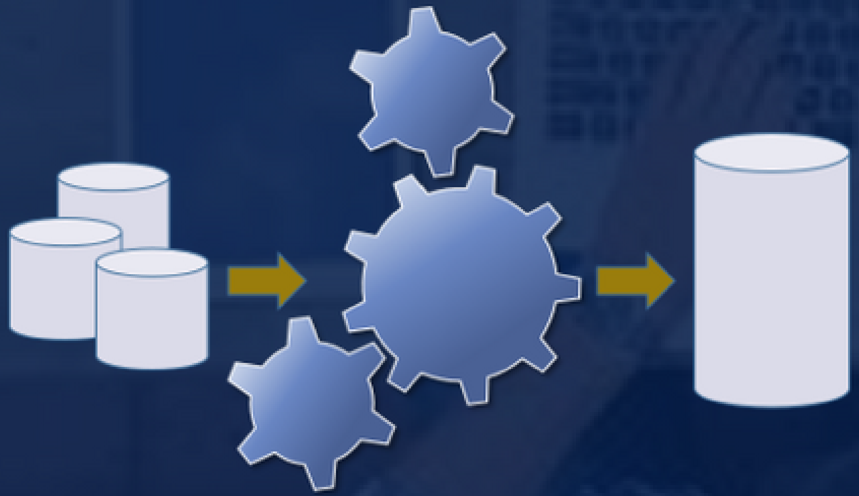
Rsquared Academy



Data Wrangling with dplyr

Agenda

In this module, we will explore a set of helper functions in order to:

- 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

Case Study

Let us look at a case study (e-commerce data) and see how we can use dplyr helper functions to answer questions we have about and to modify/transform the underlying data set. You can download the data from here or import it directly using `read_csv()` from the readr package.

Libraries

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

Data

```r
ecom <- read_csv('data/web.csv')
```

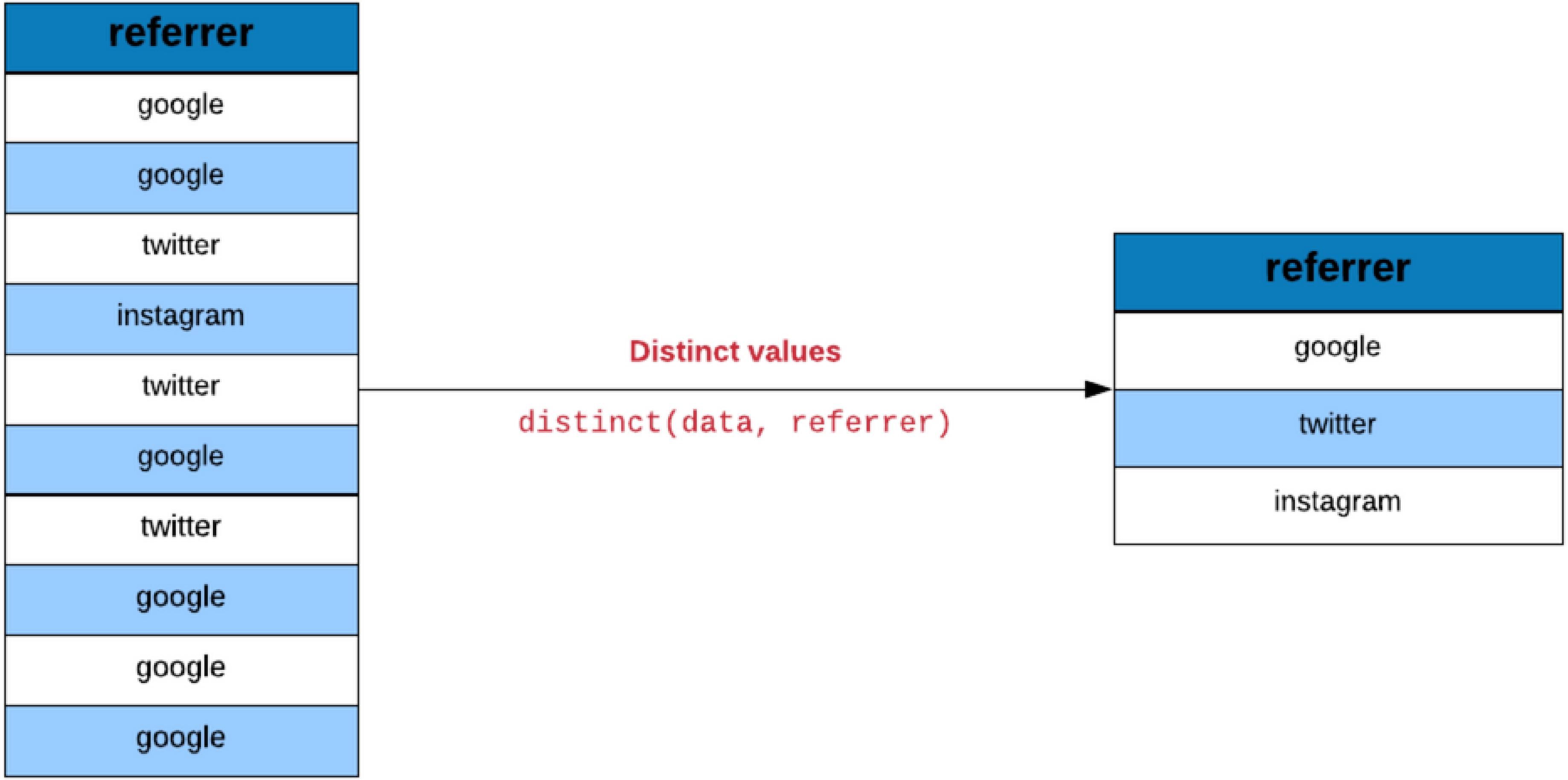
```
## # A tibble: 1,000 x 11
##       id referrer device bouncers n_visit n_pages duration country
##    <int> <chr>    <chr>  <chr>      <int>   <dbl>    <dbl> <chr>
##  1     1 google   laptop true          10    1.00      693 Czech Repu
##  2     2 yahoo    tablet true           9    1.00      459 Yemen
##  3     3 direct   laptop true           0    1.00      996 Brazil
##  4     4 bing     tablet false          3   18.0       468 China
##  5     5 yahoo    mobile true           9    1.00      955 Poland
##  6     6 yahoo    laptop false          5    5.00      135 South Afri
##  7     7 yahoo    mobile true          10    1.00       75.0 Bangladesh
##  8     8 direct   mobile true          10    1.00      908 Indonesia
##  9     9 bing     mobile false          3   19.0       209 Netherland
## 10    10 google   mobile true           6    1.00      208 Czech Repu
## # ... with 990 more rows, and 3 more variables: purchase <chr>,
## #   order_items <dbl>, order_value <dbl>
```

