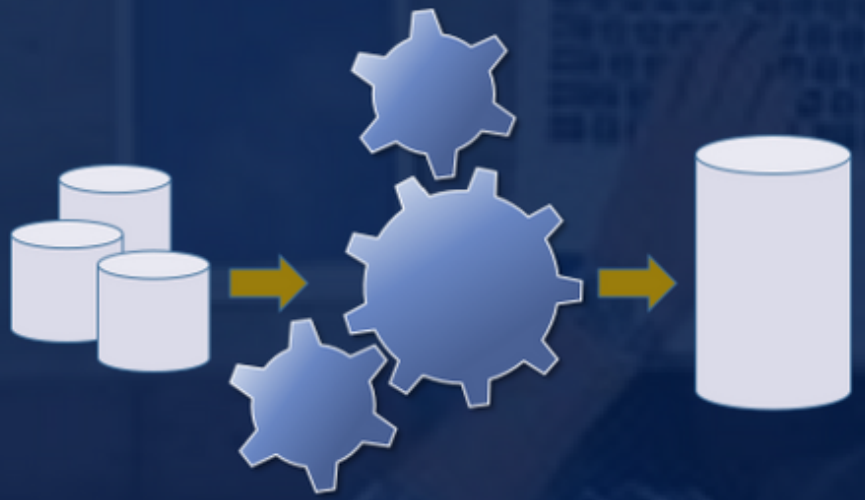
Rsquared Academy



Data Wrangling with dplyr

Agenda

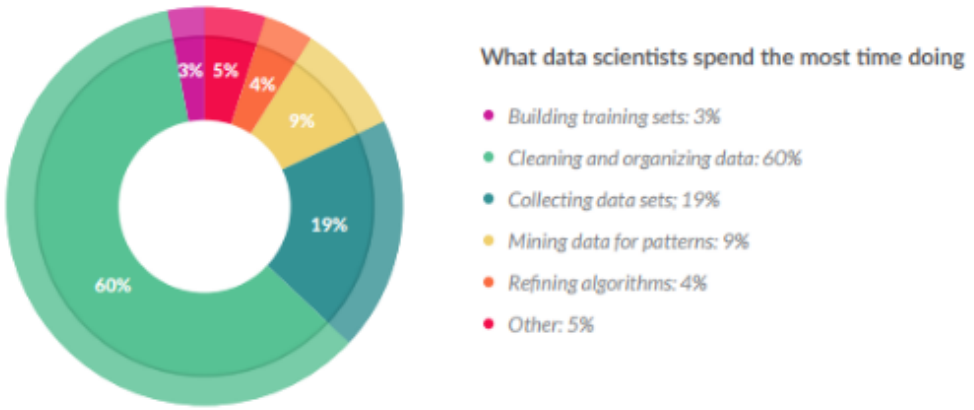
- filter rows
- select variables/columns
- sort/arrange data
- generate new variables
- create grouped summaries

Introduction

According to a survey by CrowdFlower, data scientists spend most of their time cleaning and manipulating data rather than mining or modeling them for insights. As such, it becomes important to have tools that make data manipulation faster and easier. In today's post, we introduce you to dplyr, a grammar of data manipulation.

How a Data Scientist Spends Their Day

Here's where the popular view of data scientists diverges pretty significantly from reality. Generally, we think of data scientists building algorithms, exploring data, and doing predictive analysis. That's actually not what they spend most of their time doing, however.



As you can see from the chart above, 3 out of every 5 data scientists we surveyed actually spend the most time cleaning and organizing data. You may have heard this referred to as "data wrangling" or compared to digital janitor work. Everything from list verification to removing commas to debugging databases–that time adds up and it adds up immensely. Messy data is by far the more time-consuming aspect of the typical data scientist's work flow. And nearly 60% said they simply spent too much time doing it.

