

PRE-SURVEY



Rosen Center for
Advanced Computing

Introduction to R

Ashish

Senior Research Data Scientist

Rosen Center of Advanced Computing (RCAC)



Agenda

01

Introduction to
R and RStudio

02

Basic
Operations and
Data Structures

03

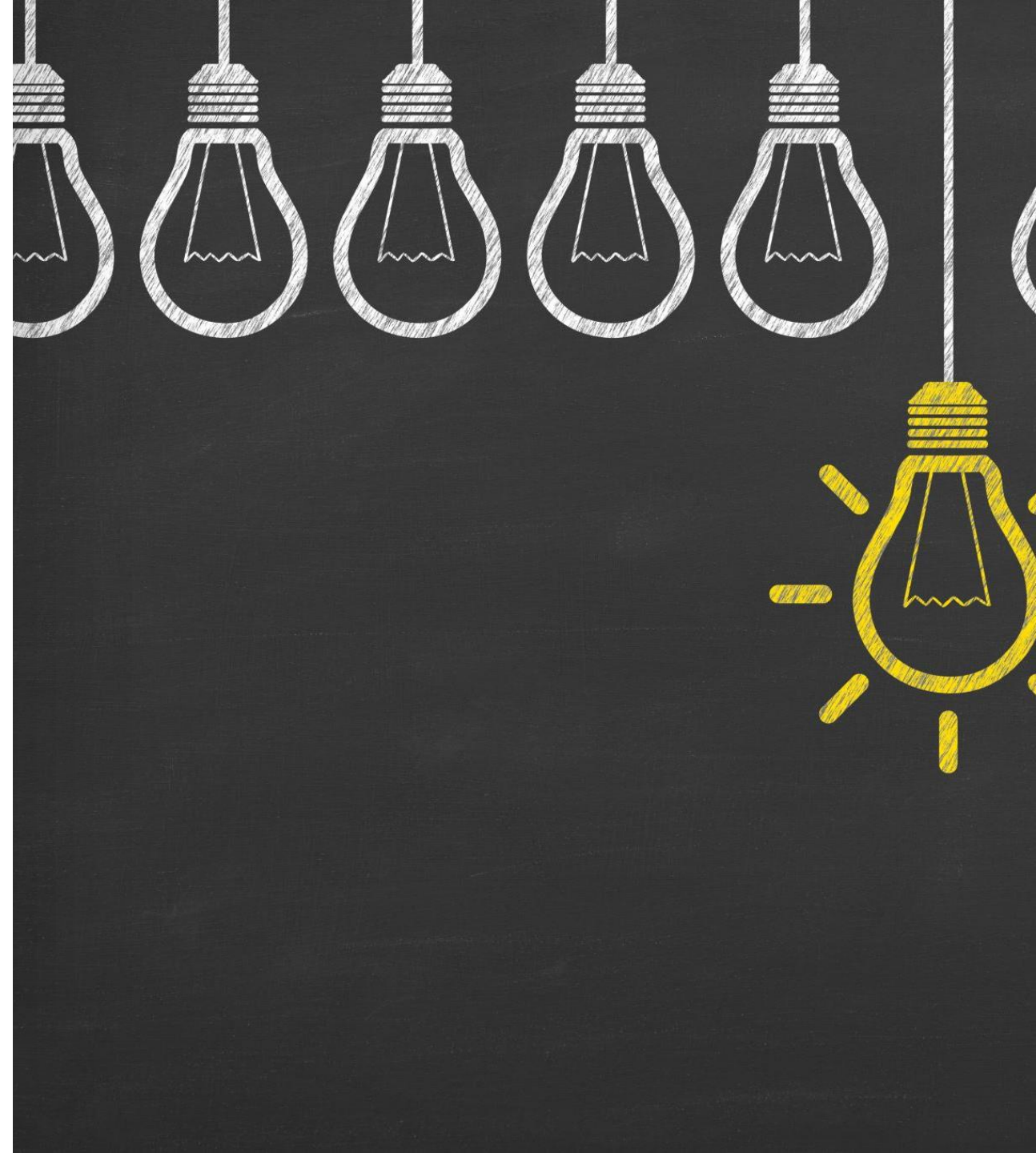
Data
Manipulation
with dplyr and
tidyverse

04

Basic and
intermediate
Data
visualization

What is R?

