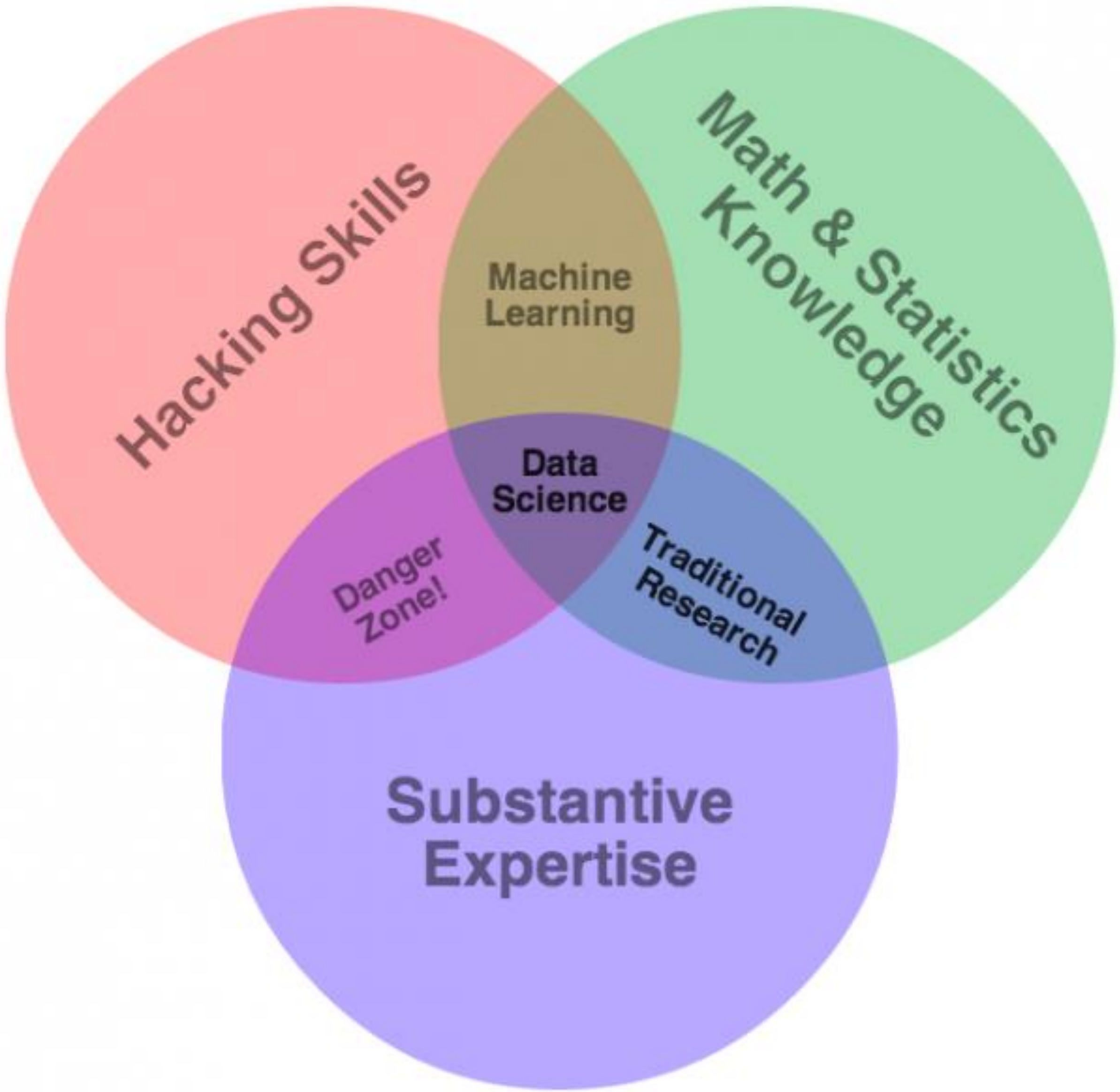
UNIVERSITY of WASHINGTON

Data Science UW Methods for Data Analysis

Introduction and Data Exploration
Lecture 1
Nick McClure


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Course Purpose

- This course isn't designed to make you an expert
- This course is designed to point you in the right direction
- Course Objectives:
  - Statistical tools for data exploration
  - The use of R to apply these tools to real data
  - Using inferential statistics to interrogate data
  - Testing and experimental design
  - Bayesian and classical statistics
- See syllabus for more information:
  - On Canvas
  - Or http://nfmcclure.github.io/DataScience350/

Course Requirements and Grading

This course will be graded by attendance, homework, and an individual project.

- Attendance: You MUST attend at least 6 out of 10 classes. This is non-negotiable, a UW requirement.
- Homework must be completed by the start of the next class. (Assigned weeks 1-8).
  - Returned as a score between 0 and 2.
    - 2 pts: The homework is well done, submitted on time.
    - 1.5-1.9 pts: The homework mostly satisfies the criteria. The major concept of the homework was done correctly with minor programming or statistical errors.
    - 1-1.4pts: The homework misses some key points of the objectives.
    - 0-1pt: The homework needs much more work in both the statistical and programming structure.
    - All late homeworks receive a 0.5 point deduction.

Individual Project

- Individual Project: Due at the start of the last class.
  - Counts as 8 points.
- Must use at least two distinct statistical methods covered in this class.
- Projects are usually heavily involved either with data gathering xOR with statistical methods.
  - Doing a project that is heavy on both is not recommended.
- By the beginning of the third class, you will have sent me a project proposal email. It is worth 0.25 points as part of homework #2.
- Ideal project schedule:
  - By the 5th class, have acquired all necessary data.
  - By the 8th class, have written all code for the project.
  - By the 9th class, have a first draft of the project write up.
  - By the 10th class, submit project.

