Northeastern University
Toronto

# ALY 6000
# Introduction to Data Analytics

Mohammad (**Shafiqul**) Islam, PhD., P.Eng.
Email:m.islam@northeastern.edu


Northeastern
University


# Welcome from Prof. Tom

# Agenda

- General Introduction
- Module Introduction
- R in Action: Chapter 1
- Introduction to Basic Statistics
- R in Action: Chapter 2
- Module Project
- Summary


- I will use this deck as a reference for this class

# Instructor's Background

Recent Professional Experiences:

- Sr. Data Scientist, Sagen Financial Inc (2019- date)
- Sr. Data Scientist, Flank Engineering (2018-2019)
- Data Scientist, Emagin (2017-2018)

Academic:

- Postdoc., MIT (2014-2017)
- PhD., UBC (2010-2013)
- M. Sc. , EuroAquae Consortium (UK, France, Germany)
- B. Sc., BUET, Bangladesh

# Class Background

## Join by Web

1. Go to **PollEv.com**
2. Enter **MISLAM933**
3. Respond to activity

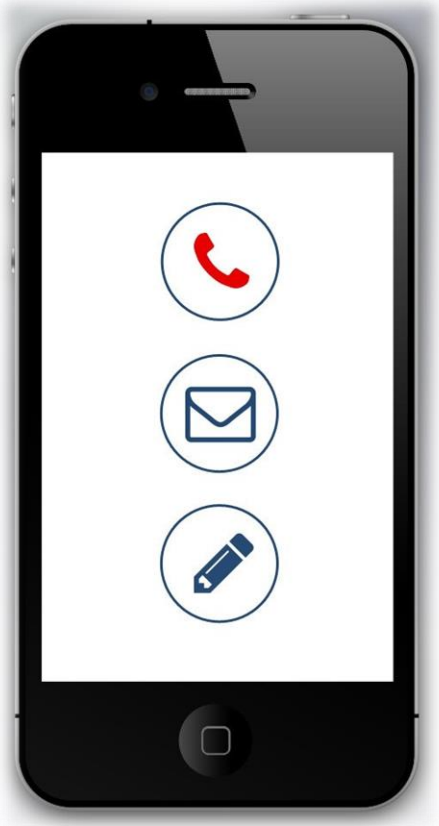
## Join by Text

1. Text **MISLAM933** to **37607**
2. Text in your message

LVX VERITAS VIRTVS

*Your participation and attendance matter*

# Contact and Session Schedule



You can reach me:

- Offline: After class
- Other time: email me please at m.islam@northeastern.edu
- https://www.linkedin.com/in/aitmsi/

**Session Schedule:**

| Day | Time | Location |
|---|---|---|
| Saturday | 12:00pm - 2:40 pm | Class Room 5 (46th Floor) |

# Other Key Resources & Contacts

**1) Public Transportation in Toronto:**

- Toronto Transit Commission (TTC) – www.ttc.ca
- PRESTO – www.prestocard.ca


**2) Student Support Services:**

- Stephanie Cochrane – s.cochrane@northeastern.edu

**3) CPS General Advising: Learner Services**

- LearnerServices@northeastern.edu
- 1-833-NU LEARN (1-833-685-3276)

**4) CPS Toronto Academic and Career Advisors**

- Tinu Olawuyi, David Gugel, (Megan Sykes)
- CPSTorontoAdvising@northeastern.edu

**5) The Office of Global Services (OGS): International Students**

- OGSCanada@northeastern.edu
- International Student Advisor: Krystal Jiang

LVX VERITAS VIRTVS

# Important Dates

- Add period for Winter courses ends on
  - **Sunday, January 16** for Part A classes
  - **Sunday, January 24** for full-term classes
- Drop period for Winter courses ends on:
  - **Sunday, January 24** to drop without a “Withdraw” grade
  - **Sunday, February 13** to drop a Part A class with a "Withdraw" grade
  - **Sunday, March 27** to drop a full-term class with a “Withdraw” grade
- Grades are due by
  - **2pm on Tuesday, February 22** for Part A classes
  - **2pm on Tuesday, April 5** for full-term classes

