Agenda

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

Resources

- Slides
- Data & Scripts
- RStudio Cloud
- Online Course
- Blog Post

Data Types

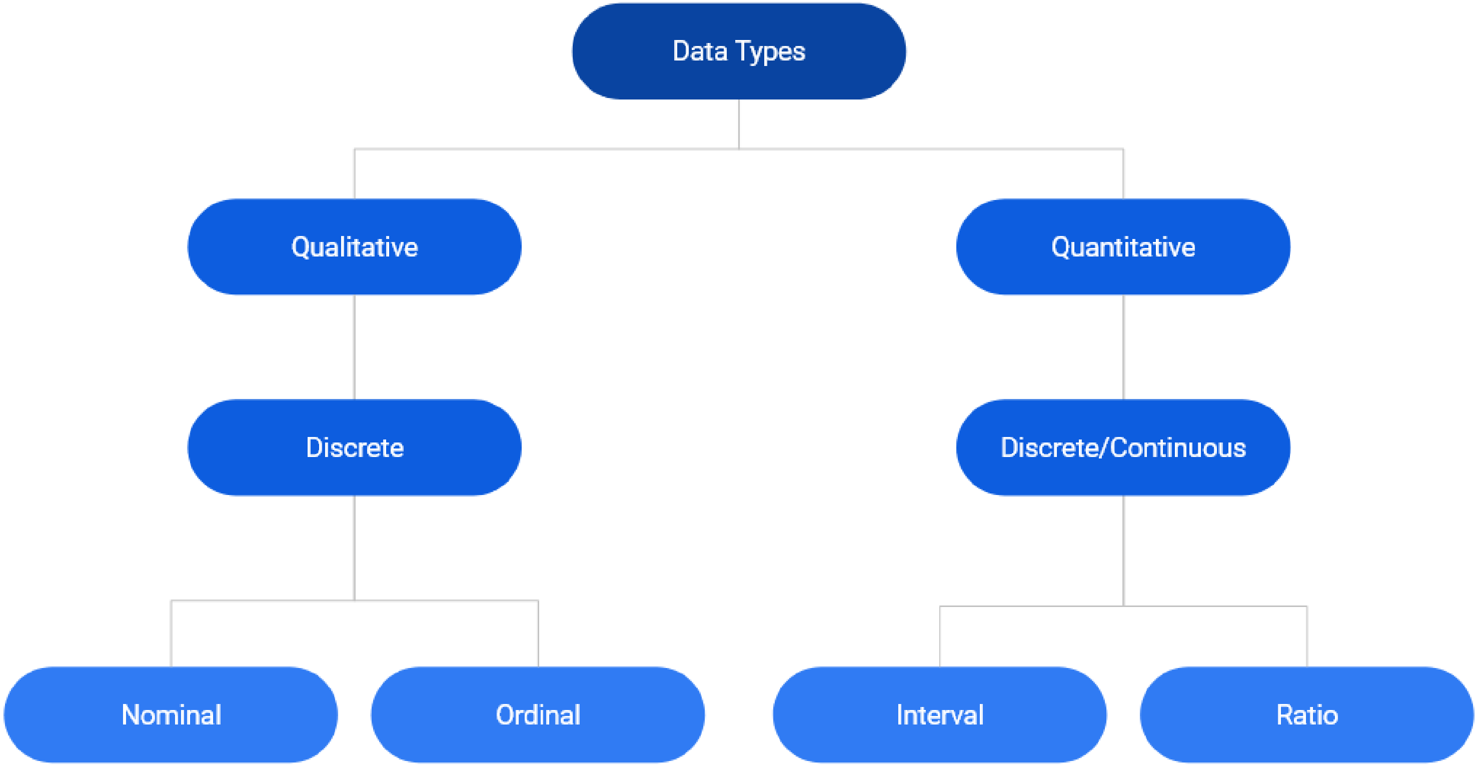


Fig 1: Data Types

Discrete

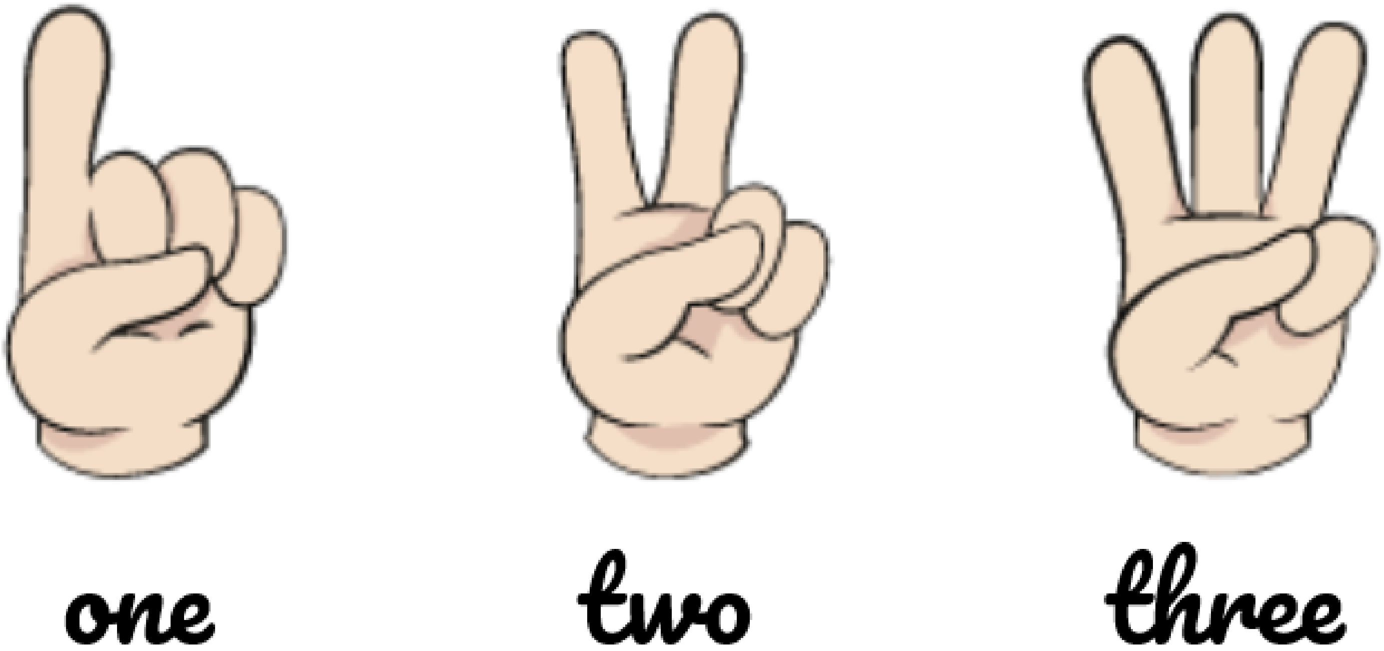


Fig 2: Discrete Data

Continuous

Fig 3: Continuous Data

Categorical Data

- is always discrete
- may be divided into groups
- consists of names or labels
- takes on limited & fixed number of possible values
- arises in situation where counting is involved
- analysis generally involves the use of tables

