

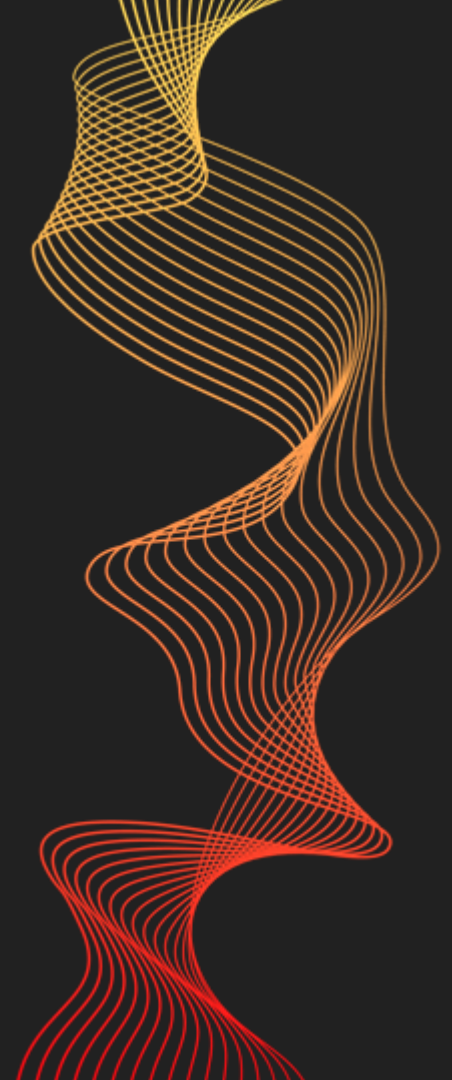


# **Coding in R For Healthcare Informatics and Health Data Science**

Kimberly Deas, MS, PhD Candidate  
BlackTIDES Informatics Lead



This presentation provides an introduction to coding in R for healthcare informatics and health data science. We'll cover the basics of R, including installation and setup, as well as data manipulation and analysis techniques. We'll end with live coding of a healthcare data set.





# Part 1: Introduction to R





## Part 1: Introduction to R

- What is R?
- Why R is Useful in the Healthcare Data Analytics Field?
- Why R is useful in the healthcare data analytics field.
- R Resources

# Introduction to R: What is R?



- R is a versatile programming language and software environment, highly regarded for statistical computing, data analysis, and data visualization.
- As an open-source tool, it's supported by a robust community and offers a vast repository of packages for statistical and data manipulation tasks.
- It excels in handling large datasets and complex analyses, while its straightforward syntax makes it accessible to users at all levels.
- R's integration with other languages and various data formats enhances its adaptability in a wide range of data scenarios.

# Introduction To R: Why R is used in Statistical and Data Analysis.



- R's popularity in statistical and data analysis is driven by its comprehensive statistical techniques, covering everything from linear and nonlinear modeling to time-series analysis.
- Its strength lies in a vast array of user-contributed packages that simplify complex tasks and adapt to evolving methodologies.
- R's advanced plotting systems like ggplot2 enable the creation of detailed and insightful visualizations.



# Introduction To R: Why R is Useful in the Healthcare Data Analytics Field?



- R plays a crucial role in healthcare data analytics with its proficiency in handling large, complex datasets common in medical research and public health.
- Its specialized packages for biostatistics and epidemiology, such as 'survival' and 'lme4', are essential tools in medical research for analyzing trends and patterns in health data.
- R excels in predictive modeling and machine learning, key for developing personalized medicine and forecasting disease outbreaks.
- Additionally, its ability to visualize health data simplifies complex information, aiding in effective communication for healthcare decision-making.

# Introduction To R : Resources

- IMO, THE definitive book on R: *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data 2nd Edition* by Hadley Wickham (Author), Mine Çetinkaya-Rundel (Author), Garrett Golemund (Author)
- R Course: <https://www.udemy.com/course/r-programming/>. Kirill is also a great Tableau instructor.
- Others?