- The R environment combines:
- effective handling of big data
- collection of integrated tools
- graphical facilities
- simple and effective programming language



Why use R?

Visualization

Open source

Data science



Platform agnostic

Computational
statistics

What is RStudio?



Graphical user interface, not just a command prompt



Great learning tool



Free for academic use



Platform agnostic



Open source

Basic syntax

The main "parts of speech" in R (syntax) include:

the **comments** # and how they are used to document function and its content

variables and **functions**

the **assignment operator** <-

the = for **arguments** in functions

NOTE: indentation and consistency in spacing is used to improve clarity and legibility

Data Types



Numeric



Character



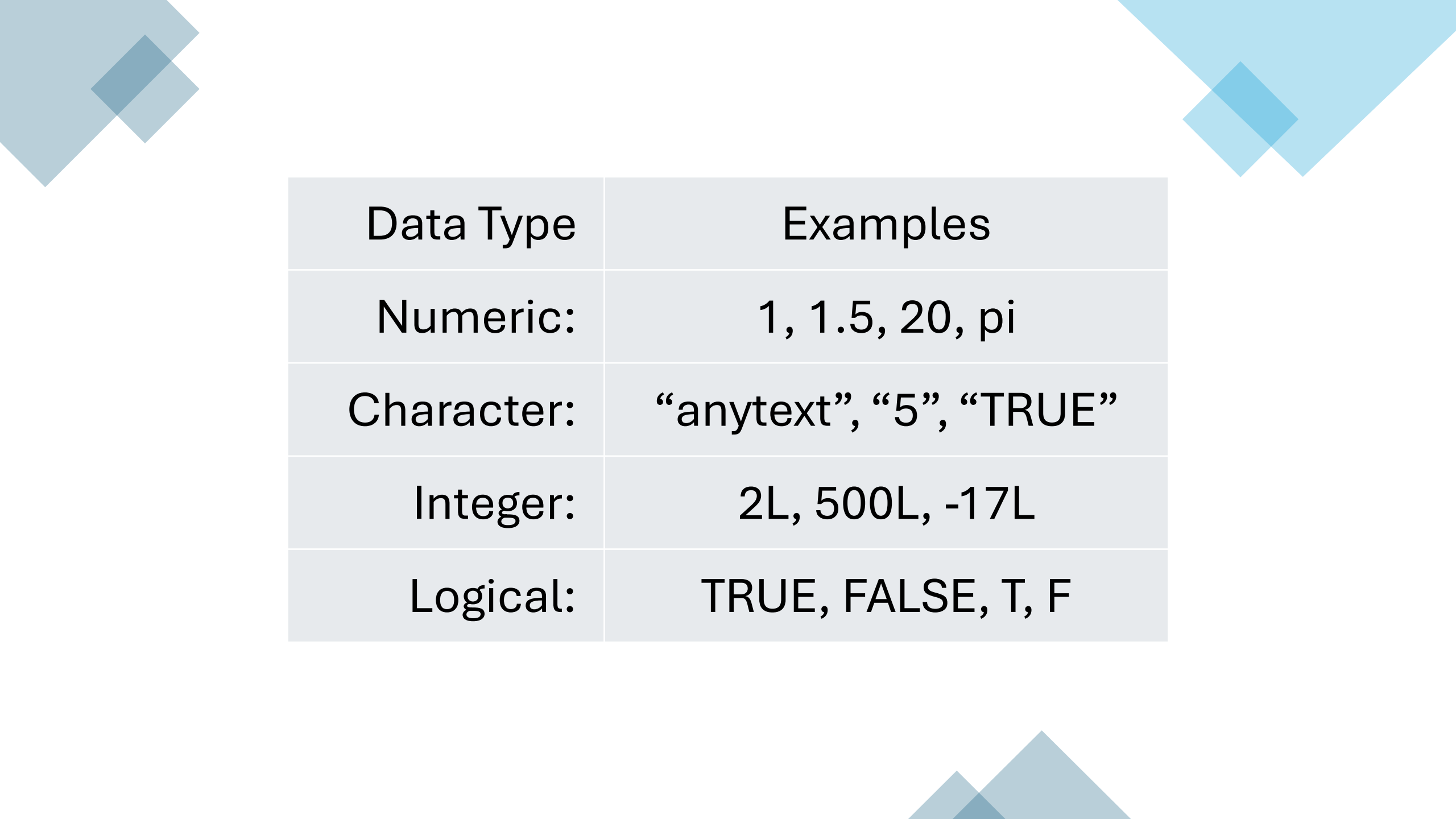
Logical



Complex



Integer



Data Type	Examples
Numeric:	1, 1.5, 20, pi
Character:	“anytext”, “5”, “TRUE”
Integer:	2L, 500L, -17L
Logical:	TRUE, FALSE, T, F