Course Requirements and Grading

There is a total of 24 possible points. (16 pts for hmk + 8 project)

> Must get 18 total points to pass.

> 4 homework assignments must be made in a production level script (every other one = 2,3,5,7).

> 4 homework assignments are regular script writing (every other one = 1,4,6,8).

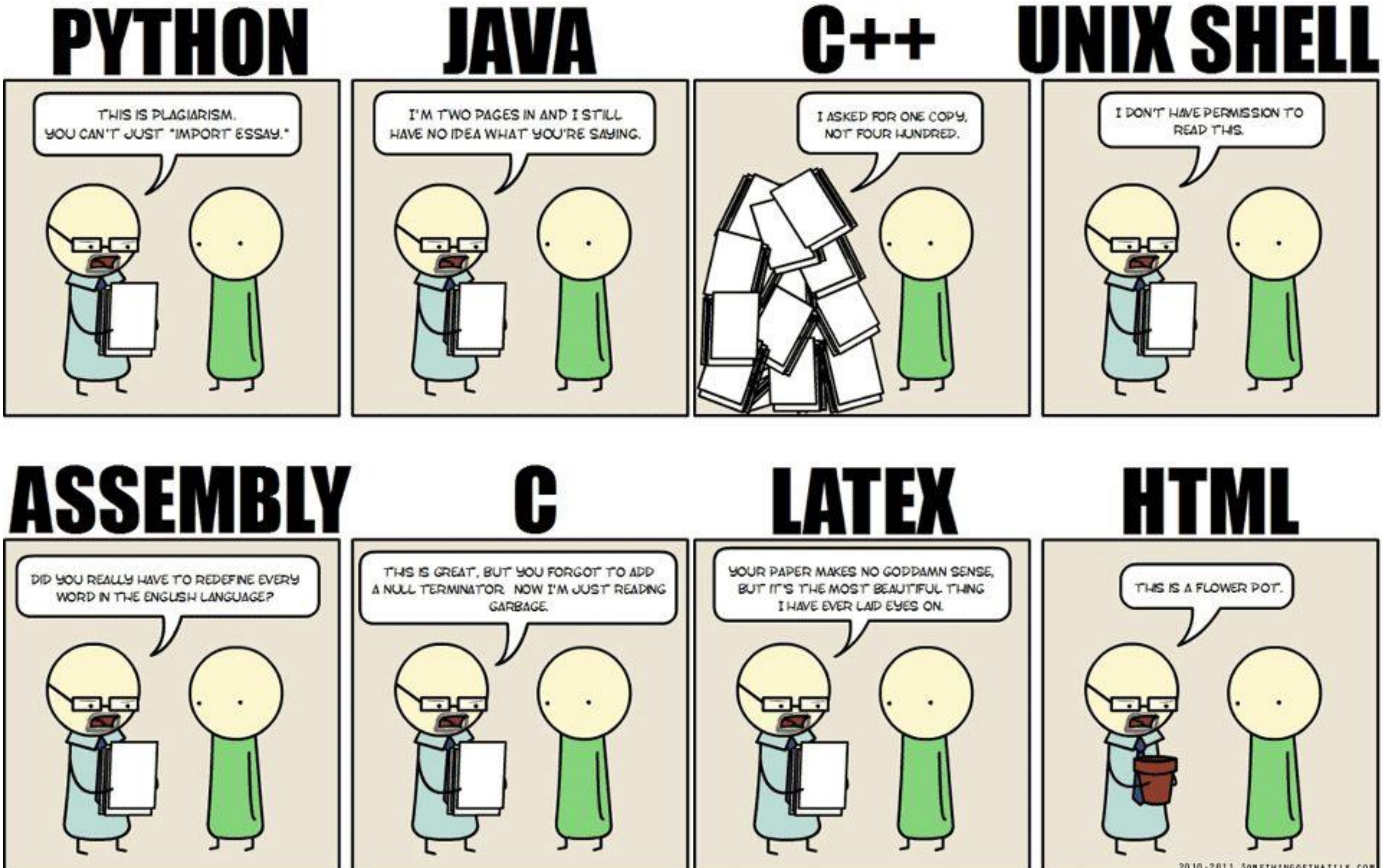
> The individual project must be production level code.

Office Hours and Contact Information

- List of ways to contact me:
  - nickmc@uw.edu (updated every hour or so)
  - nfmcclure@gmail.com (updated quite continuously)
- When I'm *usually* available:
  - Off/on for simple things during work. (M-F 8am-5pm PST)
  - Mon-Wed 7pm-10pm.
  - Sunday various afternoon/evening times.

Emergency contact: 402-980-3192

Review


PYTHON
THIS IS PLAGIARISM. YOU CAN'T JUST "IMPORT ESSAY."
JAVA
I'M TWO PAGES IN AND I STILL HAVE NO IDEA WHAT YOU'RE SAYING.
C++
I ASKED FOR ONE COPY, NOT FOUR HUNDRED.
UNIX SHELL
I DON'T HAVE PERMISSION TO READ THIS.
ASSEMBLY
DID YOU REALLY HAVE TO REDEFINE EVERY WORD IN THE ENGLISH LANGUAGE?
C
THIS IS GREAT, BUT YOU FORGOT TO ADD A NULL TERMINATOR. NOW I'M JUST READING GARBAGE.
LATEX
YOUR PAPER MAKES NO GODDAMN SENSE, BUT IT'S THE MOST BEAUTIFUL THING I HAVE EVER LAID EYES ON.
HTML
THIS IS A FLOWER POT.
2010-2011 SOMETHINGOFTHATILK.COM

Topics

- Probability and Statistics
  - Counting
  - Axioms of Probability
  - Probability Examples
- R Programming Review
  - R Resources
  - Data Exploration in R

Why Counting?

