

# 01 - Introduction to SYS 2202

01-Intro.pdf

# Preliminaries

# The Course

- ▶ This is SYS 2202: Data and Information Engineering
- ▶ Course material (including syllabus) can be found at:  
<https://mdporter.github.io/SYS2202/>

# Examples: The Perfect Job



Language

Search

Go



Contact Us

Home

About

Industries

Products

Newsroom

Clients

Careers

## Careers

:: Current Opportunities

:: Campus Recruiting

:: Life at APT

:: Puzzles

- + Number Grid
- + Simple Math
- + Cipher
- + Decoder
- + Calling Smart Candidates
- + Get to the Point
- + Prime Lighting
- + APT's Roots
- + Stop and Go
- + **The Perfect Job**
- + Fibonacci

:: Apply Now

### The Perfect Job

Recruiting season is an exciting time for students. It's filled with decisions about the future, your classmates walking around in business suits, and some lectures tossed in somewhere along the way. Important decisions are being made: what kind of opportunity to pursue, where to pursue it, and who to have with you along the way.

Luckily, you're no stranger to honing your problem solving and coding skills for just such an occasion. While others may struggle with such a decision, you know that finding the perfect job is just an optimization problem subject to certain constraints.

Put your brain to work: write some code that generates the optimal set of job assignments for the set of job offers and people below.


### The problem set

#### Jobs

Each type of job has certain benefits and drawbacks along several dimensions:

	Pay	Hours	Impact	Opportunity to Learn
Big Software Firm	6	6	2	8
Hedge Fund	8	8	4	6
Investment Bank	10	10	3	4
Startup	4	8	10	8
Grad School	1	4	3	10

# Examples: Gun Violence



WIKIPEDIA  
the free encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikipedia store

Interaction  
help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
What links here  
Related changes  
Upload file  
Special pages  
Permanent link  
Page information  
Wikidata item  
Cite this page

Print/export  
Create a book  
Download as PDF  
Printable version

Languages

## Gun violence in the United States by state

from Wikipedia, the free encyclopedia

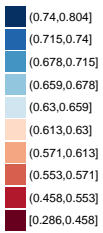
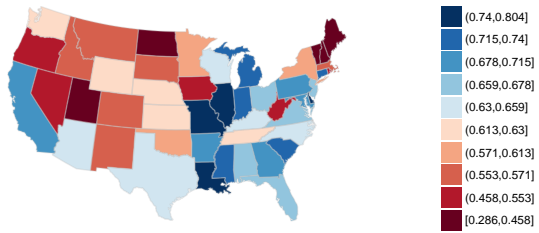
This article is a list of the U.S. states and the District of Columbia, with *population*, *murders* and *non-negligent manslaughter*, murders, gun murders, and *gun ownership* percentage, then calculated rates per 100,000. The population data is from the U.S. Census Bureau. Murder rates were calculated based on the FBI Uniform Crime Reports and the estimated 2015 population of each state. The 2015 U.S. population total was 320.9 million. The 2015 U.S. overall murder and non-negligent manslaughter rate per 100,000 inhabitants was 4.9.<sup>[1]</sup>

States [ edit ]