Data Structures

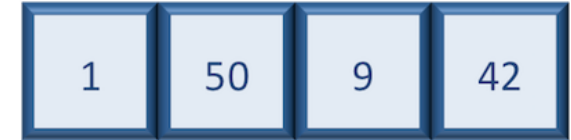
Vectors

Factors

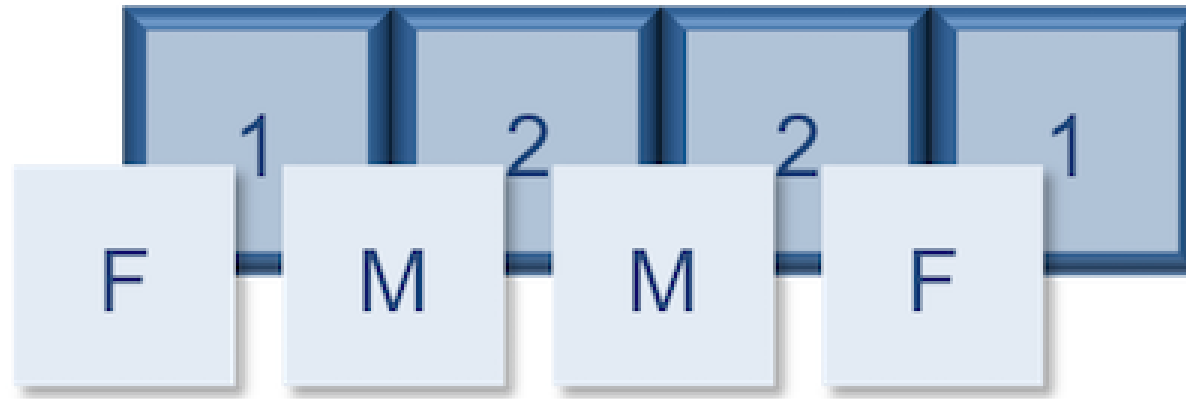
Matrix

Data frame

Lists



Vectors



FACTORS

A **factor** is a special type of vector that is used to **store categorical data**

