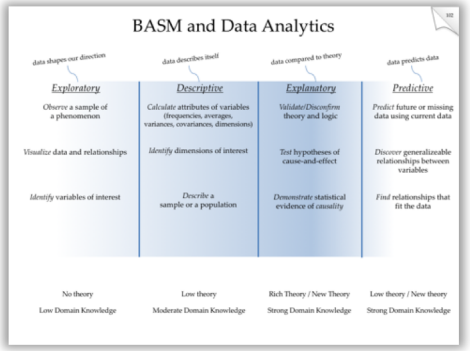
Business Analytics Using
Computational Statistics

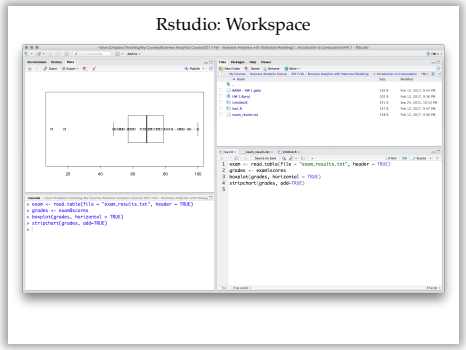
Week 1
Computation and Statistics

Week 2
Describing and Visualizing Data

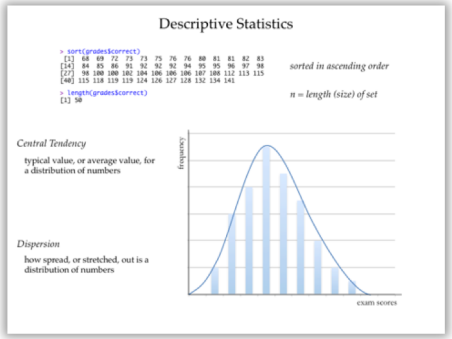
Computational Statistics



Our Tools



First Look at R



This **class is about** *programming* and *statistics*
We will **not cover** *machine learning* or *AI*
Homework + Peer Review **work every week**!
There are **many easier ways to learn R** than taking this class

Unregistered Students:

Only students who have already contacted me can add the class ☹️

Canvas

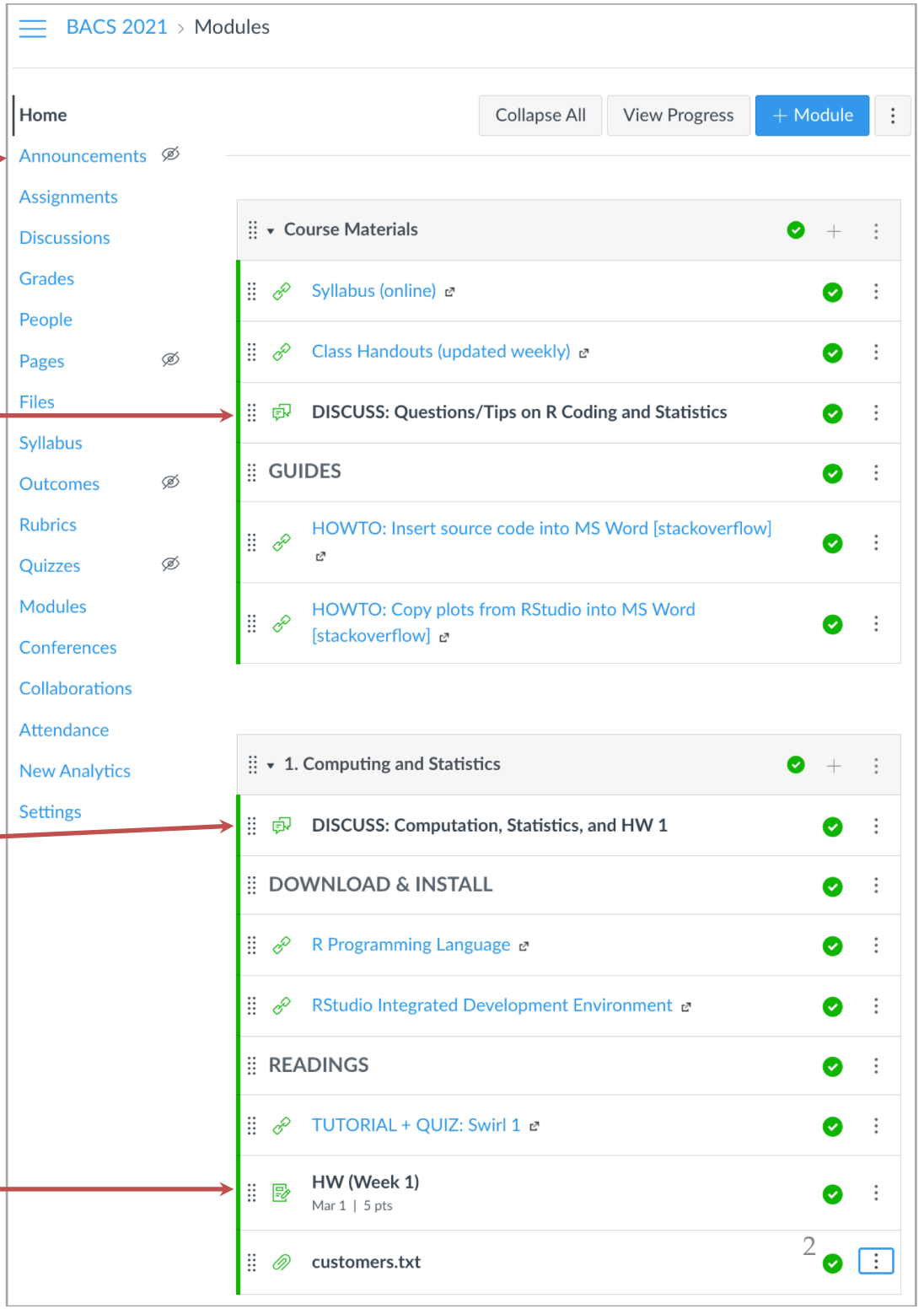
https://canvas.instructure.com/courses/2595721