Libraries

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

dplyr Verbs

- `select`
- `filter`
- `arrange`
- `mutate`
- `summarise`

CASE STUDY

Data

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

Data Dictionary

- referrer: referrer website/search engine
- device: device used to visit the website
- n_pages: number of pages visited
- duration: time spent on the website (in seconds)
- purchase: whether visitor purchased
- order_value: order value of visitor (in dollars)

- what is the average order value by device types?
- what is the average number of pages visited by purchasers and non-purchasers?
- what is the average time on site for purchasers vs non-purchasers?
- what is the average number of pages visited by purchasers and non-purchasers using mobile?

Average Order Value

Average Order Value

(AOV)

=

Total Revenue

divided by

Number of Orders Taken

AOV by Devices

```
ecom %>%
  filter(purchase) %>%
  select(device, order_value) %>%
  group_by(device) %>%
  summarise_all(list(revenue = ~sum, orders = ~n())) %>%
  mutate(
    aov = revenue / orders
  ) %>%
  select(device, aov)
```

```
## # A tibble: 3 x 2
##   device   aov
##   <fct>  <dbl>
## 1 laptop 1824.
## 2 tablet 1426.
## 3 mobile 1431.
```

FILTER

Filter

device	purchase
mobile	FALSE
tablet	FALSE
laptop	TRUE
laptop	FALSE
mobile	TRUE
laptop	TRUE
tablet	FALSE
mobile	TRUE
laptop	TRUE
laptop	FALSE

Filter data for traffic from mobile

```
filter(data, device == "mobile")
```

device	purchase
mobile	FALSE
mobile	TRUE
mobile	TRUE

Filter all visits from mobile

```
filter(ecom, device == "mobile")
```

```
## # A tibble: 344 x 8
##    referrer device n_visit n_pages duration purchase order_items
##    <fct>    <fct>    <dbl>   <dbl>    <dbl> <lgl>          <dbl>
##  1 yahoo    mobile       9       1      955 FALSE              0
##  2 yahoo    mobile      10       1       75 FALSE              0
##  3 direct   mobile      10       1      908 FALSE              0
##  4 bing     mobile       3      19      209 FALSE              0
##  5 google   mobile       6       1      208 FALSE              0
##  6 direct   mobile       9      14      406 TRUE               3
##  7 yahoo    mobile       7       1       19 FALSE              7
##  8 google   mobile       5       1      147 FALSE              0
##  9 bing     mobile       0       7      196 FALSE              4
## 10 google   mobile      10       1      338 FALSE              0
## # ... with 334 more rows, and 1 more variable: order_value <dbl>
```

Filter

device	purchase
mobile	FALSE
tablet	FALSE
laptop	TRUE
laptop	FALSE
mobile	TRUE
laptop	TRUE
tablet	FALSE
mobile	TRUE
laptop	TRUE
laptop	FALSE

Filter data for traffic from mobile devices which converted

```
filter(data, device == "mobile", purchase == TRUE)
```

device	purchase
mobile	TRUE
mobile	TRUE

Visits from mobile that converted

```
filter(ecom, device == "mobile", purchase)
```

```
## # A tibble: 36 x 8
##    referrer device n_visit n_pages duration purchase order_items
##    <fct>    <fct>    <dbl>   <dbl>    <dbl> <lgl>          <dbl>
##  1 direct   mobile       9      14      406 TRUE               3
##  2 bing     mobile       4      20      440 TRUE               3
##  3 bing     mobile       3      18      288 TRUE               6
##  4 social   mobile      10      11      242 TRUE               4
##  5 yahoo    mobile       6      14      322 TRUE               3
##  6 google   mobile       1      18      252 TRUE               3
##  7 social   mobile       7      16      352 TRUE              10
##  8 direct   mobile       4      18      324 TRUE               3
##  9 social   mobile       1      20      520 TRUE               5
## 10 yahoo    mobile       0      13      351 TRUE              10
## # ... with 26 more rows, and 1 more variable: order_value <dbl>
```