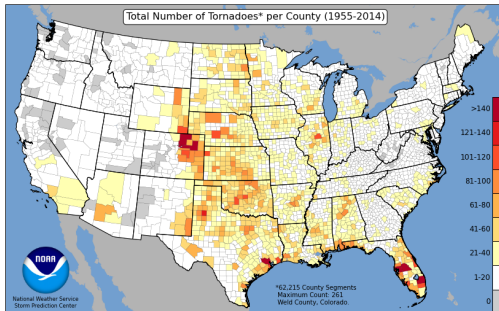
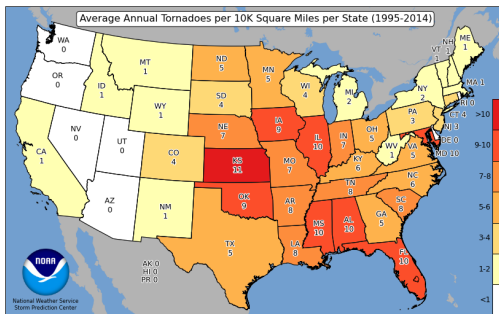
Legend								
Lowest1st Quartile2nd Quartile3rd Quartile4th QuartileHighest								
State	Population (total inhabitants) (2015) <sup>[1]</sup>	Murders and Nonnegligent Manslaughter (total deaths) (2015) <sup>[1]</sup>	Murders (total deaths) (2015) <sup>[1]</sup>	Gun Murders (total deaths) (2015) <sup>[1]</sup>	Gun Ownership (%) (2015) <sup>[4]</sup>	Murder and Nonnegligent Manslaughter Rate (per 100,000) (2015)	Murder Rate (per 100,000) (2015)	Gun Murder Rate (per 100,000) (2015)
Alabama	4,853,875	348	—[1]	—[2]	48.9	7.2	— [1]	— [1]
Alaska	737,709	59	57	39	61.7	8.0	7.7	5.3
Arizona	6,817,565	306	278	171	32.3	4.5	4.1	2.5
Arkansas	2,977,853	181	164	110	57.9	6.1	5.5	3.7
California	38,993,940	1,861	1,861	1,275	20.1	4.8	4.8	3.3
Colorado	5,448,819	176	176	115	34.3	3.2	3.2	2.1
Connecticut	3,584,730	117	107	73	16.6	3.3	3.0	2.0
Delaware	944,076	63	63	52	5.2	6.7	6.7	5.5
District of Columbia	670,377	162	121	121	25.9	24.2	24.2	18.0



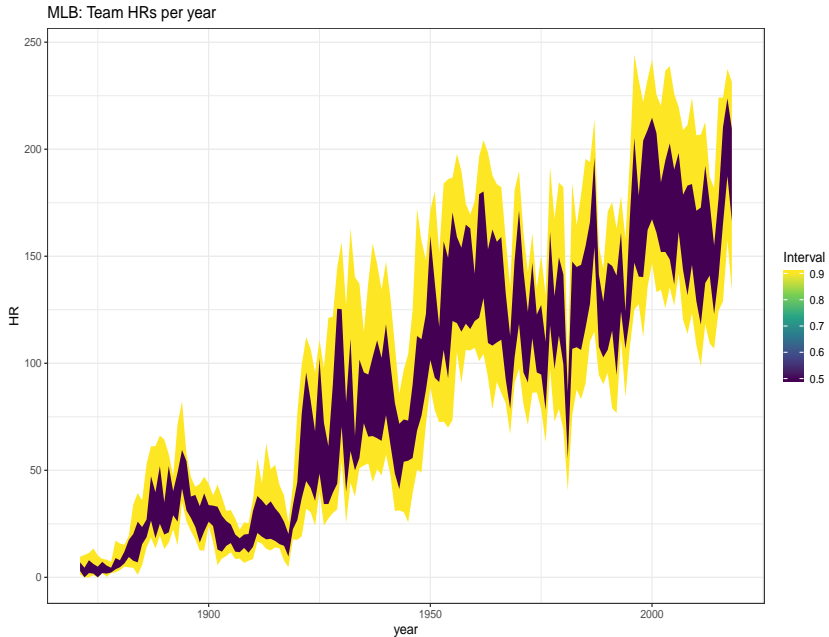
## Proportion of Murders that are Gun related



# Examples: Tornado Risk Mapping



# Examples: Fan Plots



# About you

Fill out a notecard with the following information:

1. Your name (with pronunciation hints)
2. Hometown (include country/region if far away)
3. Degree, Major and expected graduation date
4. List coding/programming experience. Language and level (Scale 1-5, 5 highest).
5. List data analysis/modeling experience (Scale 1-5, 5 highest).
6. 2 interesting things about you (to help me remember you)



# What is Analytics?

# Famous Quotes

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

*H. G. Wells*

Not many executives are information-literate. They know how to get data. But most still have to learn how to use data.

*Peter Drucker*

The greatest value of a picture is when it forces us to notice what we never expected to see.

