Data Wrangling with R - Part 2

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

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

Data

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

```
## # A tibble: 1,000 x 11
##       id referrer device bouncers n_visit n_pages durati
##    <int> <chr>    <chr>  <chr>       <int>   <dbl>    <db
##  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
##  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
## # ... with 990 more rows, and 3 more variables: purchase
## #   order_items <dbl>, order_value <dbl>
```

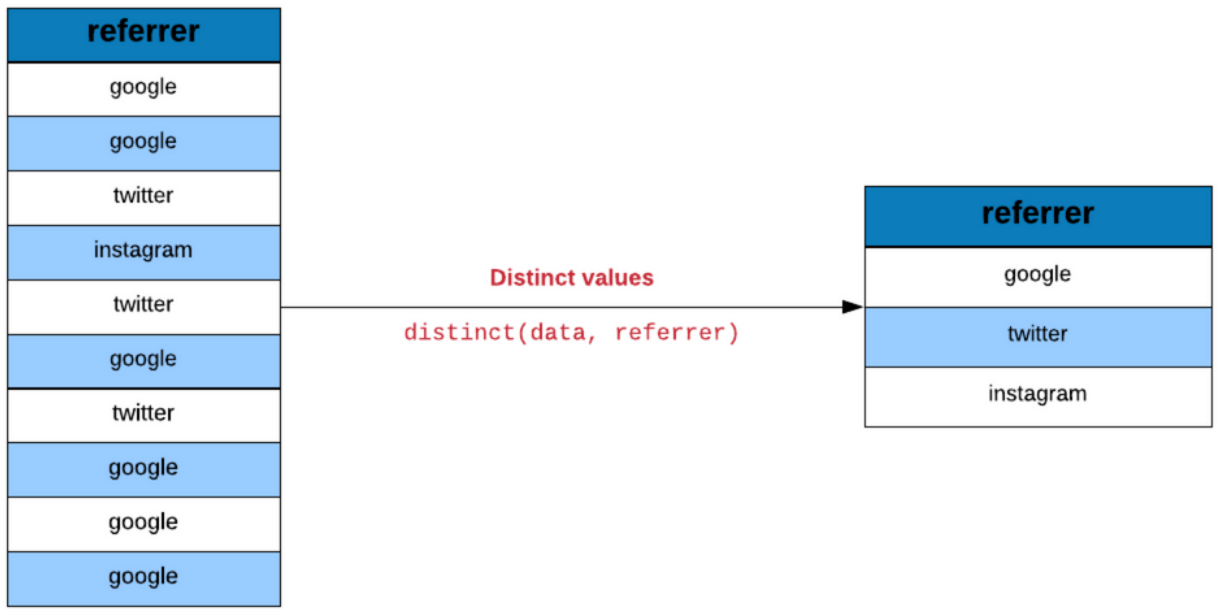
Data Dictionary

- 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)

