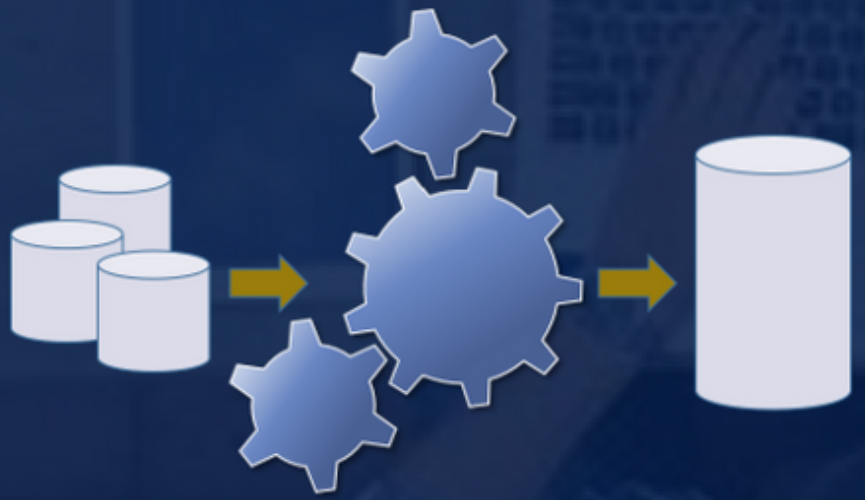
Rsquared Academy



Data Wrangling with dplyr

Agenda

- extract unique rows
- rename columns
- sample data
- extract columns
- slice rows

- arrange rows
- compare tables
- extract/mutate data using predicate functions
- count observations for different levels of a variable

Libraries

```
library(dplyr)
library(readr)
```

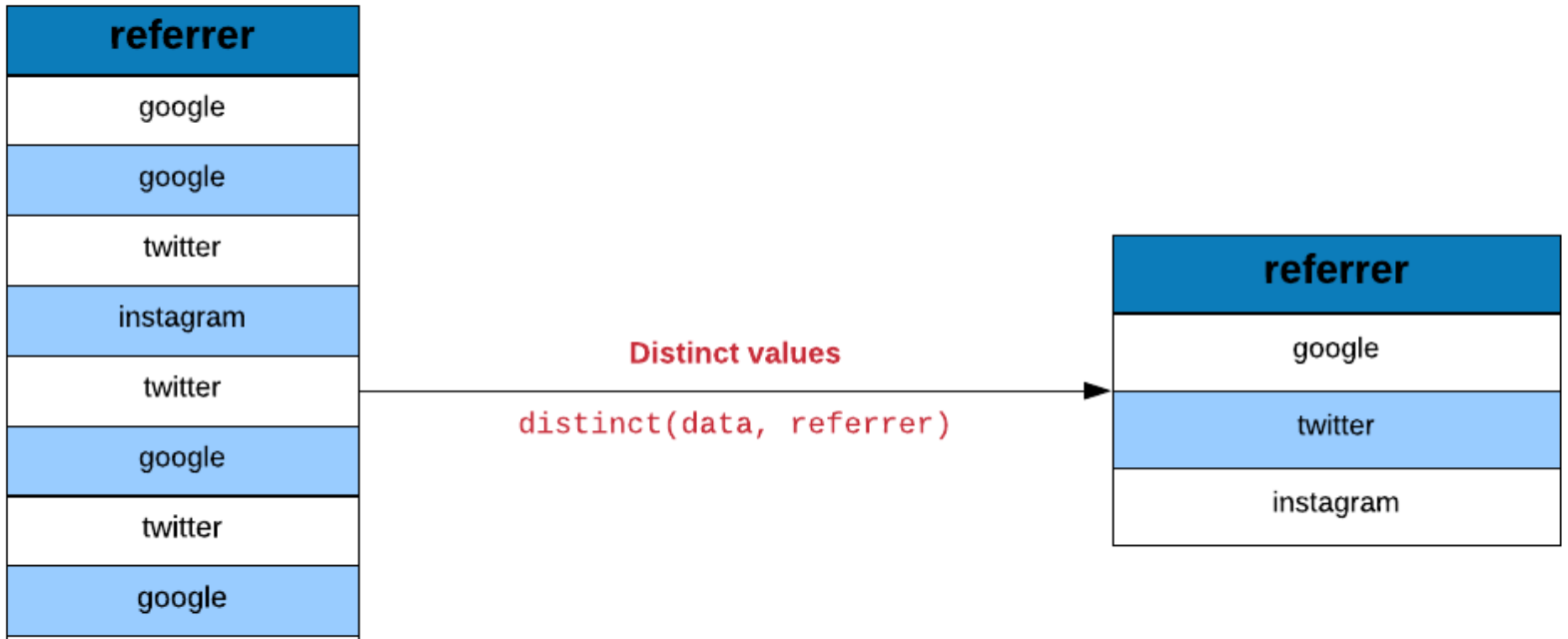
```
## # A tibble: 1,000 x 7
##    referrer device bouncers n_visit n_pages duration purchase
##    <fct>    <fct>  <lgl>      <dbl>   <dbl>    <dbl> <lgl>
##  1 google   laptop TRUE          10       1      693 FALSE
##  2 yahoo    tablet TRUE           9       1      459 FALSE
##  3 direct   laptop TRUE           0       1      996 FALSE
##  4 bing     tablet FALSE          3      18      468 TRUE
##  5 yahoo    mobile TRUE           9       1      955 FALSE
##  6 yahoo    laptop FALSE          5       5      135 FALSE
##  7 yahoo    mobile TRUE          10       1       75 FALSE
##  8 direct   mobile TRUE          10       1      908 FALSE
##  9 bing     mobile FALSE          3      19      209 FALSE
## 10 google   mobile TRUE           6       1      208 FALSE
## # ... with 990 more rows
```