- Counting is fundamental to probability theory.
- Probability is the extent or likelihood of an event or set of events.
  - Depends heavily on the ability to *count* up potential outcomes.

Counting

> This is one of the biggest areas of mathematics, called Combinatorics.

> Example:
  - Subway has 4 different breads, 5 different meats, 4 different toppings. How many sandwich combinations?
  - How many different 4-beer tasters can I have in a bar with 10 beers on tap?

> Solve these using the 'Multiplication Principle'.
  - Subway Problem:

$$\underset{\text{(\# of breads)}}{\underline{\quad 4 \quad}} * \underset{\text{(\# of meats)}}{\underline{\quad 5 \quad}} * \underset{\text{(\# of toppings)}}{\underline{\quad 4 \quad}} = 80$$

  - Beer Problem:

$$\underset{\text{(\# for 1}^{\text{st}}\text{ beer)}}{\underline{\quad 10 \quad}} * \underset{\text{(\# for 2}^{\text{nd}}\text{ beer)}}{\underline{\quad 9 \quad}} * \underset{\text{(\# for 3}^{\text{rd}}\text{ beer)}}{\underline{\quad 8 \quad}} * \underset{\text{(\# for 4}^{\text{th}}\text{ beer)}}{\underline{\quad 7 \quad}} = 5{,}040$$

Multiplication Principle

> If there are A ways of doing task a, and B ways of doing task b, then there are A*B ways of completing both tasks.

> Example:
  – If I have 5 books, how many ways can I *order* them on the bookshelf?

5 choices	*	4 choices	*	3 choices	*	2 choices	*	1 choice
1st book		2nd book		3rd book		4th book		5th book

= 5 factorial = 5! = 120

Factorials

- Factorials
  - Count # ways to order N things = N!
- Factorials get VERY large quickly.
  - 21! Is larger than the biggest long-int in 64 bit.
    - 21! = 5.1E19
    - Biggest long int (64 bit) = 9.2E18
  - Fun fact, every 52 card shuffle is highly likely to be the only time that shuffle has ever occurred.

Counting Subgroups

> Revisit: 10 beers on tap, need a sample of 4 different beers.

> Let's assume order matters, i.e., Amber-Stout-Porter-Red is different from Red-Porter-Stout-Amber.

> Use 'Permutations' (pick):

$$10 * 9 * 8 * 7 = \frac{10!}{6!} = \frac{10!}{(10-4)!} = 10P4 = P(10,4)$$

Counting Subgroups

> Now, Let's assume order doesn't matter.
> Use 'Combinations' (choose):

$$\frac{10 * 9 * 8 * 7 = \frac{10!}{6!} = \frac{10!}{(10-4)!} = 10P4 = P(10,4)}{(\# \; of \; orderings \; of \; 4 \; beers) = 4!}$$

$$= \frac{10!}{4!\,(10-4)!} = 10C4 = C(10,4) = \binom{10}{4}$$

More on Combinations

- Combinations appear on the Pascal's Triangle!
- C(N,x) appears on the Nth row, xth number (starting at 0)

$$\binom{10}{4}$$

```
                        1
                      1   1
                    1   2   1
                  1   3   3   1
                1   4   6   4   1
              1   5  10  10   5   1
            1   6  15  20  15   6   1
          1   7  21  35  35  21   7   1
        1   8  28  56  70  56  28   8   1
      1   9  36  84 126 126  84  36   9   1
    1  10  45 120 210 252 210 120  45  10   1
  1  11  55 165 330 462 462 330 165  55  11   1
1  12  66 220 495 792 924 792 495 220  66  12   1
```

Counting Examples

> There are 10 Light beers on tap, and 10 Dark beers on tap, how many ways can Rick get a 4-beer sampler that contains exactly 1 light beer? (ordering doesn't matter)

$$\frac{(\#\ of\ ways\ for\ light\ beer) \cdot (\#\ of\ ways\ for\ dark\ beer)}{(\#\ of\ ways\ to\ order\ 1L\ and\ 3D)}$$

