Introduction to Data Science
Jeffrey S. Saltz/Jeffrey M. Stanton

Week 1 Introduction

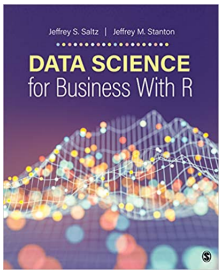
ischool.syr.edu
School of Information Studies
SYRACUSE UNIVERSITY

Outcomes For Today

- Discuss the goals for the course
- Be able to describe what "data science" is
  - Skills needed
  - Data insights
- Get started with R

A Quick Look At the Class

1. What is Data Science
2. Quant overview: Descriptive & Inferential Stats
3. R Skills: Using R functions & libraries
4. Connecting to external data sources
5. Visualization (ggplot)
6. Creating data-driven maps
7. Text Processing/Analysis
8. Linear Modeling
9. Unsupervised Data Mining
10. Supervised Data Mining



Our Backgrounds

Jeffrey S. Saltz

- Computer Science/Information Science/Business Degrees
- Started work life as a programmer
- Led visualization and high-performance computing teams
- Invested in tech companies—venture capital
- Ran credit card risk platform—credit and fraud
- Now do research on "how to do data science"

Jeffrey M. Stanton

- Computer Science/Organizational Psychology Degrees
- Started work life as a programmer
- Led embedded operating system and signal processing teams
- Worked at venture capital funded startup firms
- Developed text processing and digital audio processing platforms
- Now do research on workers and technology in organizations

Typical Structure of Lecture Classes

1. **Review & Introduce Key Concepts**
2. **Data Science Case Study / Discussion**
3. **Hands-On Activity (Using R)**

Each Week

- Hands on: Goal is to enable you to ***practice*** data science
- Bring your laptop to every lab session with access to R and R-Studio
- Read the chapter before the class!!!

How to switch lab sections:

- *We can't have one huge lab section or really small sections*
- *So, if you want to switch labs:*
  - *You need to find someone who will switch with your section (i.e. find someone to switch with)*
  - ***Both people need to come to me → with the forms for me to sign → just before the Week 2 Lecture***

Labs/Homework

- Bring your laptop to every lab
- Lab ==> Homework: The lab focuses on developing the techniques and code needed for successful completion of the homework
- Your lab instructor is your coach for helping you learn R, master the homework, and execute the term project

Grades

Lab is designed for you to practice the necessary skills in carrying out data processing, analysis, and management tasks

Homework is designed for you to practice the necessary skills in carrying out data processing, analysis, and management tasks

Participation includes professionalism in class, and actually attending class…

Quiz are weekly quick multiple choice tests

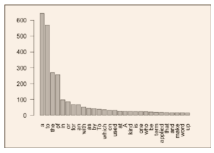
Test is designed to evaluate your mastery of concepts, methods, and tools in data analysis and management.

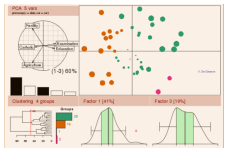
Project: For your project you work on a data set provided to you, transform the data as needed, and provide a written analysis with visualizations.

Data Science: Many Skills

Data Science: Definition

Data science: *An emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information.*





Connects strongly with areas such as databases, statistics, and computer science, but many different skills—including nonmathematical skills—are needed.

Core Goal of Data Science: Creating Actionable Insight

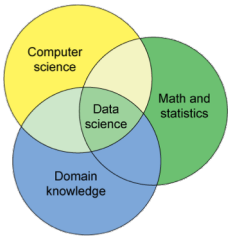
What does each of these levels mean?

What Is Data Science?

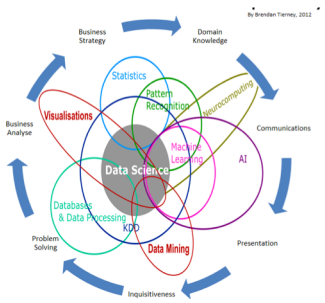
Identify which article belongs to which field:

1. Zhang, N. F. (2006). The uncertainty associated with the weighted mean of measurement data. *Metrologia*, *43*(3), 195.
2. Janssen, M., Konopnicki, D., Snowdon, J. L., & Ojo, A. (2017). Driving public sector innovation using big and open linked data (BOLD). *Information Systems Frontiers*, *19*(2), 189-195.
3. Hillar, C. J., & Lim, L. H. (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, *60*(6), 45.

