from which it comes?

Data Sources

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Primary
  - The researcher directly collects the data that have not been previously collected (e.g., interviewing)
- Secondary
  - These are sources containing data that have been collected and compiled for another purpose. (e.g., reports of government departments)

Data Collection Methods

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Observations
- Experiments
- Simulations
- Surveys

Data Collection Methods

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Observational data
  - A way of gathering data by observing behavior, events or noting characteristics of the subject of interested without ANY manipulation.
  - Can be retrospective or prospective

- Experimental data
  - Experimental data is produced by active manipulation of one or more variables to produce and measure the effect of when that variable is altered.
  - Researchers **can typically draw cause and effect** (or causal) relationships.

Data Collection Methods

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Simulations
  - Used commonly when it is too expensive or unsafe to use to obtain data from real system
- Similar systems
  - Using another system to approximate the behavior of a similar system. For example, using one vehicle to approximate behavior of another

Data Collection Methods- Example

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Quiz and Exam Score Study
  - Let's say that we want to design a study to look at the relationship between quiz-taking behavior and exam scores.
  - What type of data collection method do you suggest and why?

Data Collection Methods- Example

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- We want to decide whether Advil or Tylenol is more effective in reducing fever.
- What type of data collection method do you suggest and why?

Sampling Methods

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

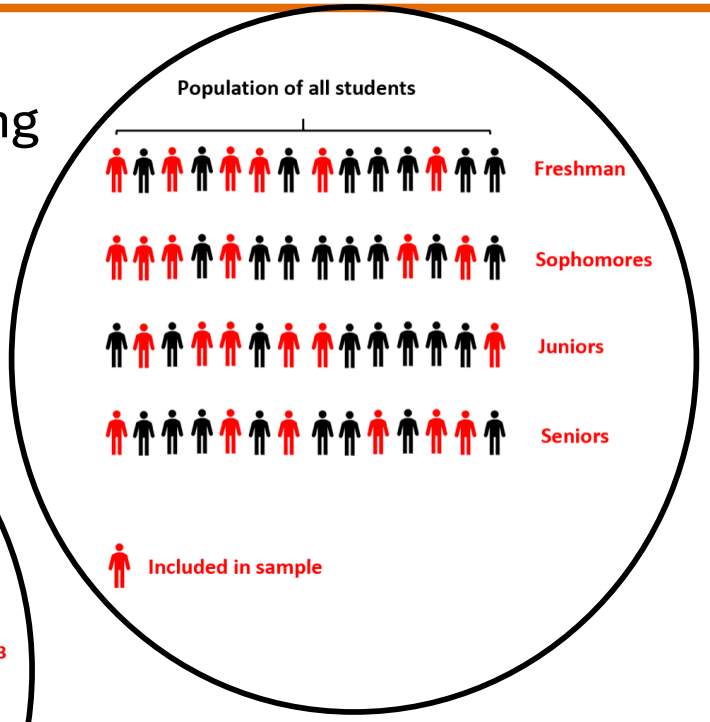
1. Non-probability methods
   - Convenience sampling
   - Snowball sampling
2. Probability methods
   - Simple random sample
   - Stratified random sample
   - Cluster sample

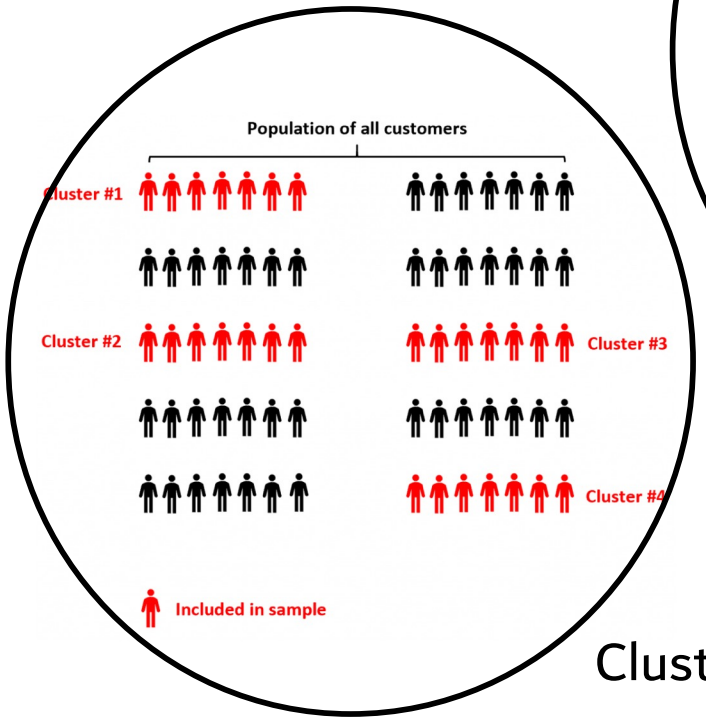
Sampling Methods

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

Stratified Sampling





Cluster Sampling

Ref: https://www.statology.org/cluster-sampling-vs-stratified-sampling/

Problems in Data Sampling

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Inadequate sample size
- Sampling errors
  - Sample not representative of population
- Non-sampling errors
  - Errors or bias in data collection process
  - Nonresponse bias
  - Selection bias

Data Sampling Methods- Example

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Assume you're the owner of a large airline company and live in NY. You want to survey your NY passengers on what they like and dislike about traveling with your airline.
  - Determine whether to use a non-probabilistic or probabilistic sampling method.
  - Determine the type of sampling.

Data Sampling Methods- Example

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

1. Since you live in NY. you go to the airport and just interview passengers as they approach your ticket counter.

2. You randomly select a set of passengers flying on your airline and question those that you have selected.

3. You group your passengers by the class they fly (first, business, economy), and then take a random sample from each of these groups.

4. You group your passengers by the class they fly (first, business, economy) and randomly select such classes from various flights and survey each passenger in that class and flight selected.

Data Types

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Distinguishing between the different types of variables is an integral part of EISE and informs the appropriate modeling technique and treatment of variables.
- Types of variables
  - Quantitative
    - Interval
    - Ratio
  - Qualitative
    - Nominal
    - Ordinal

Data Types

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Quantitative
  - Data that takes on numerical values that has a measure of distance between them. Quantitative values can be:
    - Ratio – absolute zero
    - Interval – positive and negative
    - Discrete - or “counted” as in the number of people in attendance
    - Continuous - or “measured” as in the weight or height of a person.

Data Types

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Qualitative / Categorical
  - Data that serves the function of a name only. Categorical values may be:
    - Binary – where there are two choices, e.g., Male and Female
    - Ordinal – where the names imply levels with hierarchy or order of preference, e.g., level of education
    - Nominal – where no hierarchy is implied, e.g., political party affiliation.
  - Note: For example, for coding purposes, you may assign Male as 0, Female as 1.
    - The numbers 0 and 1 stand for the two categories and there is no order between them even through they are numeric.

Data Types

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Additional examples of quantitative and qualitative variables:
  - Number of females in this class
  - A recipe for pancakes uses 3 cups of flour and 2 cups of milk
  - Nationality
  - Amount of milk in a 1-gallon container
  - Sex of students (coded as $M = 0$, $F = 1$)
  - Credit score (300-850)

Summarizing Quantitative Variables

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Measures of Central Tendency
  - A measure of central tendency is an important aspect of quantitative data. It is an estimate of a “typical” value.
  - Three of the many ways to measure central tendency are the **mean**, **median** and **mode**.
    - Mean is the average of data
    - Median is the middle value of the ordered data
    - Mode is the value that occurs most often in the data. (can be more than one)

Summarizing Quantitative Variables

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- While measures of central tendency are important, they do not tell the whole story.
- For example, suppose the mean score on a statistics exam is 80.
  - From this information, can we determine a range in which most people scored?

Summarizing Quantitative Variables

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Measures of position
  - The measures we consider here are percentiles and quartiles.
    - Percentiles
      - Give a range where a certain percentage of the data fall
      - The p^{th} percentile of the data set is a measurement such that after the data are ordered from smallest to largest, at most, p% of the data are at or below this value and at most, (100 - p)% at or above it.

Example: A common application of percentiles is their use in determining passing or failure cutoffs for standardized exams such as the GRE. If you score in the 95th percentile then you are at or above 95% of all test takers.

Summarizing Quantitative Variables

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Measures of position
  - The measures we consider here are percentiles and quartiles.
    - Quartiles
      - Quartiles are values that divide a (part of a) data table into four groups containing an approximately equal number of observations.

Summarizing Quantitative Variables

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Measures of Variability
  - Range
    - The range is the difference in the maximum and minimum values of a data set. The range is easy to calculate but it is very much affected by extreme values.
  - Interquartile range
    - The interquartile range is the difference between upper and lower quartiles and denoted as IQR.
  - Variance
    - The average squared distance from the mean
  - Standard deviation
    - Approximately the average distance the values of a data set are from the mean or the square root of the variance

Summarizing Qualitative Variables

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Once we determine that a variable is Qualitative (or Categorical), we can summarize the data by using frequencies and percentages (or proportions).
  - **Proportion:** fraction or part of the total that possesses a certain characteristic.

T-shirt color	Frequency	Percentage
White	24	80%
Blue	3	10%
Red	2	6.66%
Green	1	3.33%

Temporal Data

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Data that is associated with measurements over time.
  - For example, temperature measurement over the past month

Descriptor	Examples
Cross-sectional	Class weights
Time series	Your weight over the last month
Longitudinal / Panel	Class weights over the last month

Data Reports

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

Data description for reports includes the following elements
Source(s)
Variable names
Summary statistics
Scales
Temporal descriptors
Other: indicators of bias, errors, etc.

R Code Snippets

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

```
> summary(accident_data_20[,c("ACCDMG", "TOTKLD", "CARS", "STATION")])
     ACCDMG              TOTKLD               CARS                  STATION
 Min.   :       0   Min.   :0.00000   Min.   :  0.000   HOUSTON      :  39
 1st Qu.:   19217   1st Qu.:0.00000   1st Qu.:  0.000   BELLEVUE     :  29
 Median :   37120   Median :0.00000   Median :  0.000   NORTH PLATTE:  29
 Mean   :  161089   Mean   :0.01469   Mean   :  3.631   FORT WORTH   :  27
 3rd Qu.:  102936   3rd Qu.:0.00000   3rd Qu.:  0.000   CHICAGO      :  24
 Max.   :9270078   Max.   :4.00000   Max.   :157.000   KANSAS CITY  :  21
                                                       (Other)      :2009
```

```
> str(accident_data_20[,c("ACCDMG", "TOTKLD", "CARS", "STATION")])
'data.frame':   2178 obs. of  4 variables:
 $ ACCDMG : int  32706 19471 15000 29912 16200 37792 448394 46328 14288 58014 ...
 $ TOTKLD : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CARS   : int  0 0 1 18 20 7 2 0 0 25 ...
 $ STATION: Factor w/ 912 levels "ABBYVILLE","ABERDEEN",..: 198 4 145 397 458 82 457 664 128 181 ...
>
```

```
> class(accident_data_20$TOTKLD)
[1] "integer"
>
```

```
> var(accident_data_20$TOTKLD)
[1] 0.01999533
>
```

```
> mean(accident_data_20$TOTKLD)
[1] 0.01469238
>
```

```
> class(accident_data_20$STATION)
[1] "factor"
>
```

Summary

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- The goal of statistics is to make inferences about the population based on the sample. Therefore, knowing how the sample is obtained is essential to understand.
- How we summarize the data depends on the variable types and the problem
- Summarizing data (numerically and graphically) helps to understand data characteristics.
- The temporal aspects of data are important to consider.
- Data reports should include crucial information for stakeholders.

Next Class

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- Visualization (part 1)
- R programming (part 1)

Additional Resources and References

Agenda

- Objectives
- Data
- Data Sources
- Data Collection
- Data Sampling
- Data Types
- Temporal Data
- Summary

- https://online.stat.psu.edu/stat500
- https://www.coursera.org/lecture/six-sigma-define-measure-advanced/collecting-and-summarizing-data-pt-1-RZk0X

Graphical Analysis and Visualization (1)

Laura Barnes,

and

Julianne Quinn

Agenda

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Why visualize?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

Why Visualize?

Agenda

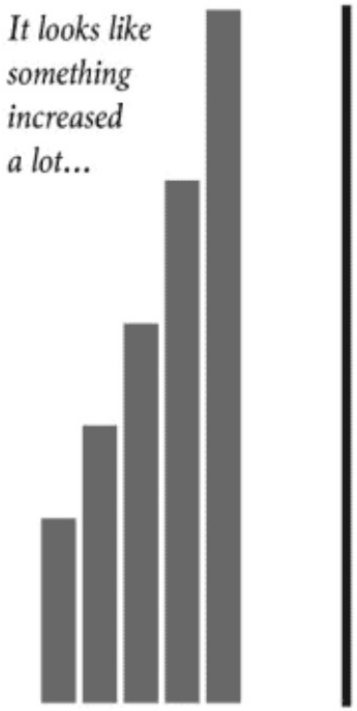
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

1. To understand data
2. Show summary statistics
3. Distribution
4. Tell a story
5. …

Visualization

Agenda

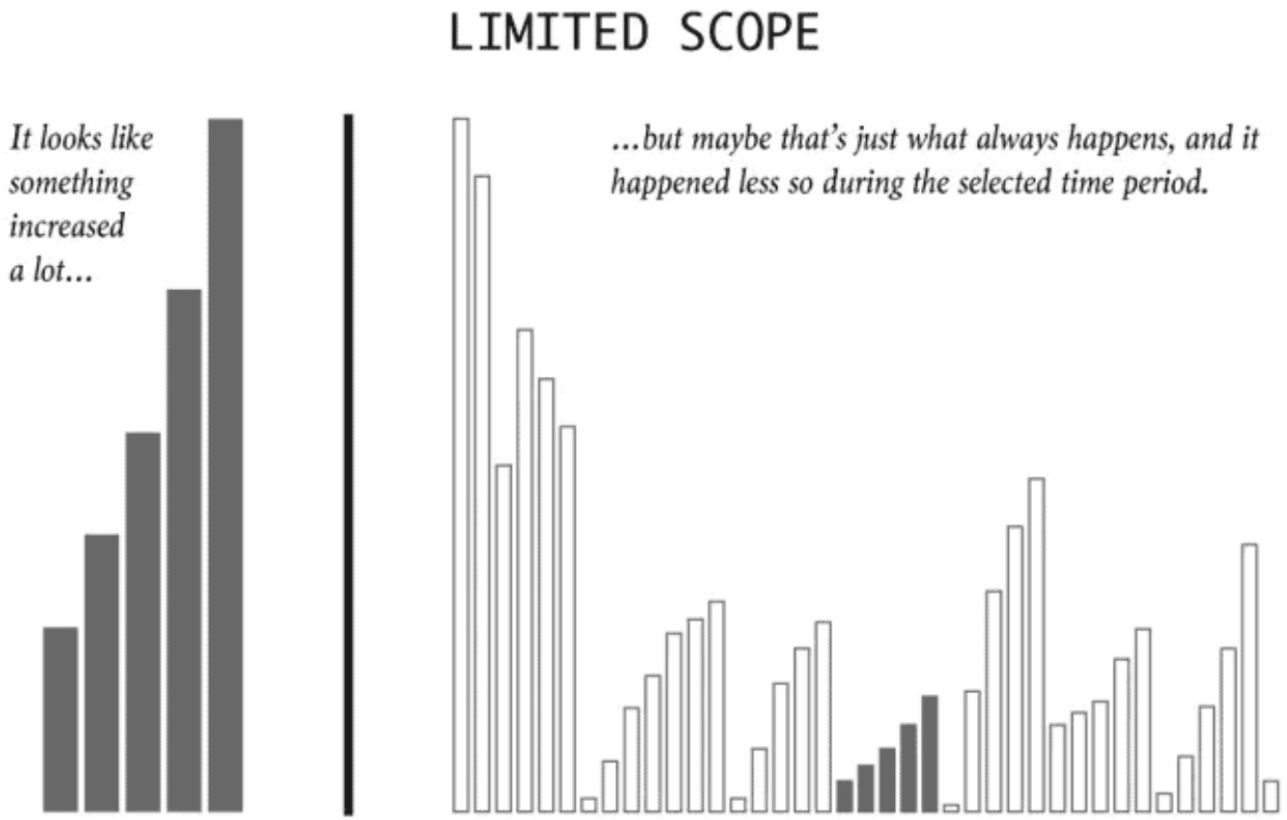
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