Data Dictionary

- referrer: referrer website/search engine
- device: device used to visit the website
- bouncers: whether a visit bounced (exited from landing page)
- duration: time spent on the website (in seconds)
- purchase: whether visitor purchased

- n_visit: number of visits
- n_pages: number of pages visited/browsed

Distinct


referrer
google
google
twitter
instagram
twitter
google
twitter
google
Distinct values
distinct(data, referrer)
referrer
google
twitter
instagram

Traffic Sources

```
distinct(ecom, referrer)
```

```
## # A tibble: 5 x 1
##   referrer
##   <fct>
## 1 google
## 2 yahoo
## 3 direct

## 4 bing
## 5 social
```

Device Types

```
distinct(ecom, device)
```

```
## # A tibble: 3 x 1
##   device
##   <fct>
## 1 laptop
## 2 tablet
## 3 mobile
```

Rename

device	order items	order value
mobile	3	267
tablet	3	297
laptop	4	378

Rename order items as items

```
rename(data, items = `order items`)
```

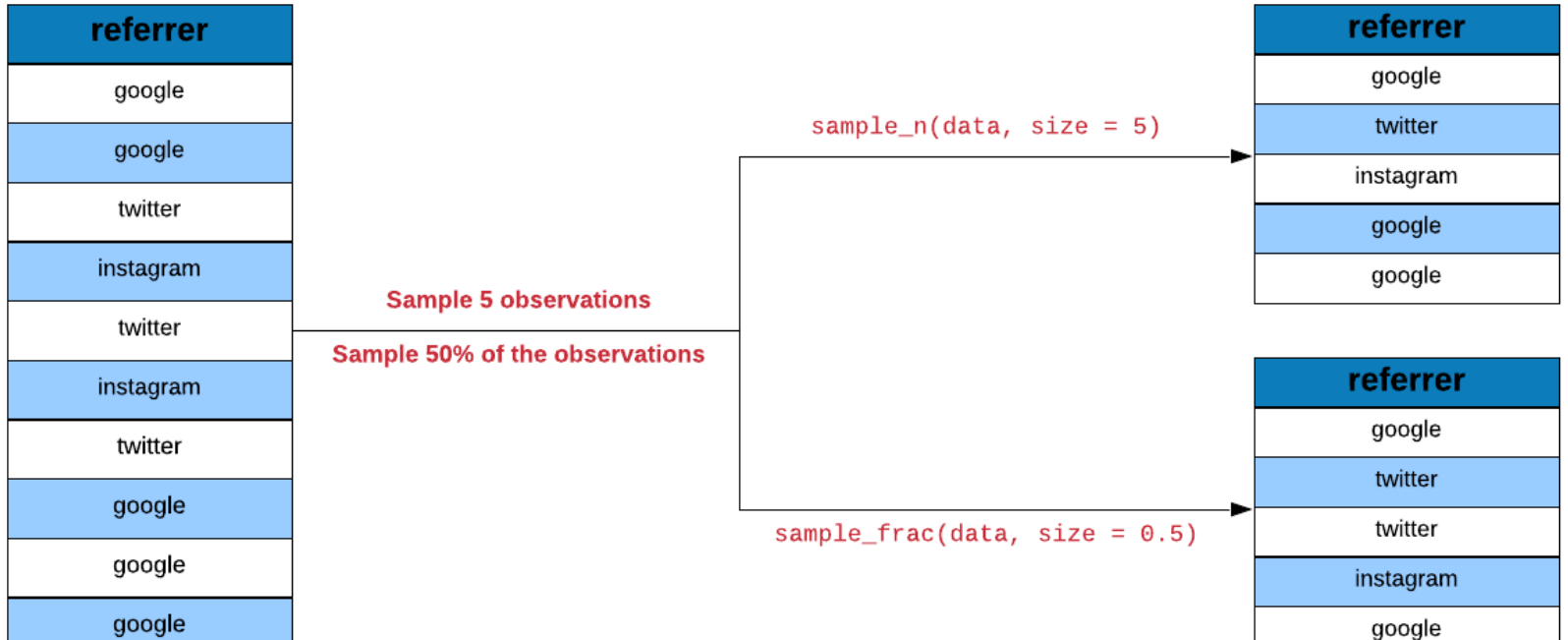
device	items	order value
mobile	3	267
tablet	3	297
laptop	4	378

Rename Columns

```
rename(ecom, time_on_site = duration)
```

```
## # A tibble: 1,000 x 7
##    referrer device bouncers n_visit n_pages time_on_site purchase
##    <fct>    <fct>  <lgl>      <dbl>   <dbl>        <dbl> <lgl>
##  1 google   laptop TRUE          10       1          693 FALSE
##  2 yahoo    tablet TRUE           9       1          459 FALSE
##  3 direct   laptop TRUE           0       1          996 FALSE
##  4 bing     tablet FALSE          3      18          468 TRUE
##  5 yahoo    mobile TRUE           9       1          955 FALSE
##  6 yahoo    laptop FALSE          5       5          135 FALSE
##  7 yahoo    mobile TRUE          10       1           75 FALSE
##  8 direct   mobile TRUE          10       1          908 FALSE
##  9 bing     mobile FALSE          3      19          209 FALSE
## 10 google   mobile TRUE           6       1          208 FALSE
## # ... with 990 more rows
```

Sampling


referrer
google
google
twitter
instagram
twitter
instagram
twitter
google
google
google
Sample 5 observations
Sample 50% of the observations
sample_n(data, size = 5)
sample_frac(data, size = 0.5)
referrer
google
twitter
instagram
google
google
referrer
google
twitter
twitter
instagram
google