Weekly Announcements

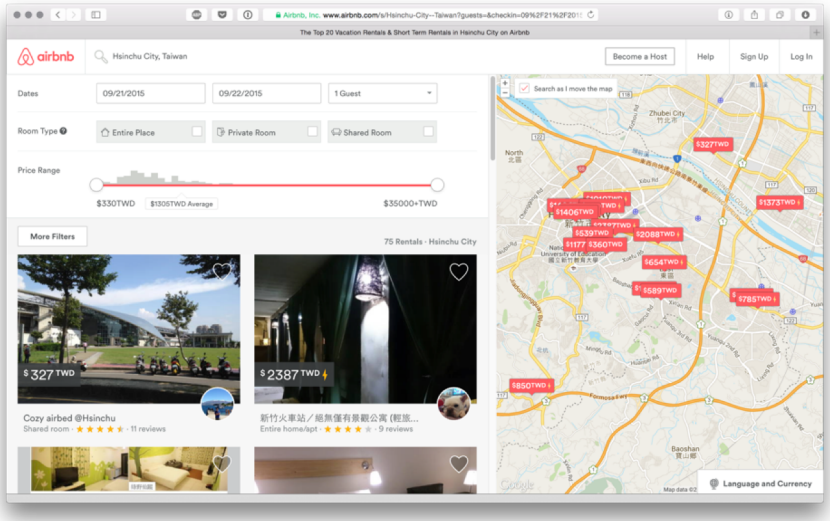
Discuss general topics on R and Statistics

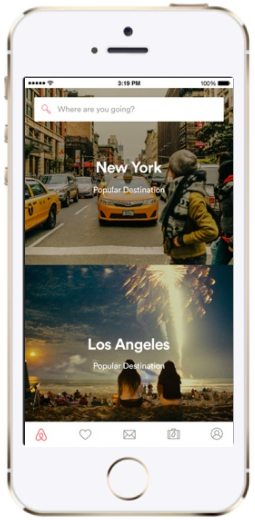
Discuss questions/solutions freely
(it's not cheating if you share openly!)

Submit assignment solutions as PDF reports
- Download data files for assignments



Data Science & Analytics





Job Openings

- **Analytics**
- **Algorithms**
- **Inference**
- **Machine Learning**

Are these all different from each other?

Why does AirBnB make these distinctions?

BACS and Data Analytics

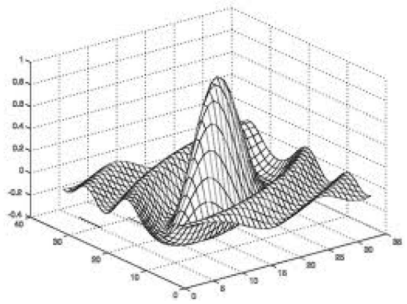
data suggests new directions	**data** describes itself	**data** compared to **theory**	**data** predicts **data**
Exploration	***Description***	***Inference***	***Prediction***
Observe a sample of a phenomenon	*Calculate* attributes of variables (frequencies, averages, variances, covariances, dimensions)	*Validate/Disconfirm* theory and logic	*Predict* future or missing data using current data
Visualize data and relationships	*Identify* dimensions of interest	*Test* hypotheses of cause-and-effect	*Discover* generalizeable relationships between variables
Identify variables of interest	*Describe* a sample's statistics	*Estimate* population statistics	*Find* relationships that might fit the data
No theory available	**Low theory**	**Rich Theory / New Theory**	**Low theory / New theory**
Low Domain Knowledge	**Moderate Domain Knowledge**	**Strong Domain Knowledge**	**Strong Domain Knowledge**

Data Scientist - **Inference**

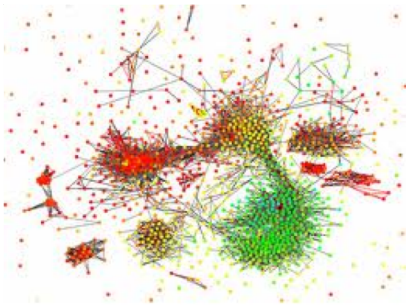
Commonly listed Needs:

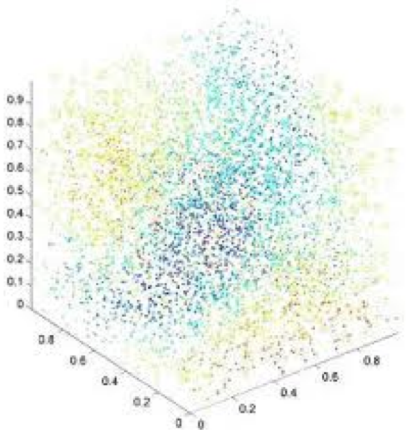
- **Define metrics** to accurately **measure** our progress
  🤔 *What is the difference between metric and measure?*
- Ensure our **understanding** of product changes is rigorous and accurate
  🤔 *Why is understanding important?*
- **Find anomalies** in transactions
  🤔 *How do companies find anomalies?*
- **Evolve our statistical models** of user lifetime value
  🤔 *What are models?*