Dichotomous

Fig 4: Dichotomous Data

Polychotomous

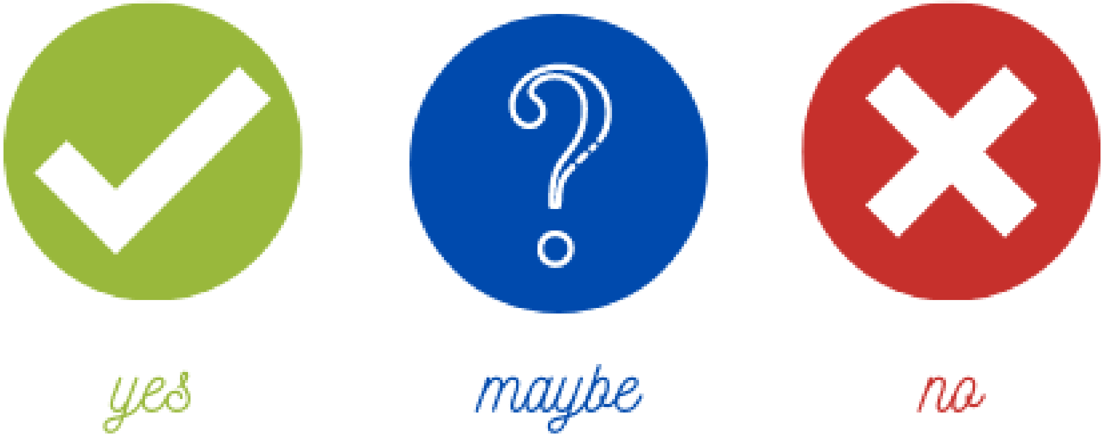


Fig 5: Polychotomous Data

Nominal

