Handling Categorical Data in R

Rsquared Academy

www.rsquaredacademy.com

Agenda

- → understand categorical data
- → case study intro
- → factor in R
- → data summarization
- → data manipulation
- → data visualization

Resources

→ Slides

→ Code & Data

→ RStudio Cloud

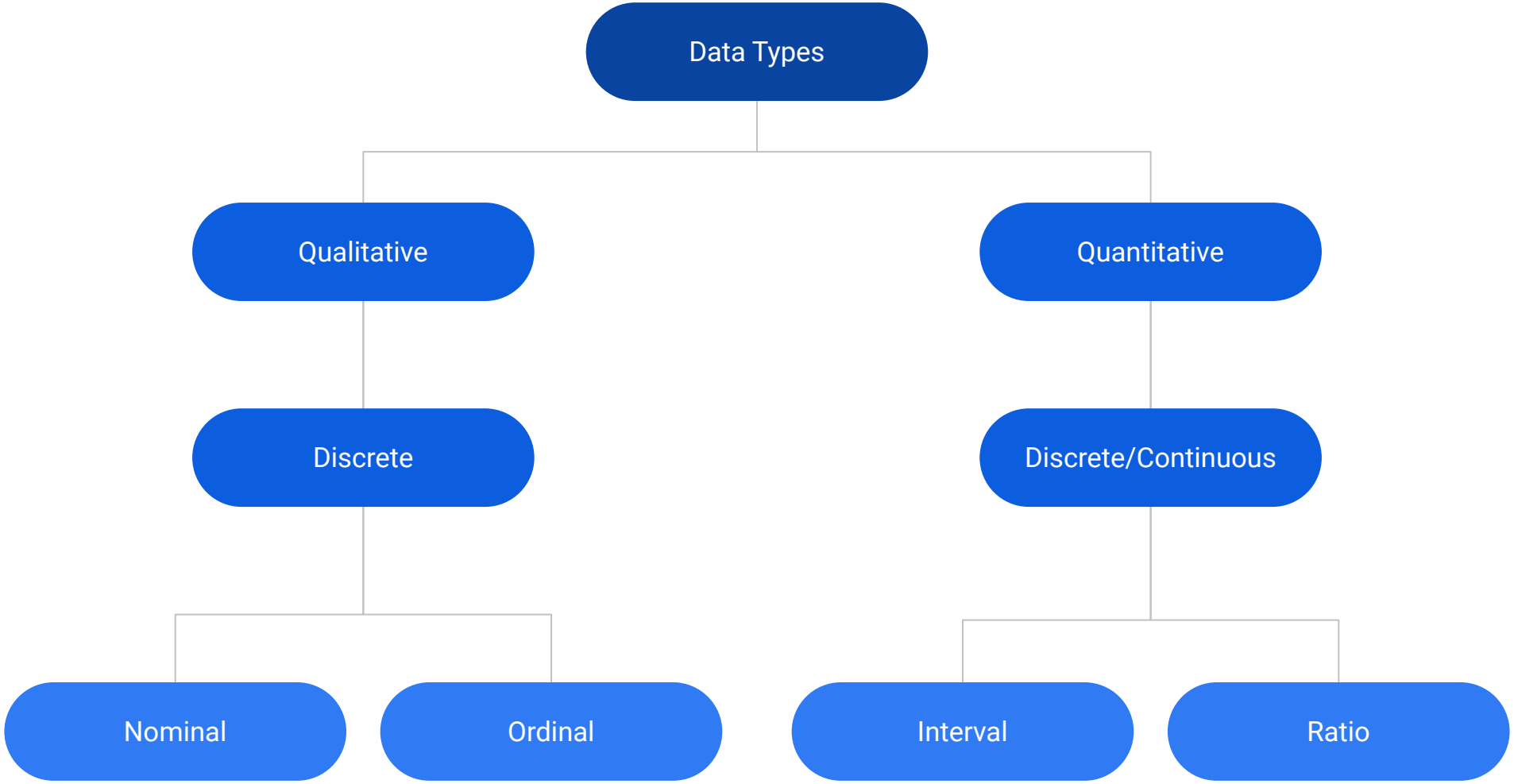
→ Online Course

→ Blog Post

Module 1

Introduction

Data Types



Discrete

one

two

three

Continuous

Categorical Data

→ it is always discrete

→ it may be divided into groups

→ consists of names or labels

→ takes on limited & fixed number of possible values