Understanding **Shape** and **Relationships** in Data

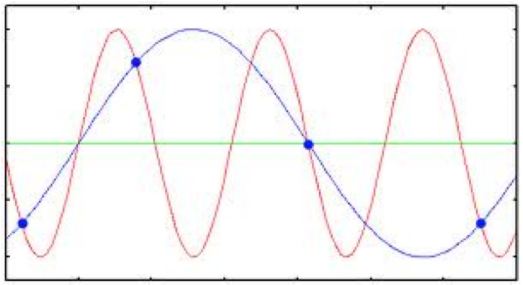


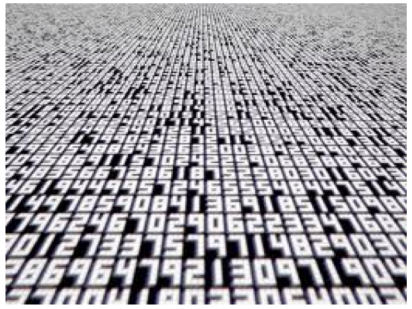
Uncovering
geometry and ***dimensionality***
of data



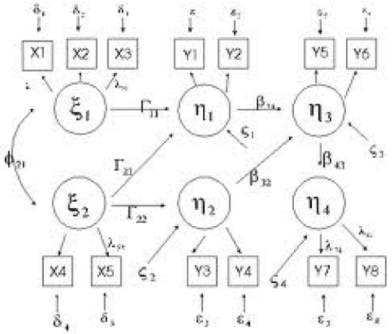


Removing the ***noise***
Finding the ***signal***





Understanding ***relationships***
Defining complex ***models***



Computational Statistics in Practice

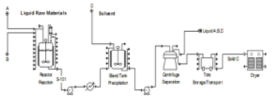
Scientific Management

Management Science

"Scientific Management"

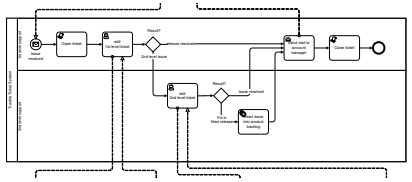
"Inventory Control"

"Total Quality Management"



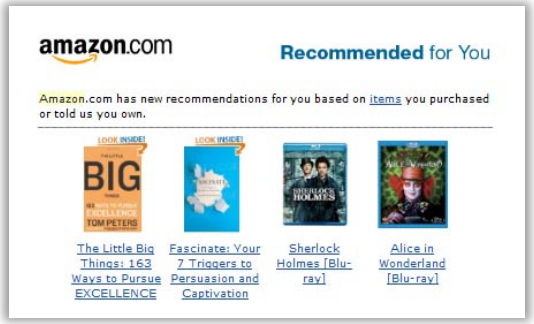
Business Process Maturity Model (BPMM)

1. Poorly defined and inconsistent practices
2. Repeatable practices at the workgroup level
3. Standard organization-wide end-to-end processes
4. **Statistically-managed and predictable processes**
5. Continuous process innovation and optimization



Service Innovation

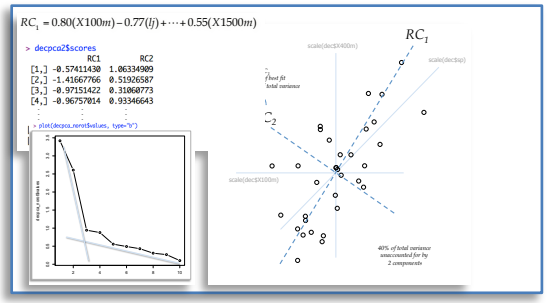
Product Tools



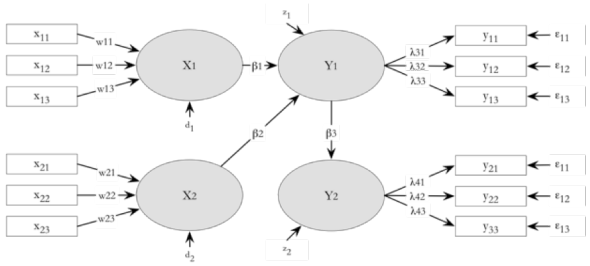
A/B Testing

Social Sciences

Data Exploration



Complex Models



Data Scientist - **Inference**

Commonly sought after abilities:

- Ability to write **clean and concise code**

  🤔 *Why is coding important? Why is "clean and concise" code important?*

- Solid **understanding of statistics** and online **experiment design**

  🤔 *Isn't data enough? Why are experiments so important?*

- Strong **analytical** and **communication skills**

  🤔 *How do we show "communication skills"?*

Computational Statistics

Traditional Statistical Methods

Central Tendency, Dispersion

Hypothesis Testing

Relationships between Variables

ANOVA

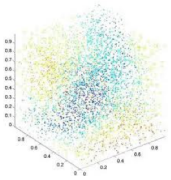
Regression

Power and Error

Dimensionality and PCA

Path Modeling

New Perspectives on Data

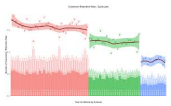


Data Simulation

Data Resampling

Geometric Interpretation

New Abilities as Analysts



Basic Programming

Advanced Visualization

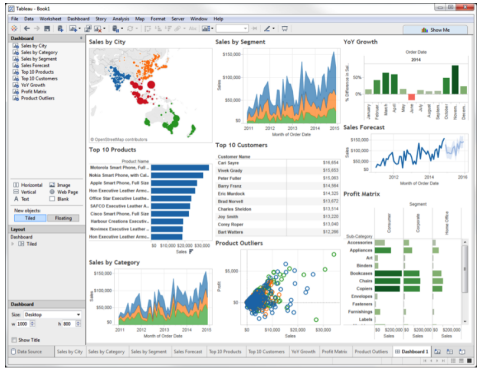
New Tools in Your Toolbelt

R programming

RStudio

Traditional Business Analytics Tools

Graphical User Interface



Easy to Learn and Use

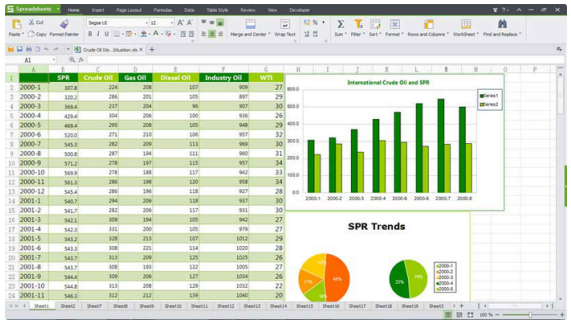
Easy to Analyze Data

Easy to Visualize and Communicate

🤔

*What are the **downsides** of being limited to GUI and spreadsheet tools?*

Spreadsheets



Familiar Metaphor (Balance Books)