Data Sanitization

- `distinct()`
- `rename()`

Distinct


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

Traffic Sources

```
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`)
```

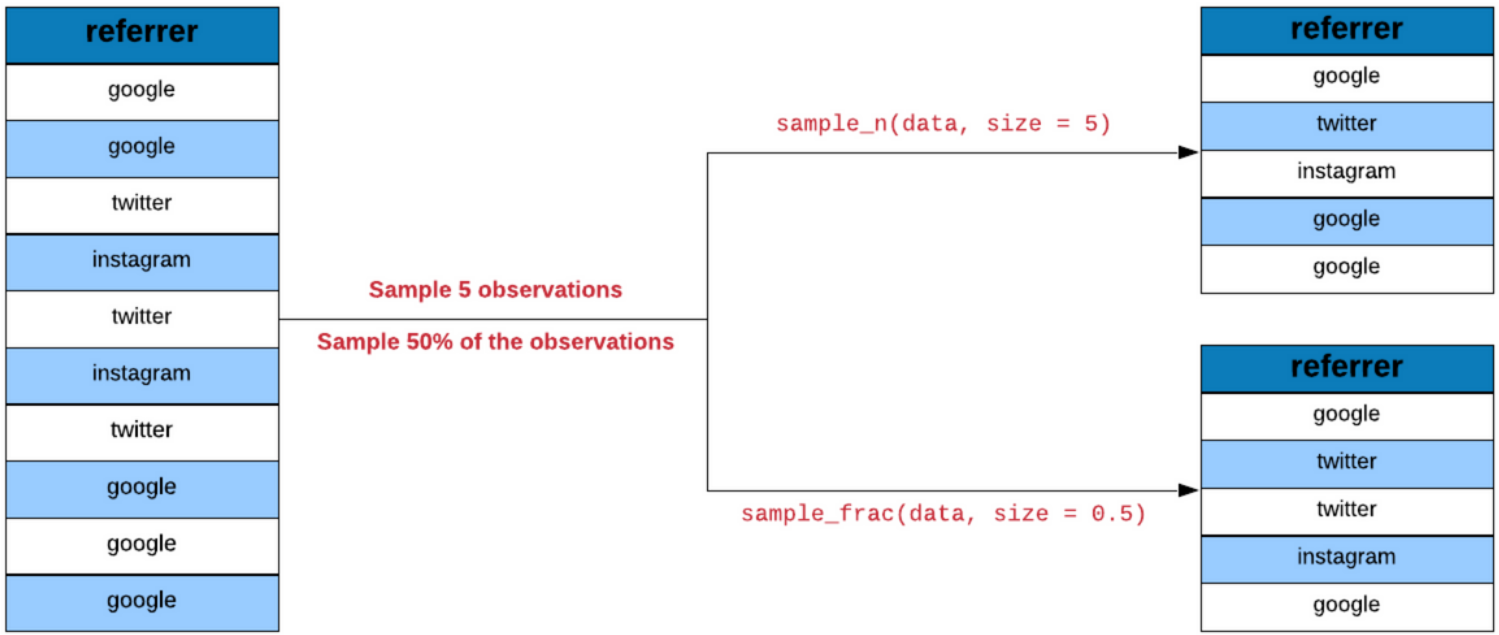
device	items	order value
mobile	3	267
tablet	3	297
laptop	4	378

Rename Columns

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

```
## # A tibble: 1,000 x 11
##       id referrer device bouncers n_visit n_pages time_
##    <int> <chr>    <chr>  <chr>       <int>   <dbl>
##  1     1 google   laptop true           10    1.00
##  2     2 yahoo    tablet true            9    1.00
##  3     3 direct   laptop true            0    1.00
##  4     4 bing     tablet false           3   18.0
##  5     5 yahoo    mobile true            9    1.00
##  6     6 yahoo    laptop false           5    5.00
##  7     7 yahoo    mobile true           10    1.00
##  8     8 direct   mobile true           10    1.00
##  9     9 bing     mobile false           3   19.0
## 10    10 google   mobile true            6    1.00
## # ... with 990 more rows, and 3 more variables: purchase
## #   order_items <dbl>, order_value <dbl>
```