→ arises in situation when counting is involved

→ analysis generally involves the use of data tables

Dichotomous

yes

no

Polychotomous

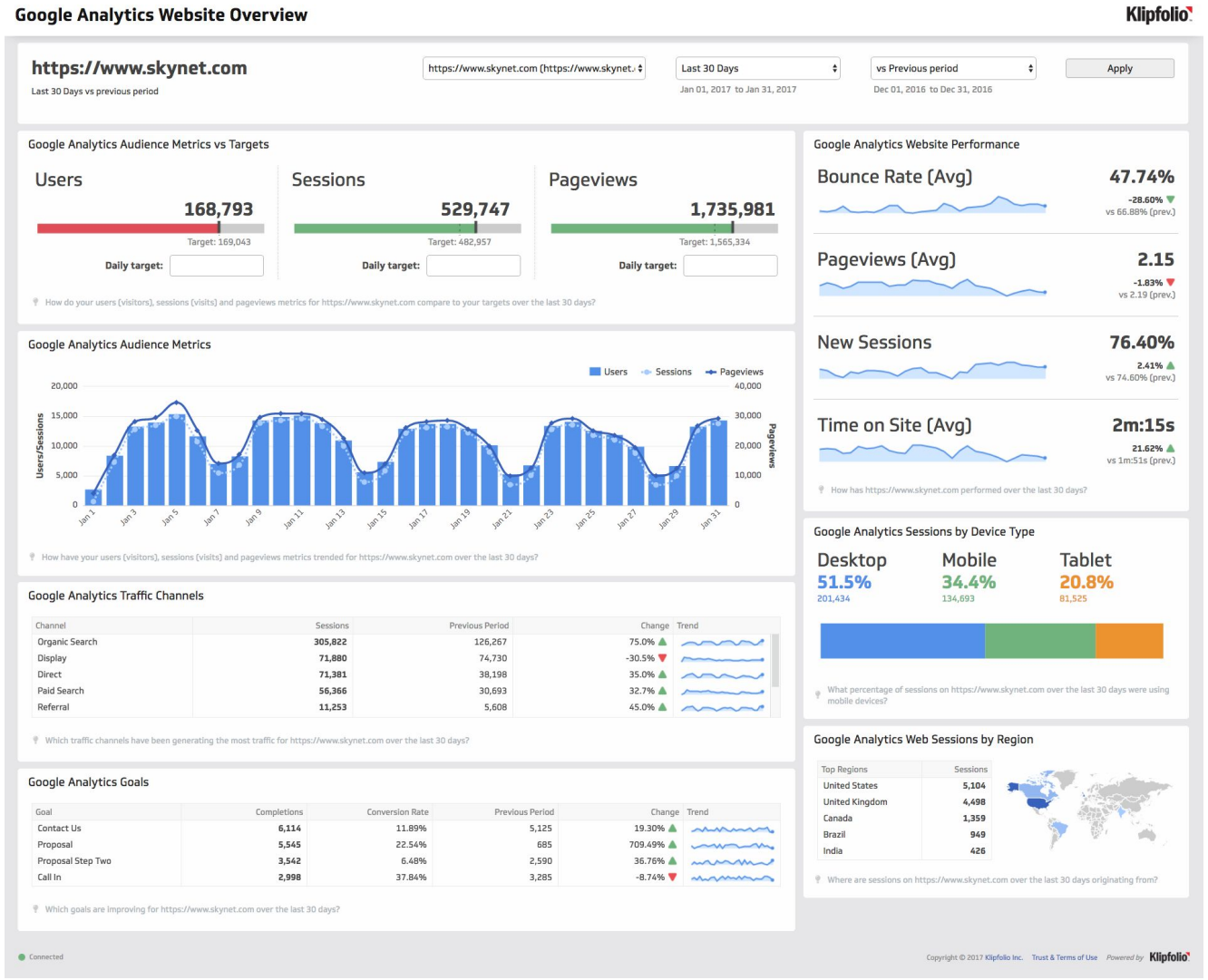
Nominal


CUSTOMER SATISFACTION
Very Poor
Poor
Average
Good
Excellent

Summary

- data can be quantitative or qualitative
- qualitative data is always discrete
- dichotomous data consists of only 2 groups/levels
- polychotomous data consists of more than 2 groups/levels
- nominal data do not have an intrinsic order
- in ordinal data
  - categories can be ordered or ranked
  - and difference between the categories cannot be determined