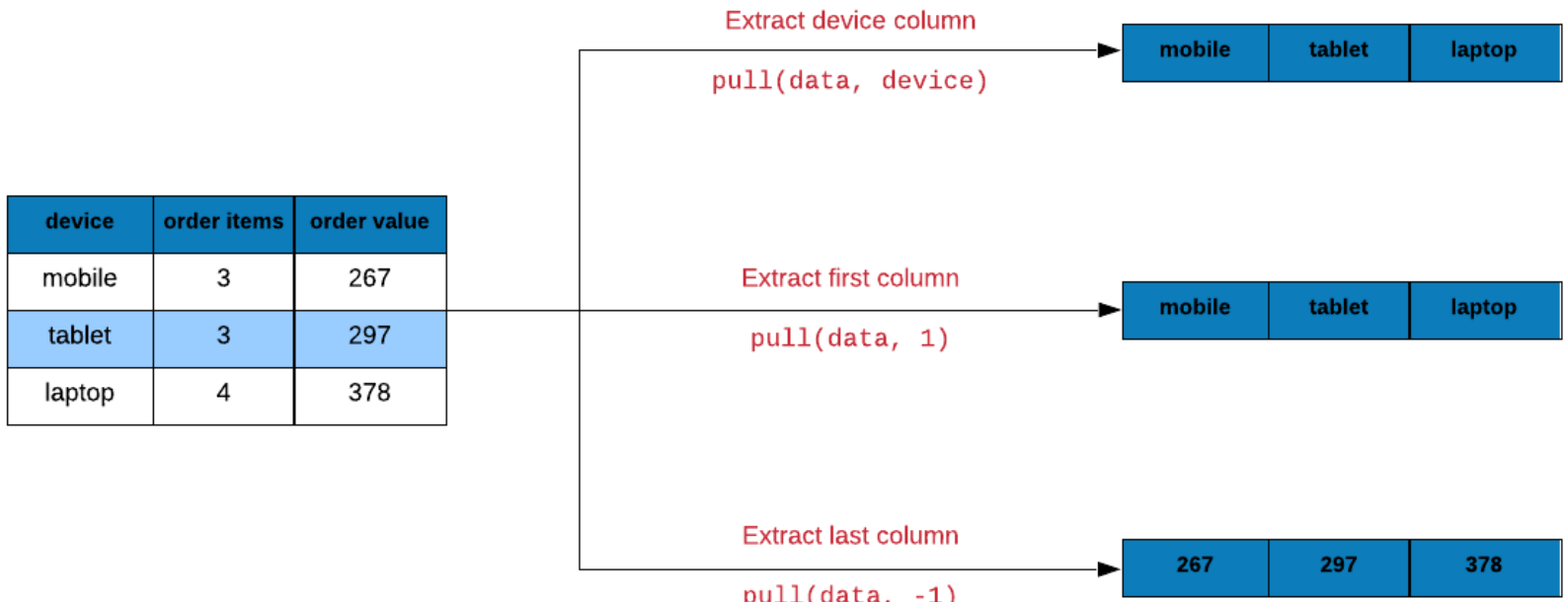
```
sample_n(ecom, size = 700)
```

```
## # A tibble: 700 x 7
##    referrer device bouncers n_visit n_pages duration purchase
##    <fct>    <fct>  <lgl>      <dbl>   <dbl>    <dbl> <lgl>
##  1 bing     tablet FALSE          2       5      150 FALSE
##  2 social   tablet TRUE           9       1      157 FALSE
##  3 yahoo    tablet TRUE           6       1       67 FALSE
##  4 direct   laptop FALSE          1      14      364 TRUE
##  5 direct   mobile FALSE          2       9      243 FALSE
##  6 direct   tablet FALSE         10       3       57 FALSE
##  7 yahoo    tablet TRUE          10       1      668 FALSE
##  8 yahoo    tablet FALSE          2      20      320 FALSE
##  9 bing     tablet TRUE           0       1      845 FALSE
## 10 yahoo    mobile FALSE          8       9      225 FALSE
## # ... with 690 more rows
```

```
sample_frac(ecom, size = 0.7)
```

```
## # A tibble: 700 x 7
##    referrer device bouncers n_visit n_pages duration purchase
##    <fct>    <fct>  <lgl>      <dbl>   <dbl>    <dbl> <lgl>
##  1 bing     tablet TRUE           6       1      567 FALSE
##  2 bing     tablet FALSE          6       9      198 FALSE
##  3 bing     laptop TRUE           3       1      271 FALSE
##  4 bing     mobile FALSE         10       1       26 FALSE
##  5 bing     mobile TRUE           5       1      751 FALSE
##  6 bing     tablet FALSE          1       8      144 FALSE
##  7 yahoo    mobile TRUE          10       1      761 FALSE
##  8 bing     laptop FALSE          8      10      260 TRUE
##  9 direct   tablet FALSE          1       3       69 FALSE
## 10 google   laptop TRUE           9       1      174 FALSE
## # ... with 690 more rows
```

Extract Columns



Sample Data

```
ecom_mini <- sample_n(ecom, size = 10)
```

Extract Device Column

```
pull(ecom_mini, device)
```

```
##  [1] mobile mobile mobile laptop mobile mobile laptop laptop tablet t
## Levels: laptop tablet mobile
```