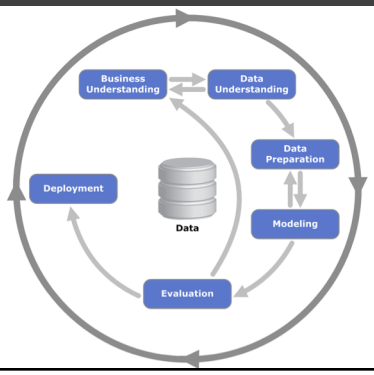
Image credit: Quora.com



Data Science is Multidisciplinary



Doing Data Science: CRISP-DM



Doing Data Science: OSEMN

The pipeline:

- O — Obtaining the data
- S — Scrubbing / Cleaning
- E — Exploring / Visualizing (patterns / trends)
- M — Modeling the data (doing predictions)
- N — Interpreting the analysis

→ Understanding the business problem?
→ This is iterative (need to keep looping)

http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

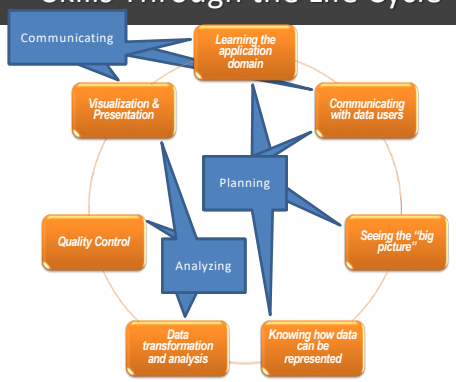
The Data Science Process

- Domain analysis/decomposition
- Identification of subject matter expert(s)
- Question/interview/observation process

Via anomalies, we learn:
- Is there an error in the system definition
- Is there an error in the system implementation
- Is there an **interesting exception** that teaches us something

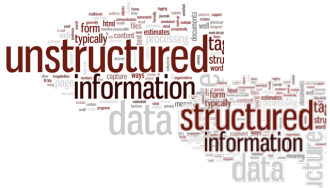
Via stories, we learn:
- What people do
- How they do it
- Information produced and consumed
- Process and information touch points
- Decisions made
- Key Challenges

Skills Through the Life Cycle



Structured & Unstructured Data

- What is the difference between structured and unstructured data?

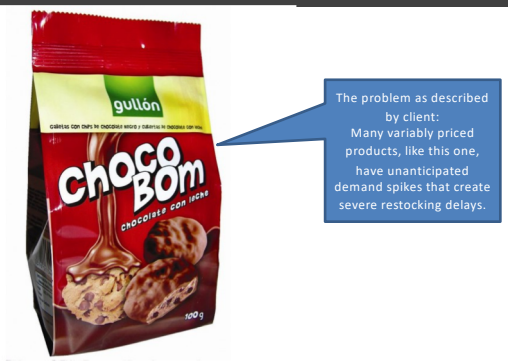


Questions

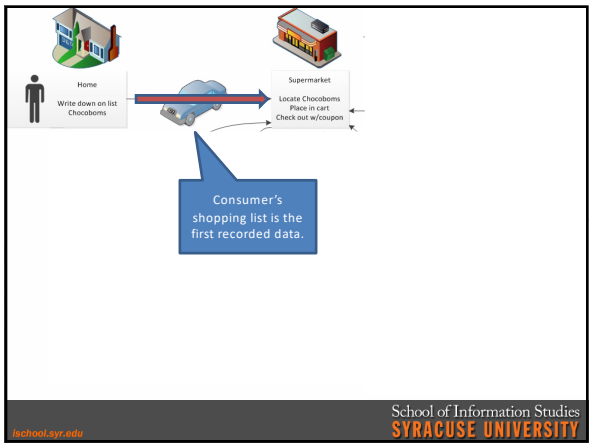
- Do data scientists mainly deal with structured data or unstructured data?
- Which kind of data is there more of?