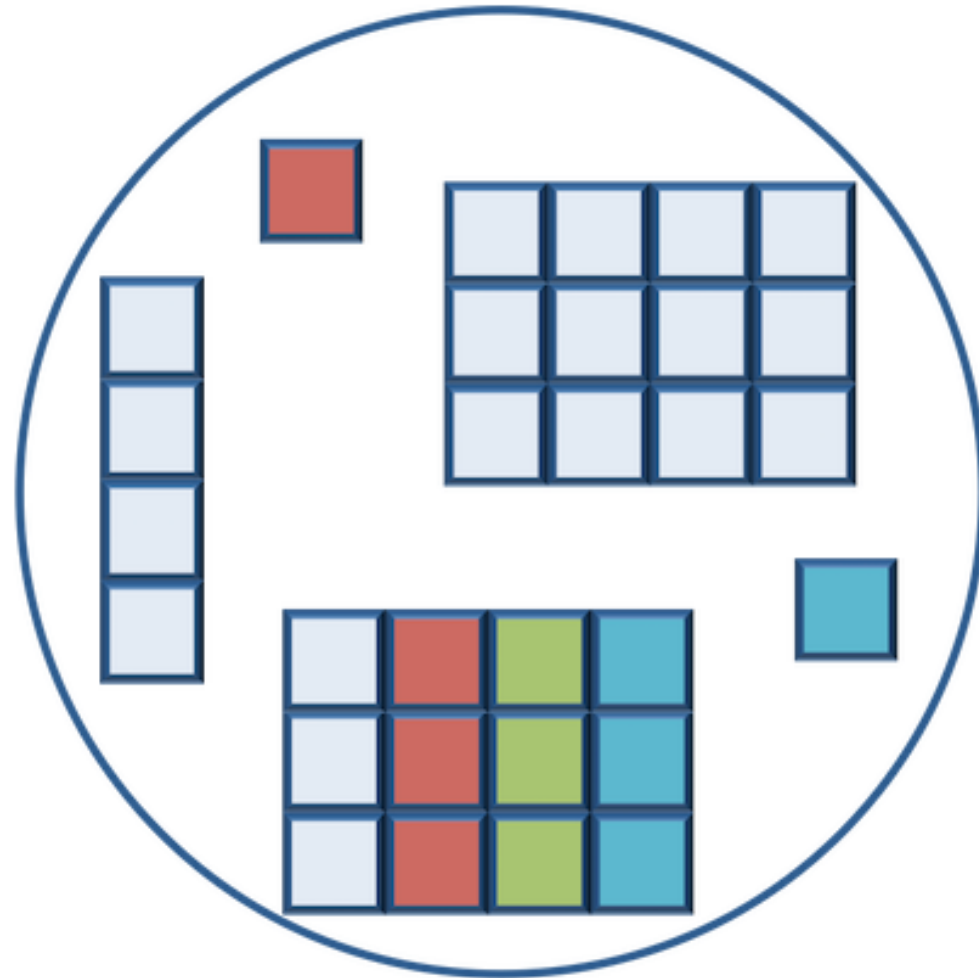


90	5	137	9
87	40	2	52
4	102	32	41

Dataframe

"A"	102	"Hela"	TRUE
"B"	40	"BHK"	F
"C"	12	"hESC"	T

Lists



Data Wrangling



how to subset vectors and
factors



use of logical operators when
subsetting vectors and factors



how to relevel factors in a
desired order

R Packages and installation



packages are collections of functions, data, and code designed to add specific features



over 10,000 user-contributed packages available



Packages for R can be installed from the [CRAN](https://cran.r-project.org/) package repository using the `install.packages` function