Visualization

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



Principles for Visualization

Agenda

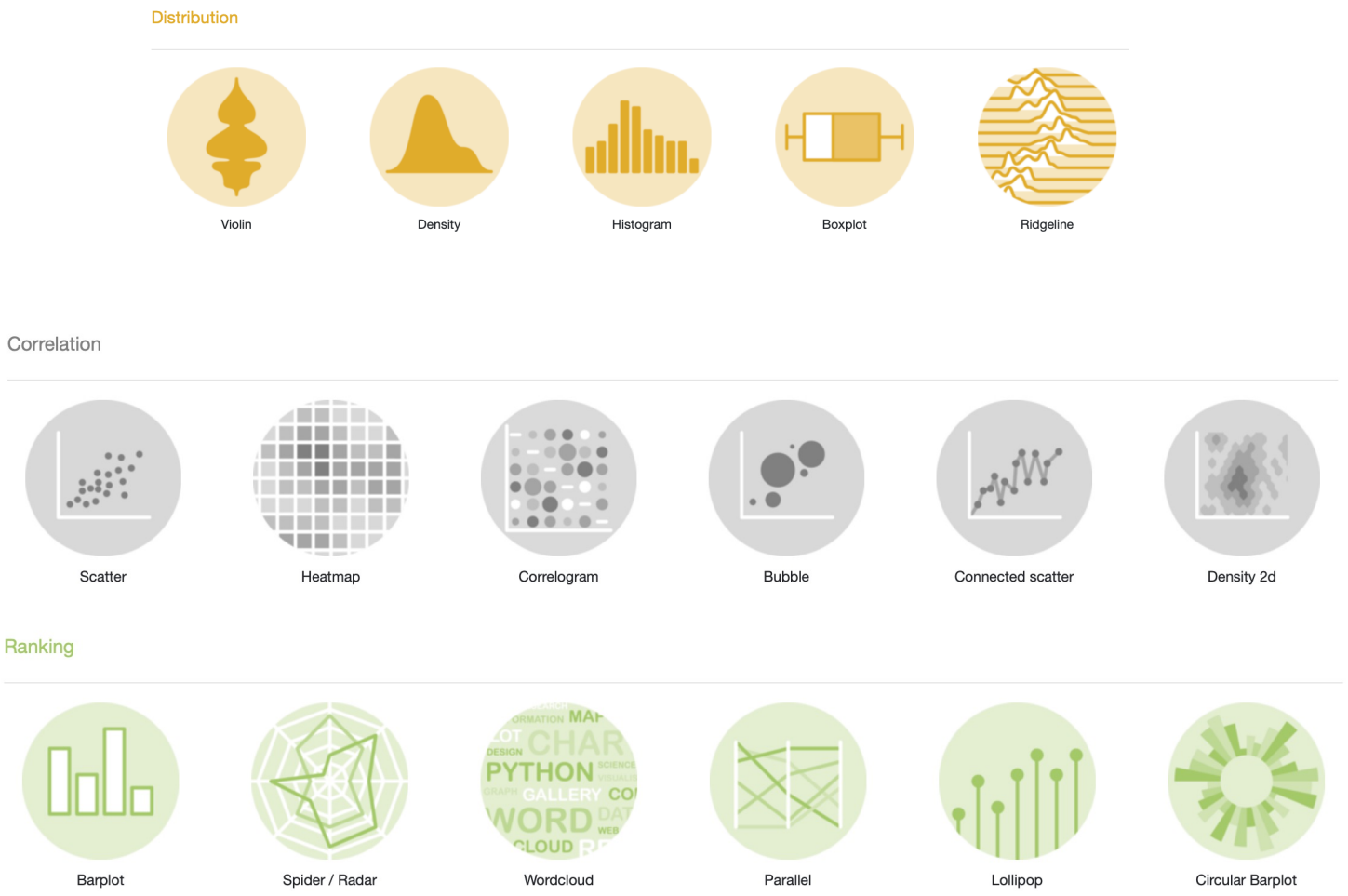
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Design to illustrate a particular point.
- Maximize information and minimize ink.
- Organize hierarchically.

Visualization with R

Agenda

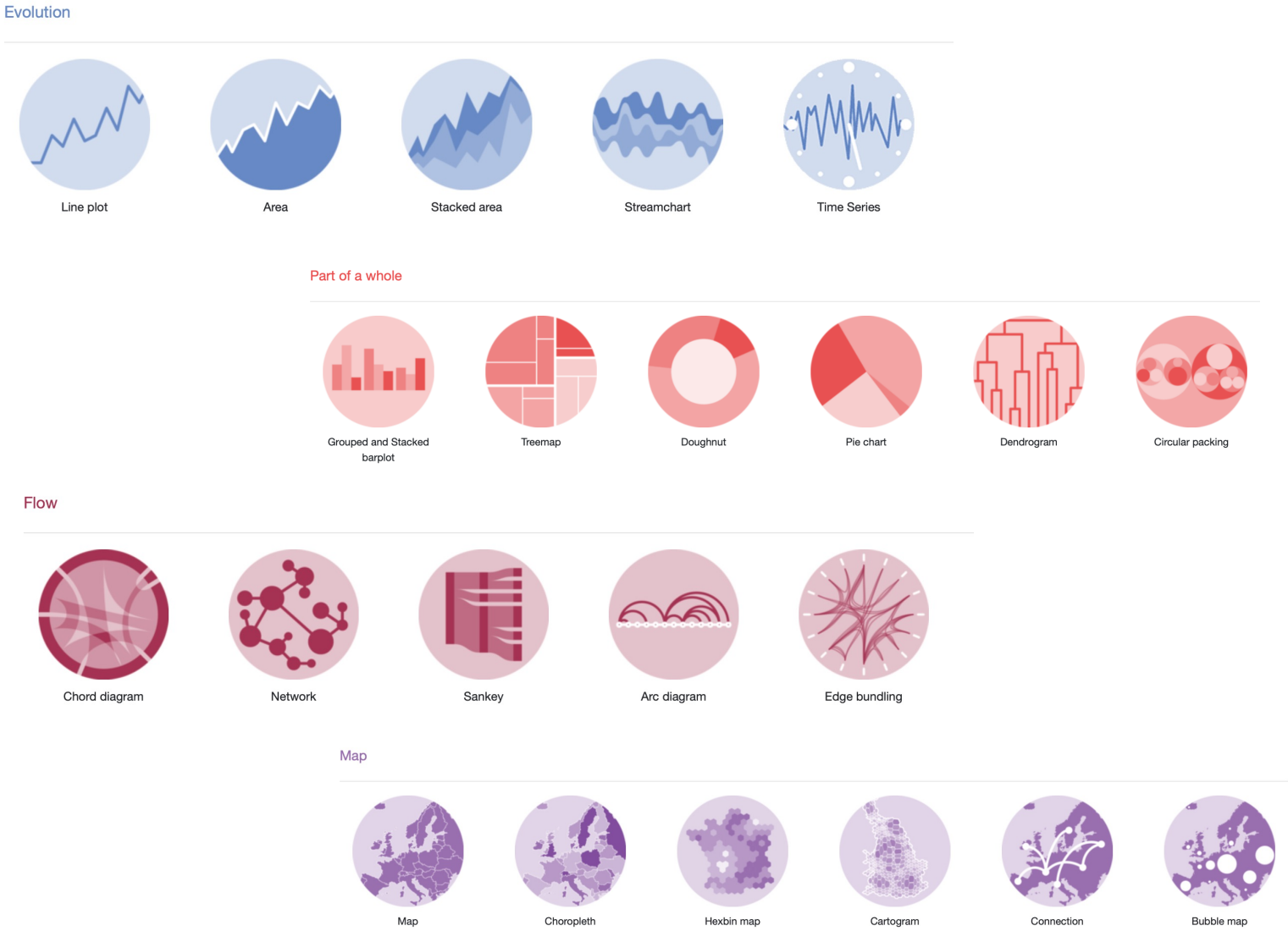
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



Visualization with R

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



Univariate Visualization

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

Univariate analysis:

- Provides visual representation of summary statistics for one variable.
- Empirical data distributions are ordered counts or values for a variable.
- Four common methods for graphical displays of data distributions.
  - Histograms
  - Boxplots
  - Density Plots
  - Bar Plots
  - QQ Plots

Histograms

Agenda

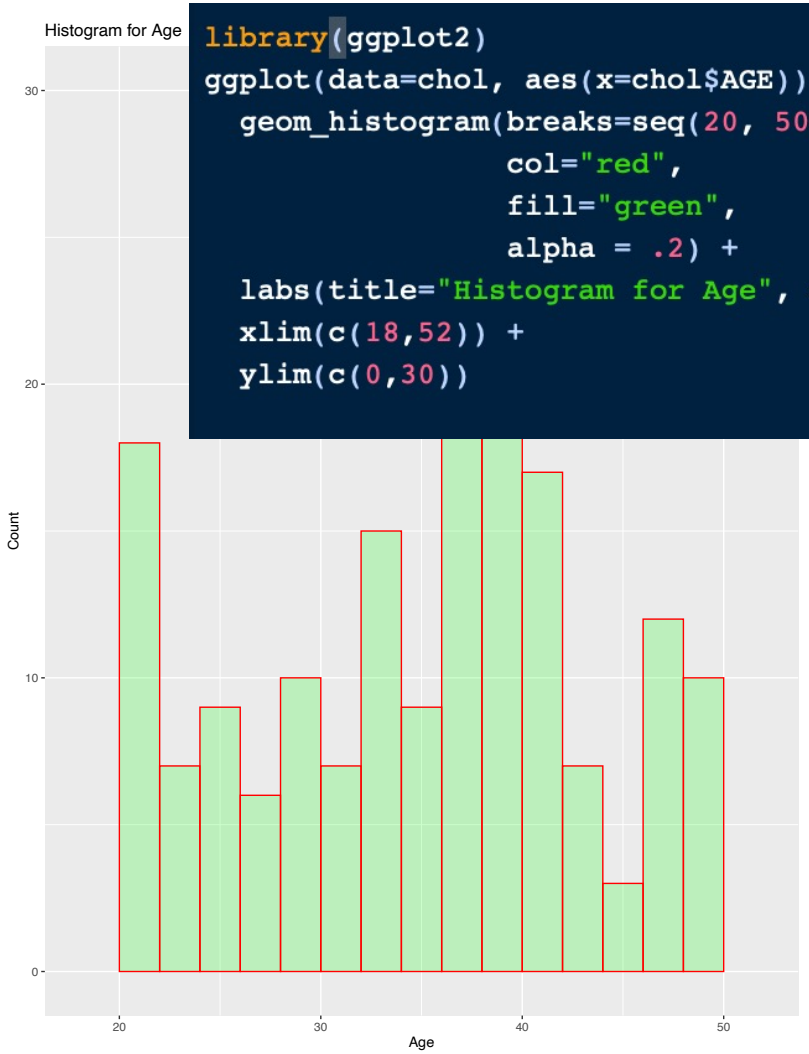
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Graphical displays of data distributions by counts in intervals or bins of the variable's values
- Example:
  - What is the frequency of different ages present in a class?
    - Ages are: 28, 24, 26, 25, 30, 20, 20, 29, … , 30

Histogram (e.g., age)

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary




```
library(ggplot2)
ggplot(data=chol, aes(x=chol$AGE)) +
  geom_histogram(breaks=seq(20, 50, by=2),
                 col="red",
                 fill="green",
                 alpha = .2) +
  labs(title="Histogram for Age", x="Age", y="Count") +
  xlim(c(18,52)) +
  ylim(c(0,30))
```

Steps to Create a Histogram

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Steps:
  1. Divide the range of the variable along the real line into N_B bins. Let h be the width of each bin and B_j for $j = 1, \ldots, N_B$ be the interval for each bin. Then the interval for each bin is

$$B_j = [x_0 + (j-1)h, x_0 + jh)$$

  2. Count how many observations of X fall into each bin and normalize by the number of observations (density estimate) or just leave the counts (frequency histogram).

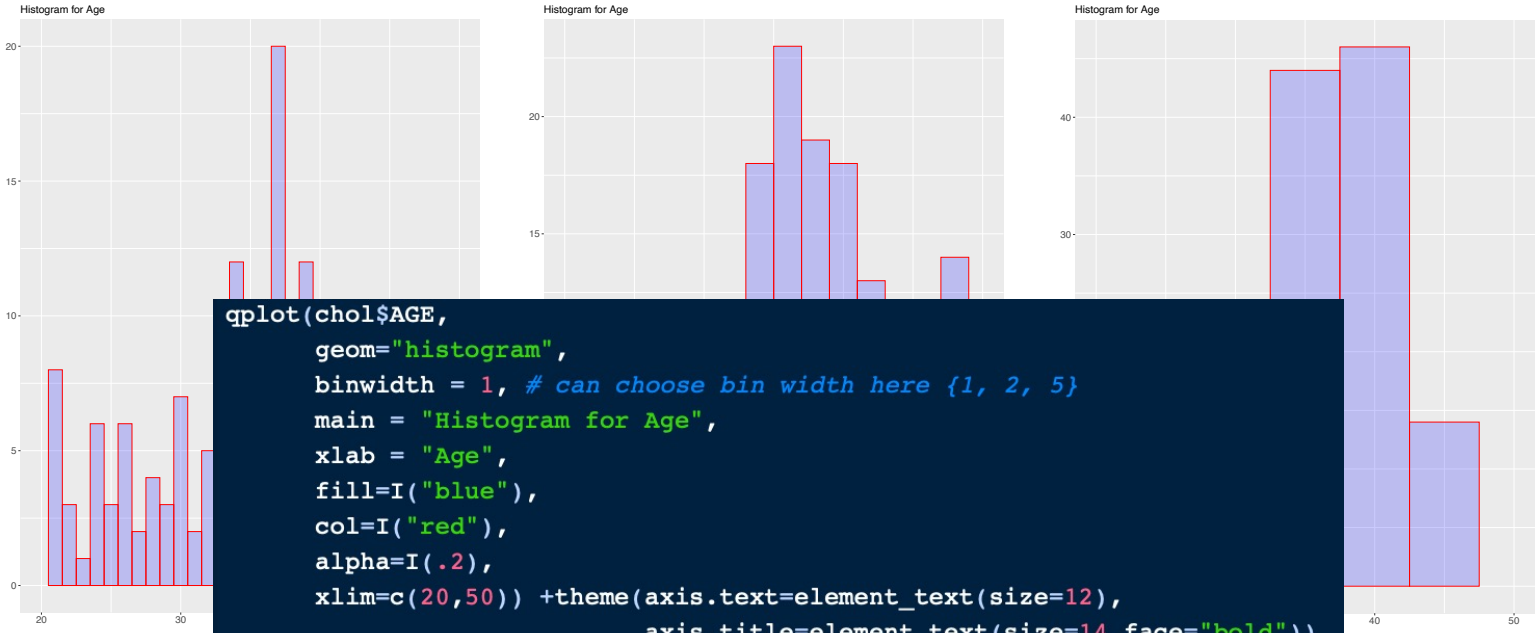
- Technically a histogram is a density estimate which integrates to one and a frequency histogram is a count.

Histogram (e.g., age)

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

Bin width = 1 | Bin width = 2 | Bin width = 5




```
qplot(chol$AGE,
      geom="histogram",
      binwidth = 1, # can choose bin width here {1, 2, 5}
      main = "Histogram for Age",
      xlab = "Age",
      fill=I("blue"),
      col=I("red"),
      alpha=I(.2),
      xlim=c(20,50)) +theme(axis.text=element_text(size=12),
                            axis.title=element_text(size=14,face="bold"))
```

$$B_j = [x_0 + (j-1)\boldsymbol{h}, x_0 + j\boldsymbol{h})$$

Histogram Bin Width

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- If you use too few bins, the histogram doesn't really portray the data very well.
- If you have too many bins, you get a broken comb look, which also doesn't give a sense of the distribution.
- There are many methods to optimally select the number of bin widths (e.g., Freedman–Diaconis rule, Sturges).

Kernel Density Plots

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- A density plot is a representation of the distribution of a numeric variable.
- It is a smoothed version of the histogram and is used in the same kind of situation.
  - Kernel density plots are usually a much more effective way to view the distribution of a variable.

What is Kernel?