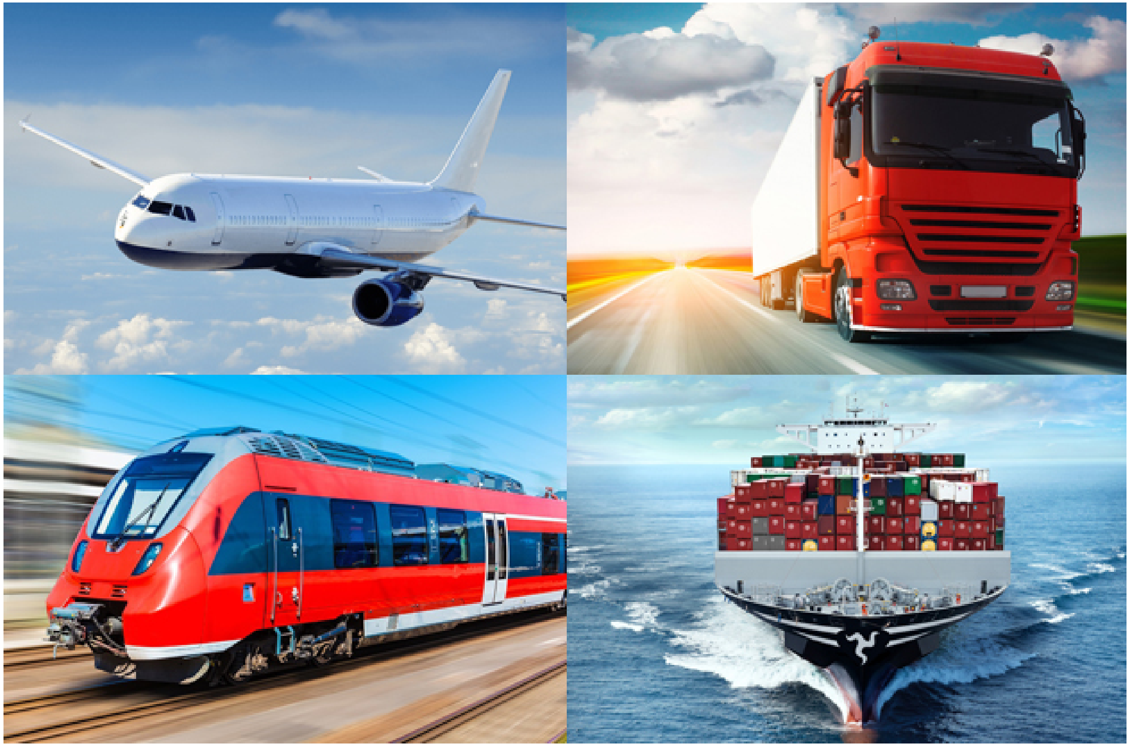


Fig 6: Nominal Data

Ordinal

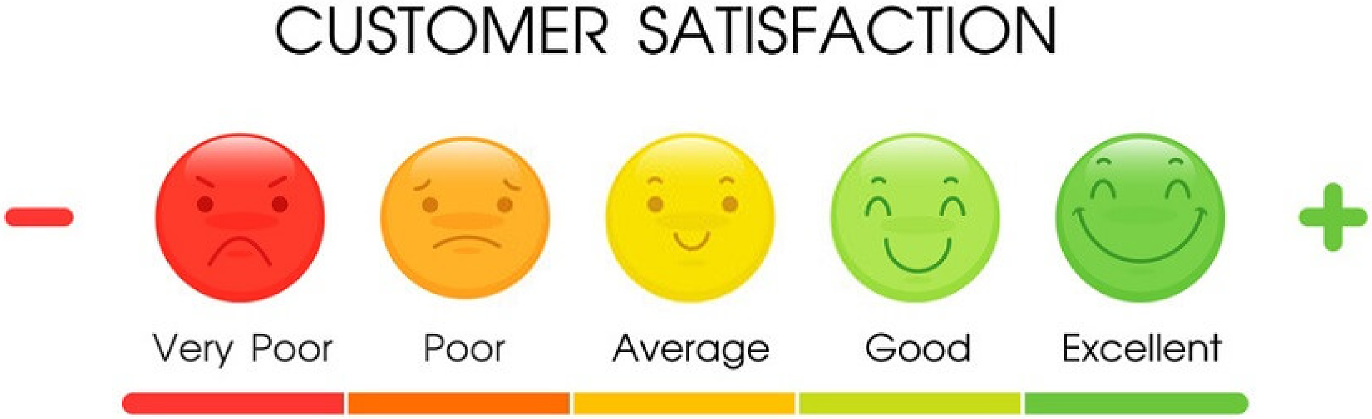
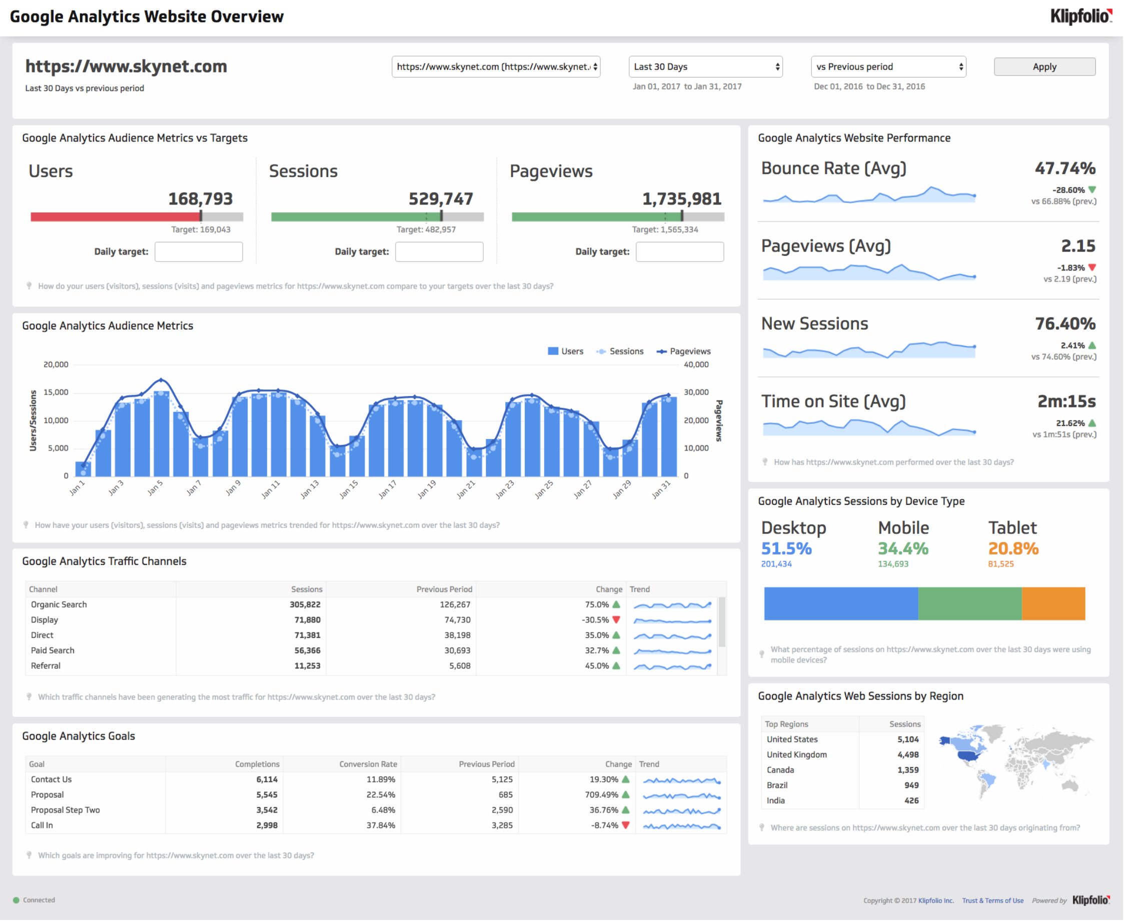