Example: `install.packages("ggplot2")`

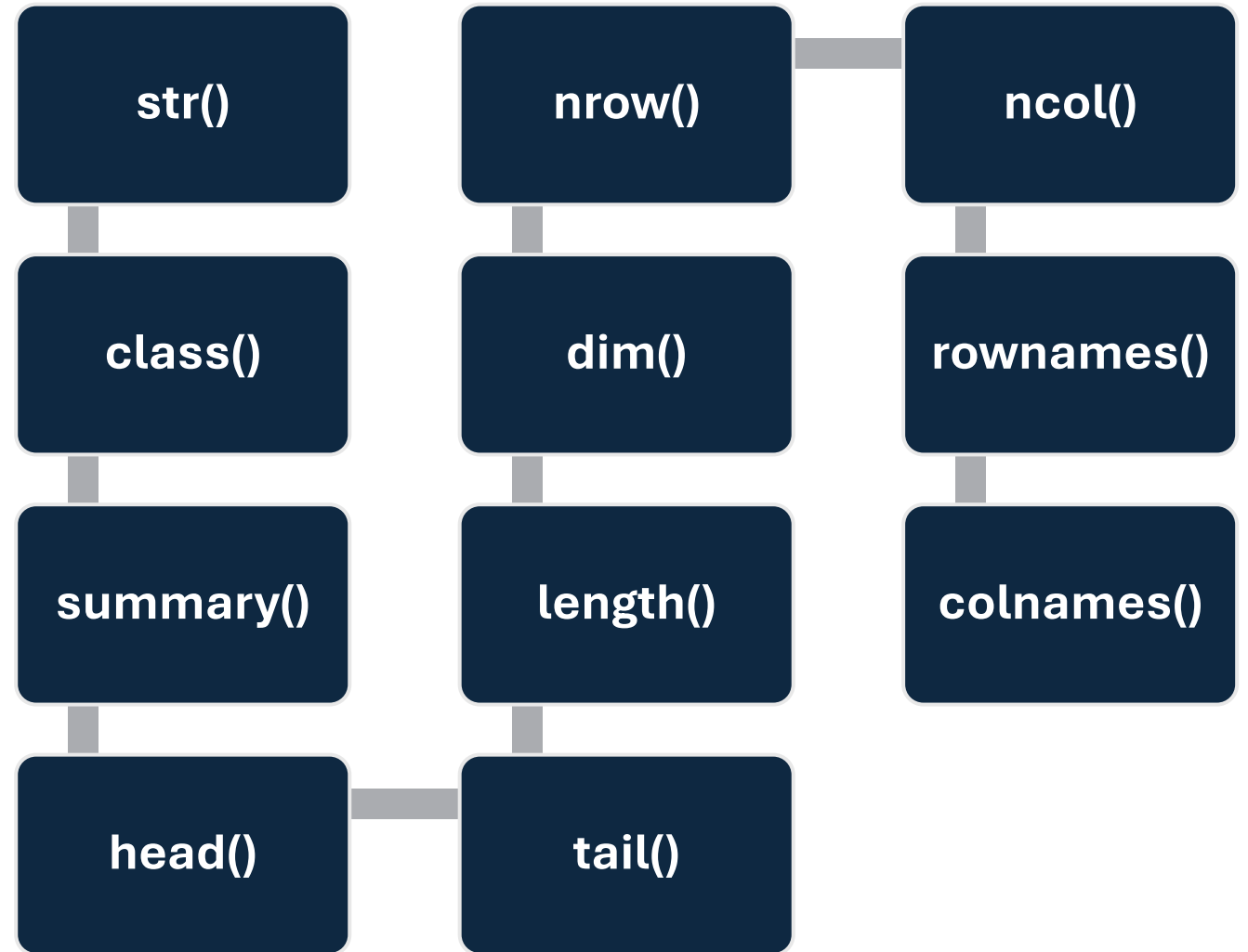


Loading: `library(ggplot2)`

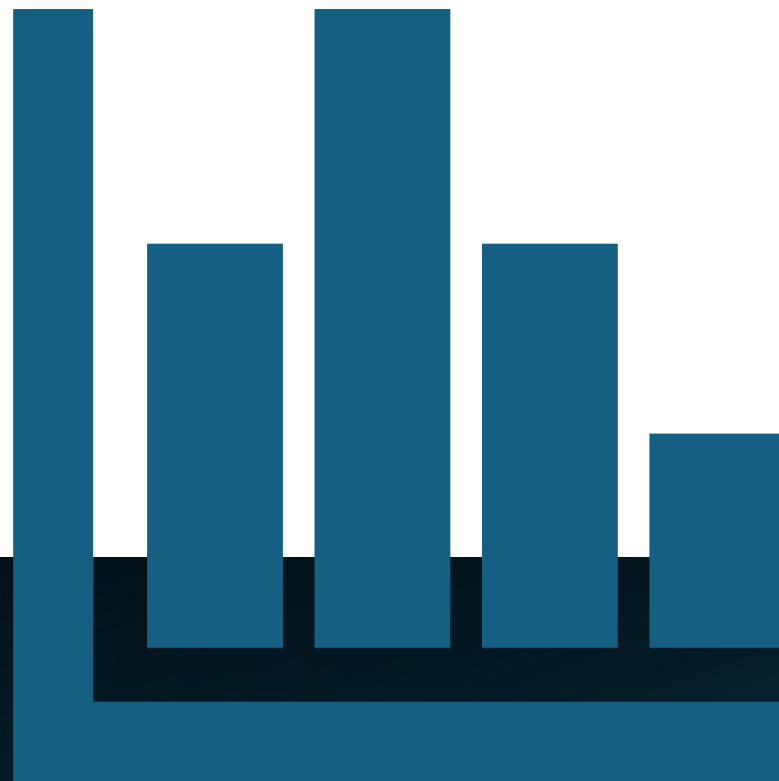
Reading files

Data type	Extension	Function	Package
Comma separated values	csv	read.csv()	utils (default)
		read_csv()	readr (tidyverse)
Tab separated values	tsv	read_tsv()	readr
Other delimited formats	txt	read.table()	utils
		read_table()	readr
		read_delim()	readr

Inspecting
the dataset
after
reading



Data Manipulation using dplyr and tidyverse



What are dplyr and tidyr?

dplyr

- Provides easy tools for common data manipulation tasks.
- Built to work directly with data frames.
- Can work with data stored in an external database.
 - This feature helps overcome R's in-memory operation limitation.
- Allows you to connect to large databases, conduct queries, and pull only the needed data into R for analysis.

tidyr

- Helps reshape data for plotting and use by different R functions.
- Facilitates moving between different data formats:
 - One row per measurement.
 - Data frames where each measurement type has its own column and rows are aggregated groups.
- Offers tools for sophisticated data manipulation.