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- A kernel is a special type of probability density function (PDF) with the added property that it must be even. Thus, a kernel is a function with the following properties:
  - Non-negative
  - Real-valued
  - Even
  - Its definite integral over its support set must equal to 1
- Some common PDFs are kernels; they include the Uniform(-1,1) and standard normal distributions.

What is Kernel Density Estimation?

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Kernel density estimation is a non-parametric method of estimating the probability density function (PDF) of a continuous random variable.
- It is non-parametric because it does not assume any underlying distribution for the variable.
- Essentially, at every datum, a kernel function is created with the datum at its center – this ensures that the kernel is symmetric about the datum.
- The PDF is then estimated by adding all kernel functions and dividing by the number of data to ensure that it satisfies the two properties of a PDF:
  1. Every possible value of the PDF, is non-negative.
  2. The definite integral of the PDF over its support set equals to 1.

Constructing a Kernel Density Estimate

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

1. Choose a kernel functions $K(x, h)$. Let h be the bandwidth that controls the amount of smoothing and $x_1, x_2, \ldots, x_n$ be the observed values for X.
2. Compute height of the kernel density function at a point x in the range of X as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Basically, the KDE smooths each data point x_i into a small density bumps and then sum all these small bumps together to obtain the final density estimate.

Constructing a Kernel Density Estimate

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

Choosing the bandwidth is a complicated topic that is better addressed in a more advanced book or paper, but here are some useful guidelines:

- A small h results in a small standard deviation, and the kernel places most of the probability on the datum. Use this when the sample size is large and the data are tightly packed.
- A large h results in a large standard deviation, and the kernel spreads more of the probability from the datum to its neighboring values. Use this when the sample size is small and the data are sparse.

et h be the
of
bserved
function

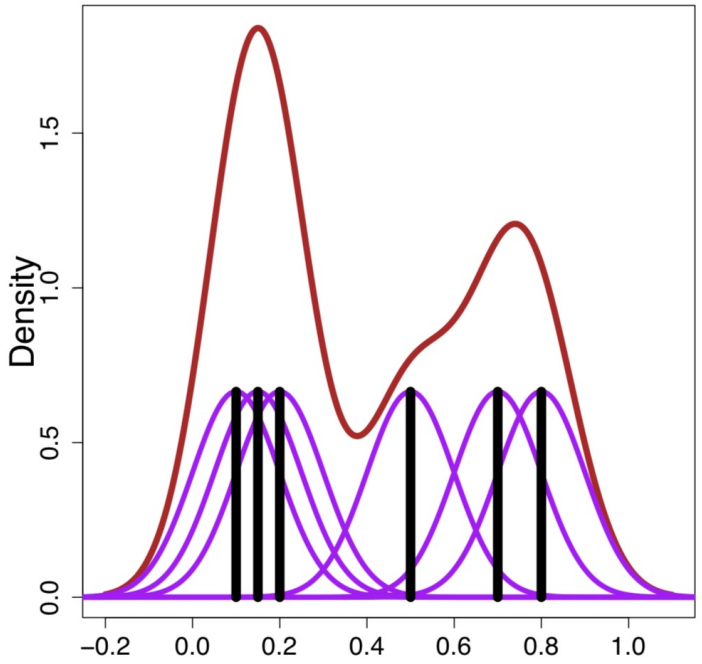
Basically, the KDE smooths each data point x_i into a small density bumps and then sum all these small bumps together to obtain the final density estimate.

Kernel Density Estimation

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- There are 6 data points located at where the black vertical segments indicate:
  - *{0.1, 0.2, 0.5, 0.7, 0.8, 0.15}*
- The KDE first smooths each data point into a purple density bump and then sum them up to obtain the final density estimate.

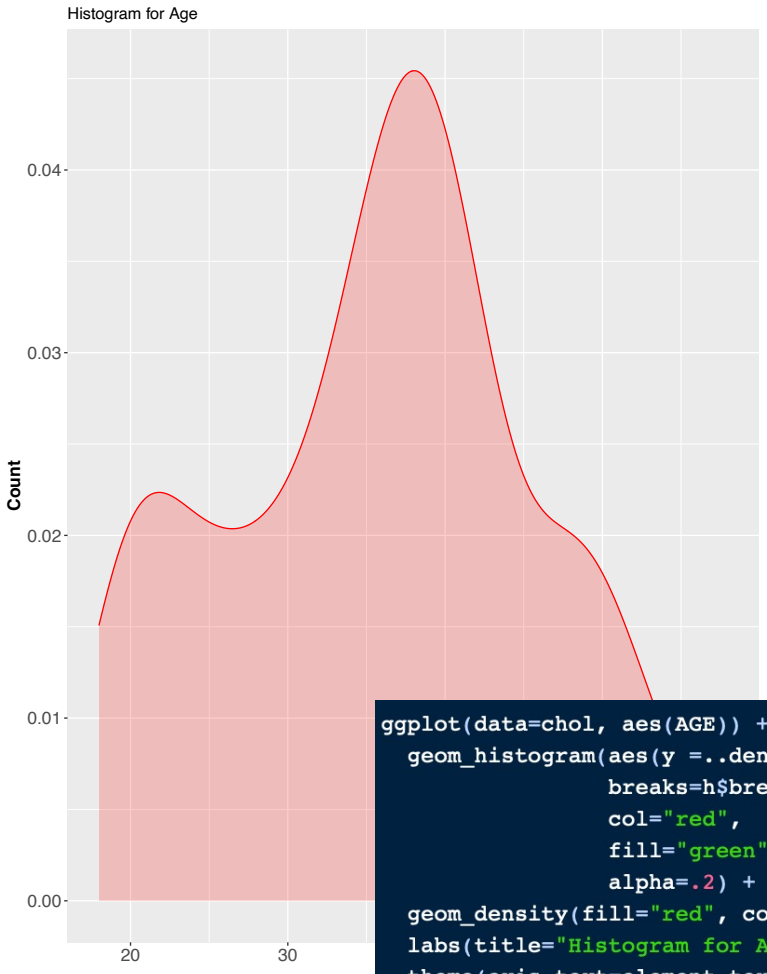


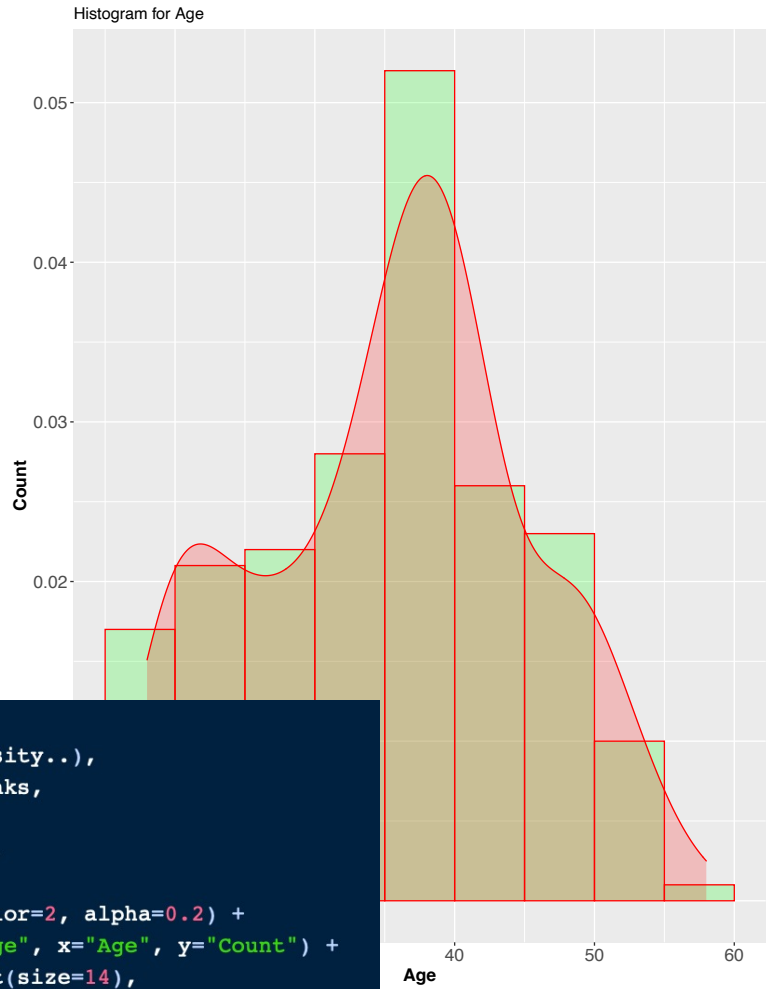
Ref: Density Estimation: Histogram and Kernel Density Estimator, Yen-Chi Chen, 2018

Kernel Density Plots (e.g., age)

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary






```
ggplot(data=chol, aes(AGE)) +
  geom_histogram(aes(y =..density..),
                 breaks=h$breaks,
                 col="red",
                 fill="green",
                 alpha=.2) +
  geom_density(fill="red", color=2, alpha=0.2) +
  labs(title="Histogram for Age", x="Age", y="Count") +
  theme(axis.text=element_text(size=14),
        axis.title=element_text(size=14,face="bold"))
```

Example Kernel Functions

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Gaussian

$$K(u) = \frac{1}{2\pi} \exp(-\frac{1}{2}u^2)$$

- Triangular

$$K(u) = 1 - |u|, |u| \leq 1$$

- Epanechnikov

$$K(u) = \frac{3}{4}(1 - u^2), |u| \leq 1$$

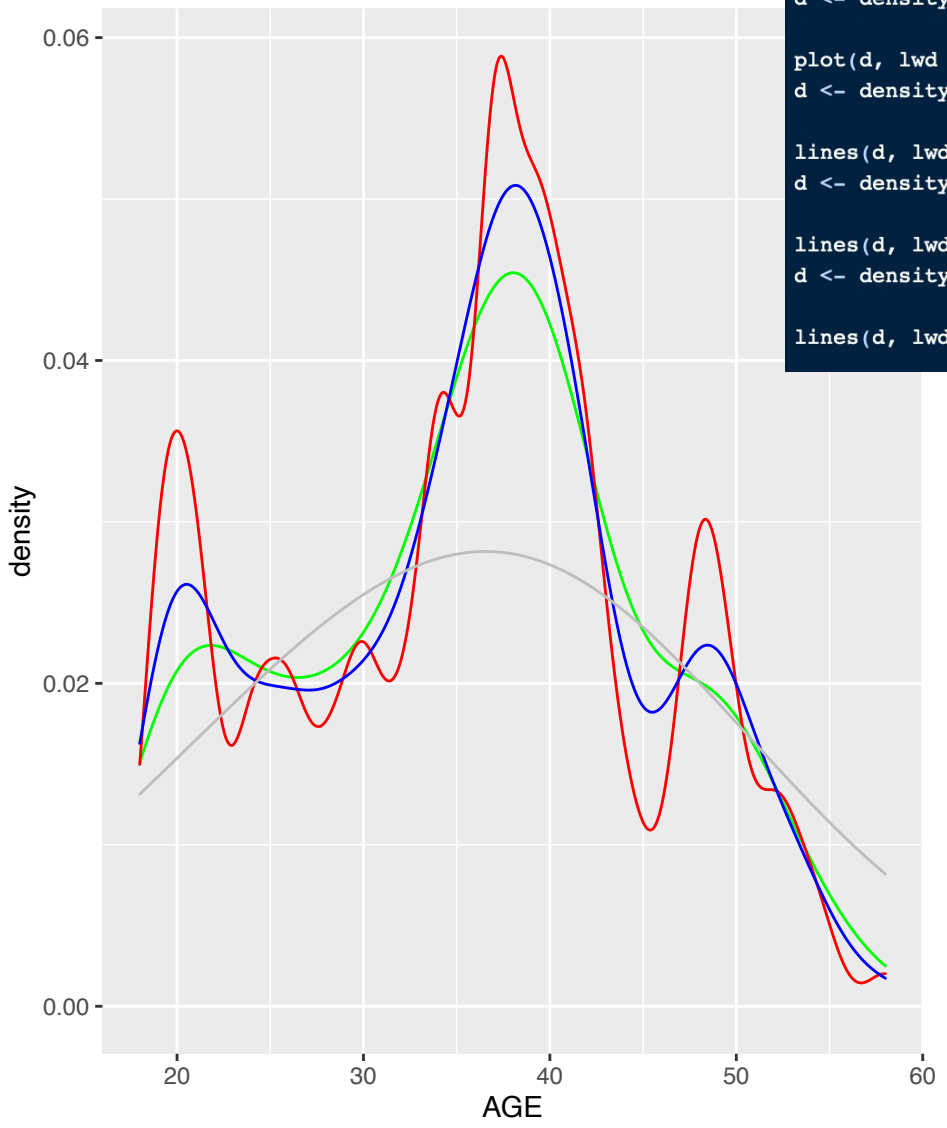
- Cosine

$$K(u) = \frac{\pi}{4} \cos(\frac{\pi}{2}u)$$

Bandwidth

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary




```
d <- density(chol$AGE,
             kernel = "gaussian", bw = 1)
plot(d, lwd = 2, col = 'blue')
d <- density(chol$AGE,
             kernel = "gaussian", bw = 2)
lines(d, lwd = 2, col = 'red')
d <- density(chol$AGE,
             kernel = "gaussian", bw = 10)
lines(d, lwd = 2, col = 'gray')
d <- density(chol$AGE,
             kernel = "gaussian")
lines(d, lwd = 2, col = 'green')
```

Bar Plots

Agenda

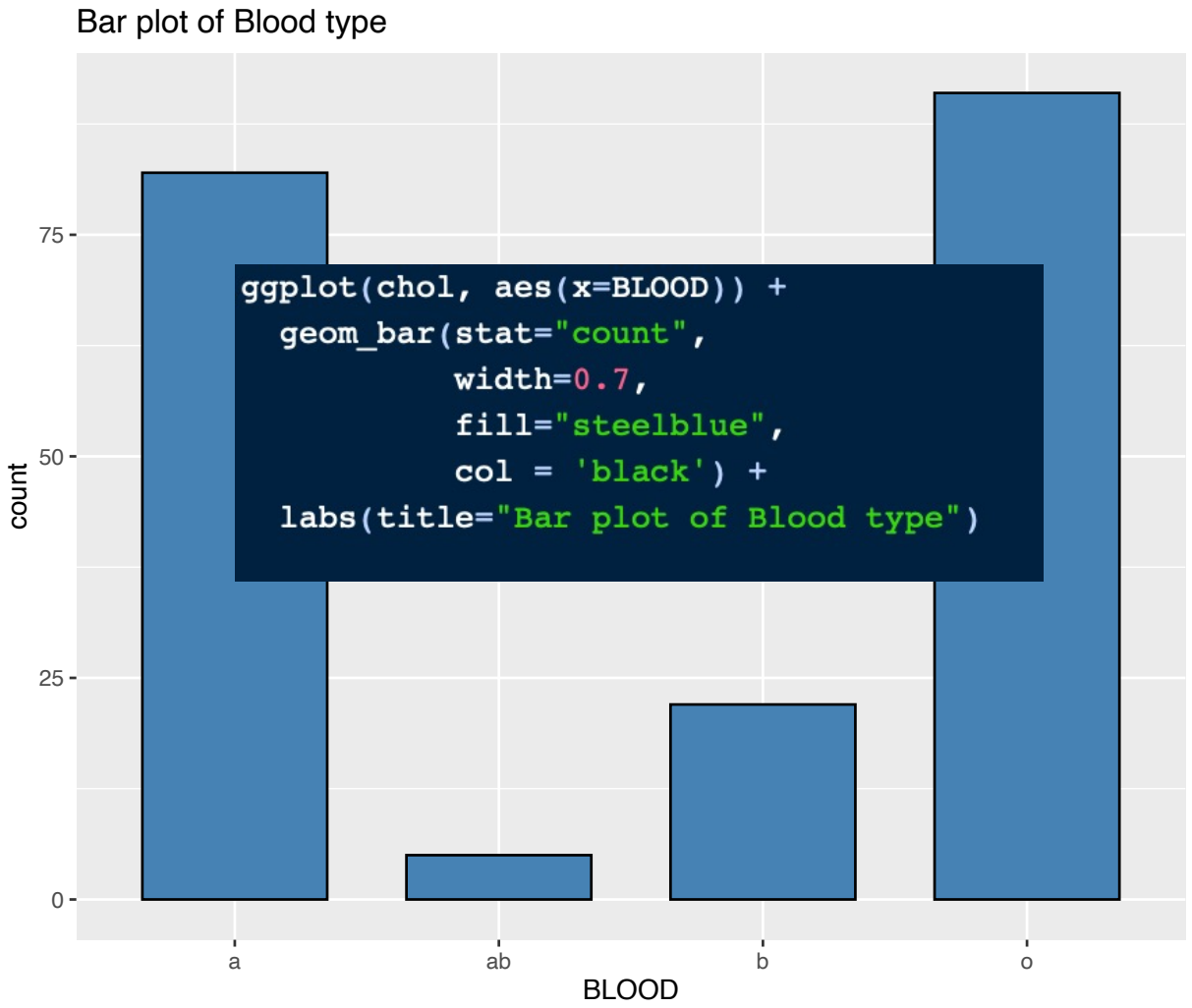
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Bar plots provide histogram or frequency count displays for categorical variables.
- Each bar shows the number or proportion of observations in each level of the categorical variable.