Fig 7: Ordinal Data

Summary

- data can be quantitative or qualitative
- qualitative data is always discrete
- dichotomous data consists of only 2 levels/categories
- polychotomous data consists of more than 2 levels/categories
- 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

Case Study



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

Google Analytics Audience Metrics vs Targets

Users	Sessions	Pageviews
168,793	529,747	1,735,981
Target: 169,043	Target: 482,957	Target: 1,565,334
Daily target:	Daily target:	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

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	Mobile	Tablet
51.5%	34.4%	20.8%
201,434	134,693	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

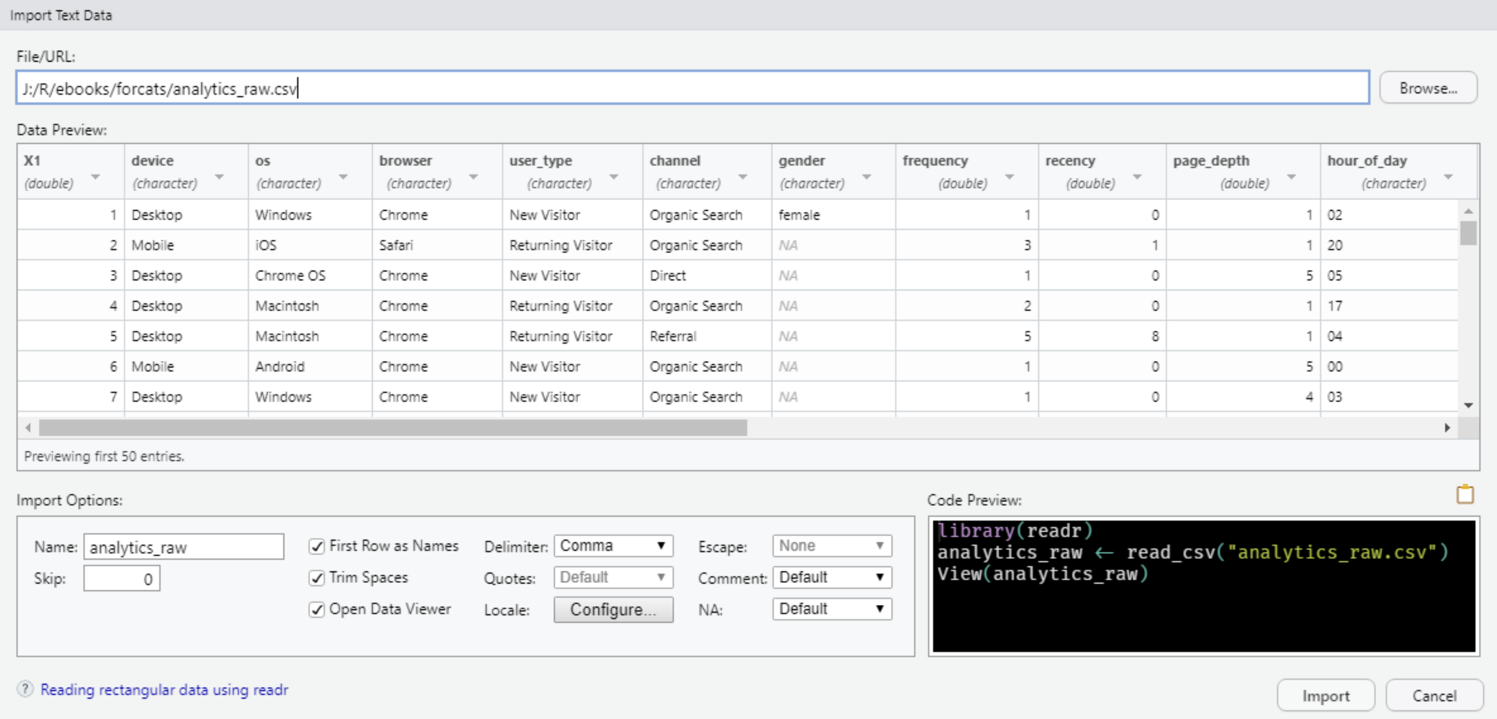
Where are sessions on https://www.skynet.com over the last 30 days originating from?