Common dplyr functions

Select

`select()`: subset columns

Filter

`filter()`: subset rows on conditions

Mutate

`mutate()`: create new columns by using information from other columns

Group

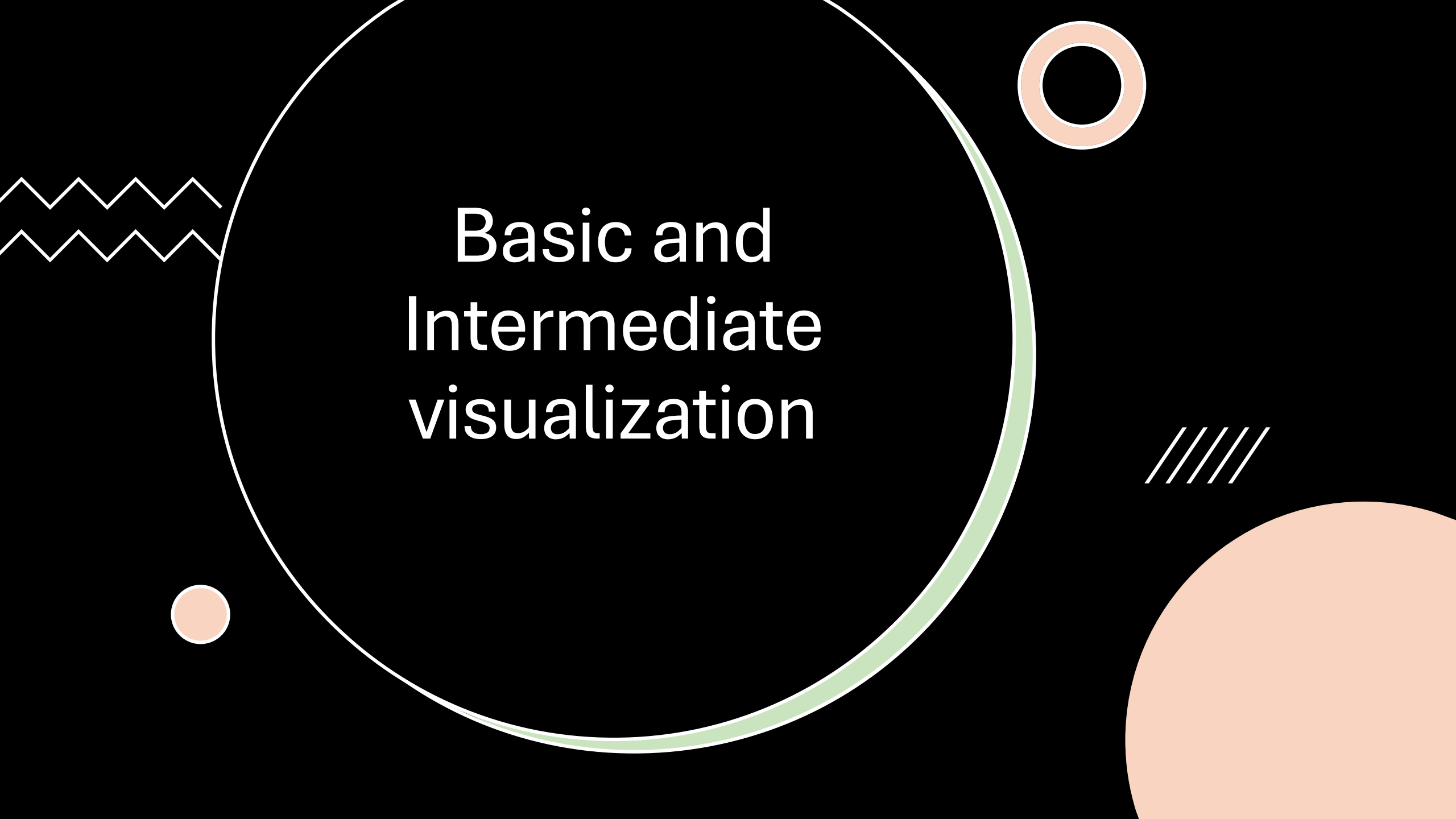
`group_by()` and `summarize()`: create summary statistics on grouped data

Arrange

`arrange()`: sort results

Count

`count()`: count discrete values

The background is black. A large circle with a thin white outline and a thick light green inner border is centered. To the left of the circle, there are two white wavy lines. Below the circle, there is a small light orange circle. To the right of the circle, there is a light orange ring. Further right, there are four white diagonal lines. In the bottom right corner, there is a large light orange semi-circle.