Data Dictionary

- id: row id
- referrer: referrer website/search engine
- os: operating system
- browser: browser
- device: device used to visit the website

- n_pages: number of pages visited
- duration: time spent on the website (in seconds)
- repeat: frequency of visits
- country: country of origin
- purchase: whether visitor purchased
- order_value: order value of visitor (in dollars)

Distinct



referrer
google
google
twitter
instagram
twitter
google
twitter
google
google
google

Distinct values

`distinct(data, referrer)`

referrer
google
twitter
instagram

Traffic Sources

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

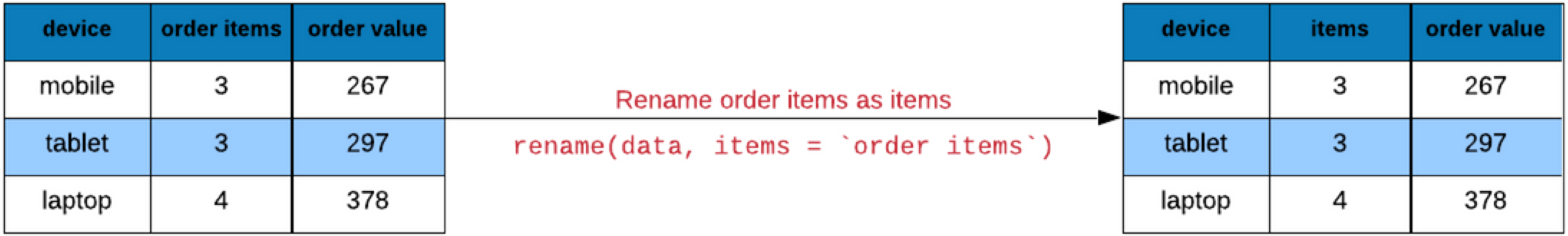
```
## # A tibble: 5 x 1
##   referrer
##   <chr>
## 1 google
## 2 yahoo
## 3 direct
## 4 bing
## 5 social
```

Device Types

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

```
## # A tibble: 3 x 1
##   device
##   <chr>
## 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`)
```