# Program Learning Outcomes (PLOs)

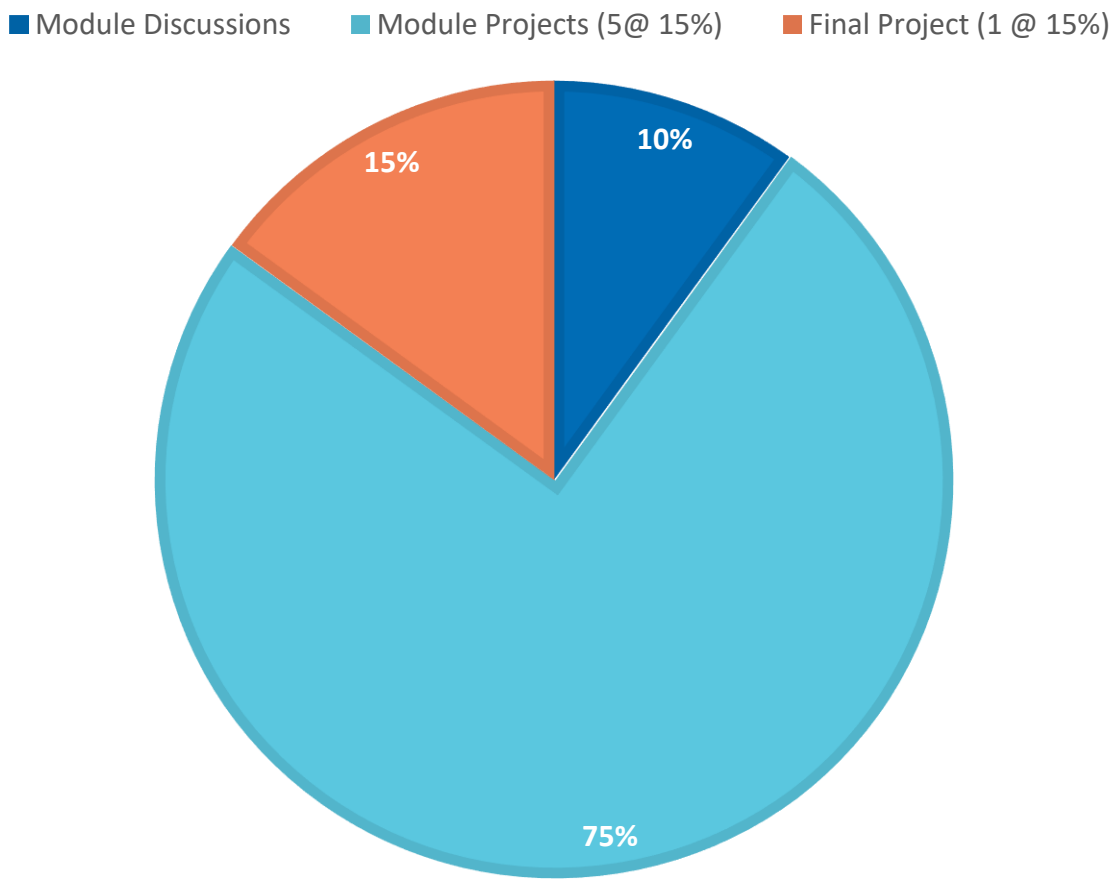
- **PLO1:** Demonstrate the foundational knowledge and skills critical to pursue data analytics as a profession in relation to statistics and math.
- **PLO3:** Demonstrate the knowledge of advanced tools in data analytics.
- **PLO4:** Articulate and effectively defend the significance of leadership, governance, and ethics in data analytics in terms of challenges and trends in a local, national or global context.
- **PLO7:** Design and deliver presentations, reports, and recommendations that effectively translate technical results/data solutions and are coherent and persuasive to different audiences.

# Course Learning Outcomes

Based on satisfactory completion of this course, a student should be able to:

- **CLO1:** Identify and apply basic probability concepts
- **CLO2:** Identify basic statistics measures of central tendency and variance
- **CLO3:** Utilize the “R” as a toolset for processing and analyzing basic data
- **CLO4:** Demonstrate how the analysis of data impacts operational and strategic decision making
- **CLO5:** Visualize data in a compelling way to enable data-driven storytelling
- **CLO6:** Describe the major steps of an analytics project, the various job functions, and who performs them with a focus on identifying where they want to participate in the data analytics ecosystem.

# Course Evaluation



# Submission of Assignments

- The 1st page is the cover-page that should include:
  - Course name and number
  - Module number
  - Student name and number
  - Date of submission

- Please staple all pages together only at the top left-hand corner (if you print)

# Expectations

- Expect 2.5 hours per week classroom activities
- A minimum 5 hours out of class work
- Module Discussions (two or more comments between Saturday and Friday)
- Module Projects (Submission before the next class)
- Practice Assignments(Submission before the next class)
- Final Project

# Late Submission

- Late submission could be subjected to a 10% penalty per calendar day unless otherwise stated.
- Students should communicate with me in advance for a due date extension as a result of exceptional circumstances.

**There will be always some updates on Canvas prior to the class**

# CPS Academic Integrity Module

## Overview

The Academic Integrity Policy at Northeastern ensures the quality and rigor of a Northeastern degree and ensures that you will have the skills and dispositions needed in today's workforce. The resources in this organization are designed to enhance your understanding of academic integrity and help you utilize Northeastern support and services to help you avoid any potential violations in your course work.


Having trouble accessing this video? Click here to open the video in a new window .

LVX VERITAS VIRTVS

# Course Materials

Textbooks:

- Bluman Elementary Statistics: A Step by Step Approach 10th edition, McGraw Hill ISBN 13: 978-1-259-755330 (textbook version) OR eBook Version
- R. Kabacoff, R in Action 2nd edition, Manning, ISBN 978-1-617-29138-8

Software:

- R Script Language (install this software first)
- Rstudio
- Rcloud (https://rstudio.cloud/)
- Python3 (optional)

# Module 1 Introduction

# Module Overview

- Data analysts need to create understandable data visualizations and clearly summarize data for a variety of uses and audiences.
- Data analytics requires a foundation in statistics, statistical computing, and graphics.
- This module will serve as an introduction to statistics and the programming language R.