Easy to Manipulate Data

Quick Results

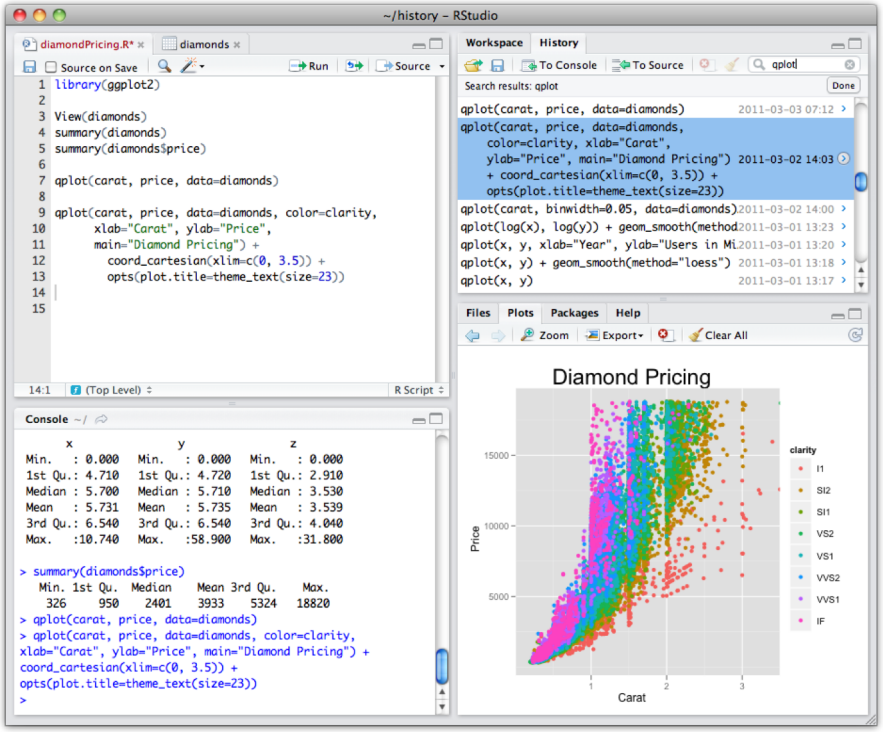
Computational Statistics

Analytics Advantage

Run simulations/scenarios
See what happens when assumptions change

Create novel solutions
Make new ways of solving problems

Generate custom visualizations
Allow others to see an understand your solution



Social Analytics Advantage

Sharable
Give your solution to others

Repeatable
Others can run your code with same results

Testable
Ensure your code produces the right results

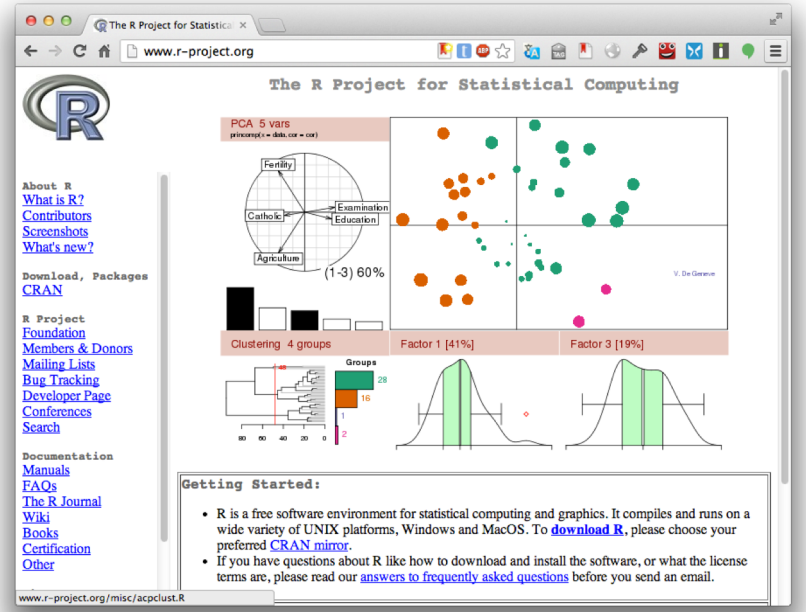
Deployable
Run your code as part of a product platform

