Datafest Workshop 2
Working with Data

Toryn Schafer

02/18/2020

Today's Topics:

- Tidyverse packages
- Long vs. wide data
- Merging multiple data sources
- Cleaning data
- Creating new variables

Tidyverse

- Packages developed at RStudio
  - ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, forcats
- Designed to make data cleaning effecient and readable
- Introduces the Pipe Operater %>%

```
## foo_foo is an instance of a little bunny
foo_foo <- little_bunny()

## adventures in base R must be read from the middle up and backwards
bop_on(
    scoop_up(
        hop_through(foo_foo, forest),
        field_mouse
    ),
    head
)

## adventures w/ pipes start at the top and work down
foo_foo %>%
    hop_through(forest) %>%
    scoop_up(field_mouse) %>%
```

Long vs Wide Data

- Long data best for analysis
- Wide data often used for display purposes
- Transition between them with a key/value pair
  - Key is a grouping variable
  - Value is a measurement

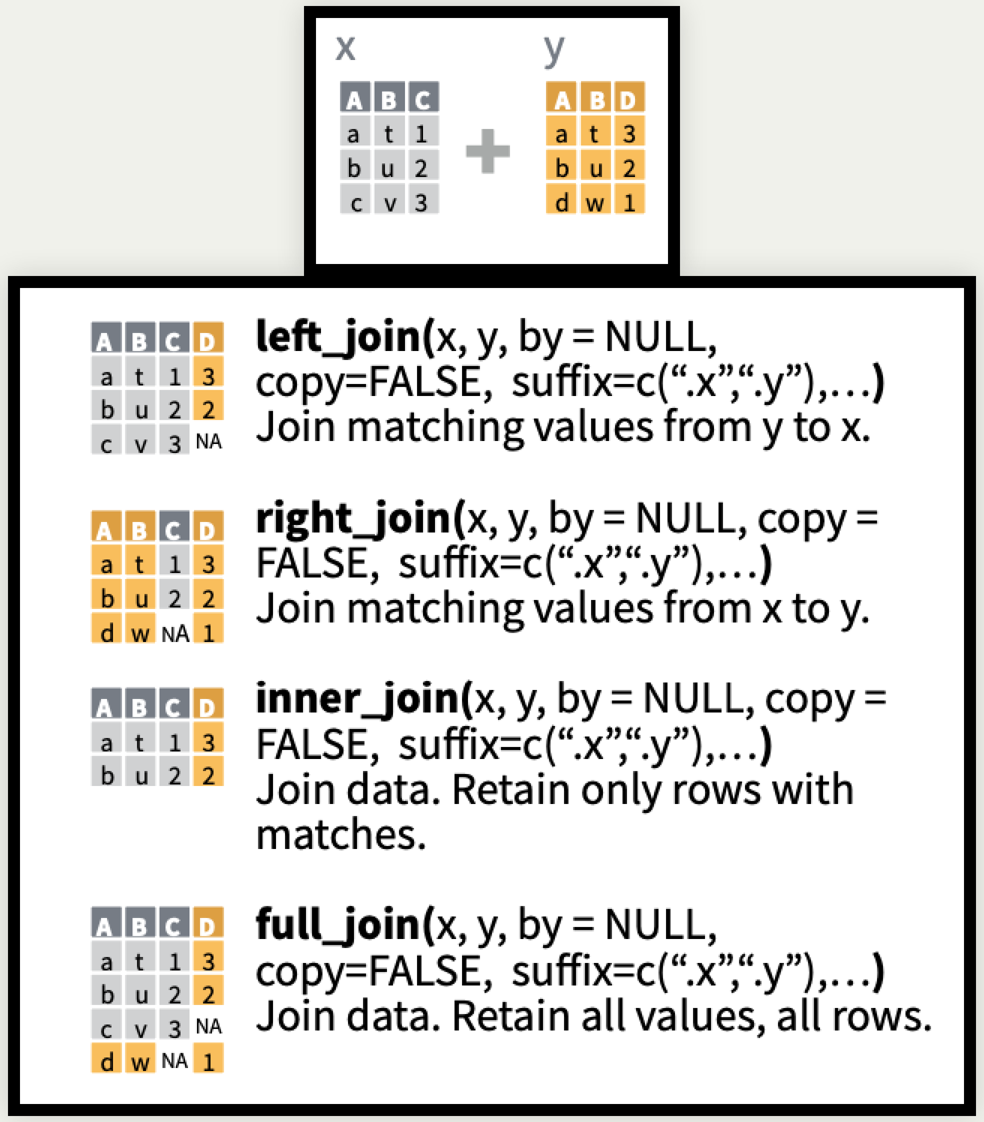
```
## [1] "WIDE"
```

```
##   row a b c
## 1   A 1 4 7
## 2   B 2 5 8
## 3   C 3 6 9
```

```
## [1] "LONG"
```

```
##   row key value
## 1   A   a     1
## 2   B   a     2
## 3   C   a     3
## 4   A   b     4
## 5   B   b     5
## 6   C   b     6
## 7   A   c     7
## 8   B   c     8
## 9   C   c     9
```