Visits from mobile & visited > 5 pages

```
filter(ecom, device == "mobile", n_pages > 5)
```

```
## # A tibble: 139 x 8
##    referrer device n_visit n_pages duration purchase order_items
##    <fct>    <fct>    <dbl>   <dbl>    <dbl> <lgl>          <dbl>
##  1 bing     mobile       3      19      209 FALSE              0
##  2 direct   mobile       9      14      406 TRUE               3
##  3 bing     mobile       0       7      196 FALSE              4
##  4 yahoo    mobile       8       9      225 FALSE              0
##  5 bing     mobile       4      20      440 TRUE               3
##  6 direct   mobile       1      13      234 FALSE              0
##  7 direct   mobile       4       8      144 FALSE              1
##  8 google   mobile       5       8      192 FALSE              1
##  9 bing     mobile       3      18      288 TRUE               6
## 10 social   mobile      10      11      242 TRUE               4
## # ... with 129 more rows, and 1 more variable: order_value <dbl>
```

```
filter(ecom, purchase)
```

```
## # A tibble: 103 x 8
##    referrer device n_visit n_pages duration purchase order_items
##    <fct>    <fct>    <dbl>   <dbl>    <dbl> <lgl>          <dbl>
##  1 bing     tablet       3      18      468 TRUE               6
##  2 direct   mobile       9      14      406 TRUE               3
##  3 bing     tablet       5      16      368 TRUE               6
##  4 social   tablet       7      10      290 TRUE               9
##  5 direct   tablet       2      19      342 TRUE               5
##  6 social   tablet       9      20      420 TRUE               7
##  7 bing     mobile       4      20      440 TRUE               3
##  8 yahoo    tablet       2      16      480 TRUE               9
##  9 bing     mobile       3      18      288 TRUE               6
## 10 yahoo    tablet       2      14      364 TRUE               6
## # ... with 93 more rows, and 1 more variable: order_value <dbl>
```

SELECT

Select

id	referrer	device	purchase	duration
VF001	google	mobile	FALSE	32
VF002	social	tablet	FALSE	56
VF003	direct	laptop	TRUE	306
VF004	facebook	laptop	FALSE	100
VF005	affiliate	mobile	TRUE	341
VF006	google	laptop	TRUE	432

Select device and purchase columns

```
select(data, device, purchase)
```

device	purchase
mobile	FALSE
tablet	FALSE
laptop	TRUE
laptop	FALSE
mobile	TRUE
laptop	TRUE

Select device and duration column

```
select(ecom, device, duration)
```

```
## # A tibble: 1,000 x 2
##    device duration
##    <fct>     <dbl>
##  1 laptop      693
##  2 tablet      459
##  3 laptop      996
##  4 tablet      468
##  5 mobile      955
##  6 laptop      135
##  7 mobile       75
##  8 mobile      908
##  9 mobile      209
## 10 mobile      208
## # ... with 990 more rows
```

Select

id	referrer	device	purchase	duration
VF001	google	mobile	FALSE	32
VF002	social	tablet	FALSE	56
VF003	direct	laptop	TRUE	306
VF004	facebook	laptop	FALSE	100
VF005	affiliate	mobile	TRUE	341
VF006	google	laptop	TRUE	432

Select all columns from referrer till purchase

```
select(data, referrer:purchase)
```

referrer	device	purchase
google	mobile	FALSE
social	tablet	FALSE
direct	laptop	TRUE
facebook	laptop	FALSE
affiliate	mobile	TRUE
google	laptop	TRUE

Select all columns from referrer to order items

```
select(ecom, referrer:order_items)
```

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

Select

id	referrer	device	purchase	duration
VF001	google	mobile	FALSE	32
VF002	social	tablet	FALSE	56
VF003	direct	laptop	TRUE	306
VF004	facebook	laptop	FALSE	100
VF005	affiliate	mobile	TRUE	341
VF006	google	laptop	TRUE	432