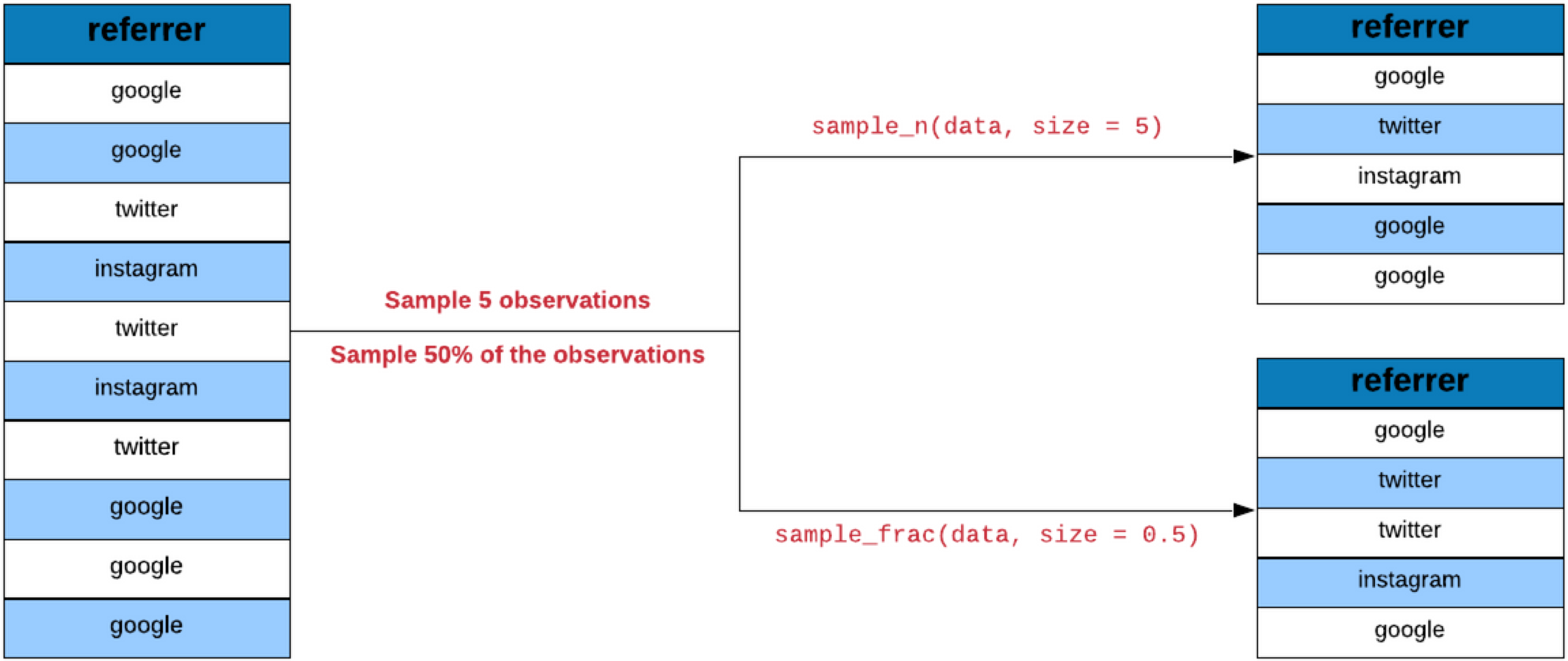
device	items	order value
mobile	3	267
tablet	3	297
laptop	4	378

Rename Columns

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

```
## # A tibble: 1,000 x 11
##       id referrer device bouncers n_visit n_pages time_on_site
##    <int> <chr>    <chr>  <chr>      <int>   <dbl>        <dbl>
##  1     1 google   laptop true          10    1.00        693
##  2     2 yahoo    tablet true           9    1.00        459
##  3     3 direct   laptop true           0    1.00        996
##  4     4 bing     tablet false          3   18.0         468
##  5     5 yahoo    mobile true           9    1.00        955
##  6     6 yahoo    laptop false          5    5.00        135
##  7     7 yahoo    mobile true          10    1.00         75.0
##  8     8 direct   mobile true          10    1.00        908
##  9     9 bing     mobile false          3   19.0         209
## 10    10 google   mobile true           6    1.00        208
##    country        purchase order_items order_value
##    <chr>          <chr>          <dbl>       <dbl>
##  1 Czech Republic false          0              0
##  2 Yemen          false          0              0
```

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)
```

referrer
google
twitter
instagram
google
google