NOW, IT'S YOUR TURN
makeameme.org

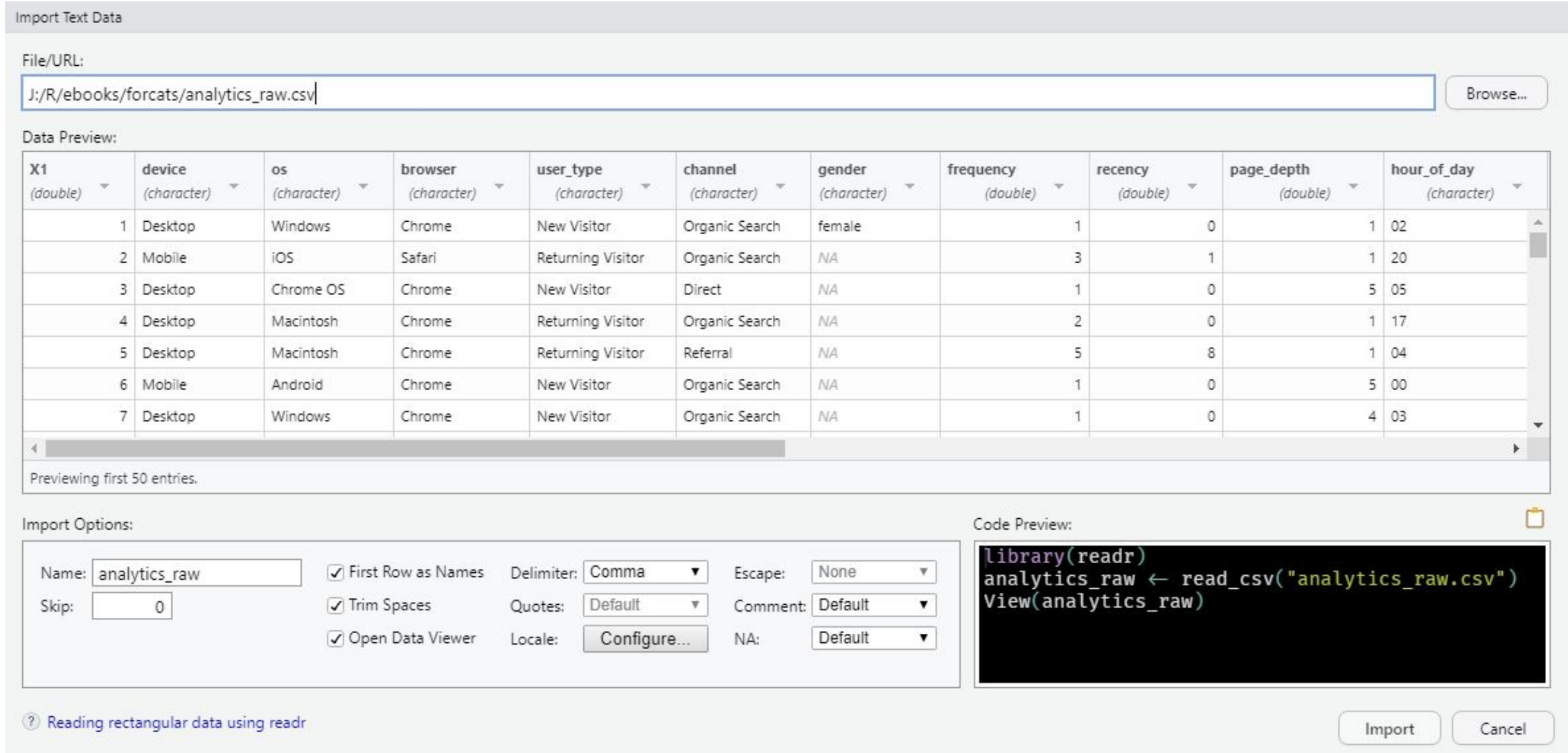

Google Analytics Website Overview
Klipfolio
https://www.skynet.com
Last 30 Days vs previous period
https://www.skynet.com (https://www.skynet.
Last 30 Days
Jan 01, 2017 to Jan 31, 2017
vs Previous period
Dec 01, 2016 to Dec 31, 2016
Apply
Google Analytics Audience Metrics vs Targets
Users
168,793
Target: 169,043
Daily target:
Sessions
529,747
Target: 482,957
Daily target:
Pageviews
1,735,981
Target: 1,565,334
Daily target:
How do your users (visitors), sessions (visits) and pageviews metrics for https://www.skynet.com compare to your targets over the last 30 days?
Google Analytics Audience Metrics
Users
Sessions
Pageviews
Users/Sessions
Pageviews
20,000
15,000
10,000
5,000
0
40,000
30,000
20,000
10,000
0
Jan 1
Jan 3
Jan 5
Jan 7
Jan 9
Jan 11
Jan 13
Jan 15
Jan 17
Jan 19
Jan 21
Jan 23
Jan 25
Jan 27
Jan 29
Jan 31
How have your users (visitors), sessions (visits) and pageviews metrics trended for https://www.skynet.com over the last 30 days?
Google Analytics Traffic Channels
Channel | Sessions | Previous Period | Change | Trend
Organic Search | 305,822 | 126,267 | 75.0%
Display | 71,880 | 74,730 | -30.5%
Direct | 71,381 | 38,198 | 35.0%
Paid Search | 56,366 | 30,693 | 32.7%
Referral | 11,253 | 5,608 | 45.0%