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

Sampling Data

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

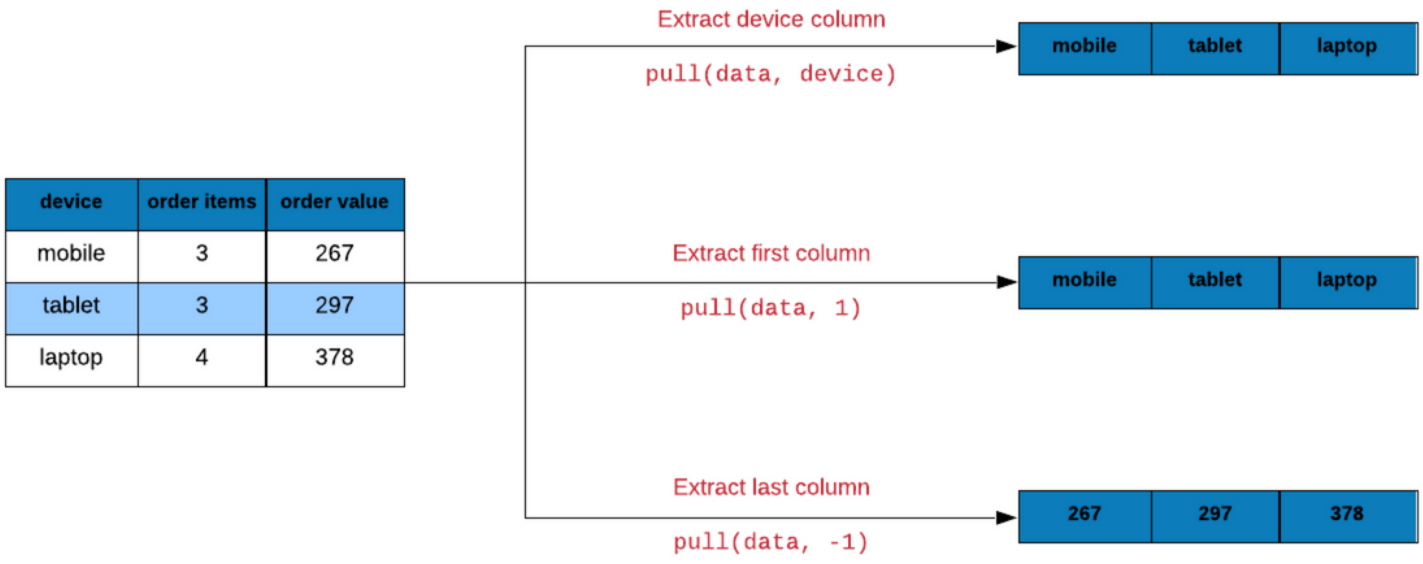
```
## # A tibble: 700 x 11
##       id referrer device bouncers n_visit n_pages durati
##    <int> <chr>    <chr>  <chr>       <int>   <dbl>    <db
##  1   133 direct   mobile false           2    4.00     76
##  2   509 yahoo    laptop true            1    1.00    991
##  3   194 google   laptop true            3    1.00    978
##  4   131 bing     tablet false           3   15.0     195
##  5   148 yahoo    laptop true            5    1.00    739
##  6   999 yahoo    mobile true            1    1.00    714
##  7     5 yahoo    mobile true            9    1.00    955
##  8   236 yahoo    tablet true            5    1.00     10
##  9   283 social   mobile true            5    1.00     22
## 10   199 yahoo    laptop false          10   11.0      110
## # ... with 690 more rows, and 3 more variables: purchase
## #   order_items <dbl>, order_value <dbl>
```

Sampling Data

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

```
## # A tibble: 700 x 11
##       id referrer device bouncers n_visit n_pages durati
##    <int> <chr>    <chr>  <chr>       <int>   <dbl>    <db
##  1   893 yahoo    laptop false           2    9.00     162
##  2   130 direct   laptop true            6    1.00     178
##  3    77 yahoo    laptop true            4    1.00     928
##  4   219 social   mobile false           1   20.0      520
##  5   876 direct   laptop false           4    2.00      44
##  6   233 google   tablet false           3    3.00      78
##  7   302 google   laptop false           6    9.00     135
##  8     3 direct   laptop true            0    1.00     996
##  9   654 direct   mobile false          10   17.0      272
## 10   109 google   tablet true            9    1.00     382
## # ... with 690 more rows, and 3 more variables: purchase
## #   order_items <dbl>, order_value <dbl>
```