● Connected

Copyright © 2017 Klipfolio Inc. Trust & Terms of Use *Powered by* Klipfolio

Fig 8: Web Analytics Dashboard

RStudio IDE



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

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

Fig 9: RStudio IDE

Case Study

```r
# load readr package
library(readr)

# import data
read_csv("data/analytics_raw.csv",
         col_types = cols_only(
           device = col_factor(levels = c("Desktop", "Tablet", "Mobil
           gender = col_factor(levels = c("female", "male", "NA")),
           user_rating = col_factor(levels = c("1", "2", "3", "4", "5
```

Case Study

```r
# load readr package
library(readr)

# import data
read_csv("data/analytics_raw.csv",
         col_types = cols_only(
           device = col_factor(levels = c("Desktop", "Tablet", "Mobil
           gender = col_factor(levels = c("female", "male", "NA")),
           user_rating = col_factor(levels = c("1", "2", "3", "4", "5
```

Case Study

```r
# load readr package
library(readr)

# import data
read_csv("data/analytics_raw.csv",
         col_types = cols_only(
           device = col_factor(levels = c("Desktop", "Tablet", "Mobil
           gender = col_factor(levels = c("female", "male", "NA")),
           user_rating = col_factor(levels = c("1", "2", "3", "4", "5
```

Case Study

```r
# load readr package
library(readr)

# import data
read_csv("data/analytics_raw.csv",
         col_types = cols_only(
           device = col_factor(levels = c("Desktop", "Tablet", "Mobil
           gender = col_factor(levels = c("female", "male", "NA")),
           user_rating = col_factor(levels = c("1", "2", "3", "4", "5
```

Case Study

```r
# load readr package
library(readr)

# import data
read_csv("data/analytics_raw.csv",
         col_types = cols_only(
           device = col_factor(levels = c("Desktop", "Tablet", "Mobil
           gender = col_factor(levels = c("female", "male", "NA")),
           user_rating = col_factor(levels = c("1", "2", "3", "4", "5
```

Case Study

```
# load readr package
library(readr)

# import data
read_csv("data/analytics_raw.csv",
         col_types = cols_only(
           device = col_factor(levels = c("Desktop", "Tablet", "Mobil
           gender = col_factor(levels = c("female", "male", "NA")),
           user_rating = col_factor(levels = c("1", "2", "3", "4", "5
```

Case Study

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## # A tibble: 6 x 3
##   device  gender user_rating
##   <fct>   <fct>  <ord>
## 1 Desktop female 4
## 2 Mobile  <NA>   5
## 3 Desktop <NA>   4
## 4 Desktop <NA>   5
## 5 Desktop <NA>   4
## 6 Mobile  <NA>   4
```

RDS Data

```r
data <- readRDS("data/analytics.rds")
head(data)
```

```
## # A tibble: 6 x 19
##   device  os     browser user_type  channel  gender frequency recency page
##   <fct>   <fct>  <fct>   <fct>      <fct>    <fct>      <dbl>   <dbl>
## 1 Desktop Windo~ Chrome  New Visit~ Organic~ female         1       0
## 2 Mobile  iOS    Safari  Returning~ Organic~ <NA>           3       1
## 3 Desktop Chrom~ Chrome  New Visit~ Direct   <NA>           1       0
## 4 Desktop Macin~ Chrome  Returning~ Organic~ <NA>           2       0
## 5 Desktop Macin~ Chrome  Returning~ Referral <NA>           5       8
## 6 Mobile  Andro~ Chrome  New Visit~ Organic~ <NA>           1       0
## # ... with 10 more variables: hour_of_day <chr>, age <dbl>, duration <dbl
## #   landing_page <fct>, exit_page <fct>, country <fct>, quantity <dbl>,
## #   revenue <dbl>, purchase_flag <lgl>, user_rating <dbl>
```

Agenda

- introduction to factor
- how to detect factor variables
- how to coerce other data types to factor
- handle missing values
- handle ordinal data
- specify orders of categories

Sample Data