*What are the **downsides** of being limited to programmatic tools?*

R and RStudio

Install **R**

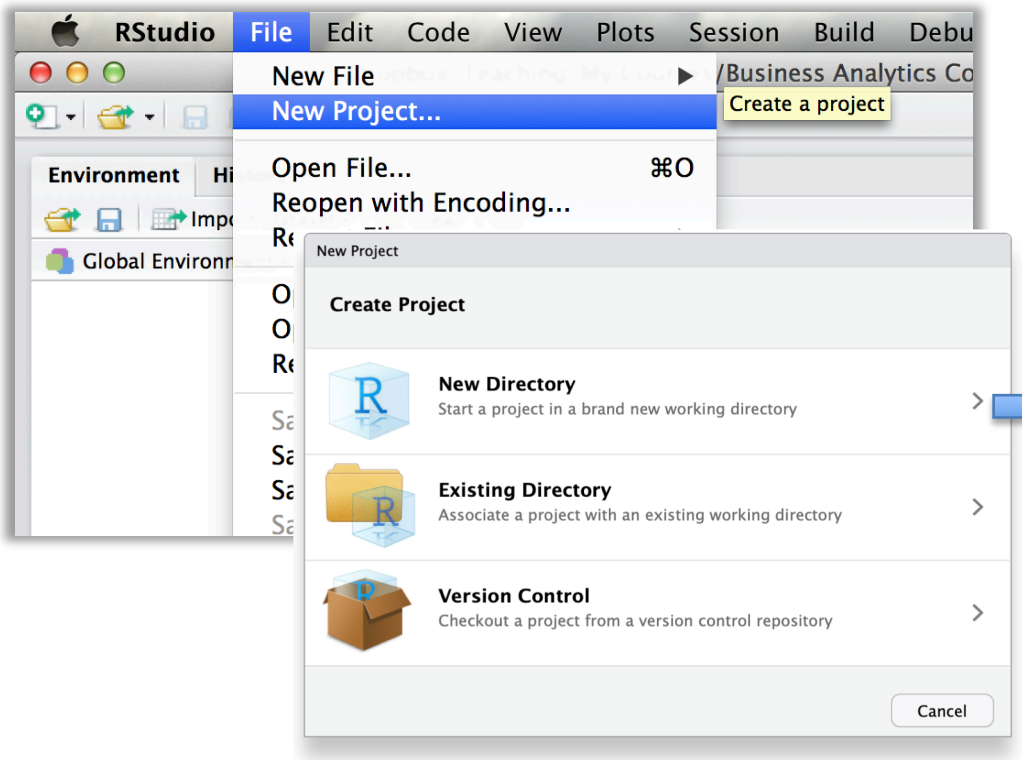
http://www.r-project.org/

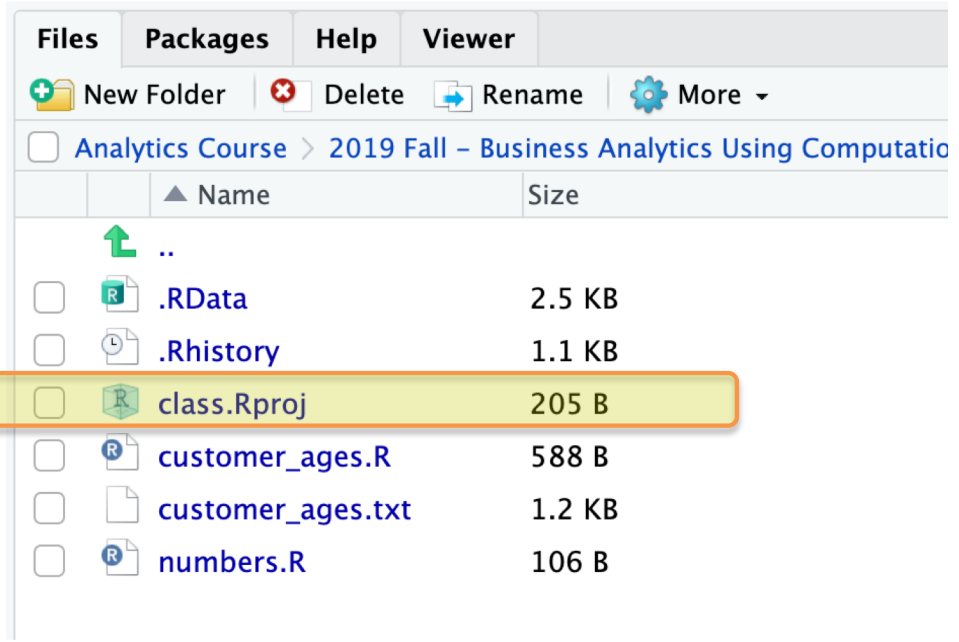


Install **RStudio**

http://www.rstudio.com/

RStudio: Using Projects



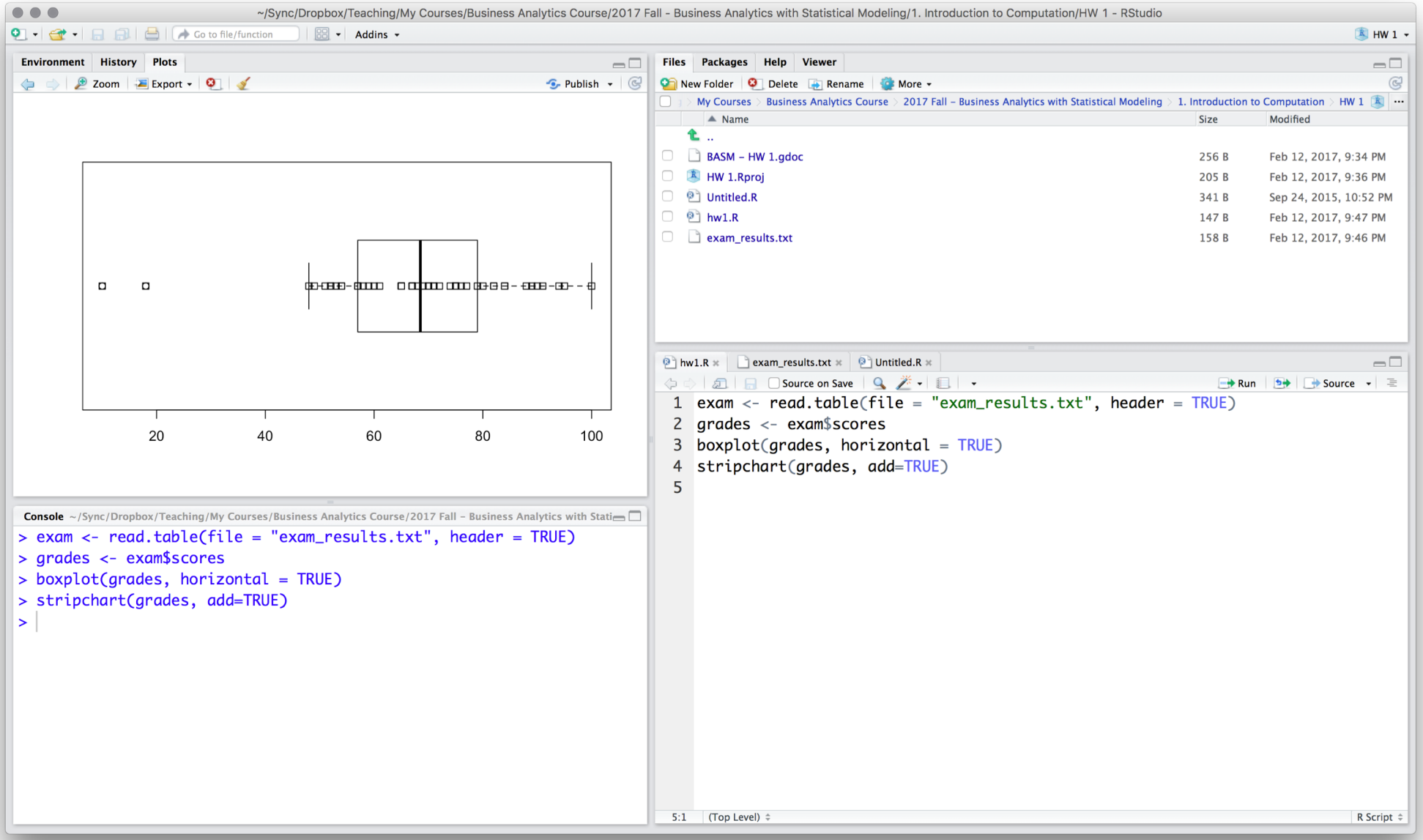


Make a new RStudio project for each:

- Homework assignment
- Research project

https://support.rstudio.com/hc/en-us/articles/200526207-Using-Projects

RStudio: Workspace



Learning R

At Home

http://swirlstats.com

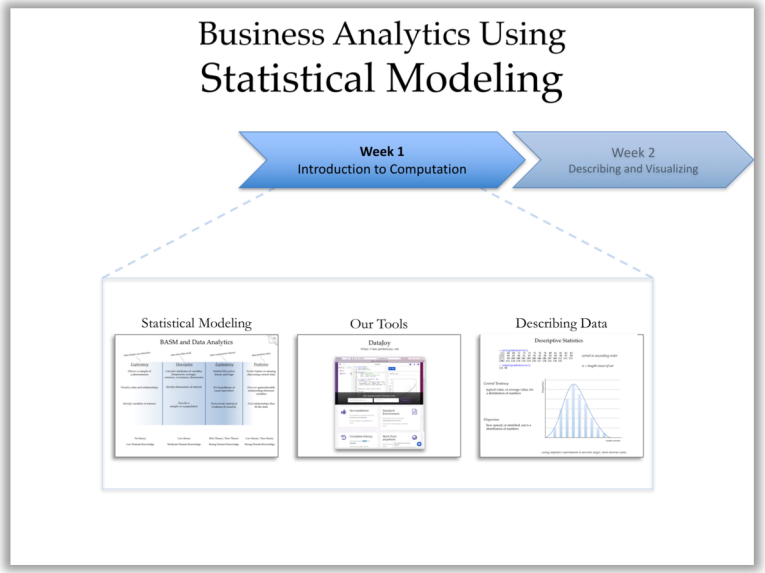
Data structures

Functions

Packages

Loops and Iterations

In Class



Different paradigms in analytic computing

Benchmarking code for performance

Writing re-readable, reusable code

Packaging your code for others

R: Basic Commands

Creating a *vector* of data

```
> numbers <- c(12, 14, 14, 25, 33, 35, 38, 38, 41, 43, 45, 50, 58, 59)
> numbers[5]