Select all columns except id and duration

```
select(data, -id, -duration)
```

referrer	device	purchase
google	mobile	FALSE
social	tablet	FALSE
direct	laptop	TRUE
facebook	laptop	FALSE
affiliate	mobile	TRUE
google	laptop	TRUE

Select all columns excluding n_pages and duration

```
select(ecom, -n_pages, -duration)
```

```
## # A tibble: 1,000 x 6
##    referrer device n_visit purchase order_items order_value
##    <fct>    <fct>    <dbl> <lgl>          <dbl>       <dbl>
##  1 google   laptop      10 FALSE              0           0
##  2 yahoo    tablet       9 FALSE              0           0
##  3 direct   laptop       0 FALSE              0           0
##  4 bing     tablet       3 TRUE               6         434
##  5 yahoo    mobile       9 FALSE              0           0
##  6 yahoo    laptop       5 FALSE              0           0
##  7 yahoo    mobile      10 FALSE              0           0
##  8 direct   mobile      10 FALSE              0           0
##  9 bing     mobile       3 FALSE              0           0
## 10 google   mobile       6 FALSE              0           0
## # ... with 990 more rows
```

```
select(ecom, device, order_value)
```

```
## # A tibble: 1,000 x 2
##    device order_value
##    <fct>        <dbl>
##  1 laptop           0
##  2 tablet           0
##  3 laptop           0
##  4 tablet         434
##  5 mobile           0
##  6 laptop           0
##  7 mobile           0
##  8 mobile           0
##  9 mobile           0
## 10 mobile           0
## # ... with 990 more rows
```

```
ecom1 <- filter(ecom, purchase)
ecom2 <- select(ecom1, device, order_value)
ecom2
```

```
## # A tibble: 103 x 2
##    device order_value
##    <fct>        <dbl>
##  1 tablet         434

##  2 mobile         651
##  3 tablet        1049
##  4 tablet        1304
##  5 tablet         622
##  6 tablet        1613
##  7 mobile         184
##  8 tablet         286
##  9 mobile         764
## 10 tablet        1667
## # ... with 93 more rows
```

GROUP BY

Group data by referrer type

```
group_by(ecom, referrer)
```

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

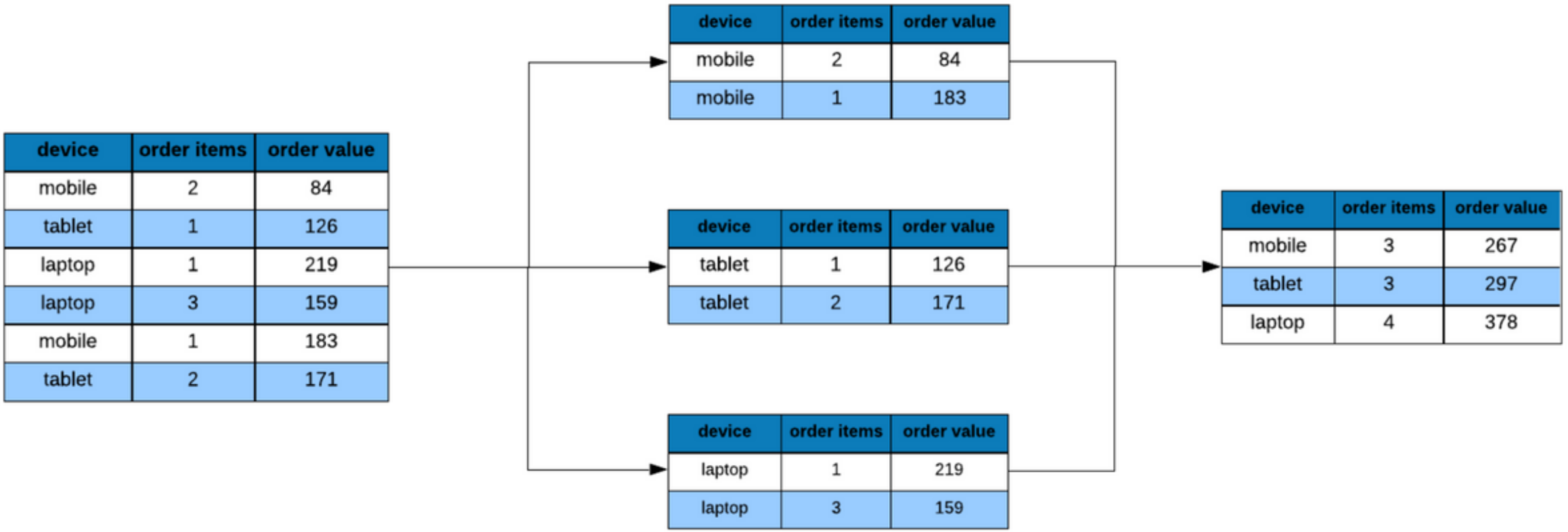
```
ecom3 <- group_by(ecom2, device)
ecom3
```

```
## # A tibble: 103 x 2
## # Groups:   device [3]
##    device order_value
##    <fct>        <dbl>
##  1 tablet         434

##  2 mobile         651
##  3 tablet        1049
##  4 tablet        1304
##  5 tablet         622
##  6 tablet        1613
##  7 mobile         184
##  8 tablet         286
##  9 mobile         764
## 10 tablet        1667
## # ... with 93 more rows
```