```
sample_frac(data, size = 0.5)
```

referrer
google
twitter
twitter
instagram
google

Sampling Data

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

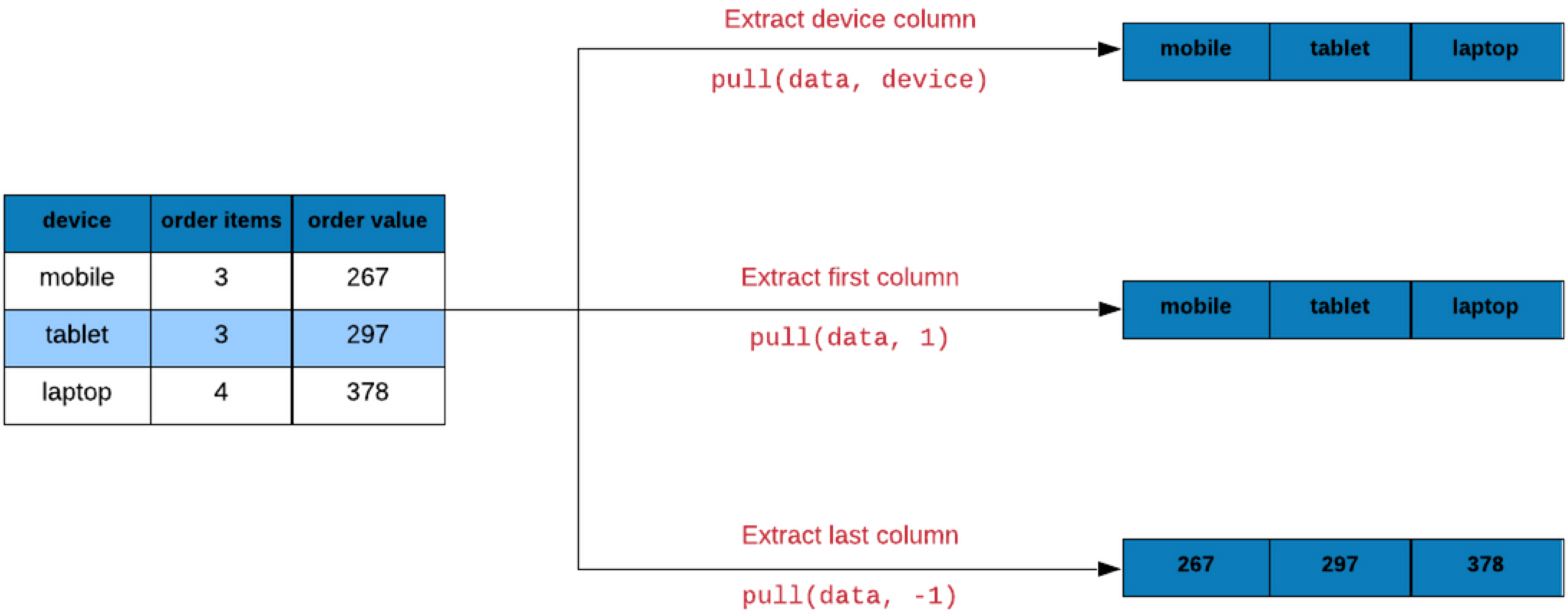
```
## # A tibble: 700 x 11
##       id referrer device bouncers n_visit n_pages duration country
##    <int> <chr>    <chr>  <chr>      <int>   <dbl>    <dbl> <chr>
##  1   876 direct   laptop false          4    2.00     44.0 United Sta
##  2   933 google   mobile false          9    6.00     96.0 Portugal
##  3   526 yahoo    tablet false         10    9.00    153   Indonesia
##  4   959 direct   mobile false          1    9.00    135   Indonesia
##  5   749 bing     laptop false          0   17.0     272   Indonesia
##  6   425 yahoo    laptop false          2    7.00    105   Cyprus
##  7   242 bing     mobile false         10   13.0     221   China
##  8   756 yahoo    tablet false          6    8.00    224   France
##  9   341 bing     mobile false          5   15.0     405   Russia
## 10   495 google   laptop false          8   18.0     522   Japan
##    purchase order_items order_value
##    <chr>          <dbl>       <dbl>
##  1 false           8.00        1801
##  2 false           8.00        1354
```

Sampling Data

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

```
## # A tibble: 700 x 11
##        id referrer device bouncers n_visit n_pages duration country
##     <int> <chr>    <chr>  <chr>      <int>   <dbl>    <dbl> <chr>
##  1   772 social   mobile false          6    9.00    126    Norway
##  2    53 social   tablet false          3   12.0     324    China
##  3   733 google   laptop true           4    1.00    164    Russia
##  4   875 direct   laptop false          2    4.00     80.0 United Sta
##  5   169 direct   laptop false          6    6.00     96.0 Albania
##  6   306 direct   mobile false          7    7.00     98.0 Philippine
##  7   442 direct   tablet true           3    1.00    632    Jamaica
##  8   217 google   mobile false          9    9.00    135    Poland
##  9   615 social   laptop true           1    1.00     10.0 Finland
## 10   684 yahoo    tablet true           5    1.00    386    China
##     purchase order_items order_value
##     <chr>          <dbl>       <dbl>
##  1 false              0           0
##  2 false              0           0
```

Extract Columns



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

Extract device column

`pull(data, device)`

mobile	tablet	laptop

Extract first column

`pull(data, 1)`

mobile	tablet	laptop

Extract last column

`pull(data, -1)`

267	297	378

Extract Device Column