Bar Plots (e.g., blood type)

Agenda

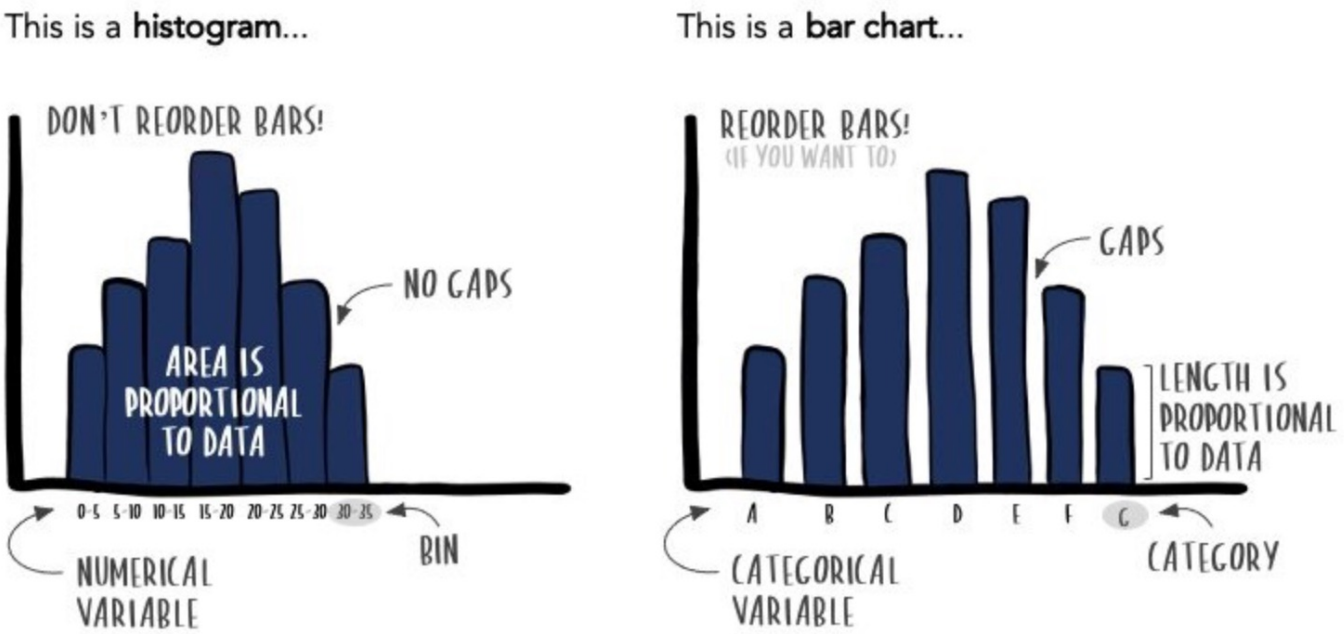
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



Histogram vs. Bar Plot

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



Ref: https://www.storytellingwithdata.com/blog/2021/1/28/histograms-and-bar-charts

Box Plots

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").
- It can tell you about your outliers and what their values are.
- It can also tell you if your data is symmetrical, how tightly your data is grouped, and
- How data is skewed.

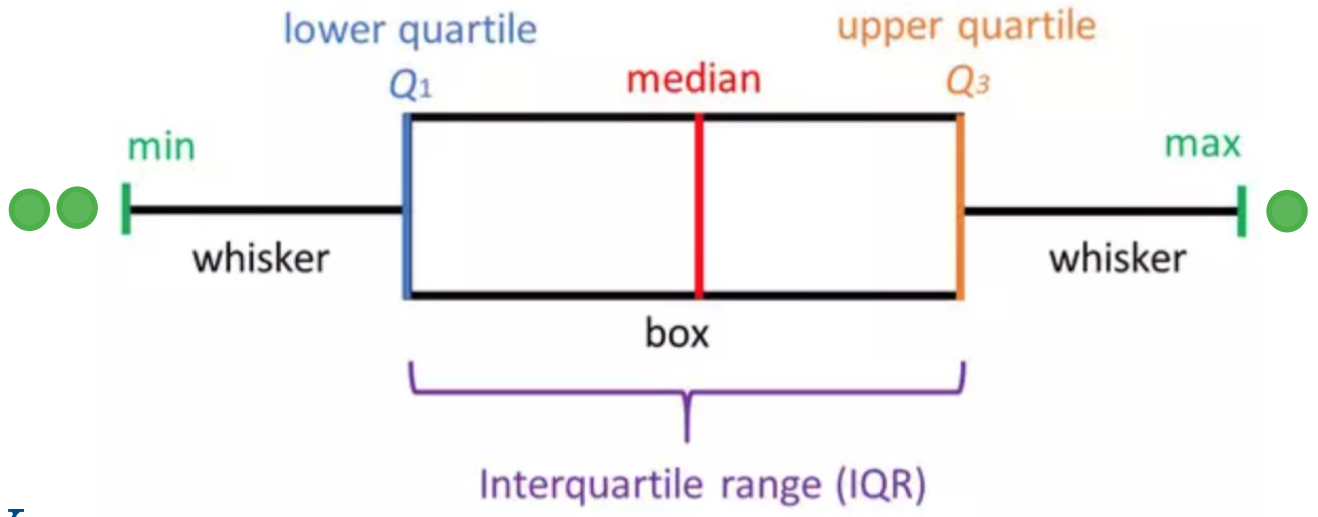
Box Plot

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- **Median (Q2/50th Percentile):**
  - Middle value of the dataset.
- **First quartile (Q1/25th Percentile):**
  - Middle number between the smallest number (not the "minimum") and the median of the dataset.
- **Third quartile (Q3/75th Percentile):**
  - Middle value between the median and the highest value (not the "maximum") of the dataset.

- **Interquartile range (IQR):**
  - 25th to the 75th percentile.
- **Whiskers**
- **Outliers** (shown as green circles)
- **"Maximum"**
  - Q3 + 1.5*IQR
- **"Minimum"**
  - Q1 -1.5*IQR



Creating a Box Plot

Agenda

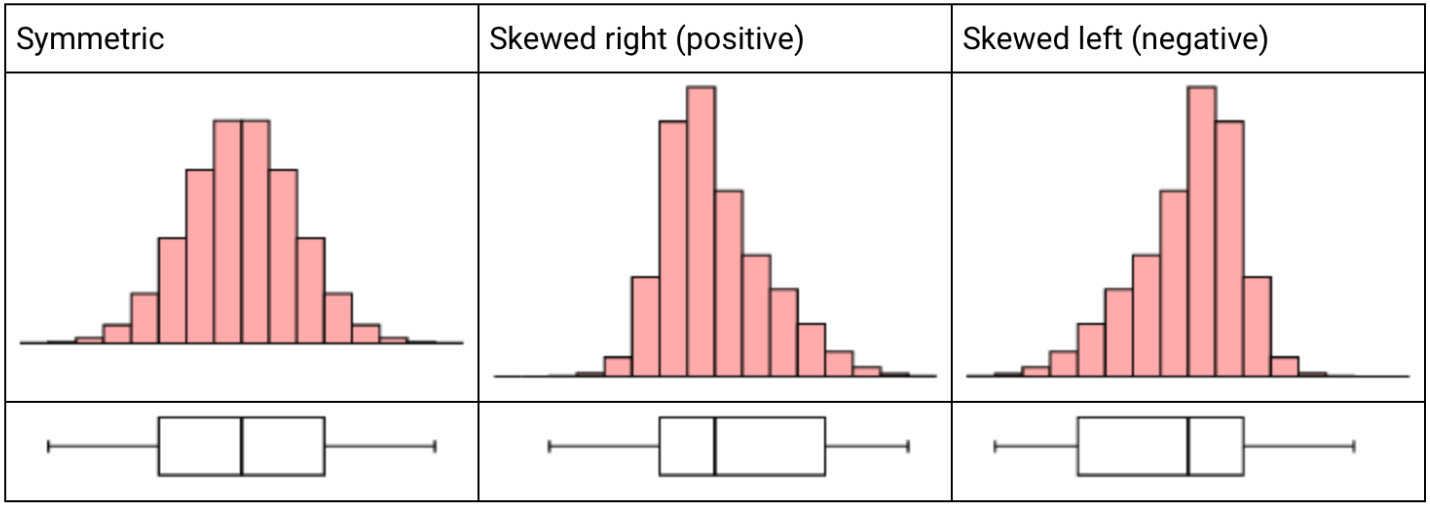
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Steps
  1. Plot a box showing the values the 25th and 75th percentiles.
  2. Draw a line in the box at the median
  3. Draw the lower whisker from the 25th percentile, P_{25}, down to the smallest observed value greater than or equal to $P_{25} - 1.5\ IQR$.
  4. Draw the upper whisker from the 75th percentile, P_{75}, up to the largest observed value less than or equal to $P_{75} + 1.5\ IQR$.
  5. Plot any values above or below the whiskers.

Box Plot

Agenda

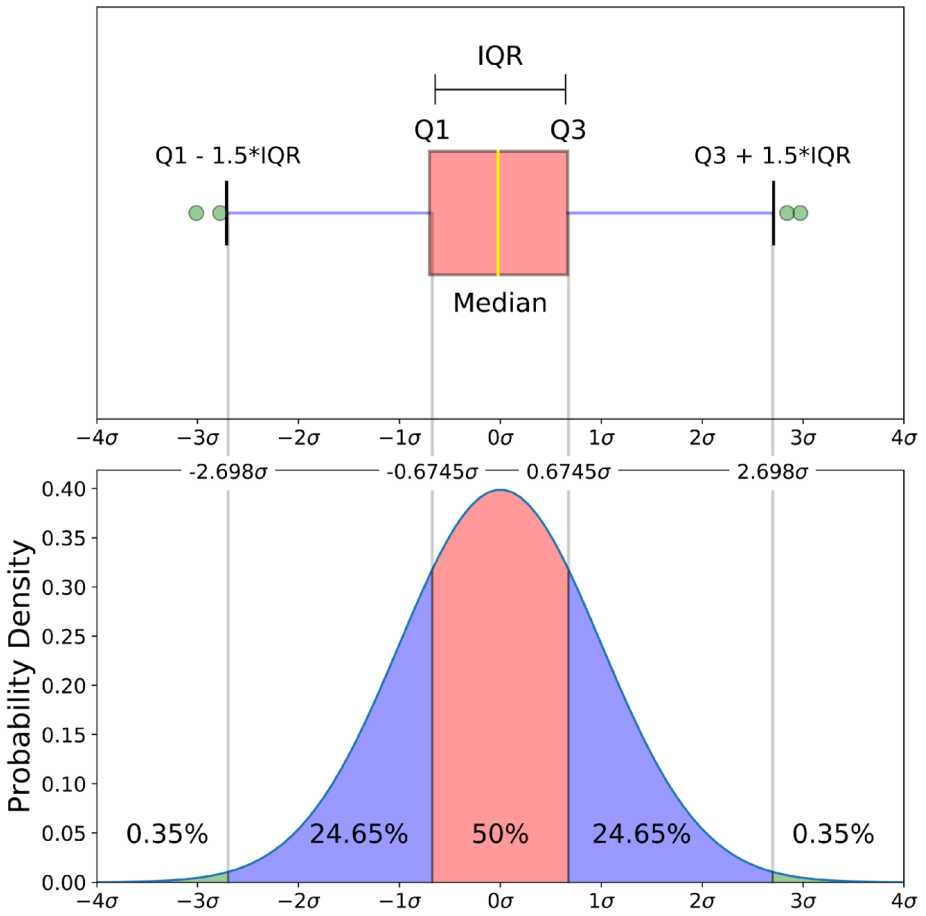
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



Box Plot (e.g., normal distribution)

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



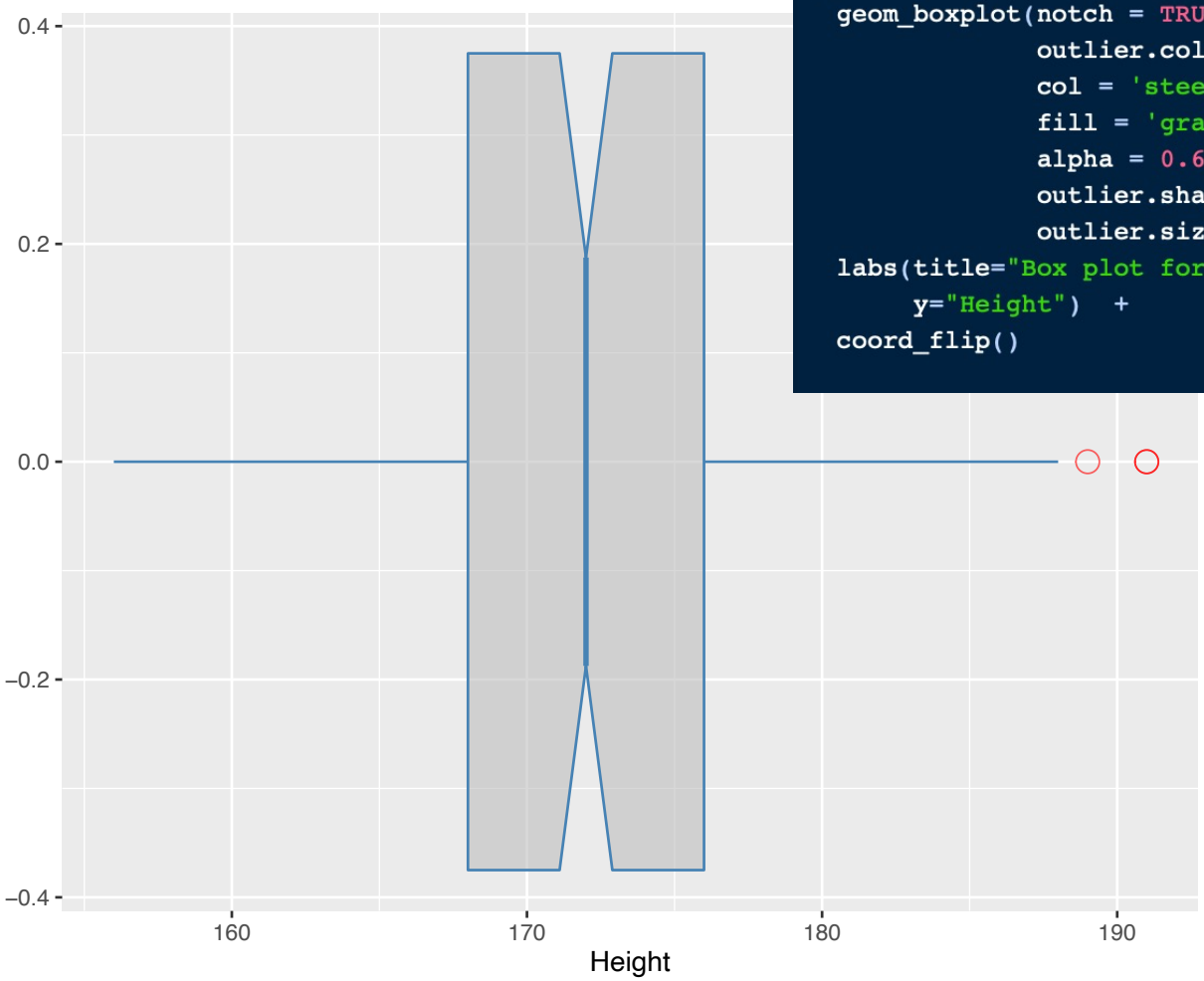
Comparison of a boxplot of a nearly normal distribution and a probability density function (pdf) for a normal distribution

Ref: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Box Plot

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary




```
ggplot(chol, aes(y=HEIGHT)) +
  geom_boxplot(notch = TRUE,
               outlier.colour="red",
               col = 'steelblue',
               fill = 'gray',
               alpha = 0.6,
               outlier.shape=1,
               outlier.size=4) +
  labs(title="Box plot for Height",
       y="Height")  +
  coord_flip()
```

QQ Plots

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- The quantile-quantile or QQ plot is an exploratory graphical plot used to check the validity of distributional assumption for a dataset.
  - QQ plots compare two probability distributions by plotting their quantiles against each other.
    - If the two distributions which we are comparing are exactly equal, then the points on the QQ plot will perfectly lie on a straight line (y = x).
- QQ plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.
- You can tell the type of distribution using the power of the QQ plot just by looking at the plot.