Extract First Column

```
pull(ecom_mini, 1)
```

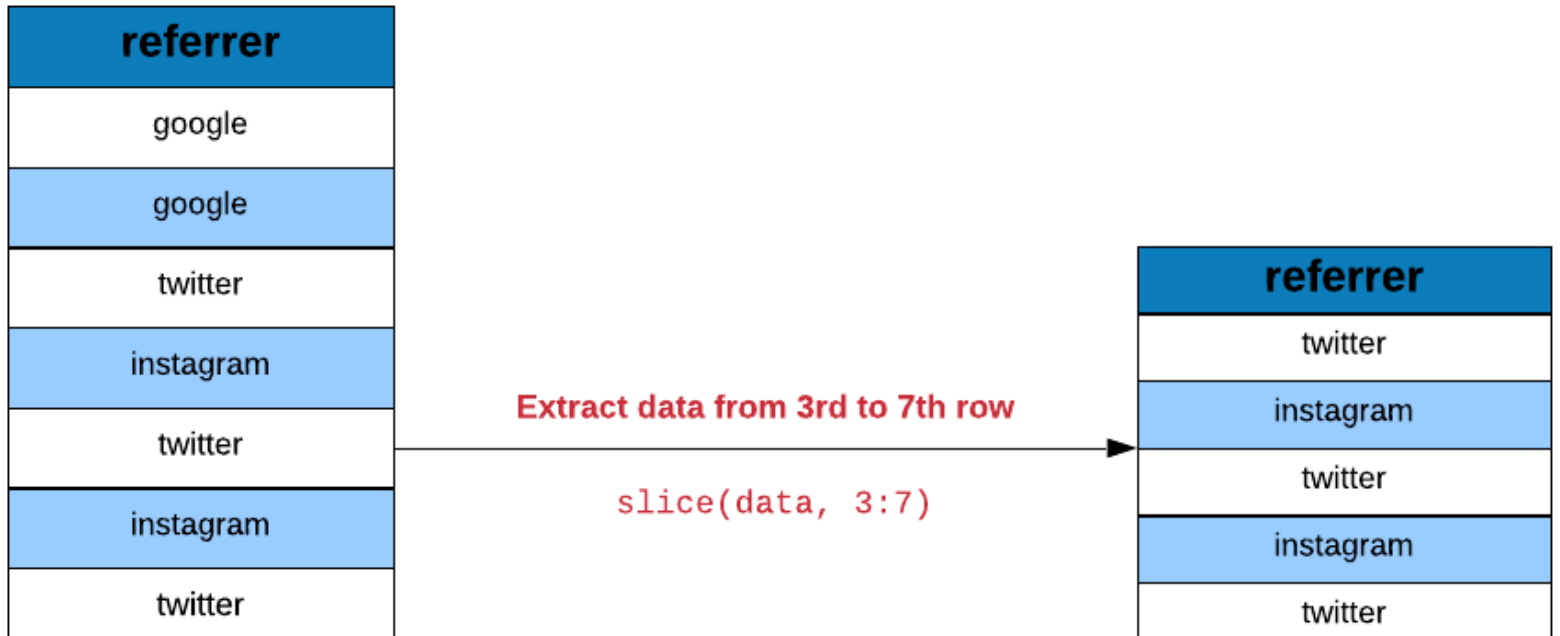
```
##  [1] yahoo  google bing   social google yahoo  social yahoo  google y
## Levels: bing direct social yahoo google
```