```r
device <- sample(c("Desktop", "Mobile", "Tablet"),
                 size = 25,
                 replace = TRUE)
```

Sample Data

```r
device <- sample(c("Desktop", "Mobile", "Tablet"),
                 size = 25,
                 replace = TRUE)
```

Sample Data

```
device <- sample(c("Desktop", "Mobile", "Tablet"),
                 size = 25,
                 replace = TRUE)
```

Sample Data

```r
device <- sample(c("Desktop", "Mobile", "Tablet"),
                 size = 25,
                 replace = TRUE)
```

Sample Data

```
device
```

```
##  [1] "Tablet"  "Desktop" "Desktop" "Desktop" "Desktop" "Tablet"  "Desktop"
##  [8] "Desktop" "Desktop" "Tablet"  "Desktop" "Tablet"  "Mobile"  "Tablet"
## [15] "Mobile"  "Desktop" "Mobile"  "Desktop" "Desktop" "Mobile"  "Tablet"
## [22] "Desktop" "Tablet"  "Mobile"  "Desktop"
```

Membership Testing

```r
is.factor(device)
```

```
## [1] FALSE
```

Coercion

```r
as.factor(device)
```

```
##  [1] Tablet  Desktop Desktop Desktop Desktop Tablet  Desktop Desktop Deskt
## [10] Tablet  Desktop Tablet  Mobile  Tablet  Mobile  Desktop Mobile  Deskt
## [19] Desktop Mobile  Tablet  Desktop Tablet  Mobile  Desktop
## Levels: Desktop Mobile Tablet
```

Coercion

```r
as_factor(device)
```

```
##  [1] Tablet  Desktop Desktop Desktop Desktop Tablet  Desktop Desktop Deskt
## [10] Tablet  Desktop Tablet  Mobile  Tablet  Mobile  Desktop Mobile  Deskt
## [19] Desktop Mobile  Tablet  Desktop Tablet  Mobile  Desktop
## Levels: Tablet Desktop Mobile
```

Factor

```
factor(device)
```

```
##  [1] Tablet  Desktop Desktop Desktop Desktop Tablet  Desktop Desktop Deskt
## [10] Tablet  Desktop Tablet  Mobile  Tablet  Mobile  Desktop Mobile  Deskt
## [19] Desktop Mobile  Tablet  Desktop Tablet  Mobile  Desktop
## Levels: Desktop Mobile Tablet
```

Specify Levels

```
factor(device,
       levels = c("Desktop", "Mobile", "Tablet"))
```

```
##  [1] Tablet  Desktop Desktop Desktop Desktop Tablet  Desktop Desktop Deskt
## [10] Tablet  Desktop Tablet  Mobile  Tablet  Mobile  Desktop Mobile  Deskt
## [19] Desktop Mobile  Tablet  Desktop Tablet  Mobile  Desktop
## Levels: Desktop Mobile Tablet
```

Specify Levels

```
factor(device,
       levels = c("Desktop", "Mobile"))
```

```
##  [1] <NA>    Desktop Desktop Desktop Desktop <NA>    Desktop Desktop Deskt
## [10] <NA>    Desktop <NA>    Mobile  <NA>    Mobile  Desktop Mobile  Deskt
## [19] Desktop Mobile  <NA>    Desktop <NA>    Mobile  Desktop
## Levels: Desktop Mobile
```

Modify Labels

```
factor(device,
       levels = c("Desktop", "Mobile", "Tablet"),
       labels = c("Desk", "Mob", "Tab"))
```

```
##  [1] Tab  Desk Desk Desk Desk Tab  Desk Desk Desk Tab  Desk Tab  Mob  Tab
## [16] Desk Mob  Desk Desk Mob  Tab  Desk Tab  Mob  Desk
## Levels: Desk Mob Tab
```

Sample Data with Missing Values

```
# sample with missing values
device <- sample(c("Desktop", "Mobile", "Tablet", NA),
                 size = 25,
                 replace = TRUE)
device
```

```
##  [1] "Tablet"  "Mobile"  NA        "Tablet"  "Mobile"  "Mobile"  NA
##  [8] "Mobile"  "Desktop" "Desktop" "Tablet"  "Tablet"  "Tablet"  "Tablet"
## [15] "Mobile"  "Mobile"  "Desktop" "Mobile"  NA        "Mobile"  "Mobile"
## [22] NA        "Desktop" "Tablet"  "Mobile"
```

NA as a Level