[1] 33

> sum(numbers)
[1] 505

> seq(3,7)
[1] 3 4 5 6 7
```

Vector: *a sequence of data elements of the same type*

Function: *a sequence of code that can be called to perform an action*

Loading data into a *data frame*

Data Frame: *collection of named vectors*

```
> customers <- read.table(file = "customers.txt", header = TRUE)

> customers$age
```

`$` *extract named vectors out of a data frame*

```
  [1] 49 69 41 73 45 71 50 43 70 32 47 77 64 50 50 45 49 47 62 50 47 72 47 63 21
 [26] 49 50 48 35 77 48 48 50 47 29 42 42 85 45 49 45 43 49 68 42 48 72 79 48 50
  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
[376] 48 18 45 62 41 71 19 73 26 75 41 46 49 49 23 74 53 23 51 71 50 50 67 74
```

```
> length(ages)
 [1] 399
```

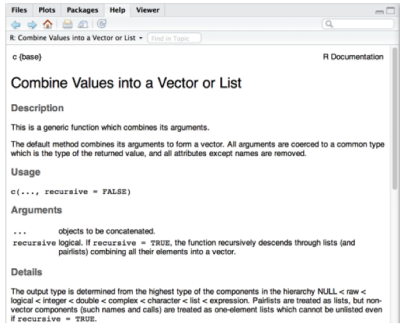
`length()` *: size of a data frame*

Data frame variable

```
> customers
     age
1     49
2     69
3     41
4     73
5     45
.    ...

398   67
399   74
```