How To Create A QQ plot

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

Steps

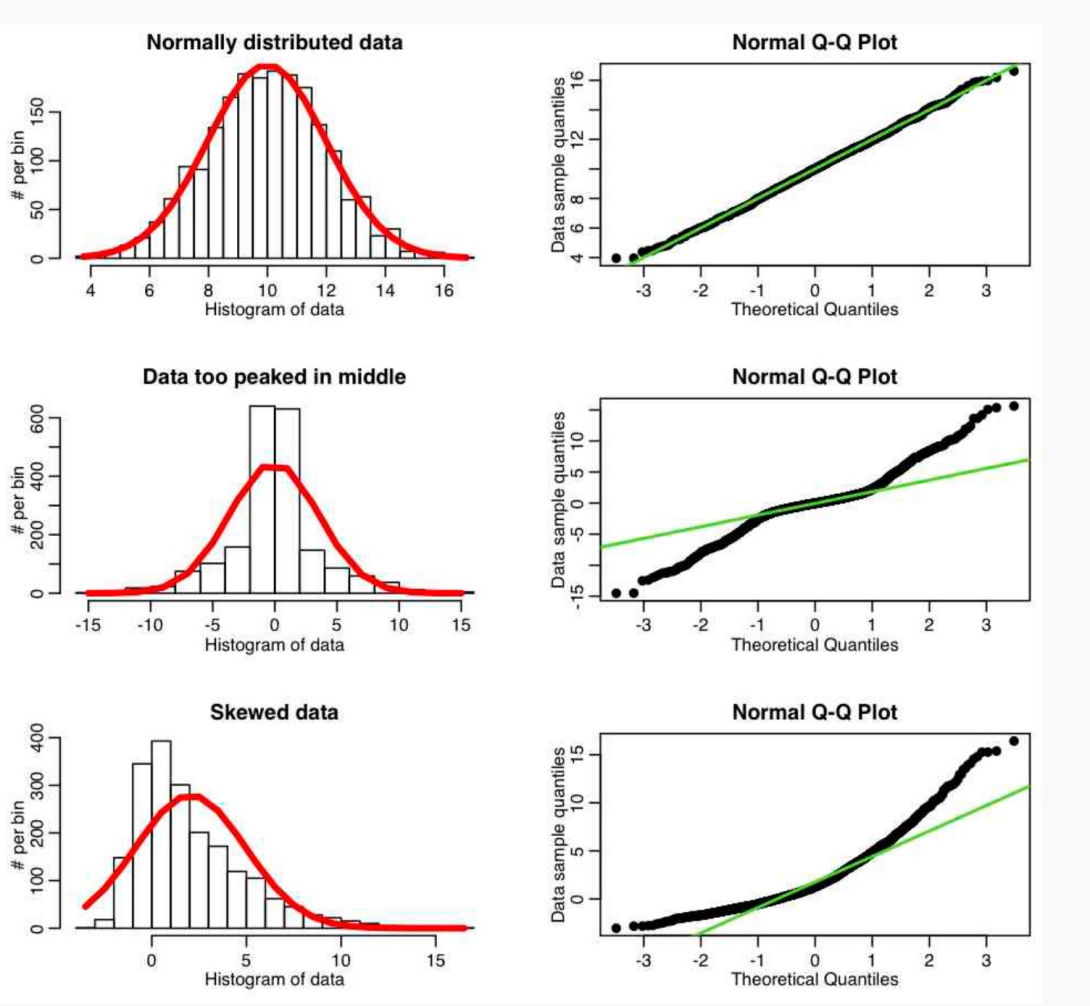
1. Plot the empirical distribution quantiles vs. the reference distribution.
2. For discrete distribution $x_i = Q_{\frac{i}{n}}, i = 1, \ldots, n$
3. For continuous distributions, modify the calculation of the fraction expressed in the quantile to account for the underlying distribution.
4. For a Gaussian distribution, use

$$x_i = Q_{\frac{i-.375}{n+.25}}$$

QQ Plots

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

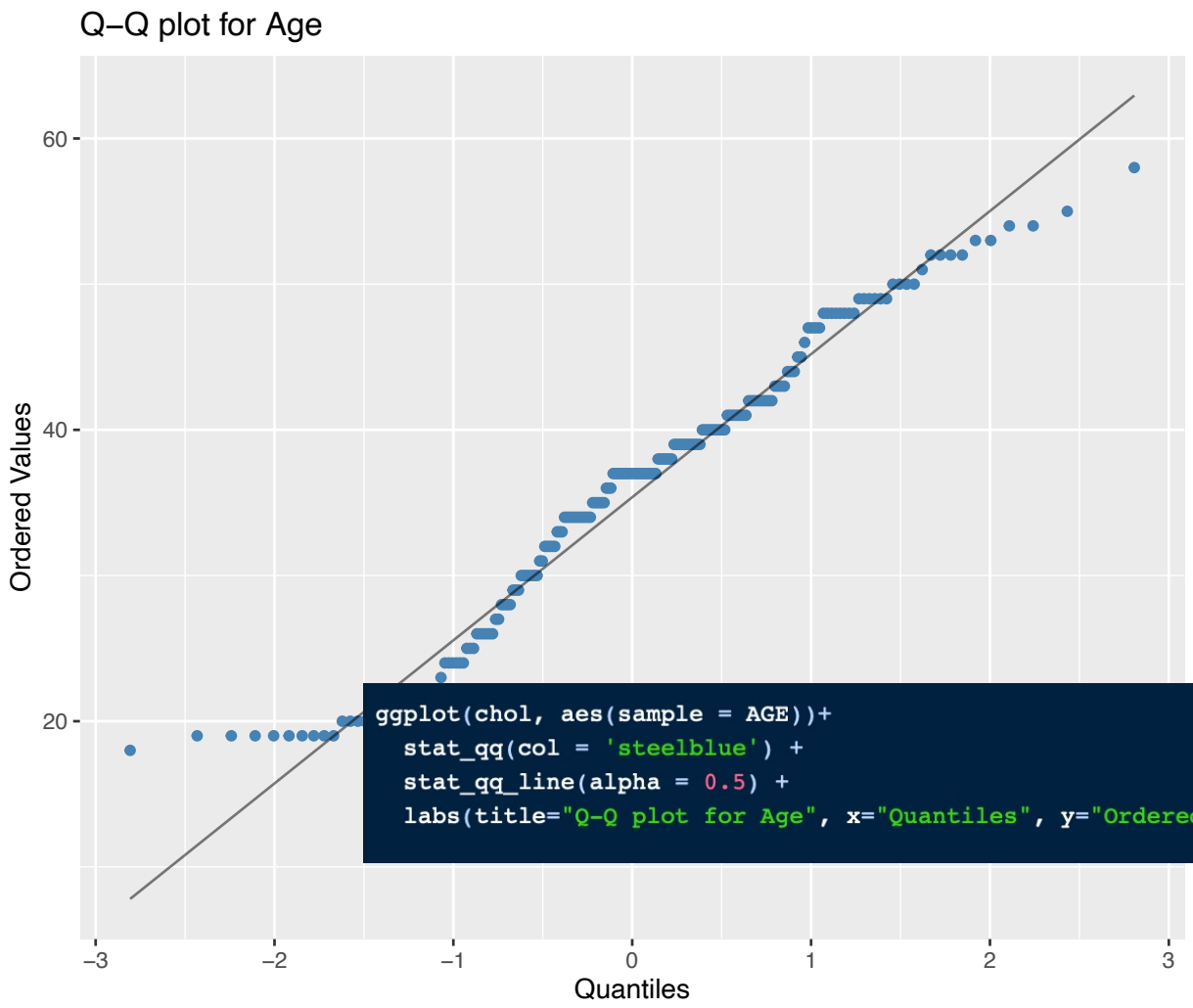


http://sherrytowers.com/2013/08/29/aml-610-fall-2013-module-ii-review-of-probability-distributions/qqplot_examples/

QQ Plot

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

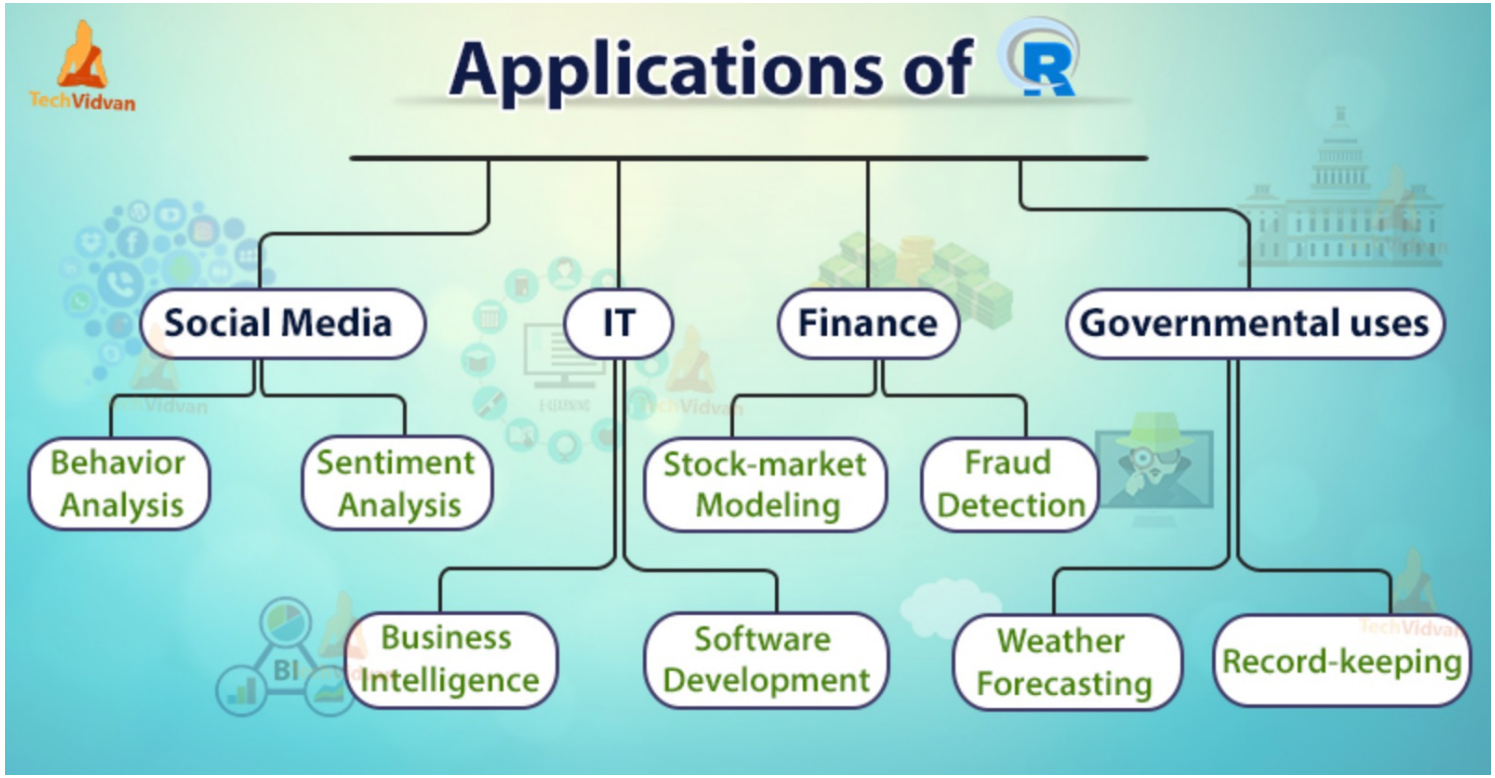



```
ggplot(chol, aes(sample = AGE))+
  stat_qq(col = 'steelblue') +
  stat_qq_line(alpha = 0.5) +
  labs(title="Q-Q plot for Age", x="Quantiles", y="Ordered Values")
```

R Session (part 1)

Agenda

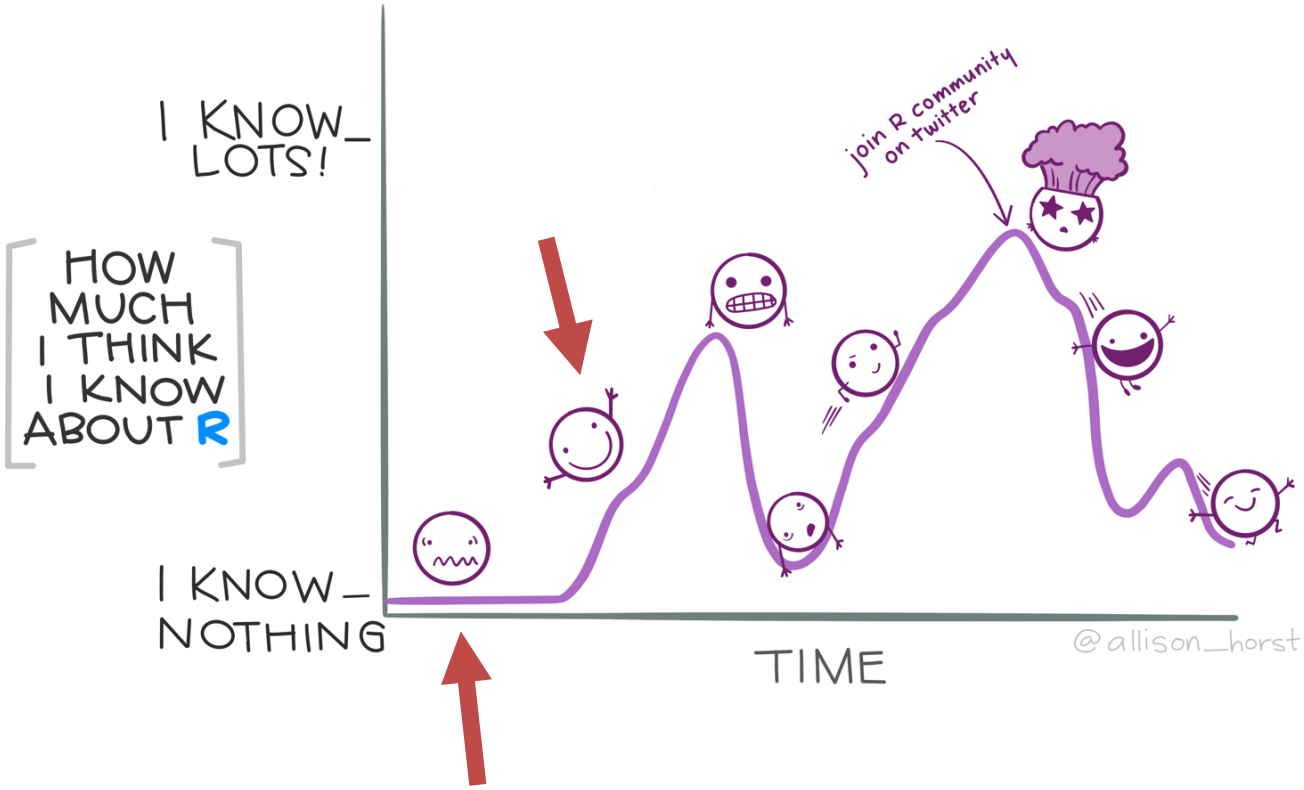
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



R Session (part 1)

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

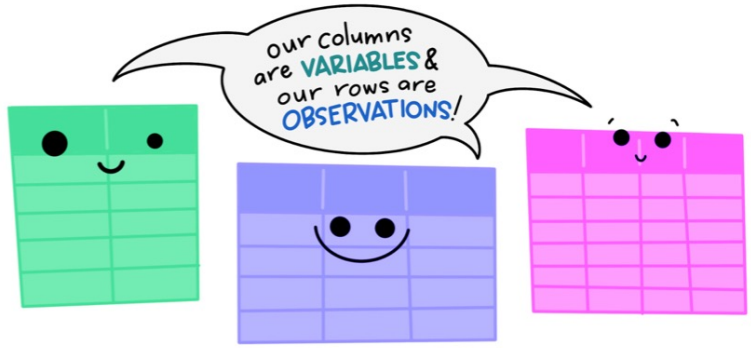


R Session (data cleaning)

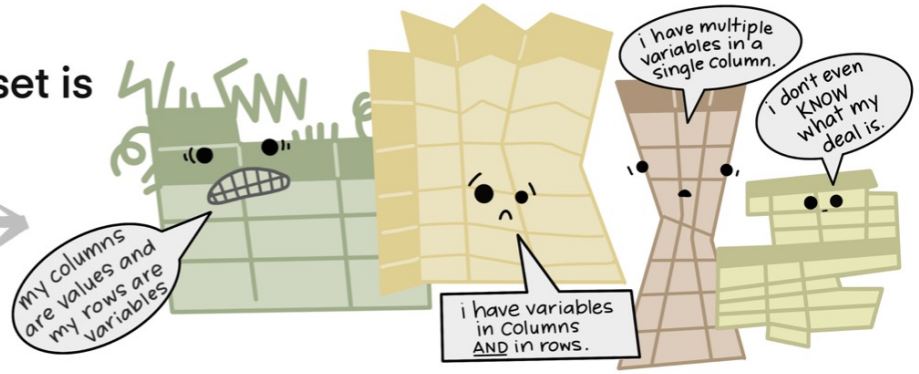
Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

The standard structure of tidy data means that "tidy datasets are all alike..."