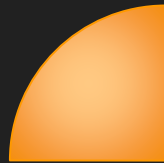


# Part 2: Basics of R

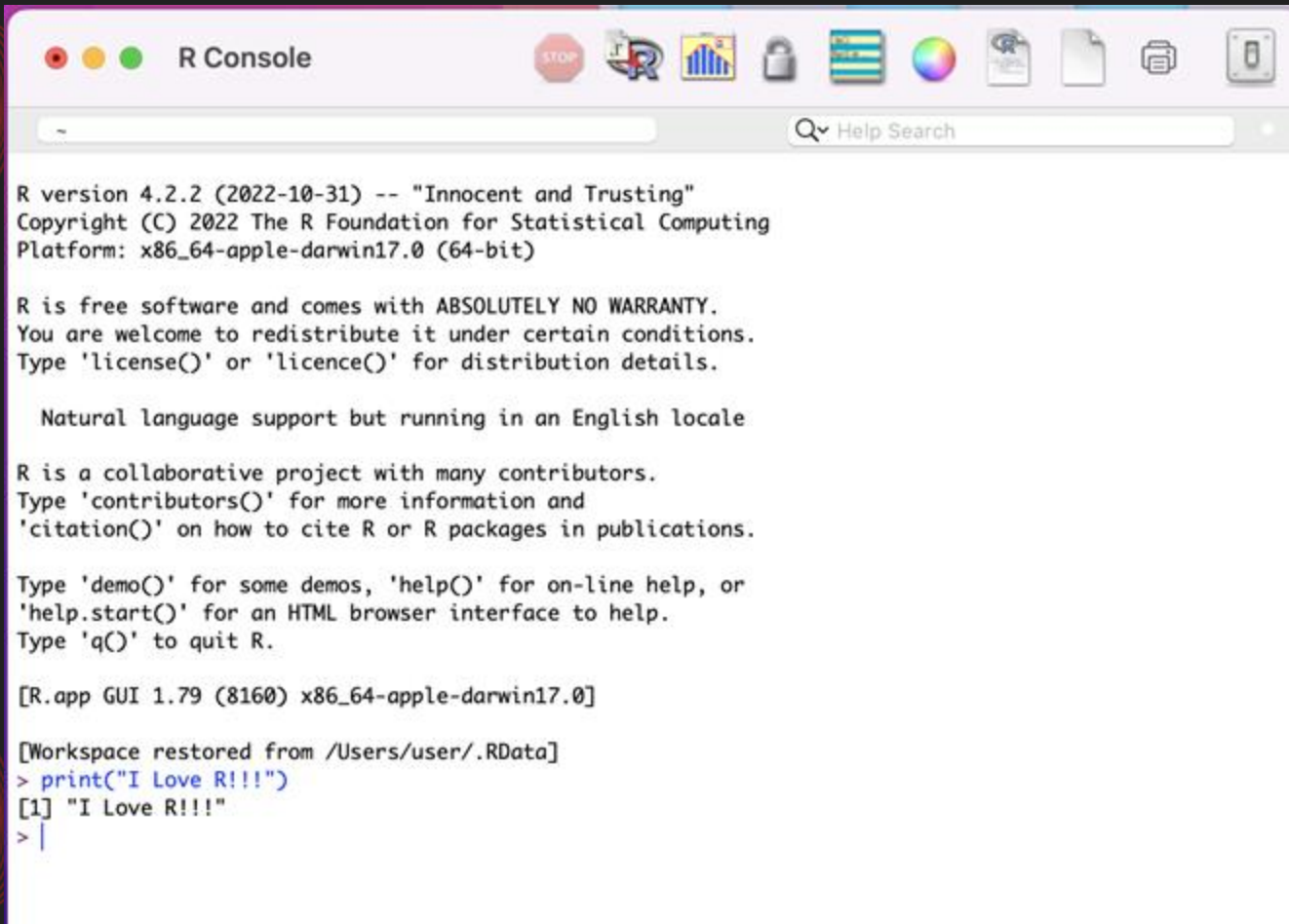


# Basics of R: Getting Started with R: Installation and Setup



- Installing R and R studio:  
<https://rstudio-education.github.io/hopr/starting.html>
  - Live demo of R and R studio
- 