SUMMARISE

Summarize


device order items order value
mobile 2 84
tablet 1 126
laptop 1 219
laptop 3 159
mobile 1 183
tablet 2 171
device order items order value
mobile 2 84
mobile 1 183
device order items order value
tablet 1 126
tablet 2 171
device order items order value
laptop 1 219
laptop 3 159
device order items order value
mobile 3 267
tablet 3 297
laptop 4 378

```
ecom4 <- summarise(ecom3, revenue = sum(order_value),
          orders = n())
ecom4
```

```
## # A tibble: 3 x 3
##   device revenue orders
##   <fct>    <dbl>  <int>
## 1 laptop   56531     31

## 2 tablet   51321     36
## 3 mobile   51504     36
```

```
ecom4 <- summarise_all(ecom3, list(revenue = ~sum, orders = ~n()))
ecom4
```

```
## # A tibble: 3 x 3
##   device revenue orders
##   <fct>    <dbl>  <int>
## 1 laptop   56531     31


## 2 tablet   51321     36
## 3 mobile   51504     36
```

MUTATE

```
ecom5 <- mutate(ecom4, aov = revenue / orders)
ecom5
```

```
## # A tibble: 3 x 4
##   device revenue orders   aov
##   <fct>    <dbl>  <int> <dbl>
## 1 laptop   56531     31 1824.


## 2 tablet   51321     36 1426.
## 3 mobile   51504     36 1431.
```

SELECT

Select Relevant Columns

```
ecom6 <- select(ecom5, device, aov)
ecom6
```

```
## # A tibble: 3 x 2
##   device   aov
##   <fct>  <dbl>
## 1 laptop 1824.


## 2 tablet 1426.
## 3 mobile 1431.
```

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

Ascending Order

```
arrange(ecom, n_pages)
```

```
## # A tibble: 1,000 x 8
##    referrer device n_visit n_pages duration purchase order_items
##    <fct>    <fct>    <dbl>   <dbl>    <dbl> <lgl>          <dbl>
##  1 google   laptop      10       1      693 FALSE              0
##  2 yahoo    tablet       9       1      459 FALSE              0
##  3 direct   laptop       0       1      996 FALSE              0
##  4 yahoo    mobile       9       1      955 FALSE              0
##  5 yahoo    mobile      10       1       75 FALSE              0
##  6 direct   mobile      10       1      908 FALSE              0
##  7 google   mobile       6       1      208 FALSE              0
##  8 direct   laptop       9       1      738 FALSE              0
##  9 yahoo    mobile       7       1       19 FALSE              7
## 10 bing     laptop       1       1      995 FALSE              0
## # ... with 990 more rows, and 1 more variable: order_value <dbl>
```