Which traffic channels have been generating the most traffic for https://www.skynet.com over the last 30 days?
Google Analytics Goals
Goal | Completions | Conversion Rate | Previous Period | Change | Trend
Contact Us | 6,114 | 11.89% | 5,125 | 19.30%
Proposal | 5,545 | 22.54% | 685 | 709.49%
Proposal Step Two | 3,542 | 6.48% | 2,590 | 36.76%
Call In | 2,998 | 37.84% | 3,285 | -8.74%
Which goals are improving for https://www.skynet.com over the last 30 days?
Google Analytics Website Performance
Bounce Rate (Avg)
47.74%
-28.60%
vs 66.88% (prev.)
Pageviews (Avg)
2.15
-1.83%
vs 2.19 (prev.)
New Sessions
76.40%
2.41%
vs 74.60% (prev.)
Time on Site (Avg)
2m:15s
21.62%
vs 1m:51s (prev.)
How has https://www.skynet.com performed over the last 30 days?
Google Analytics Sessions by Device Type
Desktop
51.5%
201,434
Mobile
34.4%
134,693
Tablet
20.8%
81,525
What percentage of sessions on https://www.skynet.com over the last 30 days were using mobile devices?
Google Analytics Web Sessions by Region
Top Regions | Sessions
United States | 5,104
United Kingdom | 4,498
Canada | 1,359
Brazil | 949
India | 426
Where are sessions on https://www.skynet.com over the last 30 days originating from?
Connected
Copyright © 2017 Klipfolio Inc.
Trust & Terms of Use
Powered by Klipfolio