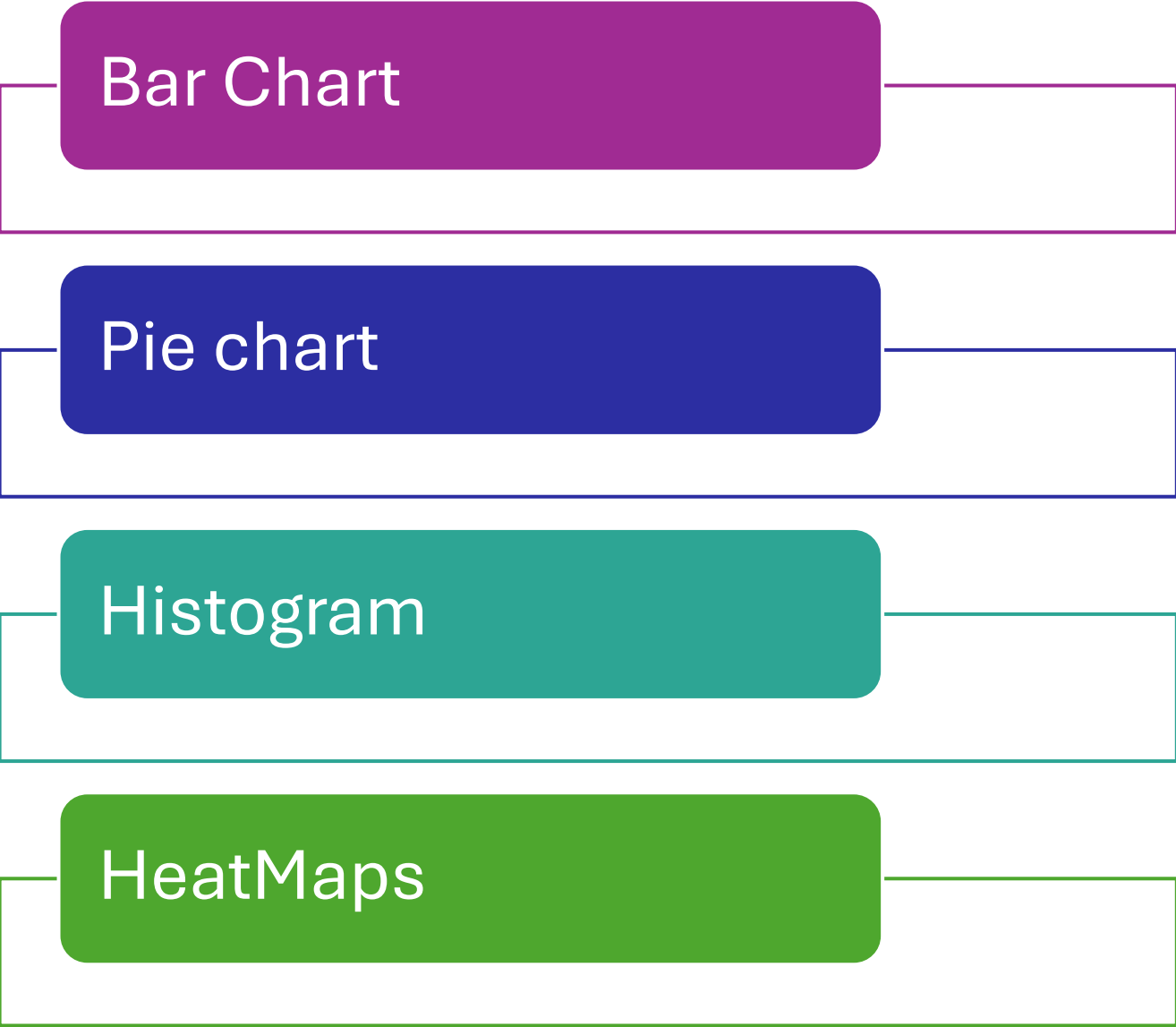
Basic and Intermediate visualization



GGPLOT2

- The functions in the **ggplot2** package build up a graph in layers. We'll build a a complex graph by starting with a simple graph and adding additional elements, one at a time.

Different type of charts



Bar Chart

Pie chart

Histogram

HeatMaps



Thank you

POST-SURVEY



Rosen Center for
Advanced Computing