"...but every messy dataset is messy in its own way."
–HADLEY WICKHAM

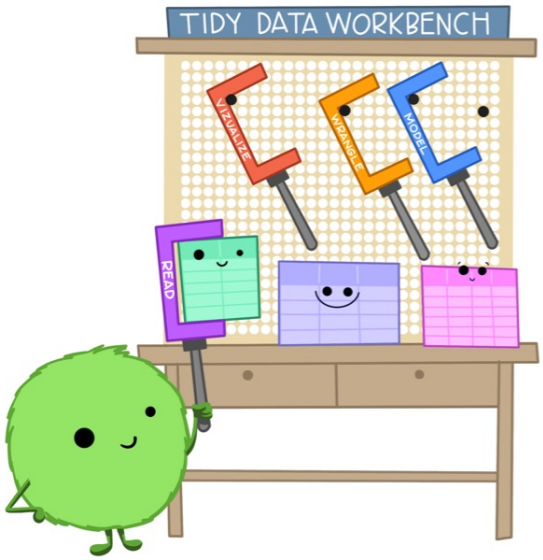


R Session (data cleaning)

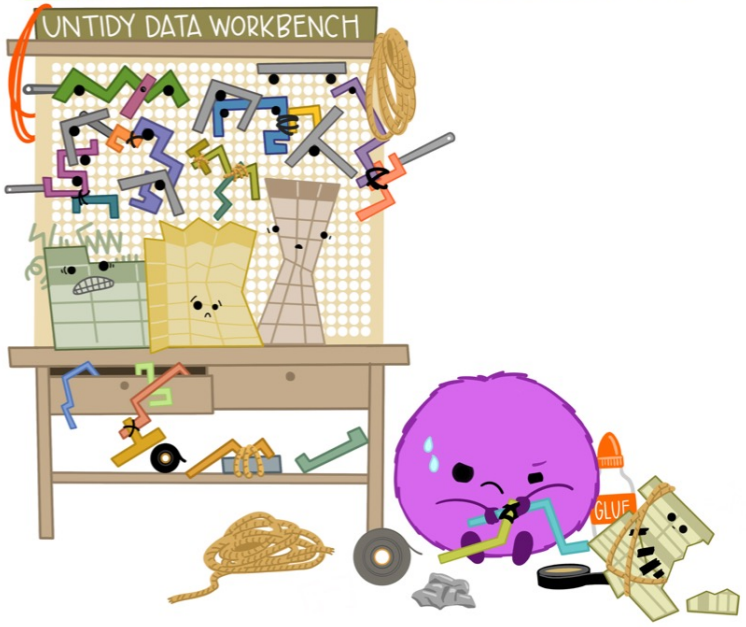
Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

When working with tidy data, we can use the same tools in similar ways for different datasets...



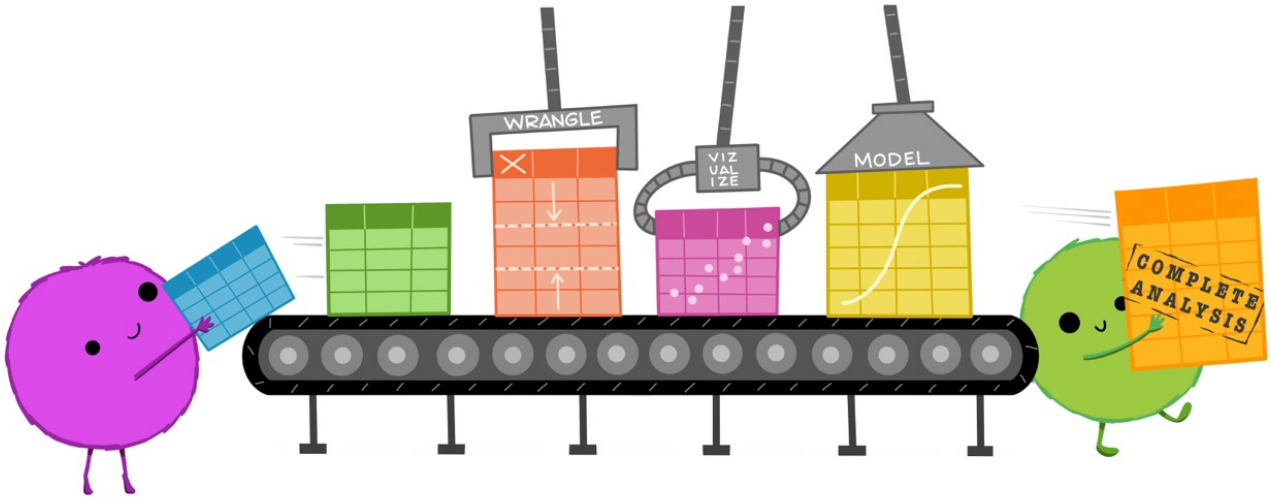
...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



R Session (data cleaning)

Agenda

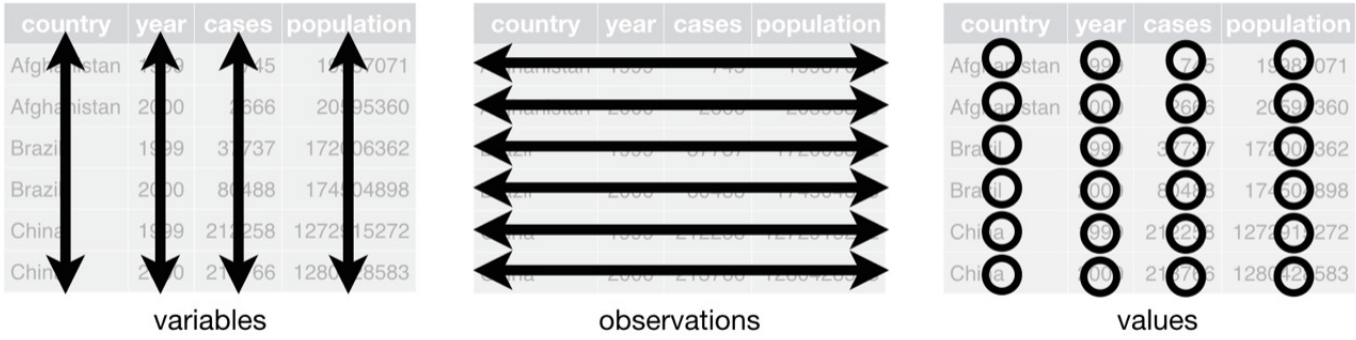
- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



R Session (data cleaning)

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

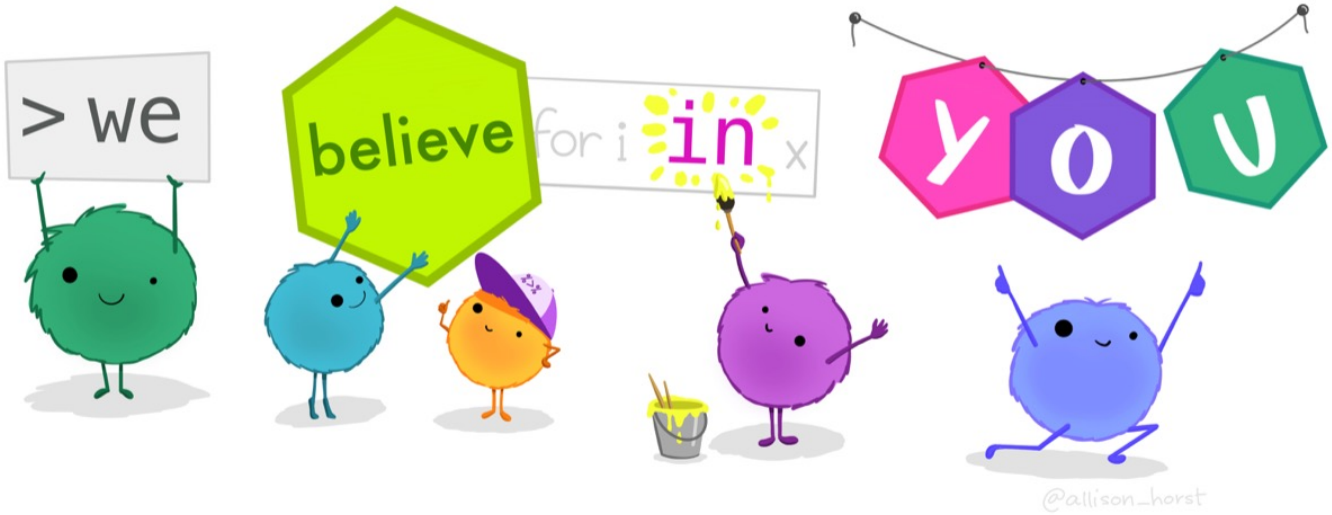


Ref: *https://r4ds.had.co.nz/tidy-data.html*

R Session (Part 1)

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary



Summary

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Visualization: *Univariate Analysis*
- R programming – part 1

Next Week

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- Class Activity
- Visualization: *Multivariate Analysis*
- R programming – Part 2

Additional Resources and References

Agenda

- Why Visualization?
- Univariate
  - Histograms
  - Density Plots
  - Bar plots
  - Box plots
  - QQ plots
- Summary

- https://www.r-graph-gallery.com/index.html
- https://www.datanovia.com/en/lessons/ggplot-scatter-plot/
- https://rb.gy/5rvsm6

Graphical Analysis and Visualization (11)

Laura Barnes

and

Julianne Quinn

Agenda

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

- Opportunity
- Why visualize?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

Why Visualize data?

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

1. To understand data
2. Show summary statistics
3. Distribution
4. Tell a story

Data analysis without data visualization is no data analysis

Univariate Visualization

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

- Univariate analysis:
  - Provides visual representation of summary statistics for one variable

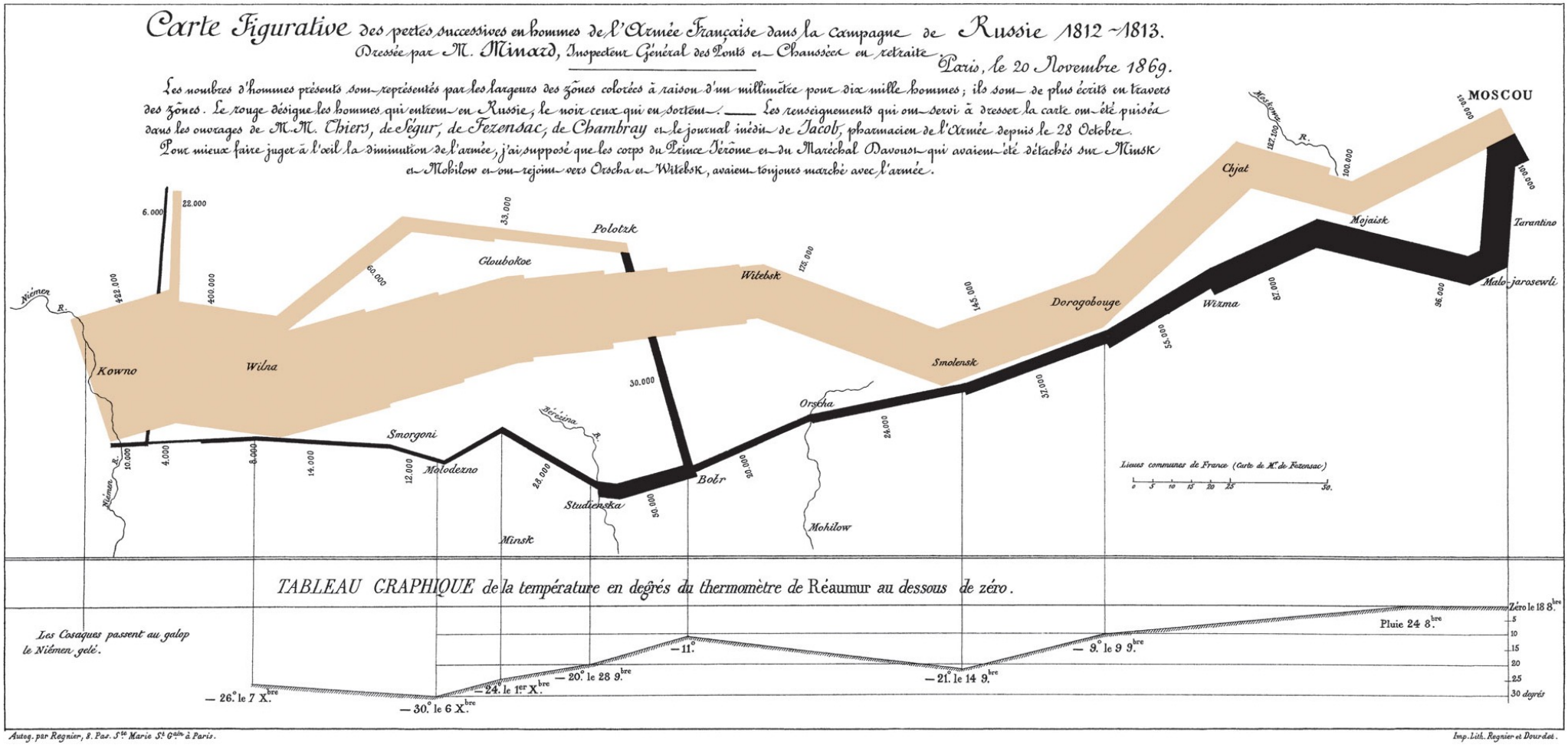
Multivariate Visualization

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

- Multivariate analysis:
  - Performed to understand relationship between 2 or more variables in a dataset