# Basics of R: R Console



The screenshot shows the R Console window on a macOS system. The title bar reads "R Console". The menu bar includes standard macOS icons (red, yellow, green buttons) and application-specific icons (STOP, R logo, bar chart, lock, R logo, rainbow circle, R logo, document, printer, and a mobile device icon). A search bar labeled "Help Search" is visible. The main content area displays the R version 4.2.2 startup screen, which includes the version number, date, copyright notice, platform information, and a list of commands for users to explore. The user has entered the command `print("I Love R!!!")`, and the console has outputted `[1] "I Love R!!!"`. The prompt `>` is visible at the bottom.

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

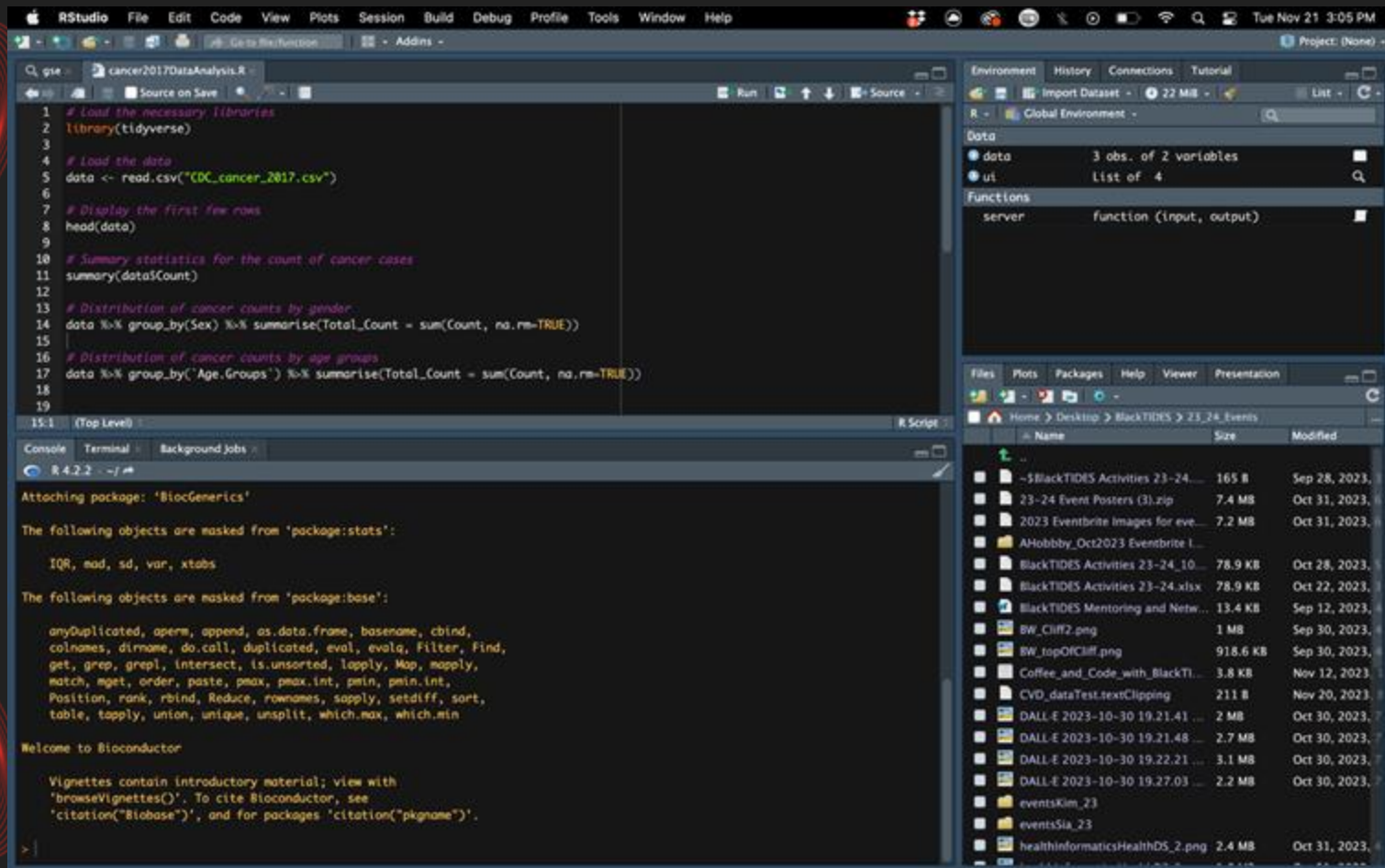
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.79 (8160) x86_64-apple-darwin17.0]