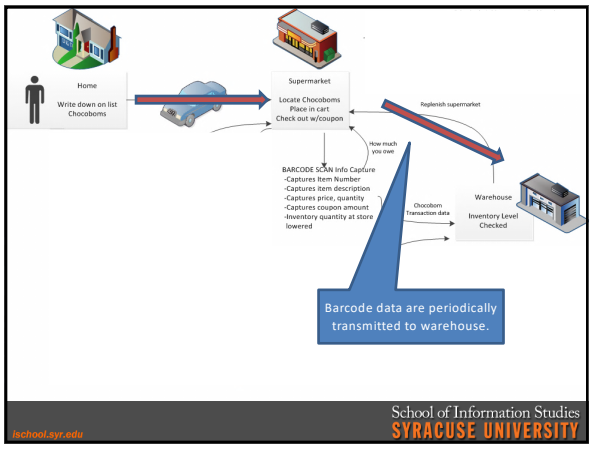
Data Science: First Example

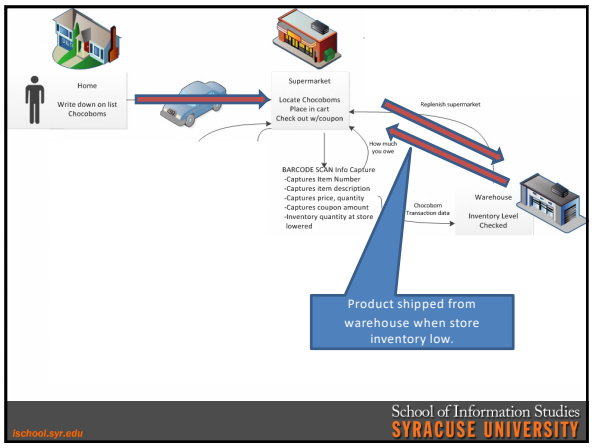
Data Science Example

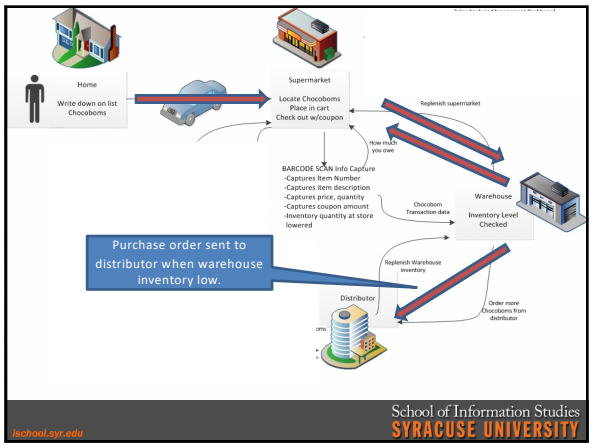


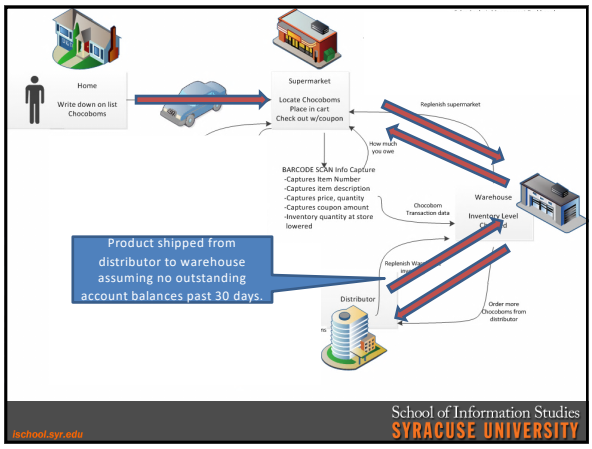
The problem as described by client:
Many variably priced products, like this one, have unanticipated demand spikes that create severe restocking delays.