$$\frac{(10) \cdot \binom{10}{3}}{4} = \frac{10 * 120}{4} = 300$$

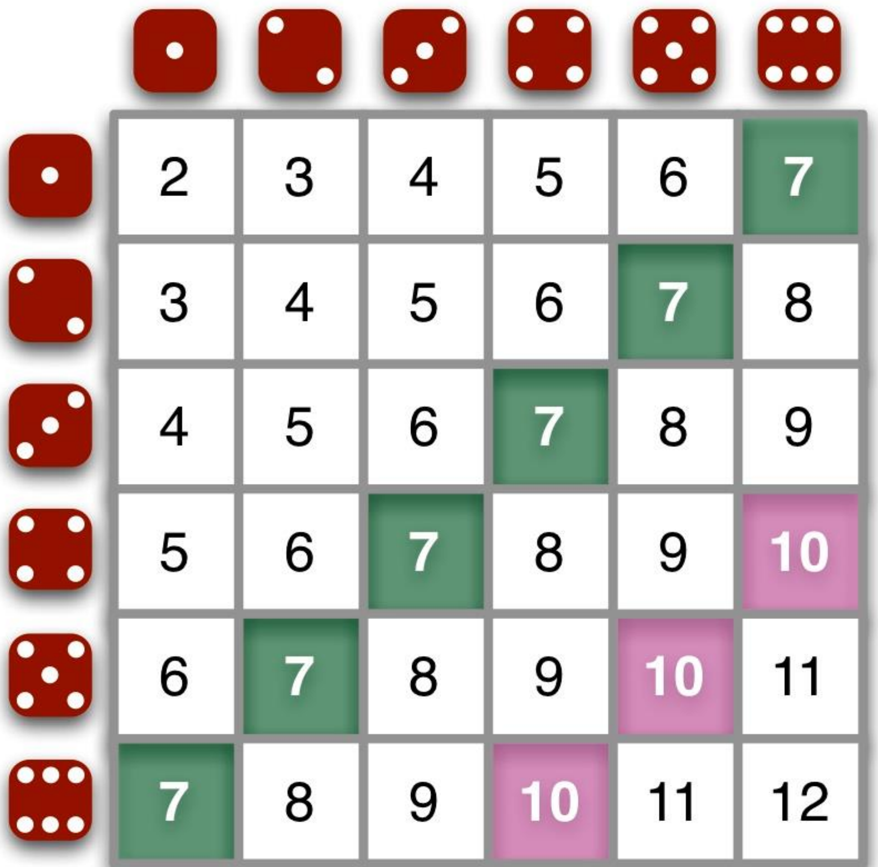
Counting Examples

> 6:5 Blackjack is dealt with a 6 shoe deck (52*6=312 cards). How many ways can someone get dealt two rank 10 cards?

$$\binom{6decks * 4ranks * 4suits}{2} = \binom{96}{2} = \frac{96!}{2!\,(94!)} = \frac{96 * 95}{2} = 4560$$

Counting Examples

> How many ways can two dice be rolled to get a sum of 10?



Counting in R

- expand.grid() – function that creates a data frame from all combinations of vectors supplied.
- R-demo

Probability

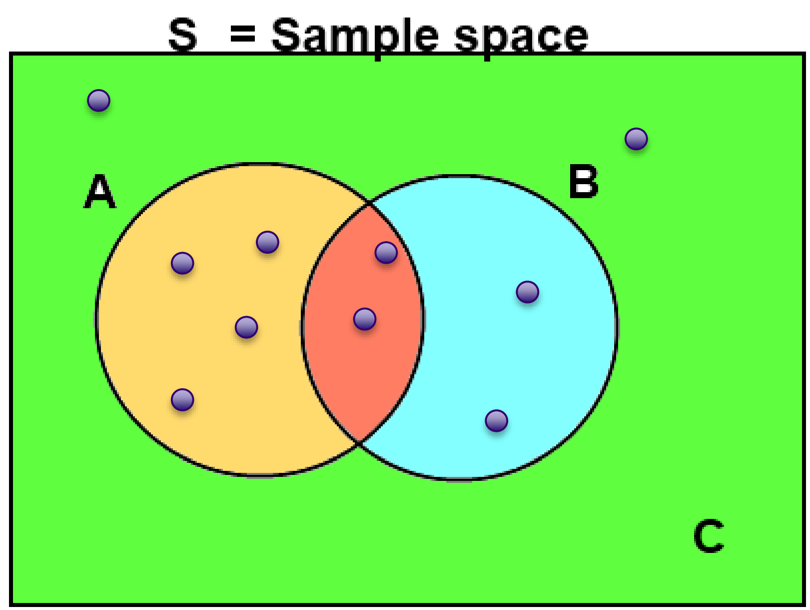
> The Probability of an event, A, is the number of ways A can occur, divided by the number of total possible outcomes in our Sample Space, S.

$$P(A) = \frac{N(A)}{N(S)}$$

> If is an event, then

$$P(A) = \frac{6}{10} = \frac{3}{5}$$

$$P(B) = \frac{4}{10} = \frac{2}{5}$$



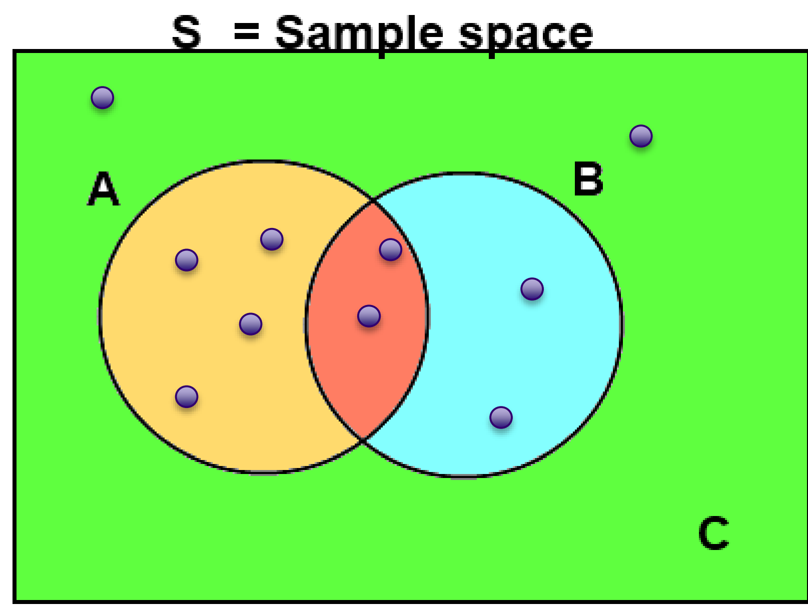
Probability

> If  is an event, then

- Intersection: $P(A \cap B) = \dfrac{2}{10} = \dfrac{1}{5}$
- Union: $P(A \cup B) = \dfrac{8}{10} = \dfrac{4}{5}$
- Negation: $P(A') = \dfrac{4}{10} = \dfrac{2}{5}$

$$P((A \cup B)') = P(C) = \frac{2}{10} = \frac{1}{5}$$

$$P(A' \cap B') = P(C) = \frac{2}{10} = \frac{1}{5}$$



Axioms of Probability

> Probability is bounded between 0 and 1.

$$0 \leq P(A) \leq 1$$

Note: “Percent” literally means per one hundred

> Probability of the Sample Space = 1.

$$P(S) = 1$$

> The probability of finite *mutually exclusive* unions is the sum of their probabilities.

$$P(A \cup B) = P(A) + P(B)$$ If A and B are M.E.