Extract Last Column

```
pull(ecom_mini, -1)
```

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

Extract Rows



Extract 10 rows starting from 15

```
slice(ecom, 5:14)
```

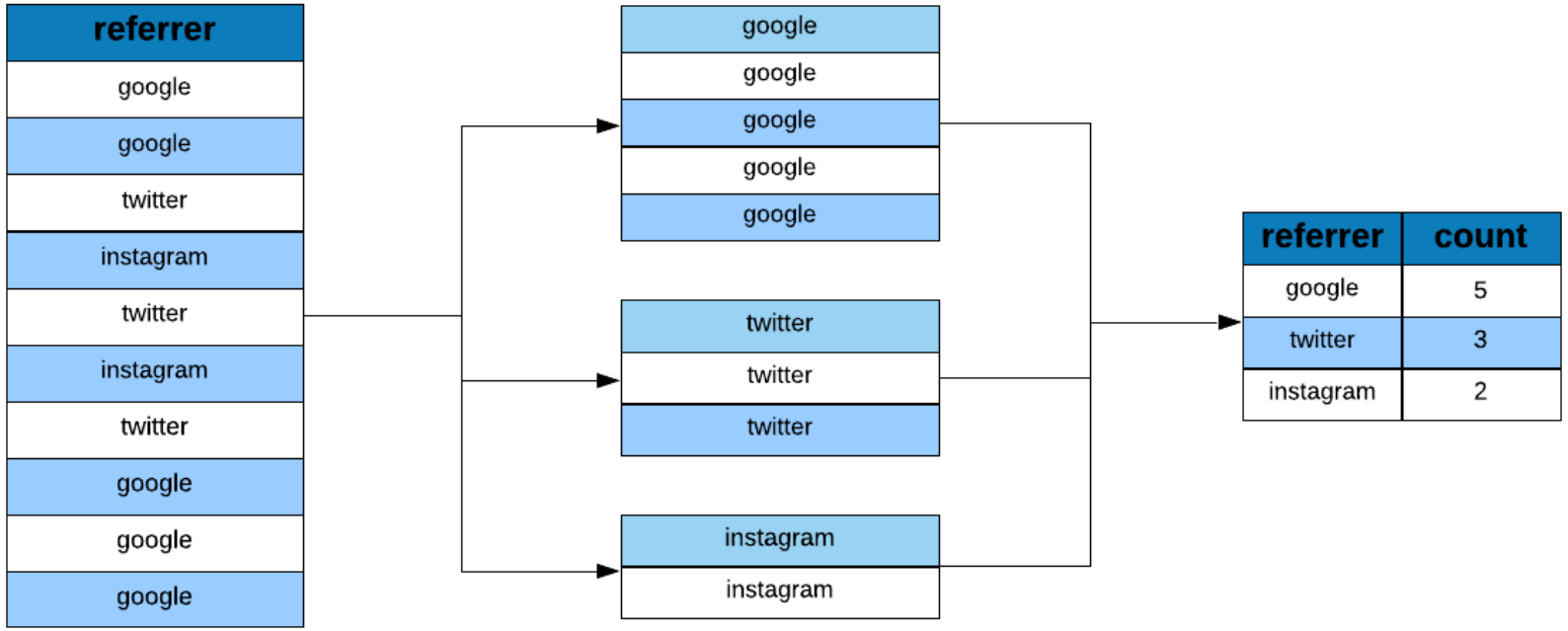
```
## # A tibble: 10 x 7
##    referrer device bouncers n_visit n_pages duration purchase
##    <fct>    <fct>  <lgl>      <dbl>   <dbl>    <dbl> <lgl>
##  1 yahoo    mobile TRUE           9       1      955 FALSE
##  2 yahoo    laptop FALSE          5       5      135 FALSE
##  3 yahoo    mobile TRUE          10       1       75 FALSE
##  4 direct   mobile TRUE          10       1      908 FALSE
##  5 bing     mobile FALSE          3      19      209 FALSE
##  6 google   mobile TRUE           6       1      208 FALSE
##  7 direct   laptop TRUE           9       1      738 FALSE
##  8 direct   tablet FALSE          6      12      132 FALSE
##  9 direct   mobile FALSE          9      14      406 TRUE
## 10 yahoo    tablet FALSE          5       8       80 FALSE
```

Extract Last Row

```
slice(ecom, n())
```

```
## # A tibble: 1 x 7
##   referrer device bouncers n_visit n_pages duration purchase
##   <fct>    <fct>  <lgl>      <dbl>   <dbl>    <dbl> <lgl>
## 1 google   mobile TRUE           9       1      269 FALSE
```

Tally


referrer
google
google
twitter
instagram
twitter
instagram
twitter
google
google
google
google
google
google
google
google
twitter
twitter
twitter
instagram
instagram
referrer
count
google
5
twitter
3
instagram
2