Data Dictionary

Column Name	Description
device	Device used to visit the website
os	Operating system of the device
browser	Browser used to visit the website
user_type	New or returning visitor
channel	Source of traffic
gender	Gender of the visitor
frequency	Number of visits to the website including the current one
recency	Number of days since the last visit to the website
page_depth	Number of pages browsed on the website
hour_of_day	Hour of day

Data Dictionary

Column Name	Description
age	Age of the visitor
duration	Time spent on the website (in seconds)
landing_page	Page on which the visitor landed (first page)
exit_page	Page from which the visitor exited the website (last page)
country	Country of the visitor
city	City of the visitor
quantity	Number of units purchased
revenue	Revenue from the visitor
purchase_flag	Whether the visitor made a transaction or not
user_rating	Rating given by customer



Import Text Data

File/URL:

J:/R/ebooks/forcats/analytics_raw.csv

Browse...

Data Preview:

X1 (double)	device (character)	os (character)	browser (character)	user_type (character)	channel (character)	gender (character)	frequency (double)	recency (double)	page_depth (double)	hour_of_day (character)
1	Desktop	Windows	Chrome	New Visitor	Organic Search	female	1	0	1	02
2	Mobile	iOS	Safari	Returning Visitor	Organic Search	NA	3	1	1	20
3	Desktop	Chrome OS	Chrome	New Visitor	Direct	NA	1	0	5	05
4	Desktop	Macintosh	Chrome	Returning Visitor	Organic Search	NA	2	0	1	17
5	Desktop	Macintosh	Chrome	Returning Visitor	Referral	NA	5	8	1	04
6	Mobile	Android	Chrome	New Visitor	Organic Search	NA	1	0	5	00
7	Desktop	Windows	Chrome	New Visitor	Organic Search	NA	1	0	4	03

Previewing first 50 entries.

Import Options:

Name: analytics_raw

Skip: 0

First Row as Names

Trim Spaces

Open Data Viewer

Delimiter: Comma

Quotes: Default

Locale: Configure...

Escape: None

Comment: Default

NA: Default

Code Preview:

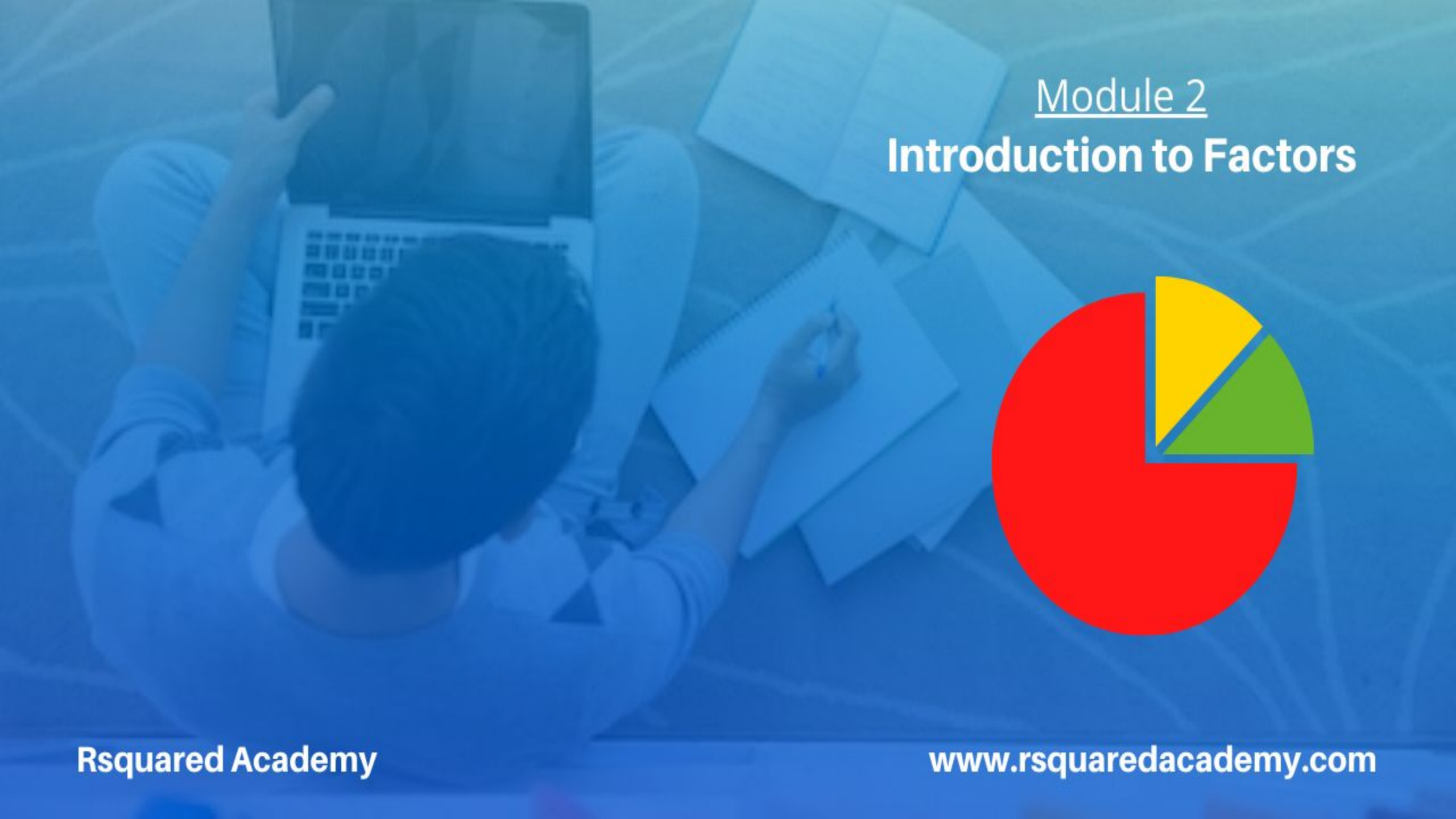
```
library(readr)
analytics_raw ← read_csv("analytics_raw.csv")
View(analytics_raw)
```

Reading rectangular data using readr

Import

Cancel

```
library(readr)
read_csv("analytics_raw.csv", col_types =
          cols_only(
          device = col_factor(levels = c("Desktop", "Tablet", "Mobile")),
          gender = col_factor(levels = c("female", "male", "NA")),
          user_rating = col_factor(levels = c("1", "2", "3", "4", "5"),
                                   ordered = TRUE)))
```



Module 2

Introduction to Factors

Agenda

- introduction to factor
- how to recognize factor variables
- how to coerce other data types to factor
- handle missing values
- handle ordinal data
- specify order of levels/categories