Data Exploration (Descriptive Statistics)

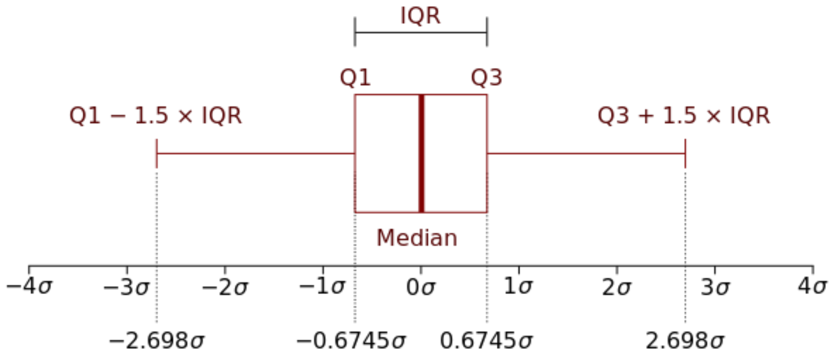
- Purpose: To gain a clear understanding of your data.
  - How large is it?
  - What columns are of interest?
  - Missing data?
  - Outliers?

Numerical Exploration

> str(): structure of the data frame
> summary(): summary of each of the columns
> head() / tail(): top / bottom of data frame
> table(): frequency table

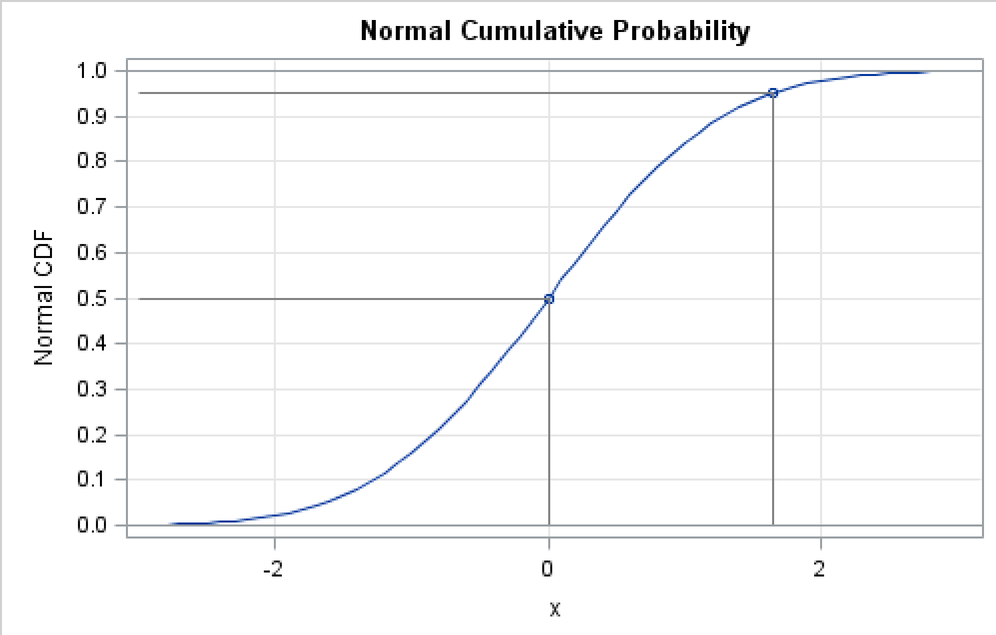
Numerical Exploration

> IQR(): inner quartile range (Q3 – Q1)



Numerical Exploration

> quantile(): quantiles of numerical vectors
  - Quantiles are inverse values of the CDF (cumulative distribution function).
  - Standard Normal: (shown in figure)
    > Quantile(0.5) = 0, means at x=0, 50% of the distribution lies to the left. (This is also the median)
    > Quantile(0.95) = 1.65



Numerical Exploration

> Relationships:

– cov(): covariances

$$cov(x, y) = E\big((x - \mu_x)(y - \mu_y)\big)$$

– Interpretation: Expected value of the differences between x and y and their corresponding mean.

– E.g. if x is above it's mean when y is also above it's mean, then they will have a high covariance.