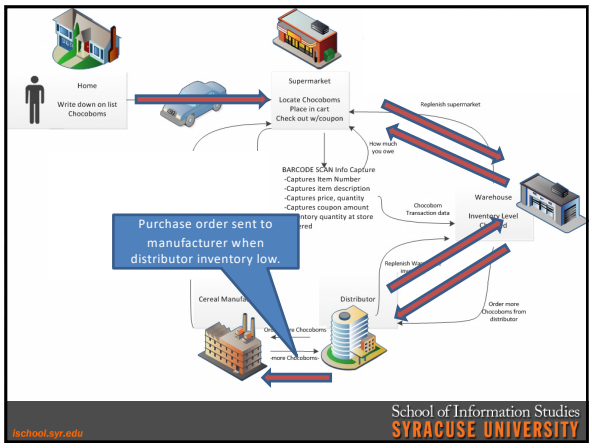


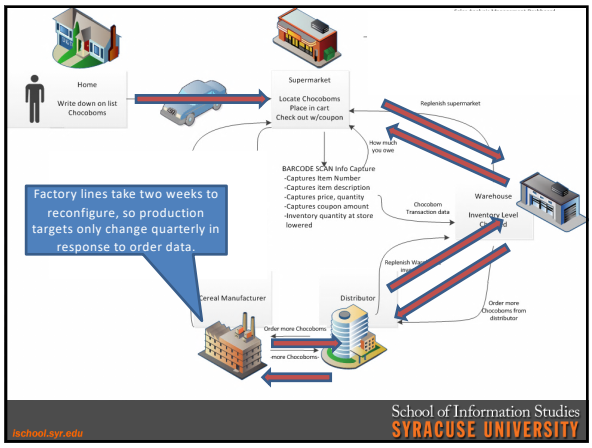


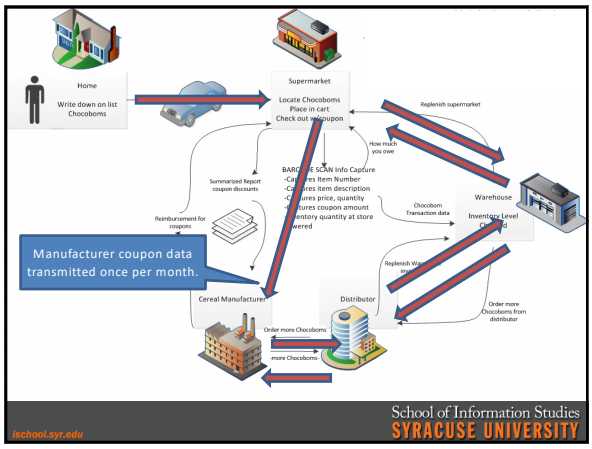


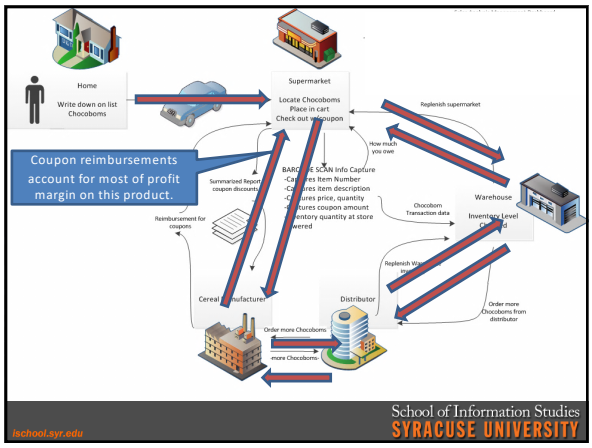






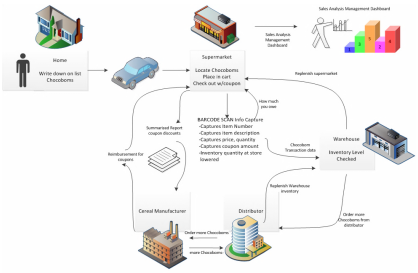


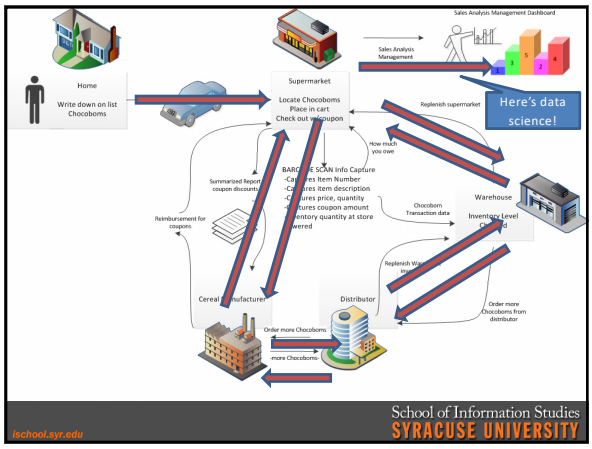




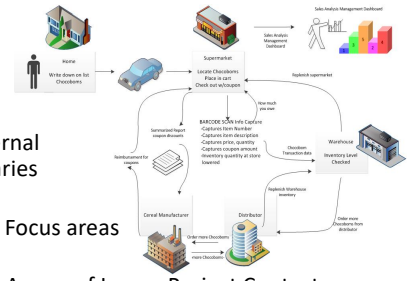
Questions

Isn't this just a supply chain issue?
Where's the data science?