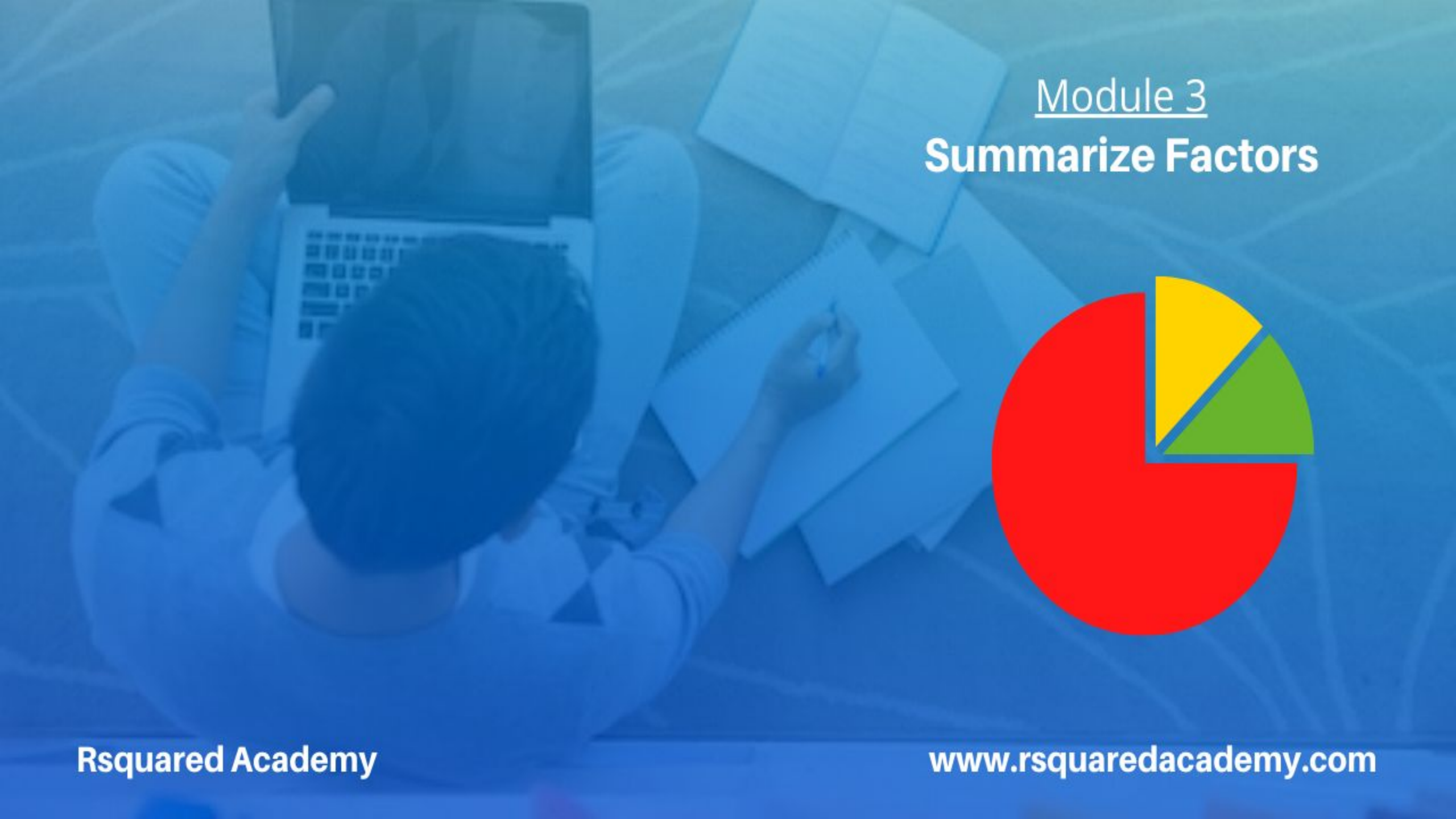
Key Functions

Function	Description
`is.factor()`	Identify
`is.ordered()`	
`as.factor()`	Convert
`as_factor()`	
`as.ordered()`	
`factor()`	Create
`ordered()`	

Summary

- R uses **`factor`** to handle categorical data
- Use **`as.factor()`** or **`as_factor()`** to coerce other data types to factor
- Use **`is.factor()`** and **`is.ordered()`** to identify factor and ordered factor respectively
- Use **`factor()`** to
  - Specify labels
  - Modify labels
  - Handle missing data
  - Create ordered factors
  - Specify order of levels
- Use **`ordered()`** to create ordered factors


NOW, IT'S YOUR TURN
makeameme.org



Module 3

Summarize Factors

References

- https://forcats.tidyverse.org/
- https://r4ds.had.co.nz/factors.html
- https://recipes.tidymodels.org/reference/discretize.html
- https://ggplot2.tidyverse.org/
- https://haleyjeppson.github.io/ggmosaic/
- https://rpkgs.datanovia.com/ggpubr/reference/ggdonutchart.html

Connect With Us

→ Website

→ Free Online R Courses

→ R Packages

→ Shiny Apps

→ Blog

→ GitHub

→ YouTube

→ Twitter

→ Linkedin

Thank You

For more information please visit our website
www.rsquaredacademy.com