# Learning Objectives

- Describe the attributes of a given data set using appropriate terms
- Distinguish between descriptive statistics and inferential statistics
- Identify a given sampling methods as simple random sample, stratified random sample or cluster random sample
- Use R to identify and display statistical information
- You are free to use Python, if you want to.

LVX VERITAS VIRTVS

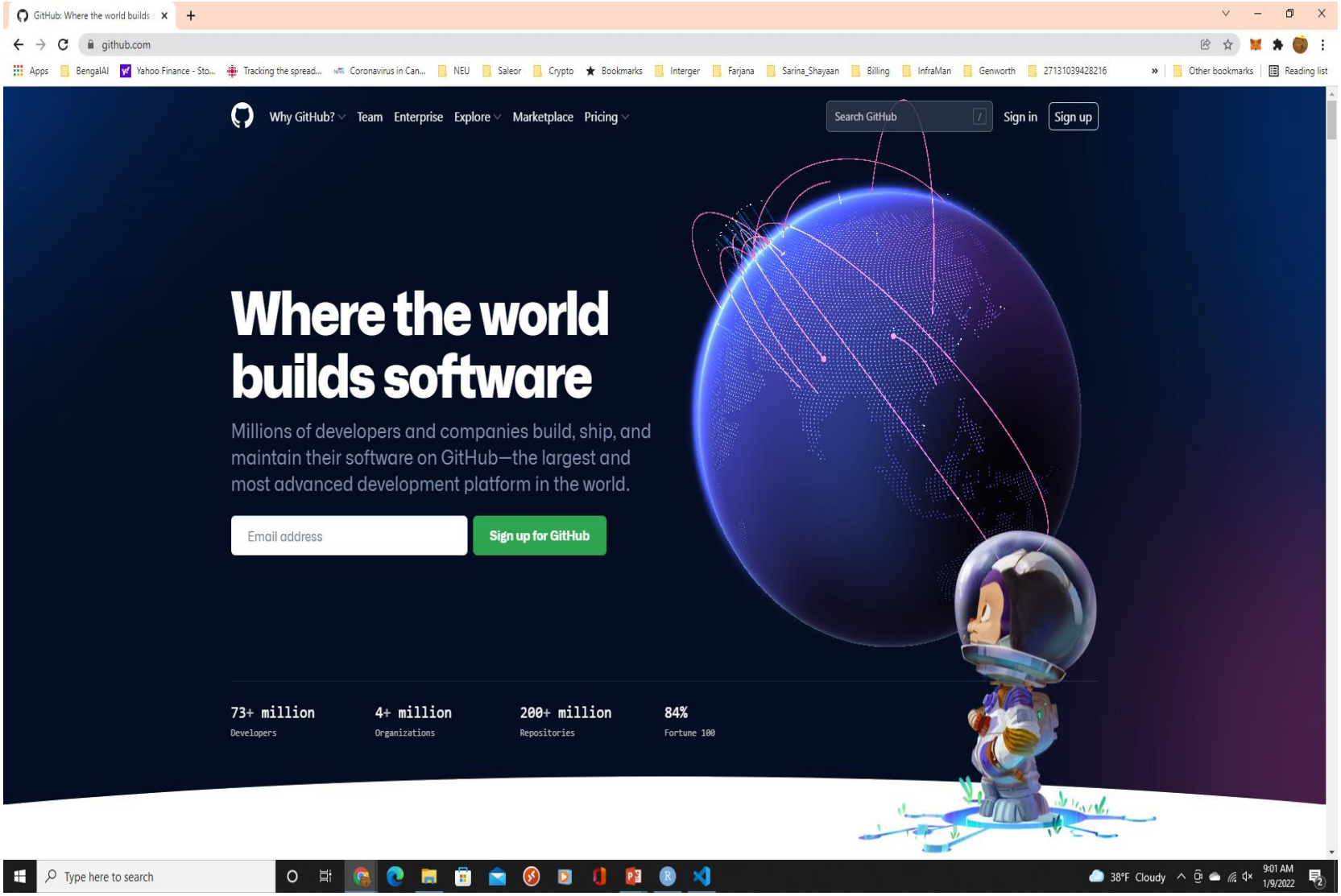
# Task List of the Week

- Download git and create github/gitlab account
- Download and install R Script Language and RStudio
- Read the Module 1 Project assignment and Discussion prompts
- Statistics (view the lessons in Canvas, do assigned reading and ungraded practice problems)
  - Bluman textbook - Chapter 1
- Complete primary Discussion post by Friday
- Getting started with R (do assigned reading, watch instructor videos, complete R practice tasks)
- Kabacoff textbook - Chapters 1 and 2
- Begin Module 1 Project - Executive Summary Report 1 - Due in Module 2