*John Tukey*

IT IS  
**TOO COMPLEX**

- ▶ This course will provide you with an introduction to **Data Science/Analytics** (using the R software).
- ▶ We will focus on the key elements of: Data Wrangling and Communication/Visualization
- ▶ **What Data Scientists Really Do, According to 35 Data Scientists. By Hugo Bowne-Anderson, HBR**
  - ▶ *Nearly all of my guests understand that working data scientists make their daily bread and butter through data collection and data cleaning; building dashboards and reports; data visualization; statistical inference; communicating results to key stakeholders; and convincing decision makers of their results.*
  - ▶ *It has been a common trope that 80% of a data scientist's valuable time is spent simply finding, cleaning, and organizing data, leaving only 20% to actually perform analysis.*

# Analytics Categories

The Institute for Operations Research and the Management Sciences (INFORMS) has proposed **three categories** of Analytics:

## 1. **Descriptive analytics**

- ▶ Prepares and analyzes historical data
- ▶ Identifies patterns from samples for reporting of trends

## 2. **Predictive analytics**

- ▶ Predicts future probabilities and trends
- ▶ Finds relationships in data that may not be readily apparent with descriptive analysis

## 3. **Prescriptive analytics**

- ▶ Evaluates and determines new ways to operate
- ▶ Targets business objectives
- ▶ Balances all constraints

# Two Phases of Data Analysis

A better way to think about analytics is in terms of **two phases**:

## 1. **Exploratory Data Analysis**

- ▶ The exploratory phase “isolates patterns and features of the data and reveals these forcefully to the analyst” (Hoaglin, Mosteller, and Tukey; 1983)
- ▶ If a model is fit to the data, exploratory analysis finds patterns that represent deviations from the model.
- ▶ These patterns lead the analyst to revise the model, and the process is repeated.

## 2. **Confirmatory Data Analysis**

- ▶ In contrast, confirmatory data analysis “quantifies the extent to which [deviations from a model] could be expected to occur by chance” (Gelman; 2004)
- ▶ Uses the traditional statistical tools of inference, significance, and confidence.

# Exploratory and Confirmatory Analysis

- ▶ **Exploratory data analysis** is sometimes compared to **detective work**: it is the process of gathering evidence.
- ▶ **Confirmatory data analysis** is comparable to a **court trial**: it is the process of evaluating evidence.
- ▶ Exploratory analysis and confirmatory analysis “can -and should- proceed side by side” (Tukey; 1977).

# The Six Divisions of Greater Data Science

1. Data Gathering, Preparation, and Exploration
2. Data Representation and Transformation
3. Computing with Data
4. Data Modeling
  - ▶ Covered in SYS 4021 (Linear Models), SYS 6018 (Data Mining), etc.
5. Data Visualization and Presentation
6. Science about Data Science

Source: David Donoho's [50 Years of Data Science](#)



# Syllabus

Let's check out the course syllabus:

<https://mdporter.github.io/SYS2202/syllabus.html>

## Course Tools

# Why R?



► <https://www.r-project.org/about.html>

## Your Turn #1

1. Download R: Go to <http://cran.r-project.org/> and click on your OS
  - ▶ Install **base** (for windows)
  - ▶ Choose your version for Mac
  - ▶ If you use linux, you don't need me to tell you what to do
  - ▶ Detailed instructions [here](#)
2. Open R
3. Use R to find:
  - ▶  $7 + 3$
  - ▶  $10 - 13$
  - ▶  $\pi * 3^2$
  - ▶  $\log(1)$
  - ▶ Try this in R:  
`plot(x=-10:10, y=(-10:10)^2, type='l')`

# Who uses R?

- ▶ Core language of almost all Statisticians
- ▶ Top data science and analytics tool
  - ▶ 49% of Data Scientists use R
  - ▶ #5 overall programming language
  - ▶ Overall popularity
- ▶ R is forefront of many on-line courses
  - ▶ DataCamp
  - ▶ Coursera: Data Science Specialization
- ▶ Many companies using R
  - ▶ Microsoft
  - ▶ Google
  - ▶ New York Times
  - ▶ Airbnb
  - ▶ General Mills, LexisNexis Risk Solutions, Novartis
  - ▶ and many, many others