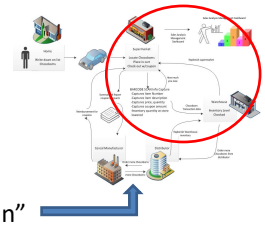


Thinking About the Problem Domain



- Define Scope
- Set Internal Boundaries
- Choose Focus areas
- Remain Aware of Larger Project Context

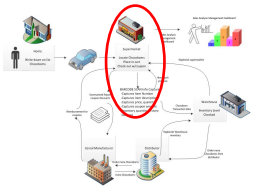
Set Overall Scope



- Problem domain
  - **Define scope**
  - Set internal boundaries
  - Choose focus areas
  - Context

Avoid "boiling the ocean"

Define Internal Boundaries



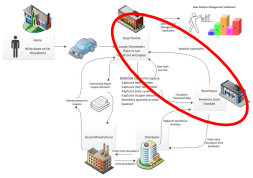
- Problem domain
  - Define Scope
  - **Set Internal Boundaries**
  - Focus area
  - Context

Avoid "boiling the ocean"

Decompose domain into manageable modules

Set Focus Area(s)

- Problem domain
  - Scope
  - Internal Boundaries
  - **Focus area(s)**
  - Context



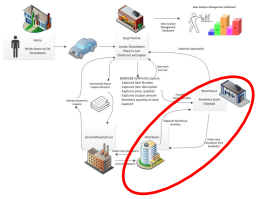
Avoid "boiling the ocean"

Decompose domain into manageable modules

Solve (at least) one important problem well

Enhance Contextual Awareness

- Problem domain
  - Scope
  - Boundaries
  - Focus area
  - **Context**



Avoid "boiling the ocean"

Decompose domain into manageable modules

Solve (at least) one important problem well

All (good) data science includes change processes

The Data Science Process

- Domain analysis/decomposition
- Identification of subject matter expert(s)
- Question/interview/observation process
  - Risks and uncertainty
  - Stories
  - Anomalies

Via anomalies, we learn:

- Is there an error in the system definition
- Is there an error in the system implementation
- Is there an **interesting exception** that teaches us something

Via stories, we learn:

- What people do
- How they do it
- Information produced and consumed
- Who owns what risks
- Process and information touch points
- Decisions made
- Challenges associated with all of the above

Key Takeaways

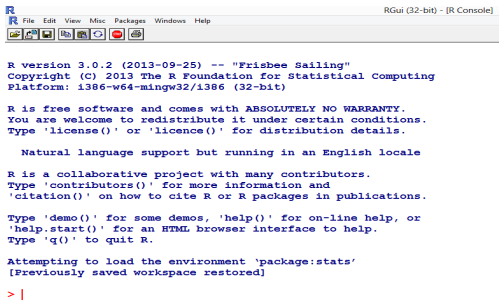
- In contrast to other fields, data science is at its heart concerned with **communicating.**
- **Domain identification** and **understanding** is essential such that we focus on the most pressing *and* tractable problem or opportunity.
- **Identify SMEs** and engage them in "storytelling" about their activities such that you can get a sense of who, what, why they do what they do to also include surfacing anomalies and risk situations.

Getting Started With R

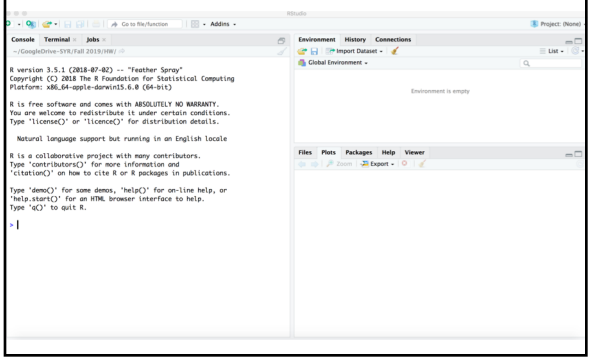
Data Science: Getting Started with R

- R is an open-source software program
  - Developed by volunteers as a service to the community of scientists, researchers, and data analysts who use it.
- R is free to download and use.
- Lots of advice is available online to help users learn R, which is good because it is a powerful and complex program.
- R is a full-featured programming language dedicated to data.