Tabulate Referrers

```
ecom %>%
  group_by(referrer) %>%
  tally()
```

```
## # A tibble: 5 x 2
##   referrer     n
##   <fct>    <int>
## 1 bing       194

## 2 direct     191
## 3 social     200
## 4 yahoo      207
## 5 google     208
```

Tabulate Referrers and Bouncers

```
ecom %>%
  group_by(referrer, bouncers) %>%
  tally()
```

```
## # A tibble: 10 x 3
## # Groups:   referrer [?]
##    referrer bouncers     n
##    <fct>    <lgl>    <int>
##  1 bing     FALSE      104
##  2 bing     TRUE        90
##  3 direct   FALSE       98
##  4 direct   TRUE        93
##  5 social   FALSE       93
##  6 social   TRUE       107
##  7 yahoo    FALSE      110
##  8 yahoo    TRUE        97
##  9 google   FALSE      101
## 10 google   TRUE       107
```

Tabulate Referrers and Purchasers

```
ecom %>%
  group_by(referrer, purchase) %>%
  tally()
```

```
## # A tibble: 10 x 3
## # Groups:   referrer [?]
##    referrer purchase     n
##    <fct>    <lgl>    <int>

##  1 bing     FALSE      177
##  2 bing     TRUE        17
##  3 direct   FALSE      166
##  4 direct   TRUE        25
##  5 social   FALSE      180
##  6 social   TRUE        20
##  7 yahoo    FALSE      185
##  8 yahoo    TRUE        22
##  9 google   FALSE      189
## 10 google   TRUE        19
```

Tabulate Referrers & Converts

```
ecom %>%
  group_by(referrer, purchase) %>%
  tally() %>%
  filter(purchase)
```

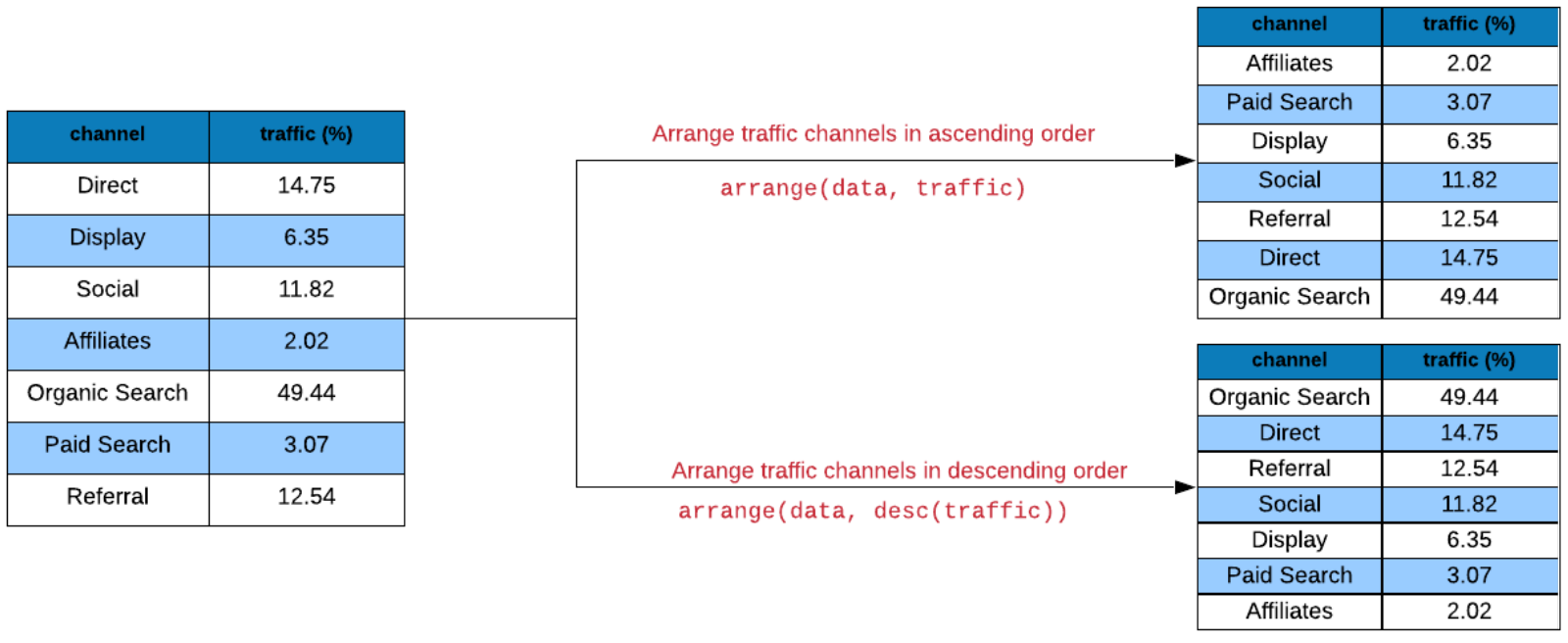
```
## # A tibble: 5 x 3
## # Groups:   referrer [5]
##   referrer purchase     n
##
##   <fct>    <lgl>    <int>
## 1 bing     TRUE        17
## 2 direct   TRUE        25
## 3 social   TRUE        20
## 4 yahoo    TRUE        22
## 5 google   TRUE        19
```