```r
pull(ecom, device)
```

```
##    [1] "laptop" "tablet" "laptop" "tablet" "mobile" "laptop" "mobile'
##    [8] "mobile" "mobile" "mobile" "laptop" "tablet" "mobile" "tablet'
##   [15] "mobile" "laptop" "tablet" "tablet" "tablet" "tablet" "laptop'
##   [22] "mobile" "mobile" "laptop" "laptop" "laptop" "tablet" "laptop'
##   [29] "mobile" "mobile" "tablet" "mobile" "laptop" "tablet" "mobile'
##   [36] "mobile" "laptop" "mobile" "mobile" "mobile" "mobile" "mobile'
##   [43] "laptop" "tablet" "laptop" "tablet" "mobile" "laptop" "mobile'
##   [50] "tablet" "mobile" "mobile" "tablet" "tablet" "mobile" "tablet'
##   [57] "laptop" "tablet" "tablet" "laptop" "laptop" "tablet" "mobile'
##   [64] "tablet" "laptop" "tablet" "tablet" "mobile" "tablet" "mobile'
##   [71] "laptop" "laptop" "tablet" "tablet" "tablet" "tablet" "laptop'
##   [78] "laptop" "mobile" "laptop" "laptop" "tablet" "mobile" "tablet'
##   [85] "tablet" "tablet" "tablet" "tablet" "mobile" "mobile" "laptop'
##   [92] "mobile" "laptop" "tablet" "tablet" "tablet" "tablet" "mobile'
##   [99] "mobile" "laptop" "tablet" "mobile" "laptop" "tablet" "mobile'
##  [106] "mobile" "mobile" "laptop" "tablet" "mobile" "tablet" "mobile'
##  [113] "tablet" "tablet" "laptop" "mobile" "tablet" "laptop" "laptop'
```

Extract First Column

```r
pull(ecom, 1)
```

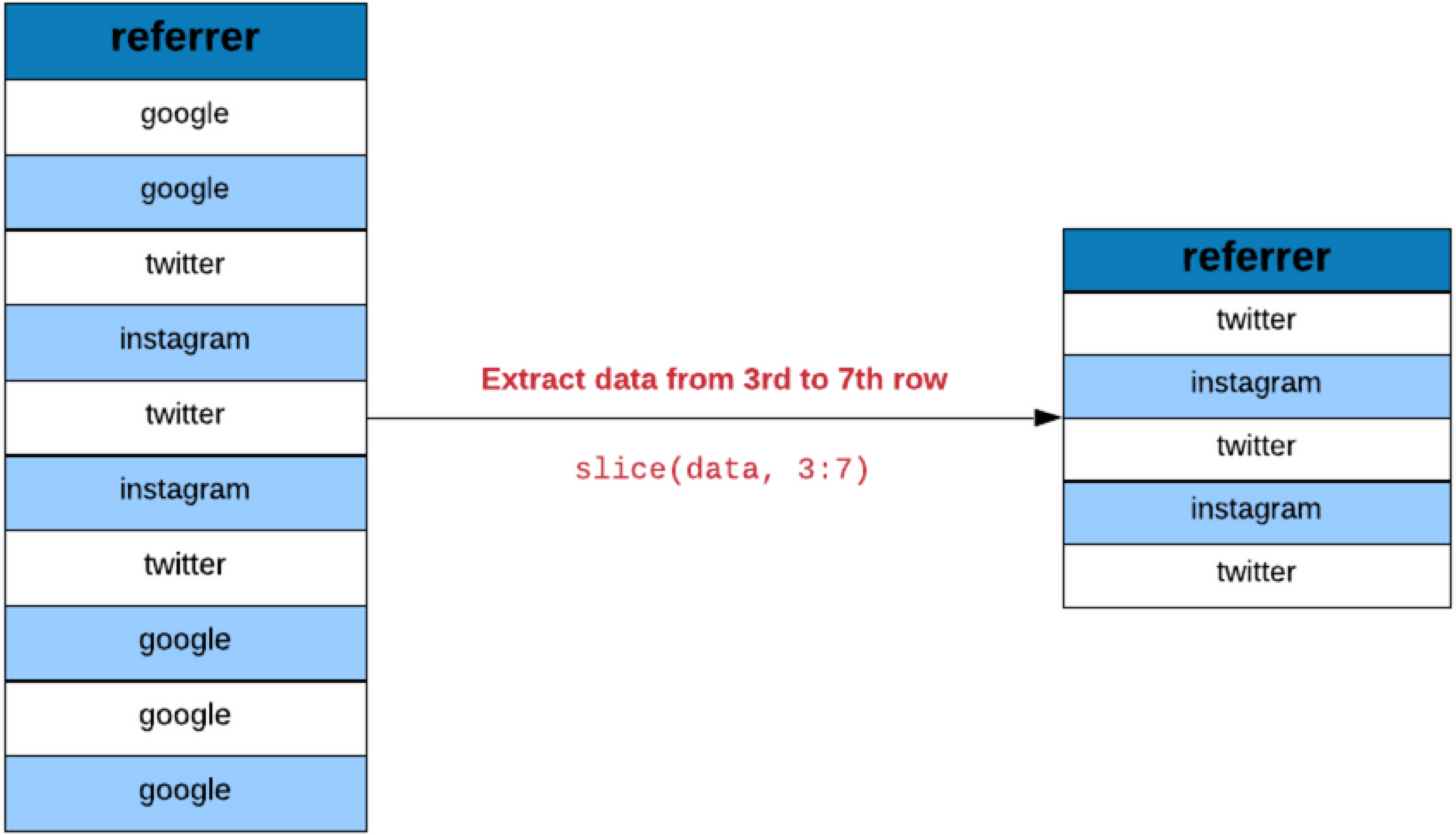
```
##    [1]   1   2   3   4   5   6   7   8   9  10  11  12
##   [14]  14  15  16  17  18  19  20  21  22  23  24  25
##   [27]  27  28  29  30  31  32  33  34  35  36  37  38
##   [40]  40  41  42  43  44  45  46  47  48  49  50  51
##   [53]  53  54  55  56  57  58  59  60  61  62  63  64
##   [66]  66  67  68  69  70  71  72  73  74  75  76  77
##   [79]  79  80  81  82  83  84  85  86  87  88  89  90
##   [92]  92  93  94  95  96  97  98  99 100 101 102 103 1
##  [105] 105 106 107 108 109 110 111 112 113 114 115 116 1
##  [118] 118 119 120 121 122 123 124 125 126 127 128 129 1
##  [131] 131 132 133 134 135 136 137 138 139 140 141 142 1
##  [144] 144 145 146 147 148 149 150 151 152 153 154 155 1
##  [157] 157 158 159 160 161 162 163 164 165 166 167 168 1
##  [170] 170 171 172 173 174 175 176 177 178 179 180 181 1
##  [183] 183 184 185 186 187 188 189 190 191 192 193 194 1
##  [196] 196 197 198 199 200 201 202 203 204 205 206 207 2
##  [209] 209 210 211 212 213 214 215 216 217 218 219 220 2
```

Extract Last Column