Data Science: Getting Started with R



Data Science: Getting Started with R



Getting Started with R

- Download R installer from: https://cran.r-project.org
- Available for Windows, Mac, and Linux
- Run installer
- Start up R
- Type commands
- End R …. q() at the command prompt
- Respond 'yes' to save your work space

→ **Or use rstudio.cloud**

An Introduction to Data in R

"Atomic" Types in R:

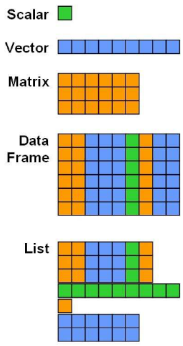
Logical – Either TRUE or FALSE; TRUE generally coded as 1 and FALSE as 0; TRUE and FALSE are reserved keywords

Integer – A number with nothing after the decimal point; sometimes displayed with an "L after it as in 95L

Double – A number that supports digits right of the decimal point

Character – Text data encoded using any of a variety of encoding systems

Rarely used types: Complex and Raw



An Introduction to Data: Type / Mode

List of integers **43, 42, 12, 8, 5**

Each integer represents the age of a family member.

Integer list is all the same "type/mode."

R refers to an atomic list as a "vector."

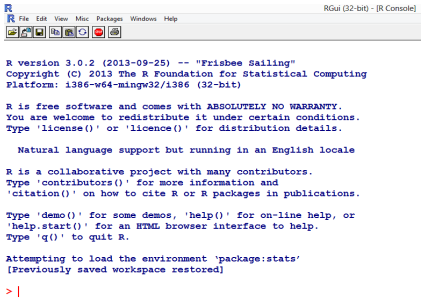
R code to create this vector looks like

c (43, 42, 12, 8, 5)

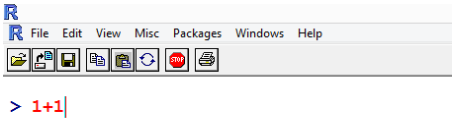
myFamilyAges <- c (43, 42, 12, 8, 5)