[Workspace restored from /Users/user/.RData]
> print("I Love R!!!")
[1] "I Love R!!!"
```

# Basics of R: R Studio



The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains an R script named `cancer2017DataAnalysis.R` with the following code:

```
1 # Load the necessary libraries
2 library(tidyverse)
3
4 # Load the data
5 data <- read.csv("CDC_cancer_2017.csv")
6
7 # Display the first few rows
8 head(data)
9
10 # Summary statistics for the count of cancer cases
11 summary(data$count)
12
13 # Distribution of cancer counts by gender
14 data %>% group_by(Sex) %>% summarise(Total_Count = sum(count, na.rm=TRUE))
15
16 # Distribution of cancer counts by age groups
17 data %>% group_by("Age.Groups") %>% summarise(Total_Count = sum(count, na.rm=TRUE))
18
19
```
- Console:** Shows the output of the R session:

```
R 4.2.2 ~ /
Attaching package: 'BiocGenerics'

The following objects are masked from 'package:stats':

  IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':

  anyDuplicated, append, as.data.frame, basename, cbind,
  colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
  get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
  match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
  Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
  table, tapply, union, unique, unsplit, which.max, which.min

Welcome to Bioconductor

Vignettes contain introductory material; view with
'browseVignettes()'. To cite Bioconductor, see
'citation("Biobase")', and for packages 'citation("pkgname")'.

> |
```
- Environment:** Shows the current environment with variables `data` (3 obs. of 2 variables) and `ui` (List of 4).
- Files:** Displays a file explorer view of the `BlackTIDES` directory, showing files like `BlackTIDES Activities 23-24.xlsx` and `BlackTIDES Mentoring and Netw...`.

# Basics of R: Basic R Syntax and Commands

```
# Variables and data types
age <- 30          # Numeric type
patient_id <- "P001" # Character type
smoker <- TRUE     # Logical type

# Basic operations
new_age <- age + 1
print(new_age)

# Data structures: vector, matrix, list, data frame
vector <- c(1, 2, 3)
matrix <- matrix(1:9, nrow=3)
list <- list(age, patient_id, smoker)
data_frame <- data.frame(ID=c(1,2), Name=c("Alice", "Bob"))
```





# Live demo R Studio - Basics







## Part 3: Overview of the Dataset



# Basics of R: Overview of the Dataset



- Loading the Dataset in R
- Exploratory Data Analysis
- Go to console



## Part 4: Data Manipulation and Analysis



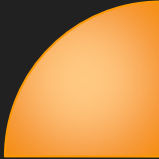
# Data Manipulation and Analysis:

## Data Cleaning

- Handling missing values
- Data Transformation
- Go to console

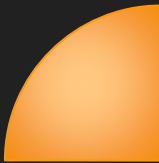


# Data Manipulation and Analysis: Basic Data Analysis Techniques

- Descriptive statistics
  - Simple Visualizations
  - Bar chart
  - Scatter plot
  - Go to console
- 

# Data Manipulation and Analysis: Using packages for Data Manipulation and Analysis



- dplyr and ggplot2
  - Data Manipulation with dplyr
  - Advanced data viz with ggplot2
  - Go to console
- 





Thank you. Please feel free to ask any questions.