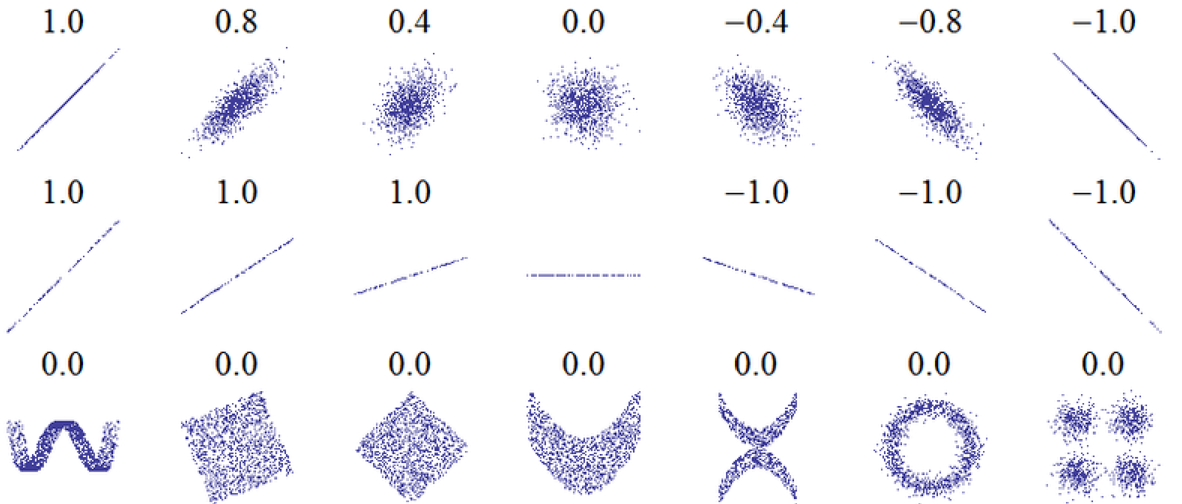
– Highly interpretable, but not bounded.

Numerical Exploration

> Relationships:
  - cor(): correlations (pearsons)

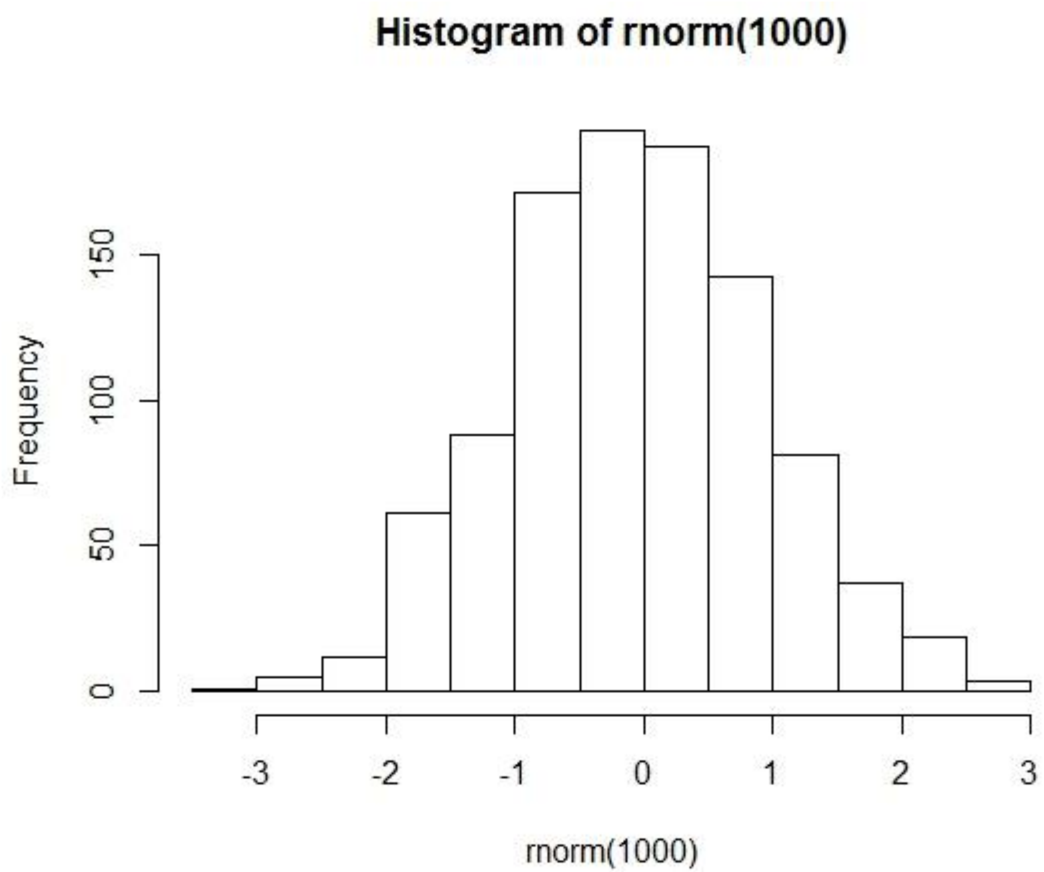
$$cor(x,y) = \frac{E\big((x-\mu_x)(y-\mu_y)\big)}{\sigma_x \sigma_y}$$

  - Bounded between -1 and 1.
  - Not as interpretable.



Visual Exploration

> Histograms:

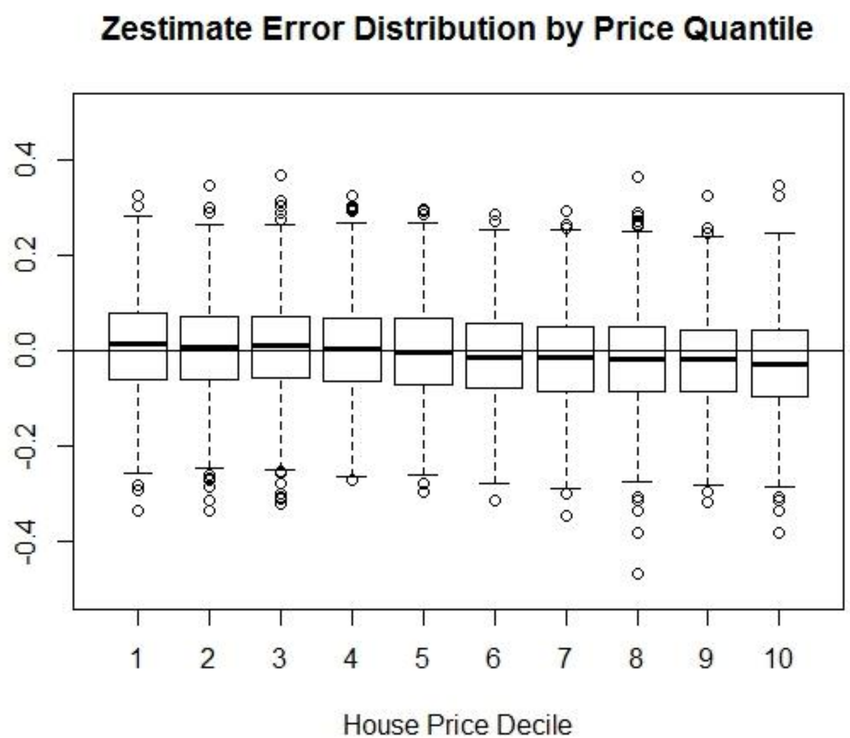


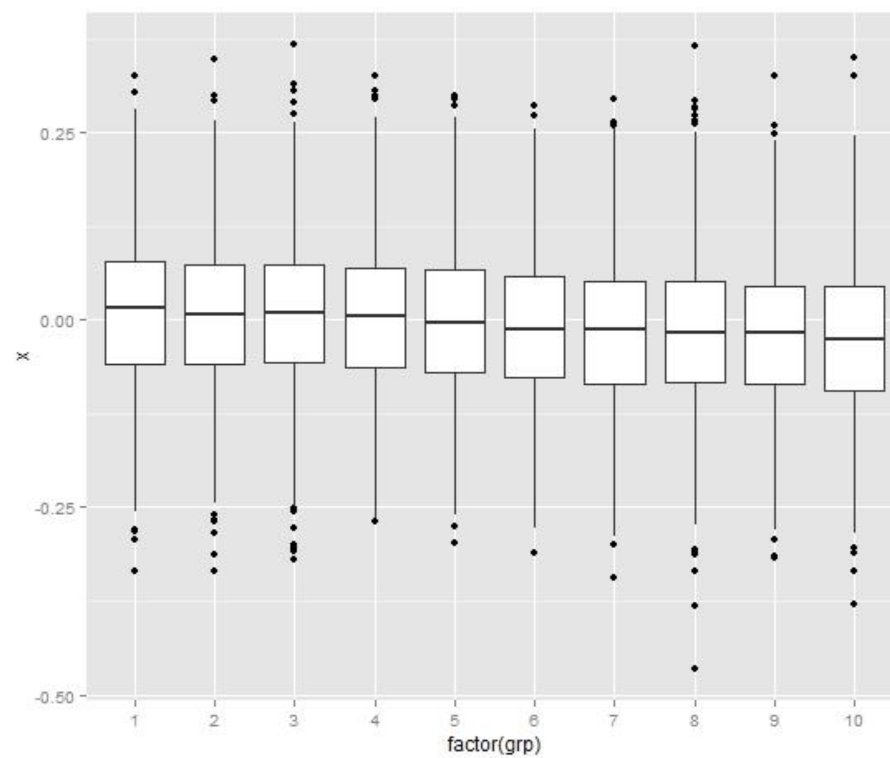
Base:
hist()

ggplot2:
+ geom_histogram()

Visual Exploration

> Boxplots:



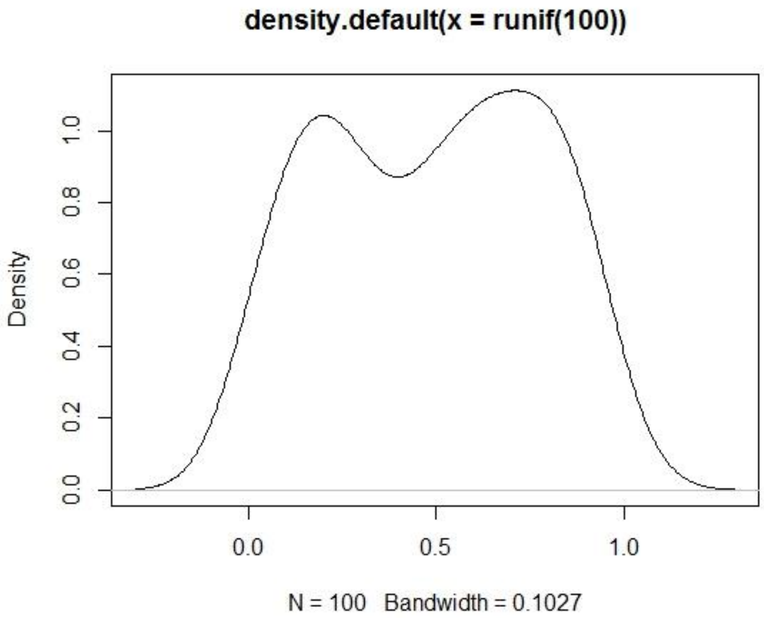


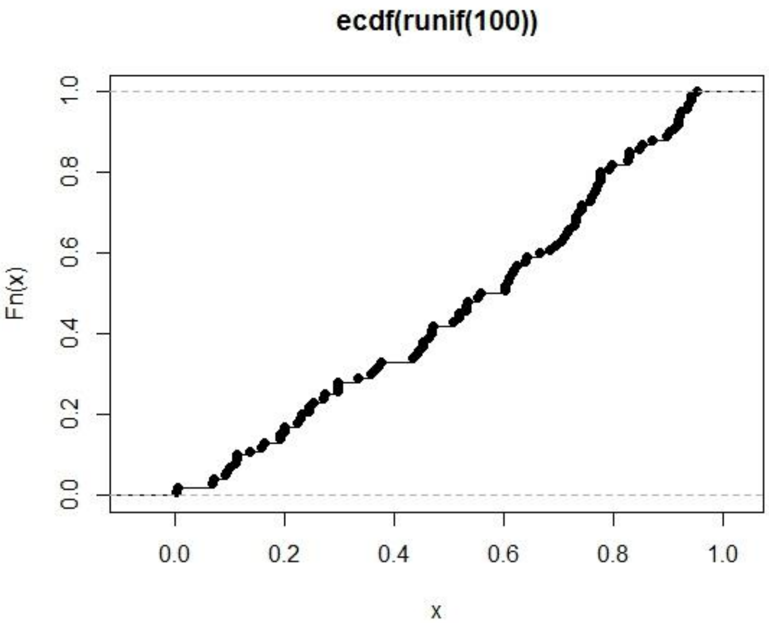
Base:
boxplot()

ggplot2:
+ geom_boxplot()

Visual Exploration

> Densities/CDFs:



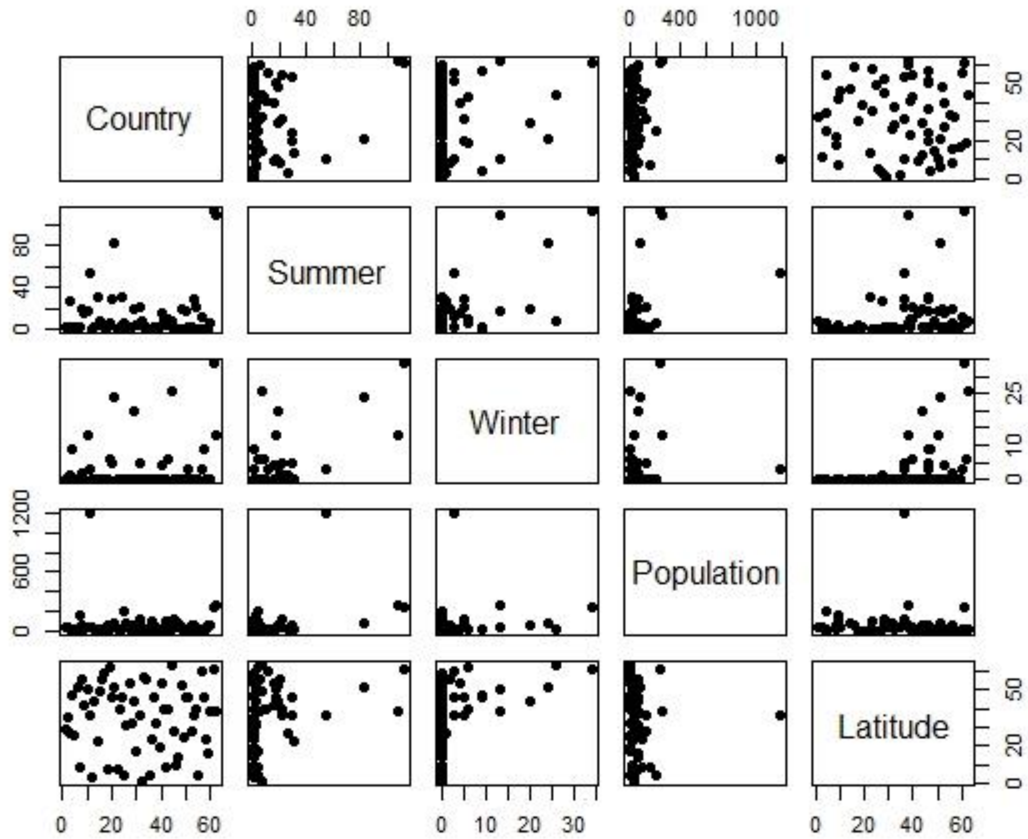


Base:
plot(density())
plot(ecdf())

ggplot2:
+ geom_density()
+ stat_ecdf()

Visual Exploration

> Scatterplots



Base:
pairs()

ggplot2:
ggpairs()

R Resources

- R page:
  - > http://www.r-project.org/other-docs.html
- Stackoverflow:
  - > http://www.stackoverflow.com
- ‘Little’ R intro:
  - > http://cran.r-project.org/doc/contrib/Rossiter-RIntro-ITC.pdf
- Quick R:
  - > http://statmethods.net/
- There are many tutorials available online, e.g.,
  - > http://cyclismo.org/tutorial/R/
- Notes from a two day course at UW:
  - > http://faculty.washington.edu/tlumley/Rcourse/
- Hadley Wickham’s Style Guide:
  - > http://adv-r.had.co.nz/Style.html
- DataCamp R Exercises:
  - > Link in Canvas announcements.

Assignment

- Go to:
  - Vote for extra topics (time permitting)
  - https://www.surveymonkey.com/r/SK6VX5T
- Complete Homework 1:
  - Explore 'JitteredHeadCount.csv', a data set from Caesar's Entertainment that has falsified/jittered table headcounts.
  - Write ***script level*** R program that shows/illustrates 3 key takeaways of your choosing from exploring the data.
  - You should submit:
    - **ONE R-script.**
    - **One word document with 3 key points.** (example next page).

Example Takeaway

> The aggregate table headcounts on the weekends are X% higher than non-weekends (figure 1). In fact, the game that has the highest difference between average highs and average low days is Gamecode AA with a difference of x.xx heads/table.

> R script Example Demo