Help function

```
> help(c)
```



1. A reference website for R:
http://cran.r-project.org/doc/manuals/R-intro.html

2. R Introduction
http://www.r-tutor.com/r-introduction

2. What is: c(…)? It makes a vector (combination of values)
http://stat.ethz.ch/R-manual/R-devel/library/base/html/c.html

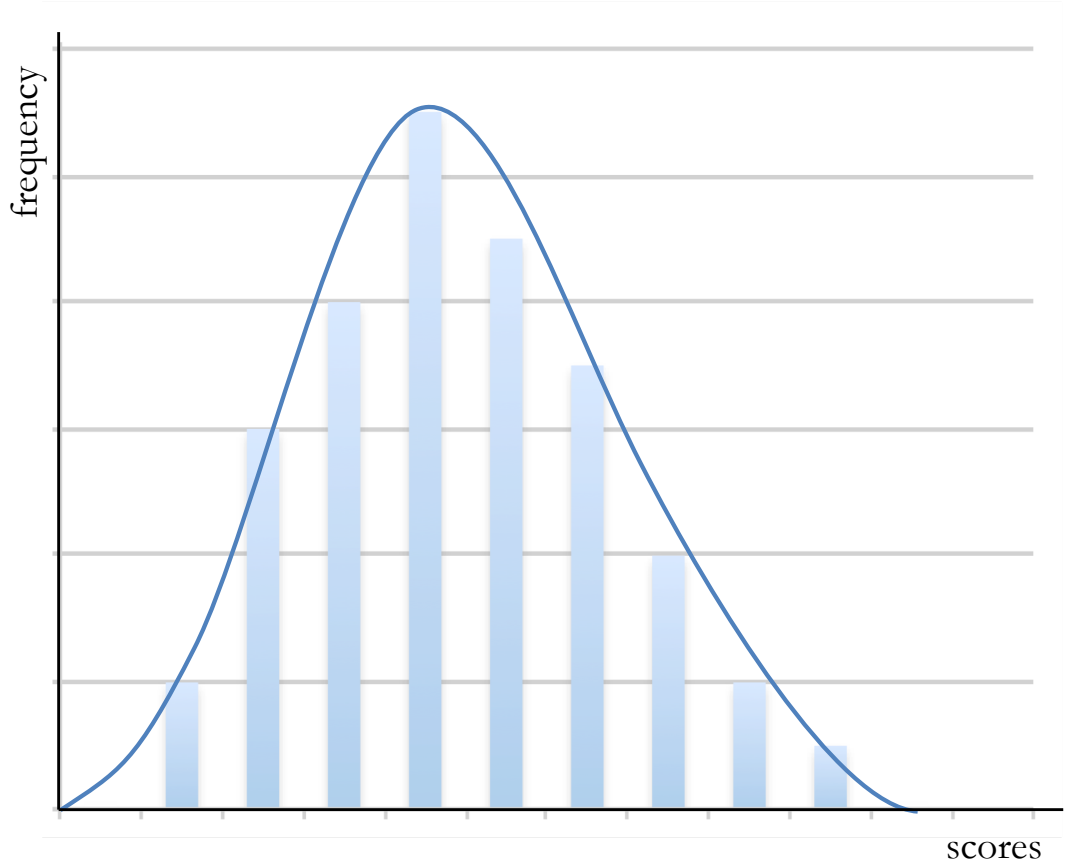
Descriptive Statistics

Central Tendency

typical value, or average value, for a distribution of numbers

Dispersion

how spread, or stretched, out is a distribution of numbers



Central Tendency

```
> ages <- customers$age
> ages
```

`<-` : *assign value to a variable*

```
  [1] 49 69 41 73 45 71 50 43 70 32 47 77 64 50 50 45 49 47 62 50 47 72 47 63 21
 [26] 49 50 48 35 77 48 48 50 47 29 42 42 85 45 49 45 43 49 68 42 48 72 79 48 50
  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
[376] 48 18 45 62 41 71 19 73 26 75 41 46 49 49 23 74 53 23 51 71 50 50 67 74
```

Population Mean : *the average of your population*

$$\bar{x} = \frac{\sum x_i}{n}$$

```
> sum(ages) / length(ages)
[1] 46.80702

> mean(ages)
[1] 46.80702
```

***Sample** elements from fixed indeces*

Sampling : *getting a subset of the full population of data*

```
  [1] 49 69 41 73 45 71 50 43 70 32 47 77 64 50 50 45 49 47 62 50 47 72 47 63 21
 [26] 49 50 48 35 77 48 48 50 47 29 42 42 85 45 49 45 43 49 68 42 48 72 79 48 50
  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
[376] 48 18 45 62 41 71 19 73 26 75 41 46 49 49 23 74 53 23 51 71 50 50 67 74
```

```
> sample_ages <- ages[c(2, 15, 28, 385)]

[1] 69 50 48 75
```

***Sample** elements from fixed indeces*

Random sample : *getting a random subset of the full population of data*

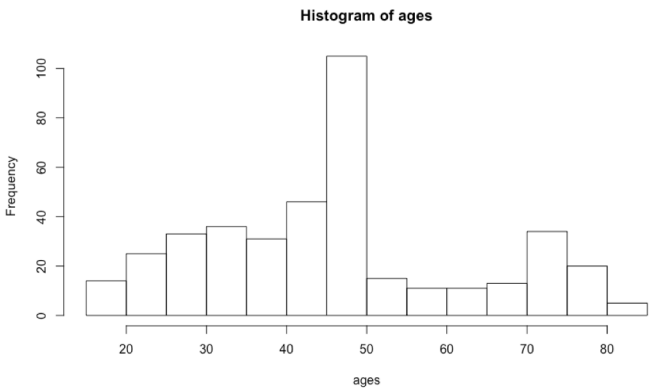
```
> sample(ages, 4)
[1] 74 43 47 34

> mean(random_ages)
 [1] 49.5
```