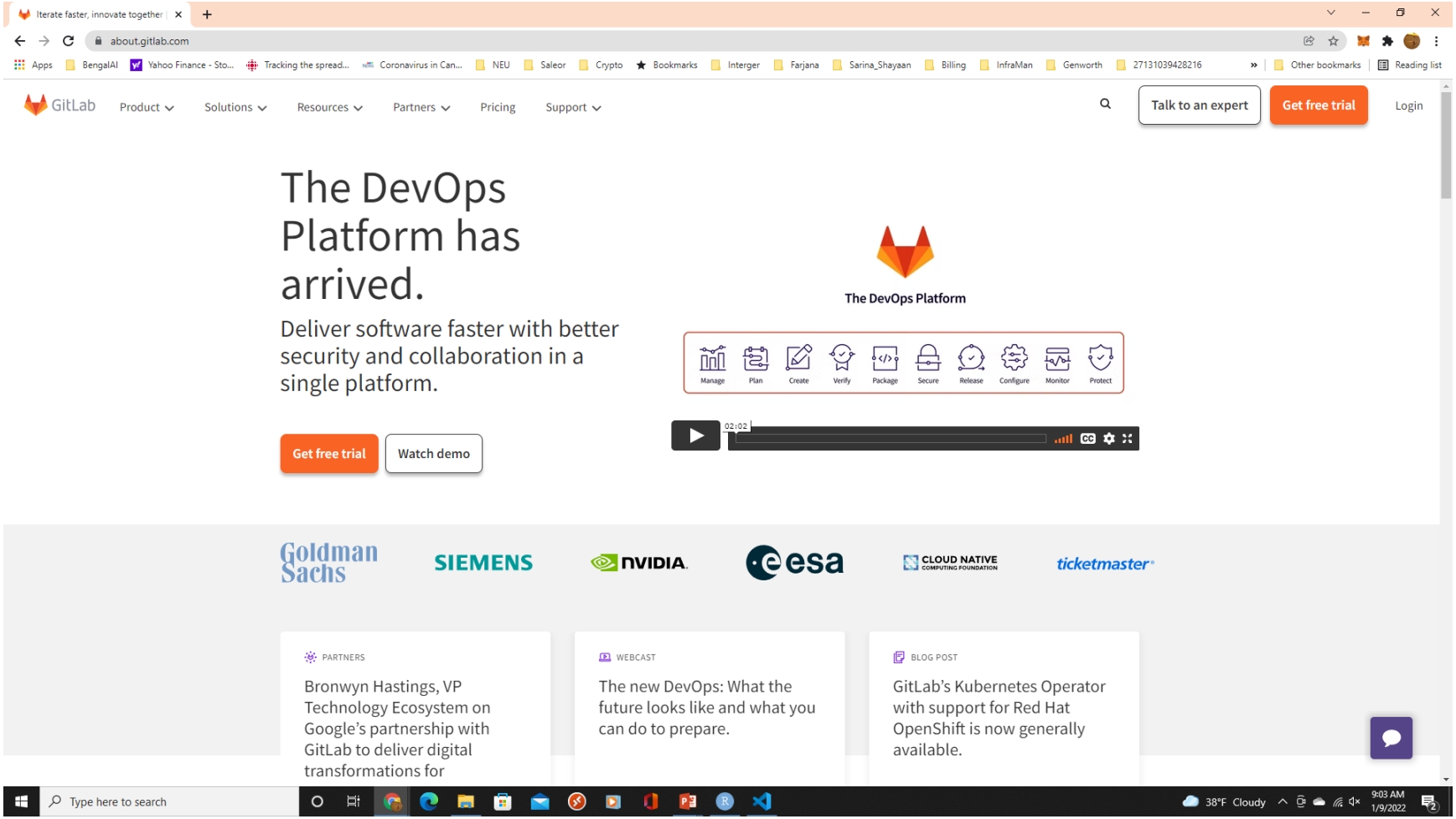
# Introduction to Git

- Git simplifies the process of working with other people and makes it easy to collaborate on projects.
- Team members can work on files and easily merge their changes in with the master branch of the project.
- This allows multiple people to work on the same files at the same time
- Git tracks the content rather than the files

# Introduction to Github



# Introduction to Gitlab



# Introduction to R

R is a language and environment for statistical computing and graphics.