Descending Order

```
arrange(ecom , desc(n_pages))
```

```
## # A tibble: 1,000 x 8
##    referrer device n_visit n_pages duration purchase order_items
##    <fct>    <fct>    <dbl>   <dbl>    <dbl> <lgl>          <dbl>
##  1 social   tablet       9      20      420 TRUE               7
##  2 bing     mobile       4      20      440 TRUE               3
##  3 yahoo    tablet       0      20      200 FALSE              0
##  4 direct   tablet       6      20      580 TRUE               5
##  5 social   mobile       1      20      520 TRUE               5
##  6 google   mobile       8      20      300 TRUE               7
##  7 social   laptop       4      20      200 FALSE              0
##  8 yahoo    mobile       3      20      480 FALSE              0
##  9 social   laptop      10      20      280 TRUE               1
## 10 yahoo    mobile       2      20      240 FALSE              0
## # ... with 990 more rows, and 1 more variable: order_value <dbl>
```

```
arrange(ecom, n_visit, desc(n_pages))
```

```
## # A tibble: 1,000 x 8
##    referrer device n_visit n_pages duration purchase order_items
##    <fct>    <fct>    <dbl>   <dbl>    <dbl> <lgl>          <dbl>
##  1 yahoo    tablet       0      20      200 FALSE              0
##  2 google   laptop       0      19      418 TRUE               2
##  3 bing     laptop       0      18      180 FALSE              0
##  4 yahoo    laptop       0      18      522 TRUE               8
##  5 direct   tablet       0      18      252 FALSE              0
##  6 social   laptop       0      17      204 FALSE              0
##  7 bing     laptop       0      17      272 TRUE               9
##  8 bing     mobile       0      16      272 FALSE              0
##  9 yahoo    mobile       0      15      255 FALSE              0
## 10 direct   laptop       0      15      255 FALSE              0
## # ... with 990 more rows, and 1 more variable: order_value <dbl>
```

```
arrange(ecom6, aov)
```

```
## # A tibble: 3 x 2
##   device   aov
##   <fct>  <dbl>
## 1 tablet 1426.
## 2 mobile 1431.
## 3 laptop 1824.
```

Average Order Value

```
ecom1 <- filter(ecom, purchase)
ecom2 <- select(ecom1, device, order_value)
ecom3 <- group_by(ecom2, device)
ecom4 <- summarise_all(ecom3, funs(revenue = sum, orders = n()))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## please use list() instead
##

## # Before:
## funs(name = f(.)
##
## # After:
## list(name = ~f(.))
## This warning is displayed once per session.
```

```
ecom5 <- mutate(ecom4, aov = revenue / orders)
ecom6 <- select(ecom5, device, aov)
ecom6
```

```
## # A tibble: 3 x 2
##   device   aov
```

AOV by Devices

```
ecom %>%
  filter(purchase) %>%
  select(device, order_value) %>%
  group_by(device) %>%
  summarise_all(list(revenue = ~sum, orders = ~n())) %>%
  mutate(
    aov = revenue / orders
  ) %>%
  select(device, aov)
```

```
## # A tibble: 3 x 2
##   device   aov
##   <fct>  <dbl>
## 1 laptop 1824.
## 2 tablet 1426.
## 3 mobile 1431.
```

Practice Questions

- what is the average number of pages visited by purchasers and non-purchasers?
- what is the average time on site for purchasers vs non-purchasers?
- what is the average number of pages visited by purchasers and non-purchasers using mobile?

Thank You

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