Named vector

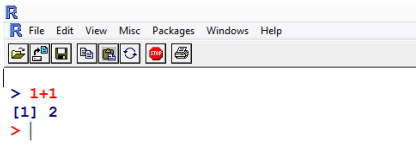
The Initial Console View



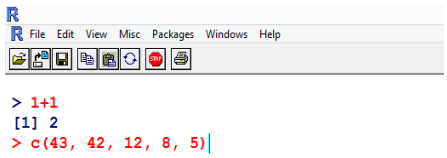
Adding Two Numbers



Adding Two Numbers

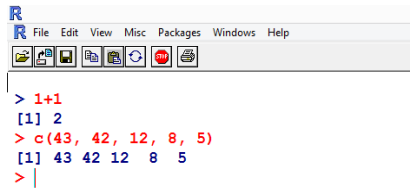


Creating a Vector

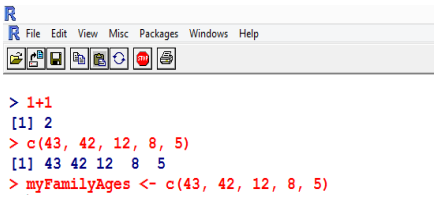


Note that the ‘c’ command stands for combine

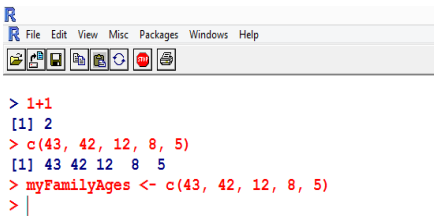
Creating a Vector



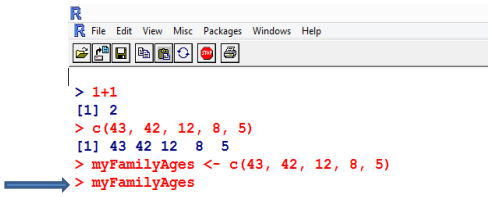
Naming/Storing a Vector



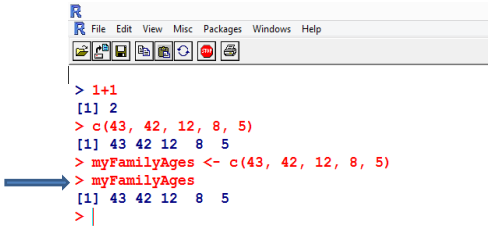
Naming/Storing a Vector



Reporting a Stored Vector

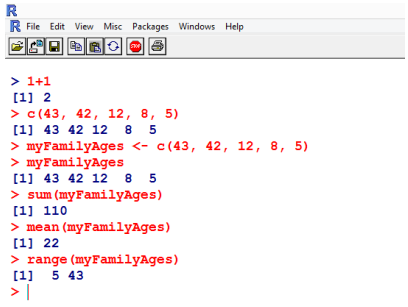


Reporting a Stored Vector

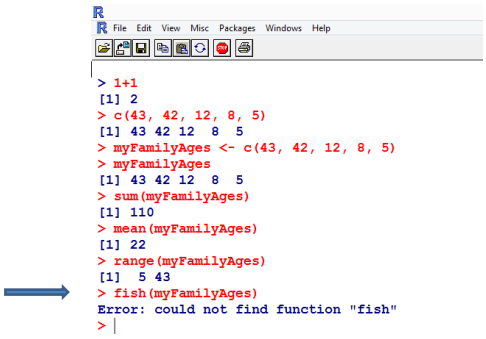


Functions Operating on a Vector

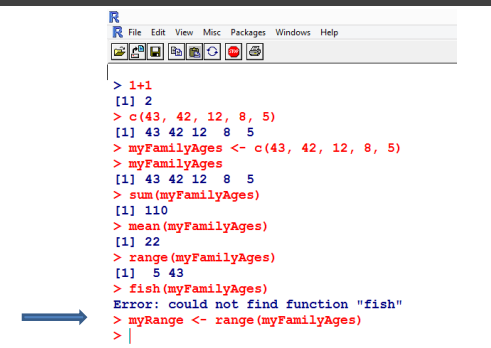
sum
mean
range



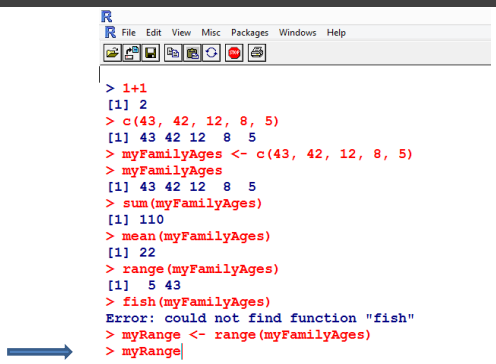
Errors in R



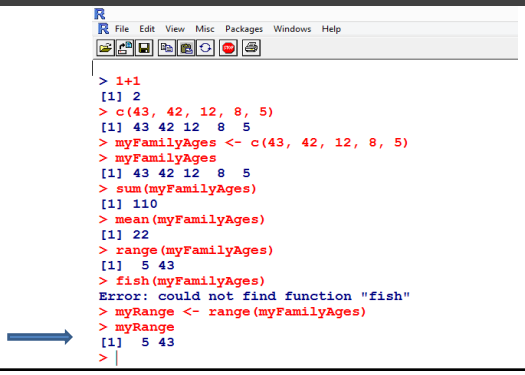
Storing Function Results in a Vector



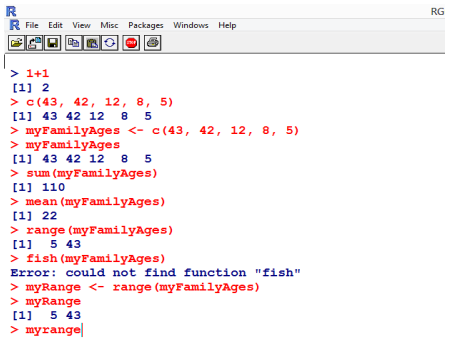
Retrieving Stored Results



Results



Capitalization is Important



Capitalization is Important