**Key Features:**

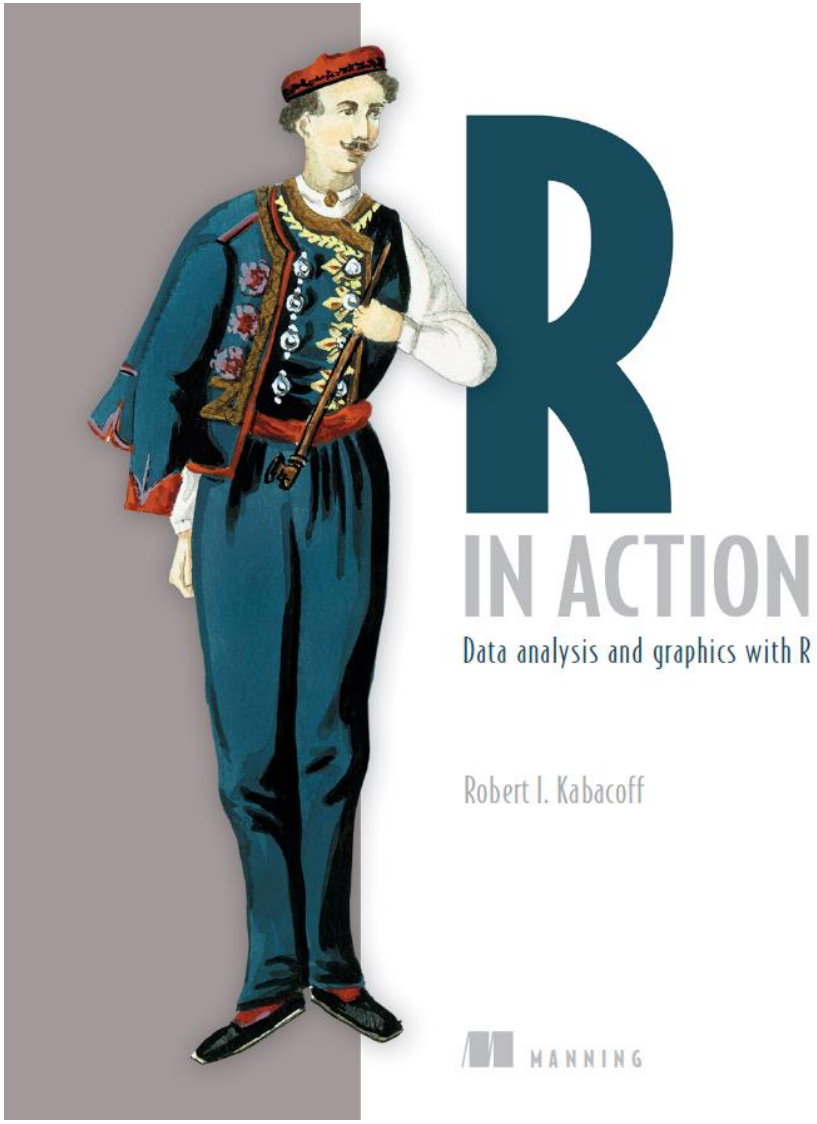
- Open Source
- Powerful! Just about any type of data analysis can be done in R.
- R has state-of-the-art graphics capabilities.
- R contains advanced statistical routines not yet available in other packages.
- R runs on a wide array of platforms, including Windows, Unix, and Mac OS X.
- CRAN and Strong Community
- Course Language

# Introduction to R

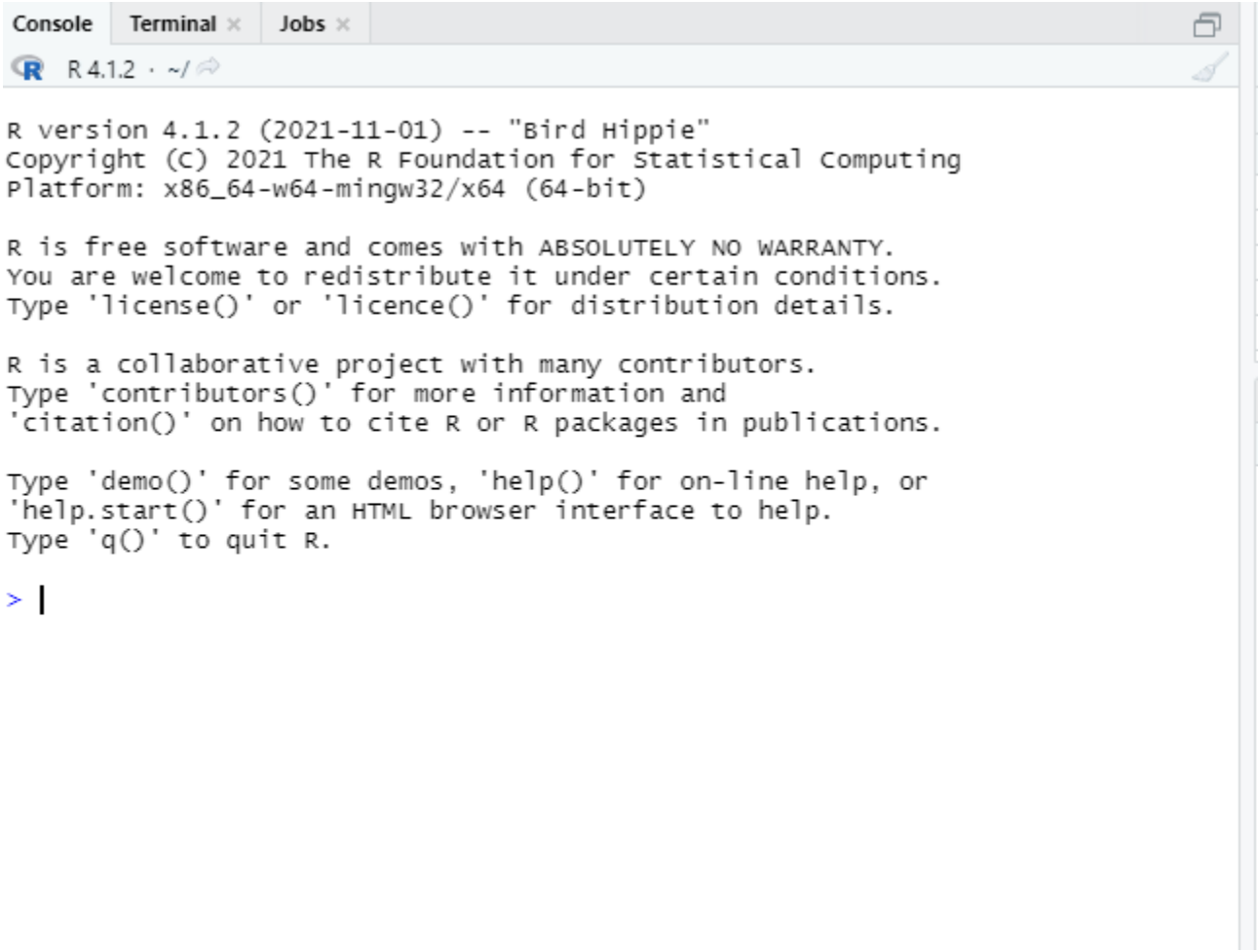
| R | Python |
| --- | --- |
| - Open-source<br>- Data stored in data frames<br>- Formulas and functions readily available<br>- Great Community<br>- Most popular among scholars and researchers | - Open-source<br>- Data stored in data frames<br>- Formulas and functions readily available<br>- Great Community<br>- Versatile language for ML and AI |
| - Great Language for data manipulation, data visualization, and statistics packages<br>- Awesome ggplot2 for visualization | - Integrates with cloud platforms like Google Cloud, Amazon Web Services, and Azure |
| - Inconsistent naming convention<br>- Handling variables may be a little complex for beginners to understand | - Many more decisions for beginners to make about data input/output, structure, variables, packages, and objects |