Multivariate Visualization



Scatter Plot

Agenda

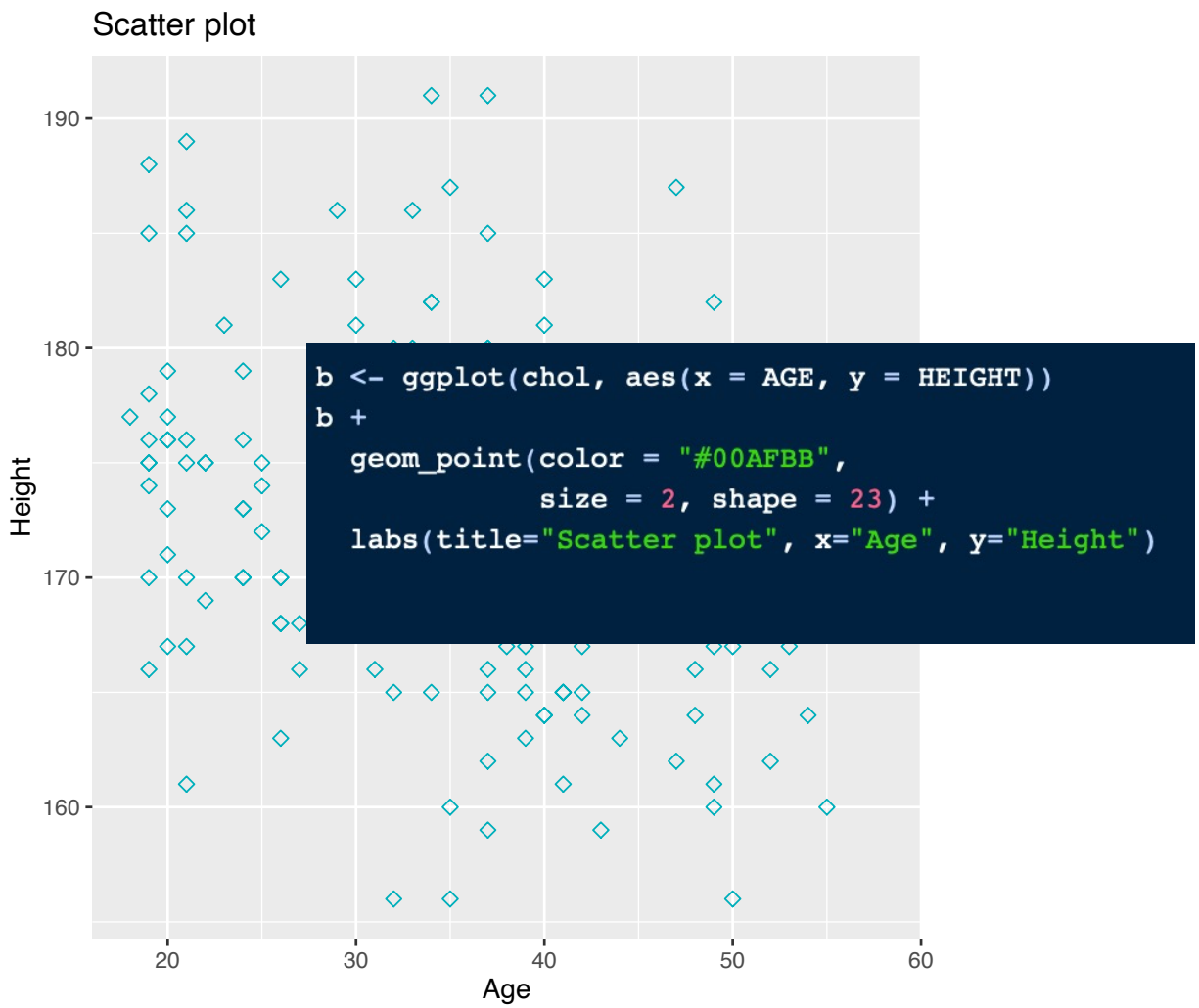
- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

- A Scatter plot (also known as X-Y plot or Point graph) is used to display the relationship between two continuous variables x and y.
- By displaying a variable in each axis, it is possible to determine if an association or a *correlation* exists between the two variables.
  - The correlation can be:
    - positive (values increase together),
    - negative (one value decreases as the other increases),
    - null (no correlation),
    - linear,
    - exponential and
    - U-shaped.

Scatter Plot (e.g., age vs height)

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

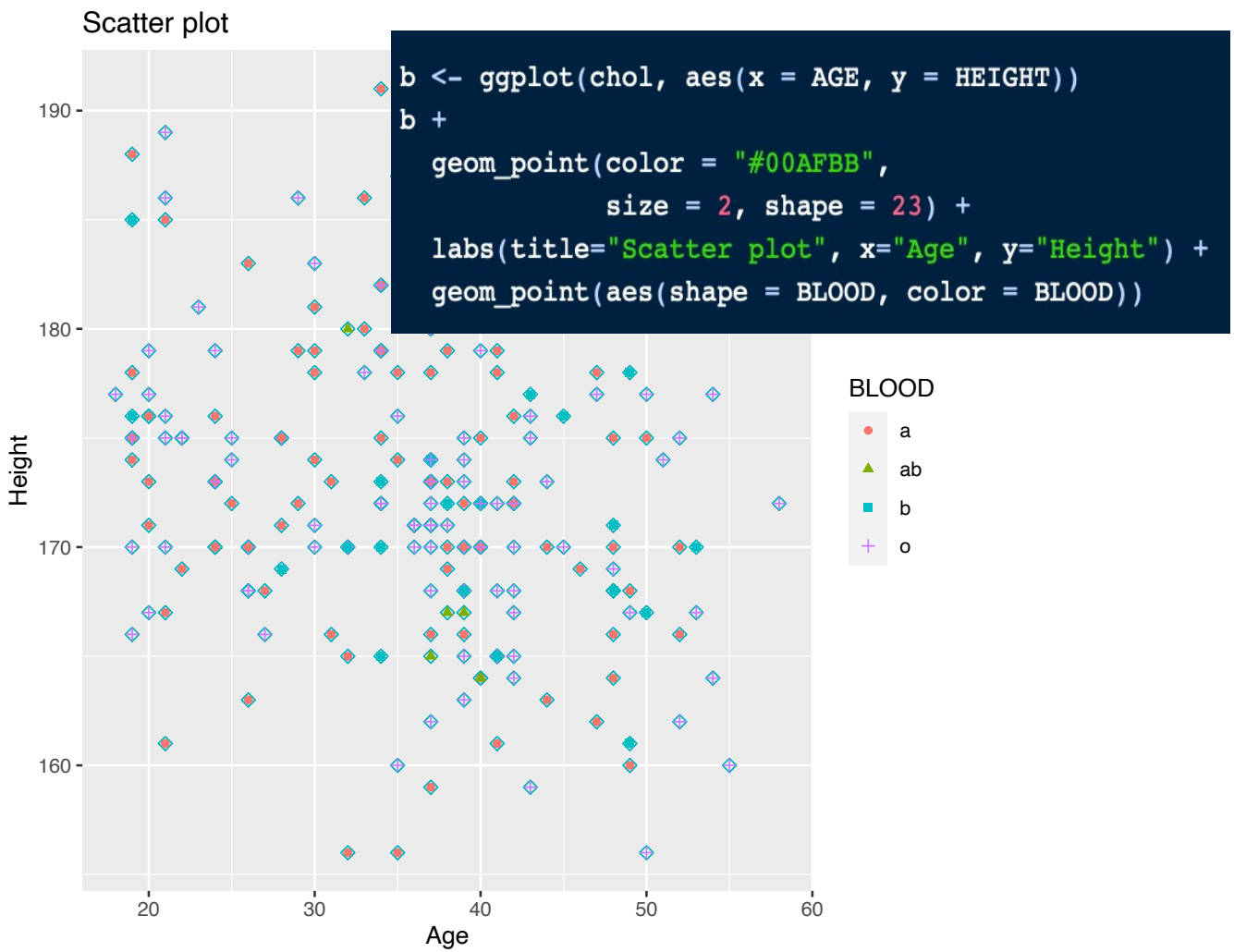



```
b <- ggplot(chol, aes(x = AGE, y = HEIGHT))
b +
  geom_point(color = "#00AFBB",
             size = 2, shape = 23) +
  labs(title="Scatter plot", x="Age", y="Height")
```

Scatter Plot (e.g., age vs height and blood)

Agenda

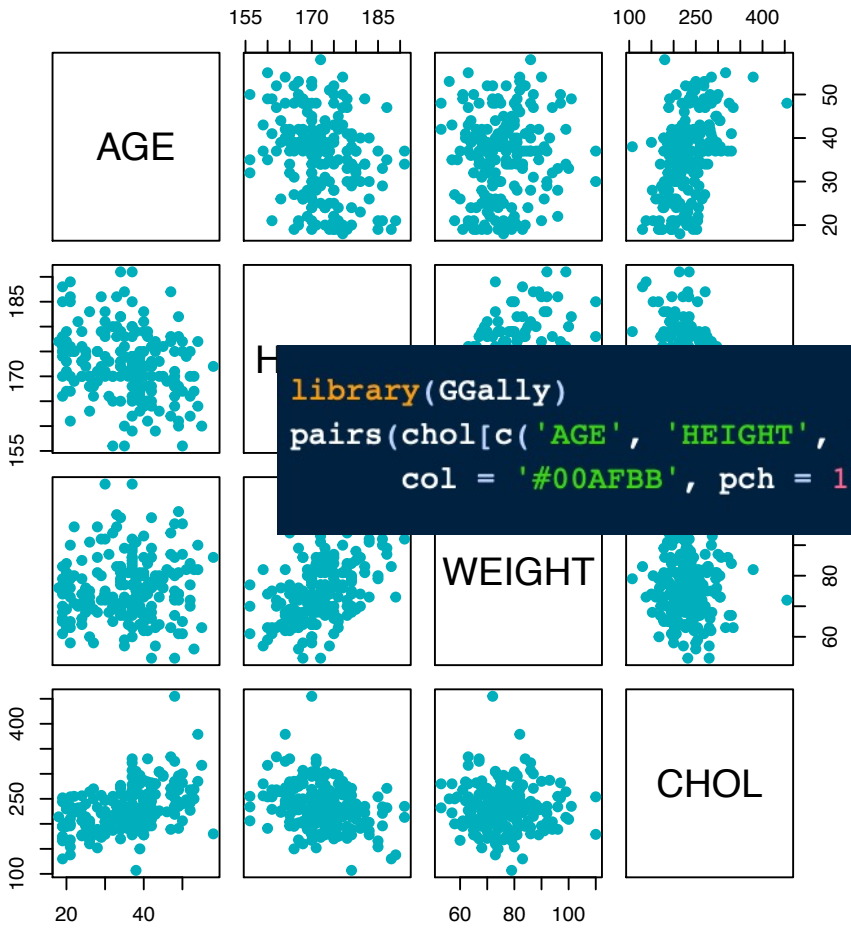
- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary



Scatter Plot Matrix

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary



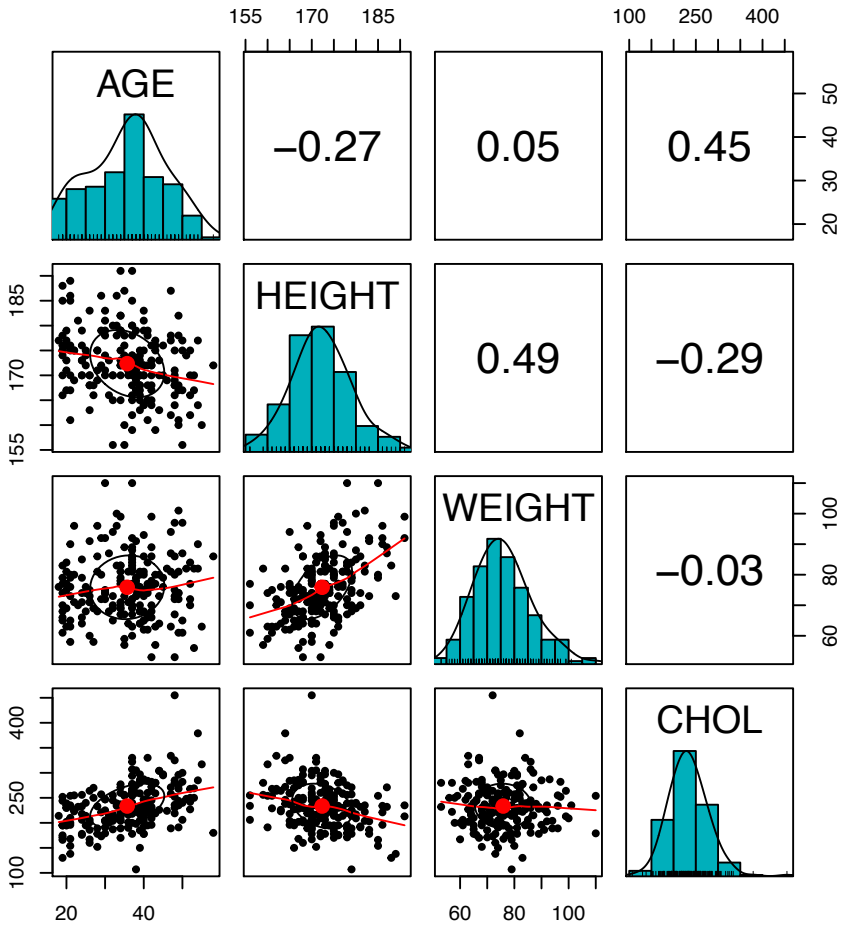

```
library(GGally)
pairs(chol[c('AGE', 'HEIGHT', 'WEIGHT', 'CHOL')],
      col = '#00AFBB', pch = 19)
```

Scatter Plot Matrix

```
library(psych)
pairs.panels(chol[c('AGE', 'HEIGHT', 'WEIGHT', 'CHOL')],
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
             )
```

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary



Example

Agenda

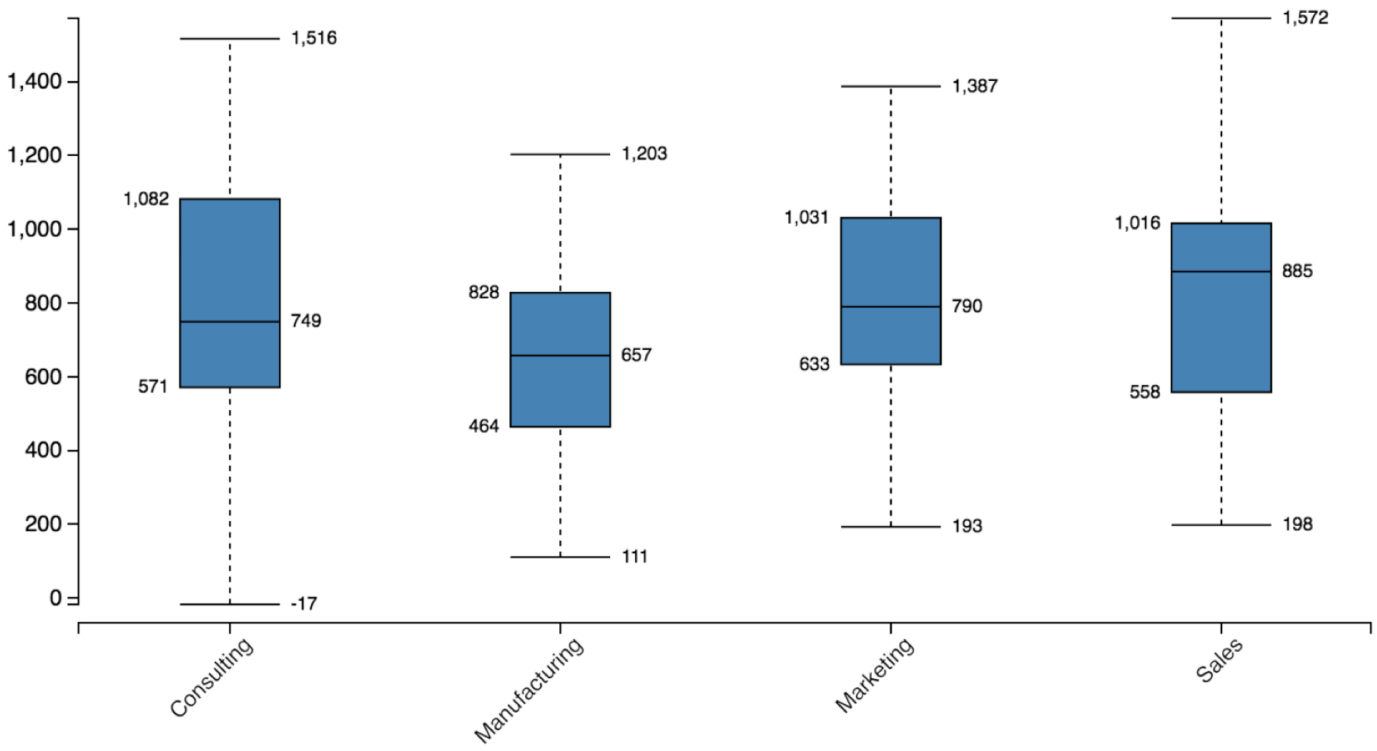
- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

- Suppose Trader Joe's is spending money on Fig Jams.
- This includes the costs in consulting, manufacturing, marketing, and sales departments.
- Now, the company wants to understand minimum and maximum values spent in different departments.
  - Which plot they should use?