# Why use R?

- ▶ \$115K average salary for R users (Dice.com 2014 survey)
- ▶ R is Good for Business
- ▶ R is Popular
- ▶ R is free!
- ▶ R is the **best** program for interactive data analysis
- ▶ R can detect credit card fraud at 1M transactions/second
- ▶ R can run on multiple platforms (Windows, Mac, Linux)
- ▶ R is open-source
  - ▶ Companies can use and modify
    - ▶ R used at Microsoft
    - ▶ R incorporated into SQL server 2016
  - ▶ You can find out what the code is actually doing
- ▶ SAS (and many other programs) allows you to run R code
- ▶ We will do a textual analysis of business analytics postings which will reveal the growing demand of R

# What can R do?

- ▶ R is a programming language so you can get it to do all sorts of things
  - ▶ but its core strength is *data analysis*
- ▶ Some basic functionality: <https://www.r-project.org/about.html>
- ▶ But strength of R is in its **packages**
  - ▶ Contributed by users
  - ▶ Over 12,000 [R packages](#)
- ▶ Task Views <http://cran.r-project.org/web/views/>
- ▶ Reproducible research



- ▶ RStudio is an integrated development environment (IDE) for R
  - ▶ It is also free, open source, and cross-platform!
  - ▶ [Download RStudio](#)
  - ▶ Install after R

<http://vimeo.com/97166163>

- ▶ We will do all "coding" in RStudio
  - ▶ R
  - ▶ Rmarkdown
- ▶ RStudio also facilitates Shiny for interactive visualizations:  
<http://shiny.rstudio.com/gallery/>
- ▶ Note: RStudio is not R. But it facilitates the use of R (and many other things).

# Class Details

- ▶ Bring laptop to class (with R and Rstudio loaded)
  - ▶ We will be doing lots of in class examples
- ▶ Class is extremely cumulative
  - ▶ If you can't come to class, spend time with the lecture notes and talk to someone who attended
  - ▶ If you are falling behind, catch up as soon as possible
- ▶ The first few weeks can be frustrating; stick with it and you will start getting the hang of it
  - ▶ Use google!
- ▶ Expected time commitment (weekly):
  - ▶ 2.5 hrs in class, 3-4 hrs homework, 1-3 hrs reading/practicing
  - ▶ Total of  $7.5 \pm 1$  hrs/week
  - ▶ Put it on the schedule; don't cheat.

## About

# About your instructor



- ▶ 1994-1998: Purdue University
  - ▶ B.S. Industrial Engineering
- ▶ 1998-2001: Chicago, IL - Sanford Markers
  - ▶ Engr/Maintenance
- ▶ 2001-2003: Vanderbilt University
  - ▶ M.S. Systems Engr
- ▶ 2003-2006: University of Virginia
  - ▶ PhD Sys and Info Engr
- ▶ 2006-2008: North Carolina State University & SAMSI
  - ▶ Postdoc (Statistics)
- ▶ 2008-2013: Spadac/GeoEye/DigitalGlobe
  - ▶ Principal Research Scientist
- ▶ 2013-2018: University of Alabama
  - ▶ Associate Prof of Statistics
- ▶ 2018-Present: University of Virginia
  - ▶ Assoc. Prof of Systems Engr, Data Science, Business (Darden)

# My Family



# About you

Fill out a notecard with the following information:

1. Your name (with pronunciation hints)
2. Hometown (include country/region if far away)
3. Degree, Major and expected graduation date
4. List coding/programming experience. Language and level (Scale 1-5, 5 highest).
5. List data analysis/modeling experience (Scale 1-5, 5 highest).
6. 2 interesting things about you (to help me remember you)

## Last Things

# To Do List

- ▶ Turn in note cards
- ▶ Start DataCamp homework (look for email)
  - ▶ HW 1 is due Sept 2
- ▶ Read R4DS 1 and 2
- ▶ Note any questions from reading and homework
- ▶ Get these [RStudio Cheatsheets](#)
  - ▶ RStudio IDE
  - ▶ Data Transformation with dplyr
  - ▶ Data Visualization with ggplot2
  - ▶ String manipulation with stringr
  - ▶ Other Cheatsheets as needed:
    - ▶ Base R
    - ▶ Regular Expressions
    - ▶ Dates and times with lubridate



- ▶ Deliberate Practice: the making of an expert  
<https://hbr.org/2007/07/the-making-of-an-expert>