tiobe-index: https://www.tiobe.com/tiobe-index/

# R in Action


R
IN ACTION
Data analysis and graphics with R
Robert I. Kabacoff
MANNING

# Introduction to R



# What are Packages?

- Packages are collections of R functions, data, and compiled code in a well-defined format.
- The directory where packages are stored on your computer is called the library.
- The function .libPaths() shows you where your library is located, and the function
- library() shows you what packages you've saved in your library.

- Installing Packages is quite easy (install.packages("ggplot2"))
- *Loading a package (library(ggplot2))*

# Common Mistakes in R Programming

- Using the wrong case —help(), Help(), and HELP() are three different functions (only the first will work)
- Forgetting to include the parentheses in a function call —for example, help() rather than help. Even if there are no options, you still need the ().
- Using the \ in a pathname on Windows; setwd("c:\mydata") generates an error instead setwd("c:/mydata") or setwd("c:\\mydata")
- Using a function from a package that's not loaded —The function order.
- Missing dependency

# Introduction to R

**Practice R using following dataset:**

| Age (mo.) | Weight (kg.) | Age (mo.) | Weight (kg.) |
|---|---|---|---|
| 01 | 4.4 | 09 | 7.3 |
| 03 | 5.3 | 03 | 6.0 |
| 05 | 7.2 | 09 | 10.4 |
| 02 | 5.2 | 12 | 10.2 |
| 11 | 8.5 | 03 | 6.1 |

Note: These are fictional data.

# Listing 1.1 - A Sample R session

```
› age <- c(1, 3, 5, 2, 11, 9, 3, 9, 12, 3)
› weight <- c(4.4, 5.3, 7.2, 5.2, 8.5, 7.3, 6, 10.4, 10.2, 6.1)
› mean(weight)
› sd(weight)
› cor(age, weight)
› plot(age, weight)
› # q()
```

# Manage the R Workspace

```
› setwd("C:/myprojects/project1")
› options()
› options(digits=3)
› x <- runif(20)
› summary(x)
› hist(x)
› savehistory()
› save.image()
› # q()
```

# Working with a new Package

› help.start()
› install.packages("vcd")
› help(package = "vcd")
› library(vcd)
› help(Arthritis)
› Arthritis
› example(Arthritis)
› # q()

# Introduction to Basic Statistics

# Important Definitions

**Data:**

- A collection of information (literally, data is the plural of datum; meaning: what is given)
- Facts and figures from which conclusions can be drawn

**Dataset**: the data that are collected for a particular study

- Elements: may be people, objects, events, or other entries
- Data and dataset often used interchangeably

**Parameter:**

A characteristic of a population (often, a numerical characteristic such as a population mean, a population variance, a population standard deviation, etc.)

**Variable**: any characteristic of an element.

**Measurement:** A way to assign a value of a variable to the element

# Data Types

There are two types of data:

- **Categorical (Qualitative):**
  - Nominal: According to Name
    Examples: Data containing names, genders, races, etc.
  - Ordinal: According to Order
    Examples: Data containing ranks, data that has been organized alphabetically, etc.
- **Numerical (Quantitative):**
  - According to the ratio scale (a possible value of zero in the data is an inherent zero)
    Examples: Data containing heights, weights, time durations, grades, etc.
  - According to the interval scale (a zero is not inherently zero)
    Example: Data containing temperatures

# Numerical (Quantitative) Data

Statistically, a numerical set of data may be discrete or continuous.

**Discrete data:**

A discrete data set is one in which the measurements take a countable set of isolated values. For example, the number of chairs, the number of patients, the number of accidents, etc., are all examples of discrete data.

**Continuous data:**

A continuous data set is one in which the measurements can take any real value within a certain range. For example, the amount of rainfall in Charlotte in January during the last 30 years or the amount of customer waiting times at a local bank are examples of continuous data sets.