```r
pull(ecom, -1)
```

```
##   [1]    0    0    0  434    0    0    0    0    0    0    0    0  6
##  [14]  362 2423    0 1049    0 1304 2077    0    0  237    0    0
##  [27]  622    0    0    0    0    0    0 1613    0 1885    0    0
##  [40]    0  184    0    0    0    0    0    0    0 1515    0    0
##  [53]    0    0    0    0 1532    0    0    0    0    0    0 2798  3
##  [66]    0    0    0    0    0    0 2216    0    0    0  632    0
##  [79]    0    0    0    0    0    0    0    0    0    0 2001    0
##  [92] 1273    0  286    0  722    0  764    0    0 1667  583    0
## [105]    0    0    0    0    0    0    0  287 1482    0 2514    0
## [118]    0    0 1772    0    0    0    0 1443    0    0    0    0
## [131]  489    0    0 2449    0    0    0    0  287    0    0    0 28
## [144]    0 2086    0 2055    0  393    0    0  907    0    0    0 16
## [157]    0 1358 1833    0    0    0    0    0    0 1155  837    0
## [170]    0    0    0  358    0    0    0 1252    0    0    0    0 24
## [183]    0    0    0    0 1286    0    0    0    0    0 1578    0
## [196]    0    0    0    0    0    0    0    0    0    0    0    0  1
## [209]    0    0    0    0    0    0    0    0 1758    0 1021    0 22
```

Slice Rows



referrer
google
google
twitter
instagram
twitter
instagram
twitter
google
google
google

Extract data from 3rd to 7th row

```
slice(data, 3:7)
```

referrer
twitter
instagram
twitter
instagram
twitter

Extract First 20 Rows