***Randomly sample** elements from a list (default: without replacement)*

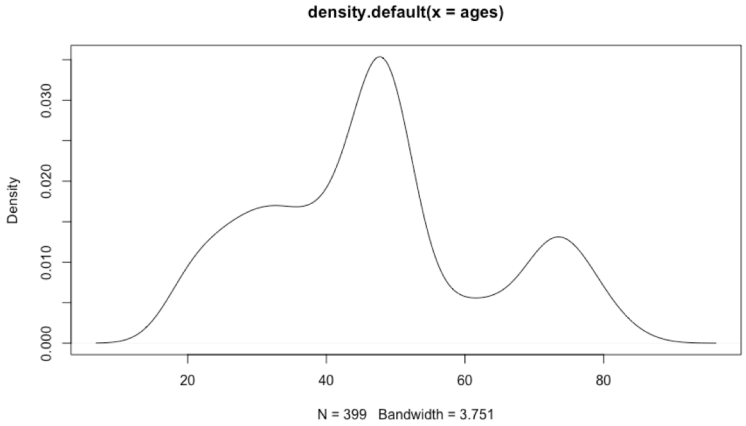
🤔 *Is the mean of my sample related to the population mean?*

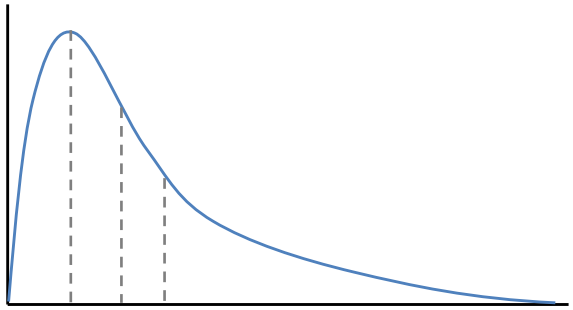
Dispersion: How the Data Deviates From Center

```
> hist(ages)
```

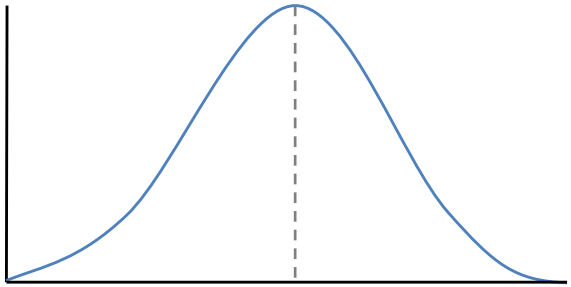



```
> plot(density(ages))
```

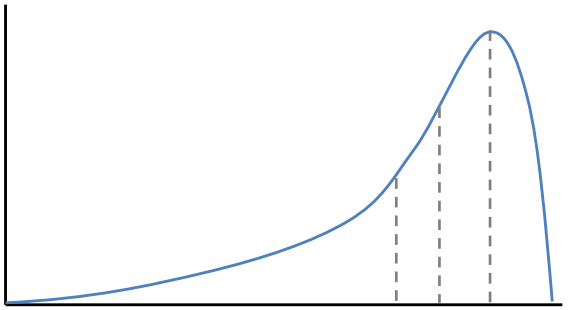




positive skew
right skew



symmetrical



negative skew
left skew

source: http://en.wikipedia.org/wiki/Kernel_density_estimation

```
> sort(ages)
  [1] 18 19 19 19 19 19 19 19 19 20 20 20 20 20 21 21 21 21 21 21 21 22 22 23 23
 [26] 23 23 23 23 24 24 24 25 25 25 25 25 25 25 26 26 26 26 26 26 26 26 26 27 27
   .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .

[376] 76 76 76 76 76 77 77 77 77 78 78 78 78 79 79 79 79 80 80 81 82 82 83 85
```

***sort**ed in ascending order*

Percentile = value at (p/100)n

10th Percentile (p=10):

```
> quantile(sorted_ages, 0.10)
26
```

25th Percentile (p=25):

```
> quantile(sorted_ages, 0.25)
34
```

Quartile = value at i(N+1)/4

1st Quantile (i=1):

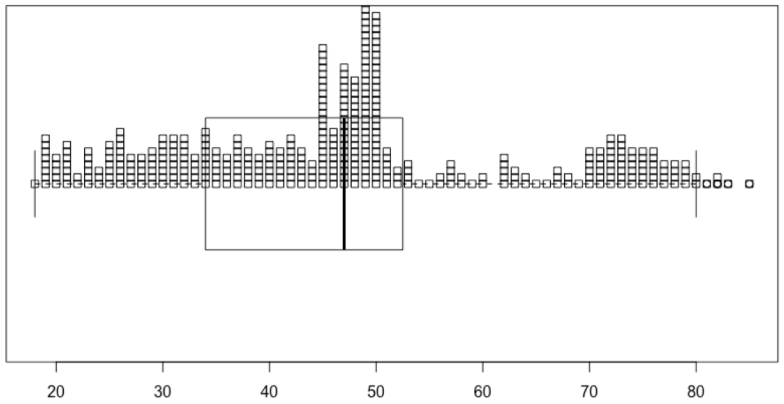
```
> quantile(sorted_ages, 0.25)
26
```

2nd Quantile (i=2):

```
> quantile(sorted_ages, 0.50)
47
```

3rd Quantile (i=3):

```
> quantile(sorted_ages, 0.75)
52.5
```




```
> boxplot(ages, horizontal = TRUE)
> stripchart(ages, method = "stack", add = TRUE)
```

```
> summary(ages)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   34.00   47.00   46.81   52.50   85.00
```