# Important Definitions

- To begin let's look at Merriam-Webster's definition of statistics. According to Merriam-Webster.com the word "statistics" has two parallel definitions.
  - It is the science of data; that is, collecting, organizing, and analyzing data
  - It is the plural of the word "Statistic“

Statistic:  A characteristic of a sample (such as a sample mean or a standard deviation, etc.)

# Types of Statistics

- Descriptive statistics
- Inferential Statistics

**Descriptive statistics** is used to describe a set of data graphically or numerically

- **Graphical Descriptive Statistics**
  Describes a set of data graphically by creating bar graphs, pie charts, histograms, line plots, scatter plots, etc.

- **Numerical Descriptive Statistics**
  There are a number of particular characteristics of data that are often the focus of interest to the data analyst, like mean, mode, median.

# Numerical Descriptive Statistics

- Measures describing the center of data
  - mean (arithmetic average), median. mode, and the weighted mean
- Measures describing the variability (spread or dispersion) of data
  - the range, the variance, and the standard deviation of data
- Measures of location
  - Examples of such measures are the percentile ranking and the z-score. These measures describe where a particular measurement stands compared to the rest of the data.
- Measures describing the shape of the distribution of data
  - Skewness and kurtosis are two measures that describe the shape of the distribution of a data set

# Inferential Statistics

The process of utilizing one or more random samples in order to gain insight about the population from which those samples were selected.

There are three main methods of inferential statistics:

- Constructing Confidence Intervals: This is to estimate a population parameter to within two limits: a lower limit and an upper limit
- Performing Hypothesis Testing: This is to verify or to reject hypotheses or claims
- Modeling or Testing Relationships between Data sets

# Population

- A set of all elements about which we wish to draw conclusions
- Measurement of the variable of interest for each and every population unit
- For example, annual starting salaries of all graduates from last year's MBA program
- Sometimes called observations
- If the population is too large, analyze a subset

# Samples

A sample is a subset of elements from the set of individuals with one or more common features, known as the population, which has been selected for the study. The number of elements in a sample is denoted by n.

- n : number of elements in a sample
- N: number of elements in the population
- $n < N$

Samples are necessary to learn about populations, because in most real-world examples, it is impossible to measure a characteristic from every member of a population.

Here are some examples of large populations from which it would be too difficult to measure characteristics:

- The population of humans on Earth is more than 7 billion
- The population of the US is more than 300 million
- There are 1.8 billion bottles of Coca-Cola sold each day
- There may be more than 1 million pigeons in New York City alone

# Methods of Sampling

Four Methods of Sampling:

In order to eliminate bias, statisticians use four basic methods of sampling that are designed to ensure that each member of a population has an equal probability of being selected for the sample. These four sampling techniques are called

- Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling

# Statistical Estimation

- In inferential statistical analysis, we use samples to make generalizations about the populations from which they were selected.
- Statistical estimation is a specific type of inferential statistical analysis where we use statistics calculated from data measured from random samples to estimate population parameters.
- Statistical estimation is also used to quantify the uncertainty in these estimates.

# Statistical Notations

Common Statistical Notations:

$\overline{X}$: sample mean

$\mu$: population mean

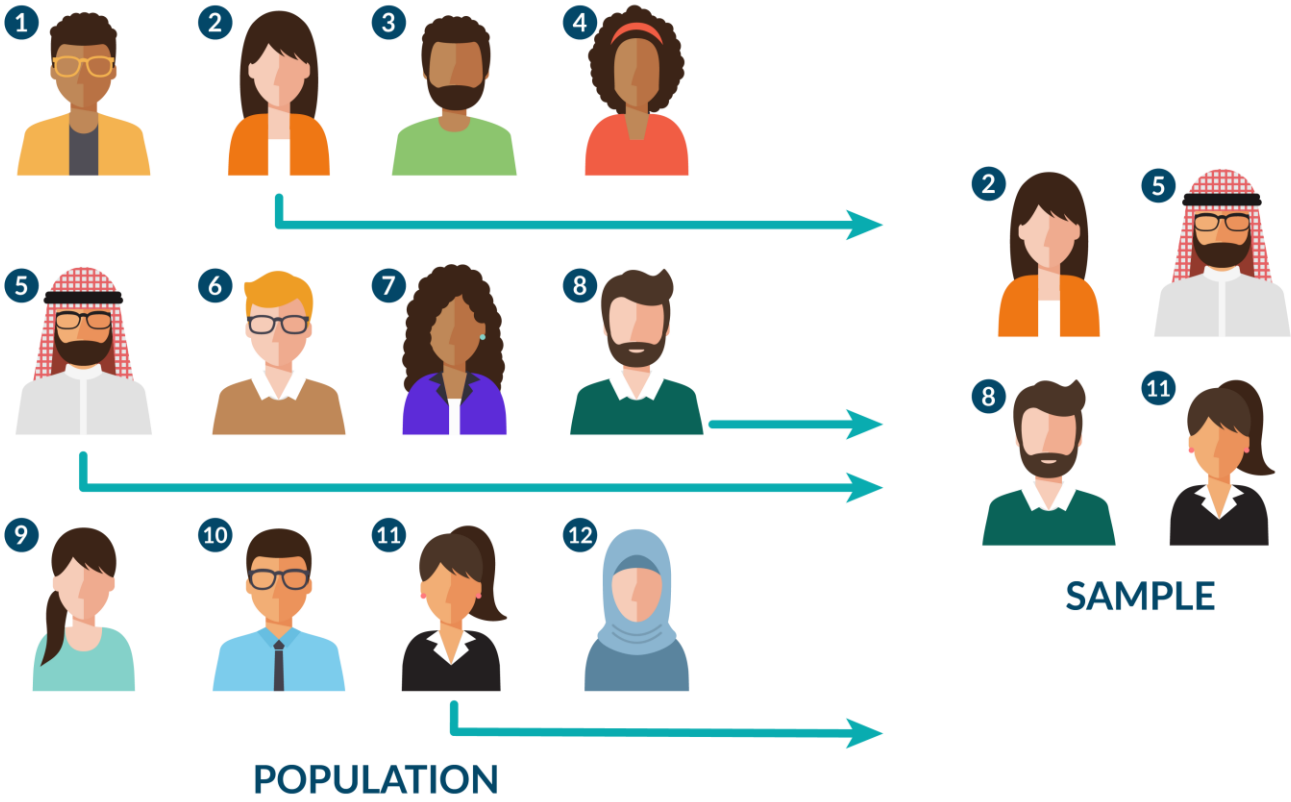
$s^2$: variance of a sample

$\sigma^2$ (sigma squared): variance of a population

$s$: standard deviation of a sample

$\sigma$: standard deviation of the population

# Random Sampling



# Random Sampling

A random sample is defined as a subset of a population, where the subset is chosen in such a way that every member of the population has an equal chance of being selected for the subset or sample. The number of members of a random sample is also denoted by n.

- Consider the following example to illustrate the concepts of random sampling and representative samples:

  population:
  {1, 2, 3, 4, 5}
  $N$=5
  $n$=2

- 10 possible **random samples** *without* replacement
  {1, 2,} {1, 3,} {1, 4}, {1, 5}, {2, 3}, {2, 4}, {2, 5}, {3, 4}, {3, 5}, {4, 5}
  The population is integers 1 through 5.
- We collect a random sample without replacement from the population of size n equals two. There are ten distinct random samples that can be drawn from this population, and they are each equally likely.

# Systematic Sampling

A systematic sample is a sample obtained by selecting every kth member of the population where k is a counting number.

- Where k is a counting number.

Example:

Suppose a bottling company would like to test the machines that are filling the bottles by selecting a sample of filled bottles and measuring the amount of product that the machine is putting in the bottles. The company statistician goes to the end of the bottling line and selects every 20th bottle and removes it for testing. This “system” has thus generated a systematic sample

# Stratified Sampling

# Stratified Sampling

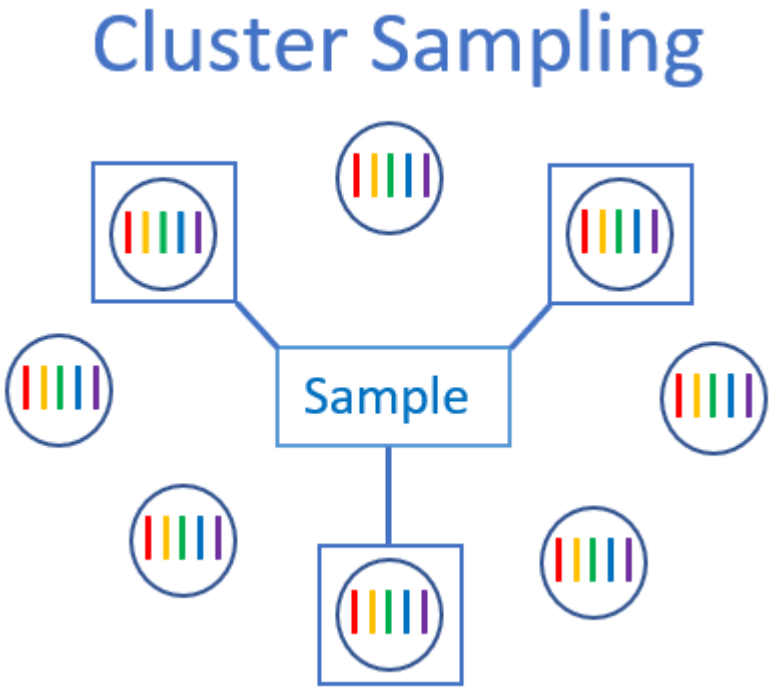
In stratified sampling, you divide the population into separate groups, called strata, and then do a simple random sample from each stratum.

- Population N is divided into H strata by stratification
- Each population member is assigned to exactly one stratum

# Cluster Sampling

In cluster sampling, the population is divided into separate groups or clusters, and then a set of these clusters, is randomly selected as the final sample.

- Not necessary to obtain samples from all clusters as each cluster reflects the entire population, and their homogeneity makes them interchangeable
- we can divide rural communities into similar groups and pick a random sample of communities.




Source: statisticsbyjim.com

# Review of Module Project

- Module 1 Project Instructions
- Executive Summary Report 1

# Summary

- Introduction
- Reviewed Course Objectives
- Introduced Git and R
- Took our first look at data and became familiar with different types of data and statistics.
- Introduced to different forms of sampling which will become more important as you progress.
- Reviewed Module Project

# Q &A