```r
slice(ecom, 1:20)
```

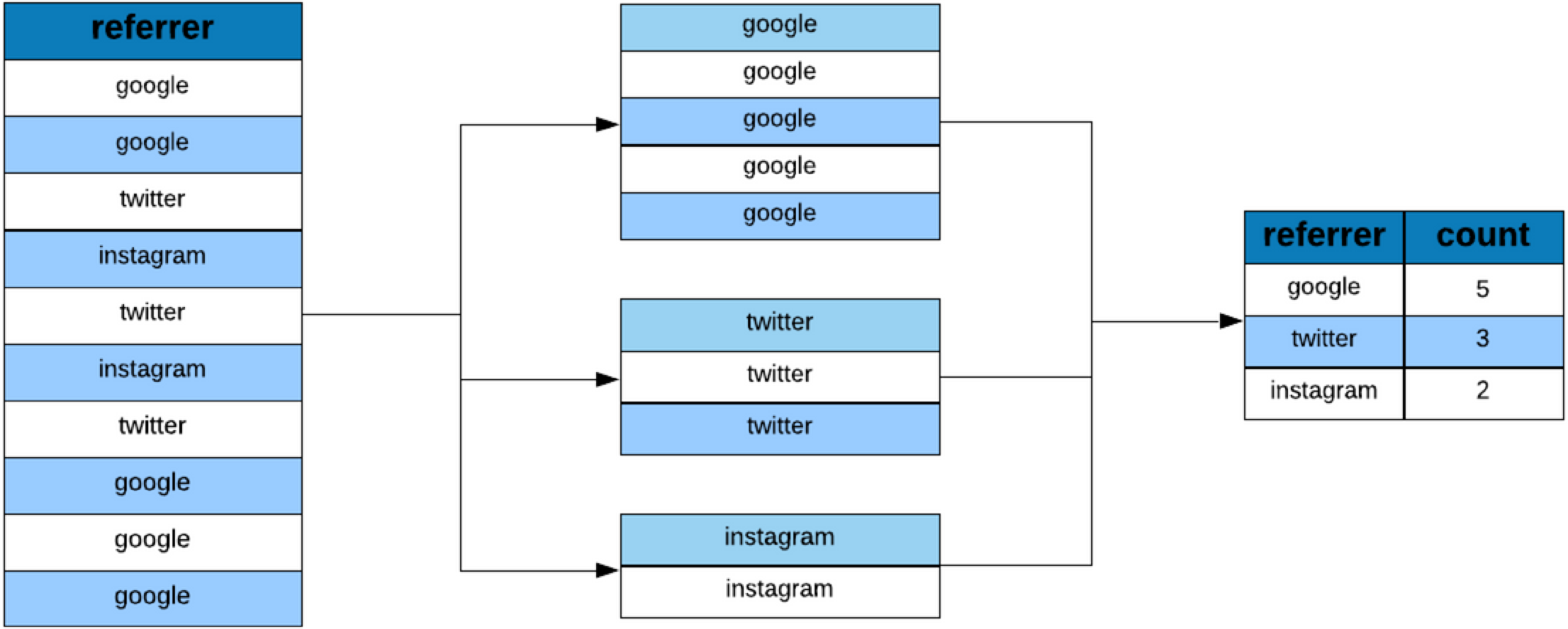
```
## # A tibble: 20 x 11
##        id referrer device bouncers n_visit n_pages duration country
##    <int> <chr>    <chr>  <chr>       <int>   <dbl>    <dbl> <chr>
##  1     1 google   laptop true          10    1.00    693   Czech Repu
##  2     2 yahoo    tablet true           9    1.00    459   Yemen
##  3     3 direct   laptop true           0    1.00    996   Brazil
##  4     4 bing     tablet false          3   18.0     468   China
##  5     5 yahoo    mobile true           9    1.00    955   Poland
##  6     6 yahoo    laptop false          5    5.00    135   South Afri
##  7     7 yahoo    mobile true          10    1.00     75.0 Bangladesh
##  8     8 direct   mobile true          10    1.00    908   Indonesia
##  9     9 bing     mobile false          3   19.0     209   Netherland
## 10    10 google   mobile true           6    1.00    208   Czech Repu
## 11    11 direct   laptop true           9    1.00    738   Jamaica
## 12    12 direct   tablet false          6   12.0     132   Estonia
## 13    13 direct   mobile false          9   14.0     406   Ireland
## 14    14 yahoo    tablet false          5    8.00     80.0 Philippine
```

Extract Last Row

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

```
## # A tibble: 1 x 11
##      id referrer device bouncers n_visit n_pages duration country pur
##   <int> <chr>    <chr>  <chr>      <int>   <dbl>    <dbl> <chr>   <ch
## 1  1000 google   mobile true           9    1.00      269 China   fal
##   order_items order_value
##         <dbl>       <dbl>
## 1           0           0
```

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
##   <chr>    <int>
## 1 bing       194
## 2 direct     191
## 3 google     208
## 4 social     200
## 5 yahoo      207
```

Tabulate referrers and bouncers

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

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

Tabulate referrers and purchasers

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

```
## # A tibble: 10 x 3
## # Groups:   referrer [?]
##    referrer purchase     n
##    <chr>    <chr>    <int>
##  1 bing     false      177
##  2 bing     true        17
##  3 direct   false      166
##  4 direct   true        25
##  5 google   false      189
##  6 google   true        19
##  7 social   false      180
##  8 social   true        20
##  9 yahoo    false      185
## 10 yahoo    true        22
```