Joining Multiple Data Sets



x

A	B	C
a	t	1
b	u	2
c	v	3

+

y

A	B	D
a	t	3
b	u	2
d	w	1

A	B	C	D
a	t	1	3
b	u	2	2
c	v	3	NA

left_join(x, y, by = NULL, copy=FALSE, suffix=c(".x",".y"),...**)**
Join matching values from y to x.

A	B	C	D
a	t	1	3
b	u	2	2
d	w	NA	1

right_join(x, y, by = NULL, copy = FALSE, suffix=c(".x",".y"),...**)**
Join matching values from x to y.

A	B	C	D
a	t	1	3
b	u	2	2

inner_join(x, y, by = NULL, copy = FALSE, suffix=c(".x",".y"),...**)**
Join data. Retain only rows with matches.

A	B	C	D
a	t	1	3
b	u	2	2
c	v	3	NA
d	w	NA	1

full_join(x, y, by = NULL, copy=FALSE, suffix=c(".x",".y"),...**)**
Join data. Retain all values, all rows.

RstudioCheatsheets

Challenge Problems

1. Read in the purchase (approved_data_purchase-v5.csv) and user (approved_ga_data_v2.csv) data sets.
   - Make sure to use read_csv as the files are quite large

2. From the user (ga) data set create a contingency table with the following
    - Grouping variables: clickinfo_slot and device_operatingsystem
    - Only consider the following operating systems:
        - Android
        - iOS
        - Windows
        - Macintosh
    - remove NA category for clickinfo_slot
    - Display the median of totals_timeonsite
        - Hint: You will need to use na.rm = T for median

Full Join the following aggregated datasets by event_id

1. Purchase data set (final dimensions is 27747x3)
   - Remove Parking and future events as in the workshop code
   - Calculate the following summaries by event_id
     - Third quartile of trans_face_val_amt
     - First day of event_dt
       - Hint: check out ?first
2. GA data set (final dimensions is 18592x4)
   - Keep only events happening in the subcontinent ‘Northern America’
   - Summarize the following by event_id and device_devicecategory
     - Count of observations
     - Mean of total_hits

Contingency Table Answer

```
## # A tibble: 2 x 5
##   clickinfo_slot Android   iOS Macintosh Windows
##   <chr>            <dbl> <dbl>     <dbl>   <dbl>
## 1 RHS                253   272       233    310.
## 2 Top                240   218       419    422.
```

Merged data set Answer

```
## # A tibble: 42,532 x 6
##    event_id             face_val_Q3 start_day  device       n tot_hits_
##    <chr>                      <dbl> <date>     <chr>    <int>         <
##  1 0000e75ff4d477a1ea12        35   2015-12-22 <NA>        NA         N
##  2 00016a474558940e2b5e       190   2012-12-06 <NA>        NA         N
##  3 0004a552022180768fb0       200.  2013-07-10 <NA>        NA         N
##  4 000594247e4d6ae97bd9        60   2012-10-22 <NA>        NA         N
##  5 00071bfcbb27802045b2       135   2015-07-04 desktop      9
##  6 00071bfcbb27802045b2       135   2015-07-04 mobile      24
##  7 00071bfcbb27802045b2       135   2015-07-04 tablet       7
##  8 0007822f6e5ce8882118        45   2015-06-26 <NA>        NA         N
##  9 000a141a26dc783c2258        79.5 2015-07-10 desktop     69
## 10 000a141a26dc783c2258        79.5 2015-07-10 mobile     164
## # … with 42,522 more rows
```