Count

```
count(ecom, referrer, purchase)
```

```
## # A tibble: 10 x 3
##    referrer purchase     n
##    <fct>    <lgl>    <int>
##  1 bing     FALSE      177
##  2 bing     TRUE        17
##  3 direct   FALSE      166
##  4 direct   TRUE        25
##  5 social   FALSE      180
##  6 social   TRUE        20
##  7 yahoo    FALSE      185
##  8 yahoo    TRUE        22
##  9 google   FALSE      189
## 10 google   TRUE        19
```

Arrange


channel
traffic (%)
Direct
14.75
Display
6.35
Social
11.82
Affiliates
2.02
Organic Search
49.44
Paid Search
3.07
Referral
12.54
Arrange traffic channels in ascending order
arrange(data, traffic)
Arrange traffic channels in descending order
arrange(data, desc(traffic))
channel
traffic (%)
Affiliates
2.02
Paid Search
3.07
Display
6.35
Social
11.82
Referral
12.54
Direct
14.75
Organic Search
49.44
channel
traffic (%)
Organic Search
49.44
Direct
14.75
Referral
12.54
Social
11.82
Display
6.35
Paid Search
3.07
Affiliates
2.02

Top 2 Referrers by Orders

```
ecom %>%
  count(referrer, purchase) %>%
  filter(purchase) %>%
  arrange(desc(n)) %>%
  top_n(n = 2)
```

```
## Selecting by n
```

```
## # A tibble: 2 x 3
##   referrer purchase     n
##   <fct>    <lgl>    <int>
## 1 direct   TRUE        25
## 2 yahoo    TRUE        22
```

Between

```
ecom_sample <- sample_n(ecom, 30)
ecom_sample %>%
  pull(n_pages) %>%
  between(5, 15)
```

```
##  [1] FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALS
## [12] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRU
## [23] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
```

Case When

```
ecom %>%
  mutate(
    repeat_visit = case_when(
      n_visit > 0 ~ TRUE,
      TRUE ~ FALSE
    )
  ) %>%
  select(n_visit, repeat_visit)
```

```
## # A tibble: 1,000 x 2
##    n_visit repeat_visit
##      <dbl> <lgl>
##  1      10 TRUE
##  2       9 TRUE
##  3       0 FALSE
##  4       3 TRUE
##  5       9 TRUE
##  6       5 TRUE
##  7      10 TRUE
##  8      10 TRUE
##  9       3 TRUE
## 10       6 TRUE
## # ... with 990 more rows
```

Select First Observation

```
ecom %>%
  pull(referrer) %>%
  nth(1)
```

```
## [1] google
## Levels: bing direct social yahoo google
```

```
ecom %>%
  pull(referrer) %>%
  first()
```

```
## [1] google
## Levels: bing direct social yahoo google
```

Select 1000th Observation

```
ecom %>%
  pull(referrer) %>%
  nth(1000)
```

```
## [1] google
## Levels: bing direct social yahoo google
```

Select Last Observation

```
ecom %>%
  pull(referrer) %>%
  last()
```

```
## [1] google
## Levels: bing direct social yahoo google
```

Thank You

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