Tabulate Referrers & Converts

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

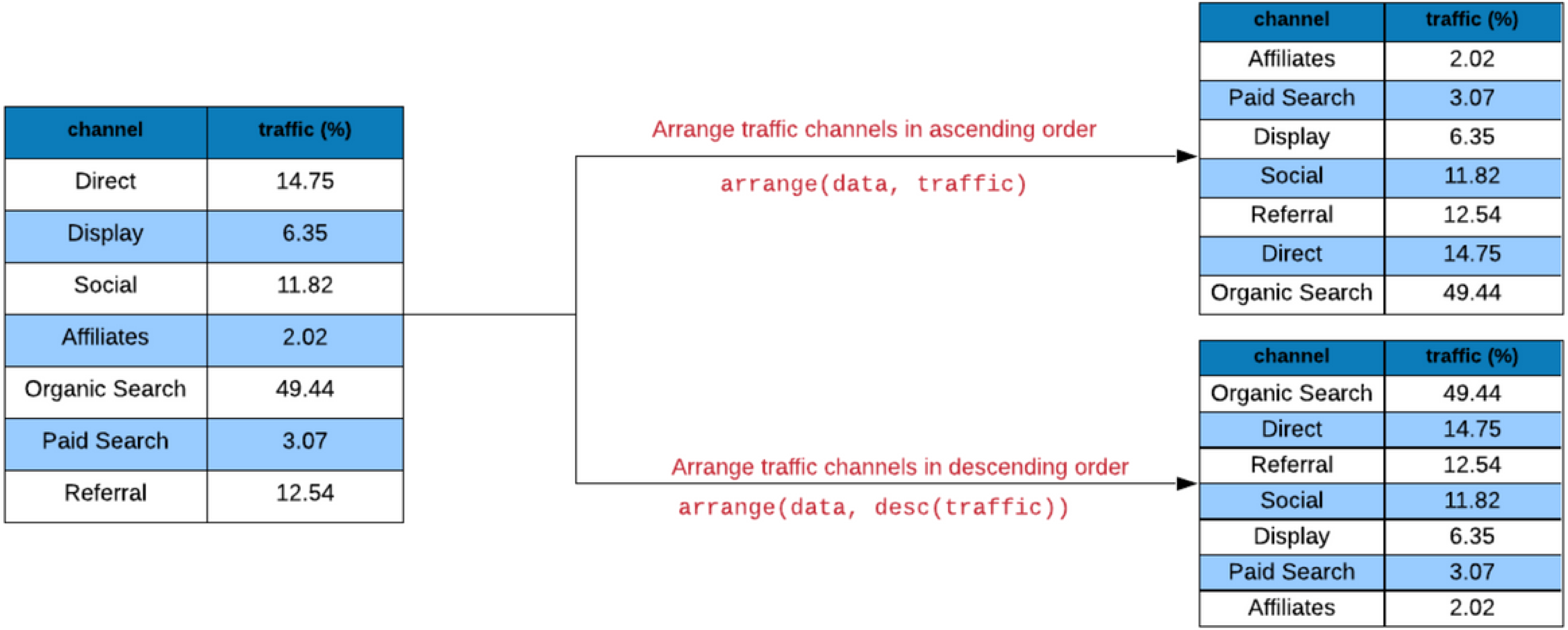
```
## # A tibble: 5 x 3
## # Groups:   referrer [5]
##   referrer purchase     n
##   <chr>    <chr>    <int>
## 1 bing     true        17
## 2 direct   true        25
## 3 google   true        19
## 4 social   true        20
## 5 yahoo    true        22
```

Count

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

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

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)
```

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

Arrange traffic channels in descending order

```
arrange(data, desc(traffic))
```

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 == 'true') %>%
  arrange(desc(n)) %>%
  top_n(n = 2)
```

```
## Selecting by n
```

```
## # A tibble: 2 x 3
##   referrer purchase     n
##   <chr>    <chr>    <int>
## 1 direct   true        25
## 2 yahoo    true        22
```

Between

```r
ecom %>%
  pull(n_pages) %>%
  between(5, 15)
```

```
##    [1] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FA
##   [12]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FA
##   [23]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FA
##   [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  T
##   [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FA
##   [56] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  T
##   [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FA
##   [78] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  T
##   [89]  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FA
##  [100] FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  T
##  [111] FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE FA
##  [122] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FA
##  [133] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FA
##  [144] FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FA
##  [155]  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FA
##  [166] FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FA
##  [177] FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FA
```

Case When

```
mtcars %>%
  select(mpg, disp, cyl, gear, carb) %>%
  mutate(
    type = case_when(
      disp > 300 ~ 'High',
      cyl == 8 ~ 'Eight',
      TRUE ~ 'True'
    )
  )
```

```
##     mpg  disp cyl gear carb  type
## 1  21.0 160.0   6    4    4  True
## 2  21.0 160.0   6    4    4  True
## 3  22.8 108.0   4    4    1  True
## 4  21.4 258.0   6    3    1  True
## 5  18.7 360.0   8    3    2  High
## 6  18.1 225.0   6    3    1  True
## 7  14.3 360.0   8    3    4  High
## 8  24.4 146.7   4    4    2  True
## 9  22.8 140.8   4    4    2  True
## 10 19.2 167.6   6    4    4  True
## 11 17.8 167.6   6    4    4  True
## 12 16.4 275.8   8    3    3 Eight
## 13 17.3 275.8   8    3    3 Eight
## 14 15.2 275.8   8    3    3 Eight
## 15 10.4 472.0   8    3    4  High
## 16 10.4 460.0   8    3    4  High
```

Select First Observation

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

```
## [1] "google"
```

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

```
## [1] "google"
```

Select 1000th Observation

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

```
## [1] "google"
```

Select Last Observation

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

```
## [1] "google"
```

Thank You

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