Extract Columns



Extract Device Column

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

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

Extract First Column

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

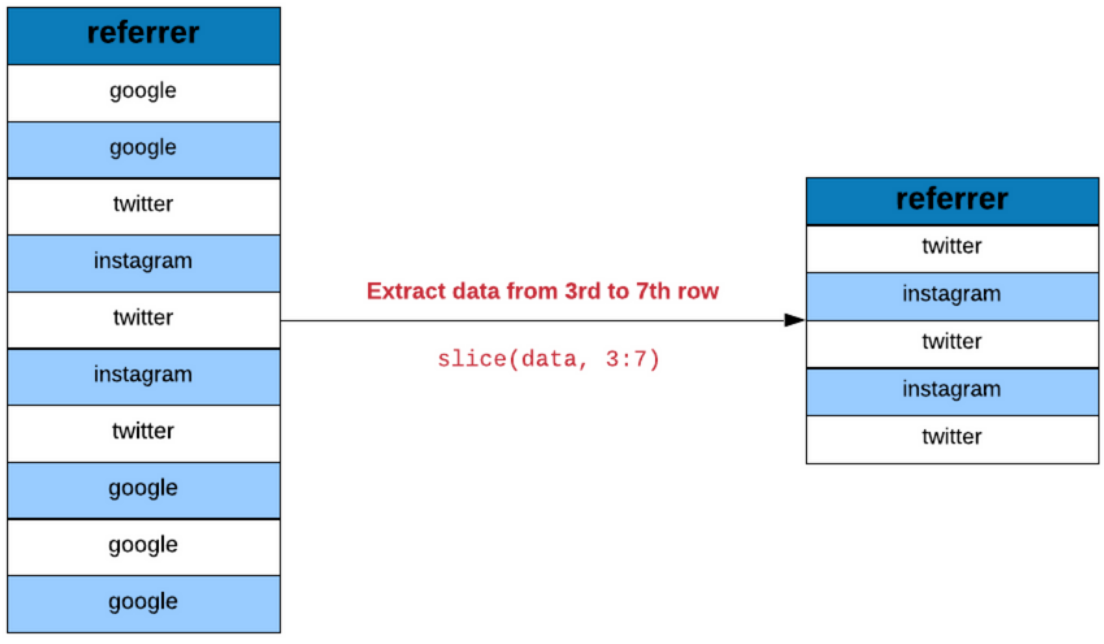
```
##     [1]   1   2   3   4   5   6   7   8   9  10
##    [14]  14  15  16  17  18  19  20  21  22  23
##    [27]  27  28  29  30  31  32  33  34  35  36
##    [40]  40  41  42  43  44  45  46  47  48  49
##    [53]  53  54  55  56  57  58  59  60  61  62
##    [66]  66  67  68  69  70  71  72  73  74  75
##    [79]  79  80  81  82  83  84  85  86  87  88
##    [92]  92  93  94  95  96  97  98  99 100 101
##   [105] 105 106 107 108 109 110 111 112 113 114
##   [118] 118 119 120 121 122 123 124 125 126 127
##   [131] 131 132 133 134 135 136 137 138 139 140
##   [144] 144 145 146 147 148 149 150 151 152 153
##   [157] 157 158 159 160 161 162 163 164 165 166
##   [170] 170 171 172 173 174 175 176 177 178 179
##   [183] 183 184 185 186 187 188 189 190 191 192
##   [196] 196 197 198 199 200 201 202 203 204 205
```

Extract Last Column

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

```
##    [1]    0    0    0  434    0    0    0    0    0    0
##   [14]  362 2423    0 1049    0 1304 2077    0    0  237
##   [27]  622    0    0    0    0    0    0 1613    0 1885
##   [40]    0  184    0    0    0    0    0    0    0 1515
##   [53]    0    0    0    0 1532    0    0    0    0    0
##   [66]    0    0    0    0    0    0 2216    0    0    0
##   [79]    0    0    0    0    0    0    0    0    0    0
##   [92] 1273    0  286    0  722    0  764    0    0 1667
##  [105]    0    0    0    0    0    0    0  287 1482    0
##  [118]    0    0 1772    0    0    0    0 1443    0    0
##  [131]  489    0    0 2449    0    0    0    0  287    0
##  [144]    0 2086    0 2055    0  393    0    0  907    0
##  [157]    0 1358 1833    0    0    0    0    0    0 1155
##  [170]    0    0    0  358    0    0    0 1252    0    0
##  [183]    0    0    0    0 1286    0    0    0    0    0
##  [196]    0    0    0    0    0    0    0    0    0    0
```