```r
# store as categorical data
factor(device)
```

```
##  [1] Tablet  Mobile  <NA>    Tablet  Mobile  Mobile  <NA>    Mobile  Deskt
## [10] Desktop Tablet  Tablet  Tablet  Tablet  Mobile  Mobile  Desktop Mobil
## [19] <NA>    Mobile  Mobile  <NA>    Desktop Tablet  Mobile
## Levels: Desktop Mobile Tablet
```

NA as a Level

```
factor(device,
       exclude = NULL)
```

```
##  [1] Tablet  Mobile  <NA>    Tablet  Mobile  Mobile  <NA>    Mobile  Deskt
## [10] Desktop Tablet  Tablet  Tablet  Tablet  Mobile  Mobile  Desktop Mobil
## [19] <NA>    Mobile  Mobile  <NA>    Desktop Tablet  Mobile
## Levels: Desktop Mobile Tablet <NA>
```

Satisfaction Rating Sample Data

```r
rating <- sample(c("Dislike", "Neutral", "Like"),
                 size = 25,
                 replace = TRUE)
rating
```

```
##  [1] "Like"    "Like"    "Neutral" "Like"    "Like"    "Like"    "Dislike"
##  [8] "Dislike" "Neutral" "Like"    "Dislike" "Like"    "Dislike" "Neutral"
## [15] "Neutral" "Like"    "Dislike" "Dislike" "Dislike" "Dislike" "Neutral"
## [22] "Like"    "Neutral" "Dislike" "Neutral"
```

Membership Testing

```
is.ordered(rating)
```

```
## [1] FALSE
```

Coercion

```r
as.ordered(rating)
```

```
##  [1] Like    Like    Neutral Like    Like    Like    Dislike Dislike Neutr
## [10] Like    Dislike Like    Dislike Neutral Neutral Like    Dislike Disli
## [19] Dislike Dislike Neutral Like    Neutral Dislike Neutral
## Levels: Dislike < Like < Neutral
```

Ordered Factor

```
factor(rating,
       ordered = TRUE)
```

```
##  [1] Like    Like    Neutral Like    Like    Like    Dislike Dislike Neutr
## [10] Like    Dislike Like    Dislike Neutral Neutral Like    Dislike Disli
## [19] Dislike Dislike Neutral Like    Neutral Dislike Neutral
## Levels: Dislike < Like < Neutral
```

Modify Order of Levels

```
factor(rating,
       levels = c("Dislike", "Neutral", "Like"),
       ordered = TRUE)
```

```
##  [1] Like    Like    Neutral Like    Like    Like    Dislike Dislike Neutr
## [10] Like    Dislike Like    Dislike Neutral Neutral Like    Dislike Disli
## [19] Dislike Dislike Neutral Like    Neutral Dislike Neutral
## Levels: Dislike < Neutral < Like
```

Ordered

```
ordered(rating)
```

```
##  [1] Like    Like    Neutral Like    Like    Like    Dislike Dislike Neutr
## [10] Like    Dislike Like    Dislike Neutral Neutral Like    Dislike Disli
## [19] Dislike Dislike Neutral Like    Neutral Dislike Neutral
## Levels: Dislike < Like < Neutral
```

Ordered

```
ordered(rating,
        levels = c("Dislike", "Neutral", "Like"))
```

```
##  [1] Like    Like    Neutral Like    Like    Like    Dislike Dislike Neutr
## [10] Like    Dislike Like    Dislike Neutral Neutral Like    Dislike Disli
## [19] Dislike Dislike Neutral Like    Neutral Dislike Neutral
## Levels: Dislike < Neutral < Like
```

Key Functions

- `is.factor()`
- `is.ordered()`
- `as.factor()`
- `as_factor()`
- `as.ordered()`
- `factor()`
- `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()` or `is.ordered()` for membership testing
- use `factor()` function to
    - specify labels
    - modify labels
    - handle missing data
    - create ordered factors
    - specify order of levels
- use `ordered()` function to create ordered 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
- eBook
- GitHub
- YouTube
- Twitter
- Linkedin