Categorical Variable Box Plot

Agenda

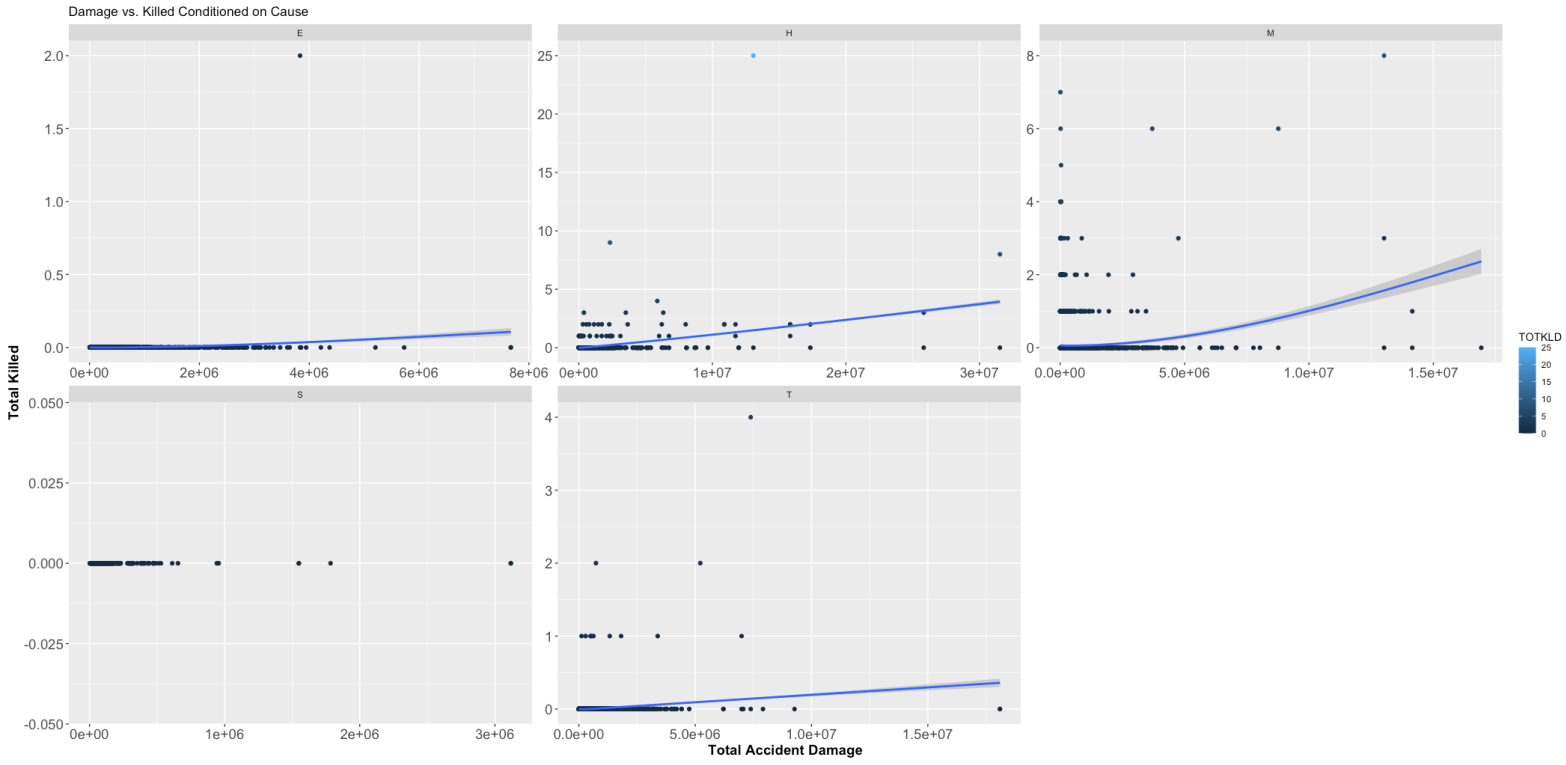
- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary



Conditional Scatter Plot

Agenda

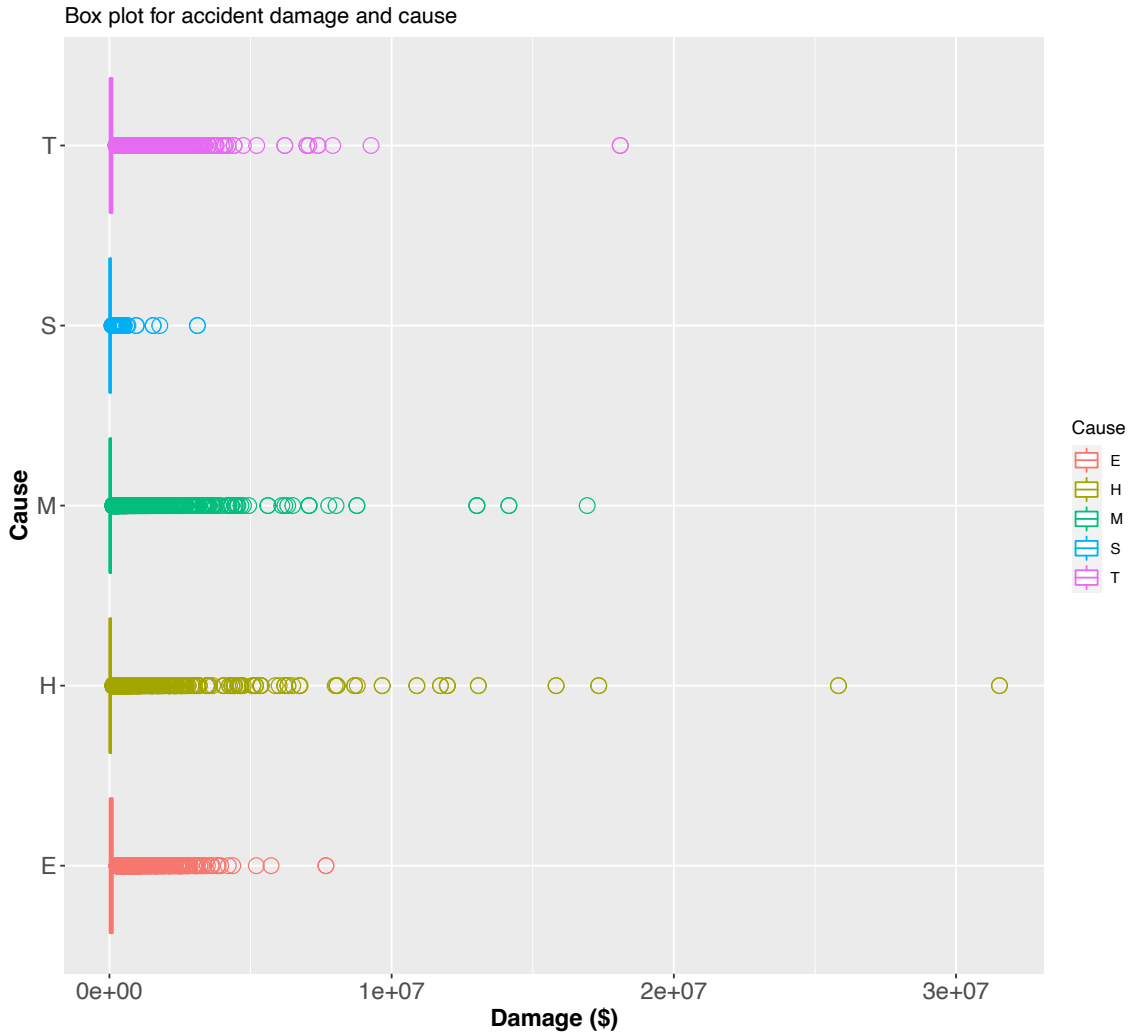
- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary



Conditional Boxplot

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

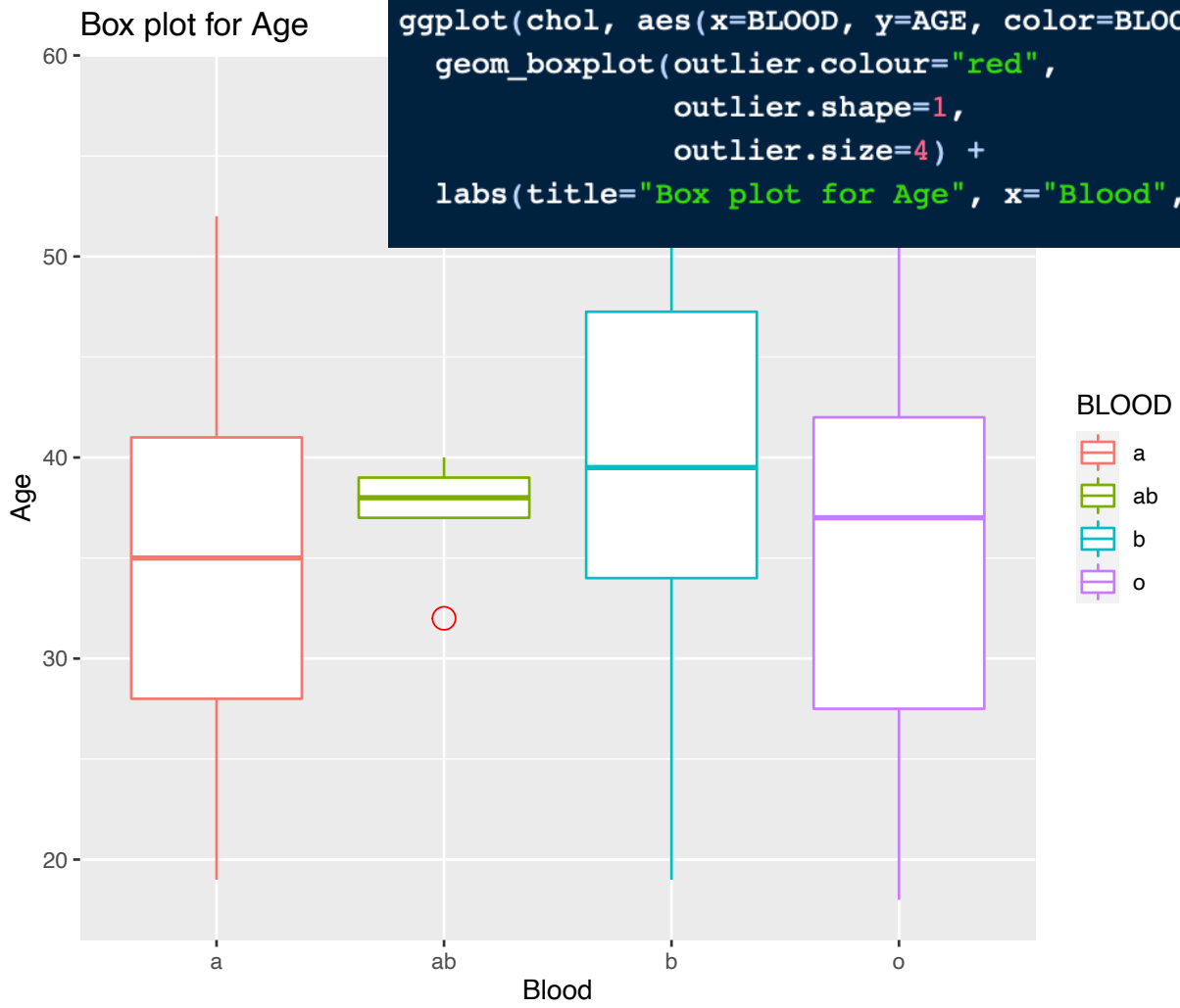


Conditional Box Plot

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

```
ggplot(chol, aes(x=BLOOD, y=AGE, color=BLOOD)) +
  geom_boxplot(outlier.colour="red",
               outlier.shape=1,
               outlier.size=4) +
  labs(title="Box plot for Age", x="Blood", y="Age")
```



Other Methods

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

- 3-D scatter plot (scatterplot3d)
- Rotate and spin (e.g., using rgl or aplpack libraries)
- Parallel coordinates plots (parcoord in MASS)
- Glyphs
- …

Summary

Agenda

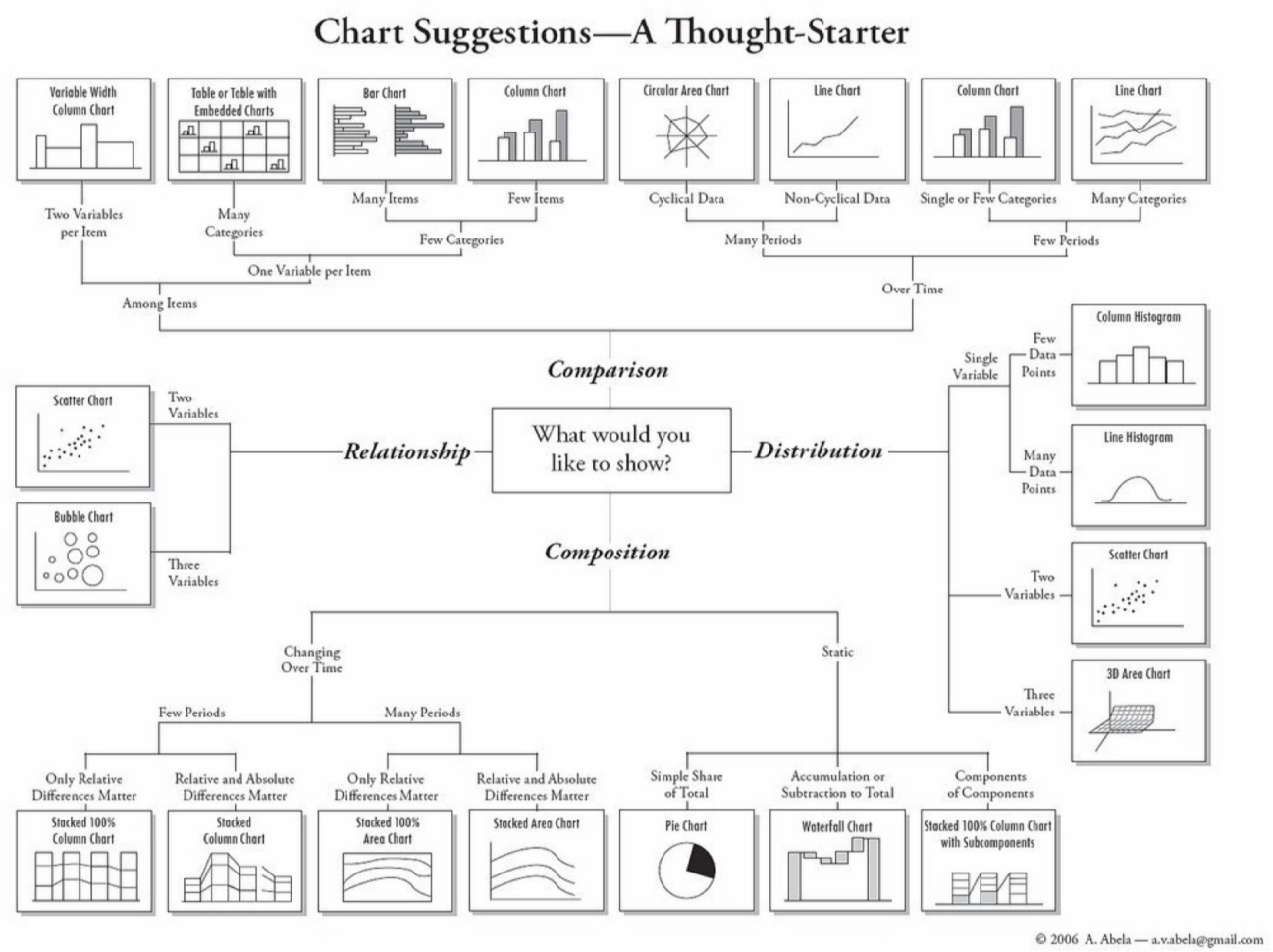
- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

- Univariate plots provide a quick and convenient to gain understanding about the characteristics of single variable distributions
- Whenever we use visualization we must always guard against seeing what we expect to see.
- Viewing multiple dimensional objects and data in two dimensions has provided a challenge that humans have worked on for 10's of thousands of years.
- Multidimensional data give us insight into problem solving

Summary

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary



Additional Resources and References

Agenda

- Why visualization?
- Univariate
- Multivariate
  - Scatter plots
  - Categorical plots
  - Others
- Summary

- https://www.r-graph-gallery.com/index.html
- https://www.datanovia.com/en/lessons/ggplot-scatter-plot/
- https://rb.gy/5rvsm6