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

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

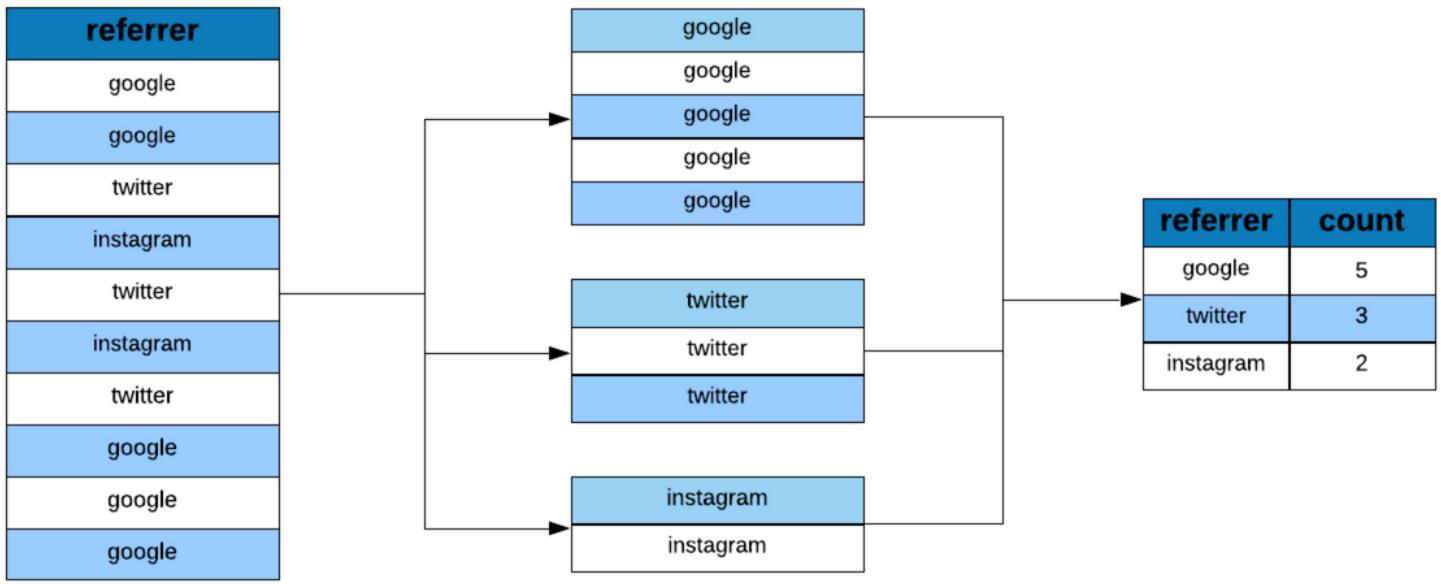
```
## # A tibble: 20 x 11
##       id referrer device bouncers n_visit n_pages durati
##    <int> <chr>    <chr>  <chr>       <int>   <dbl>    <db
##  1     1 google   laptop true           10    1.00     693
##  2     2 yahoo    tablet true            9    1.00     459
##  3     3 direct   laptop true            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
##  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
## 11    11 direct   laptop true            9    1.00     738
## 12    12 direct   tablet false           6   12.0      132
## 13    13 direct   mobile false           9   14.0      406
## 14    14 yahoo    tablet false           5    8.00      80
```

Extract Last Row

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

```
## # A tibble: 1 x 11
##      id referrer device bouncers n_visit n_pages duratio
##   <int> <chr>    <chr>  <chr>       <int>   <dbl>    <dbl
## 1  1000 google   mobile true            9    1.00       26
## # ... with 2 more variables: order_items <dbl>, order_va
```

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

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

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

Case When

```
mtcars %>%
  select(mpg, disp, cyl, gear, carb) %>%
  mutate(
    type = case_when(
      disp > 200 ~ '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 High
## 5   18.7 360.0   8    3    2 High
## 6   18.1 225.0   6    3    1 High
## 7   14.3 360.0   8    3    4 High
## 8   24.4 146.7   4    4    2 True
```

Select First Observation

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

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

Select First Observation

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

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

Select 1000th Observation

```
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