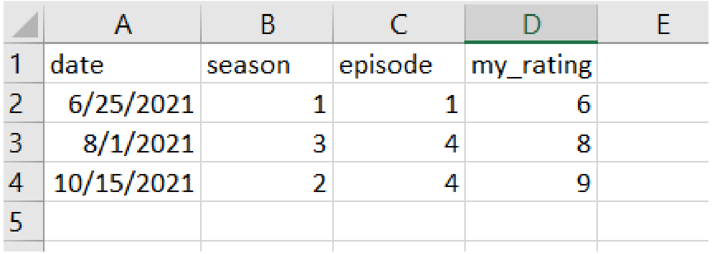
How to do VLOOKUP in R

R-Ladies NYC Lightning Talks

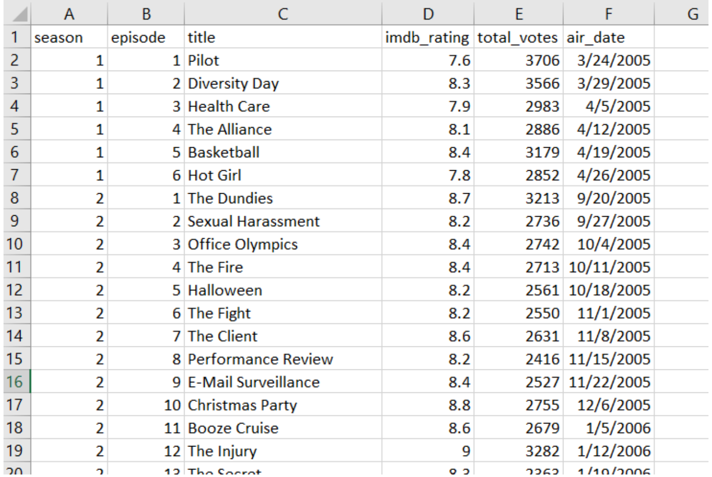
Alejandra Gerosa

October 19th, 2021

What is VLOOKUP?

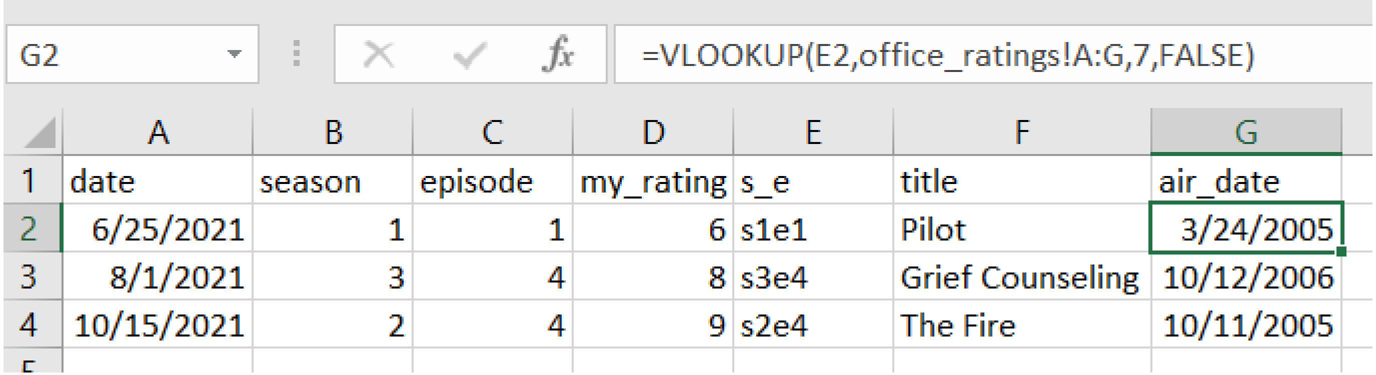


	A	B	C	D	E
1	date	season	episode	my_rating	
2	6/25/2021	1	1	6	
3	8/1/2021	3	4	8	
4	10/15/2021	2	4	9	
5					



	A	B	C	D	E	F	G
1	season	episode	title	imdb_rating	total_votes	air_date	
2	1	1	Pilot	7.6	3706	3/24/2005	
3	1	2	Diversity Day	8.3	3566	3/29/2005	
4	1	3	Health Care	7.9	2983	4/5/2005	
5	1	4	The Alliance	8.1	2886	4/12/2005	
6	1	5	Basketball	8.4	3179	4/19/2005	
7	1	6	Hot Girl	7.8	2852	4/26/2005	
8	2	1	The Dundies	8.7	3213	9/20/2005	
9	2	2	Sexual Harassment	8.2	2736	9/27/2005	
10	2	3	Office Olympics	8.4	2742	10/4/2005	
11	2	4	The Fire	8.4	2713	10/11/2005	
12	2	5	Halloween	8.2	2561	10/18/2005	
13	2	6	The Fight	8.2	2550	11/1/2005	
14	2	7	The Client	8.6	2631	11/8/2005	
15	2	8	Performance Review	8.2	2416	11/15/2005	
16	2	9	E-Mail Surveillance	8.4	2527	11/22/2005	
17	2	10	Christmas Party	8.8	2755	12/6/2005	
18	2	11	Booze Cruise	8.6	2679	1/5/2006	
19	2	12	The Injury	9	3282	1/12/2006	
20	2	13	The Secret	8.3	2363	1/19/2006	

=VLOOKUP([value to look for], [table to look in], [column nr], [approx. match ok?])



=VLOOKUP(E2,office_ratings!A:G,7,FALSE)

	A	B	C	D	E	F	G
1	date	season	episode	my_rating	s_e	title	air_date
2	6/25/2021	1	1	6	s1e1	Pilot	3/24/2005
3	8/1/2021	3	4	8	s3e4	Grief Counseling	10/12/2006
4	10/15/2021	2	4	9	s2e4	The Fire	10/11/2005
5							

Let's do this in R

```r
my_ratings
```

```
## # A tibble: 3 x 4
##   date       season episode my_rating
##   <chr>       <dbl>   <dbl>     <dbl>
## 1 2021-06-25      1       1         6
## 2 2021-08-01      3       4         8
## 3 2021-10-15      2       4         9
```

```r
office_ratings
```

```
## # A tibble: 188 x 6
##    season episode title             imdb_rating total_votes air_date  
##     <dbl>   <dbl> <chr>                   <dbl>       <dbl> <date>    
##  1      1       1 Pilot                     7.6        3706 2005-03-24
##  2      1       2 Diversity Day             8.3        3566 2005-03-29
##  3      1       3 Health Care               7.9        2983 2005-04-05
##  4      1       4 The Alliance              8.1        2886 2005-04-12
##  5      1       5 Basketball                8.4        3179 2005-04-19
##  6      1       6 Hot Girl                  7.8        2852 2005-04-26
##  7      2       1 The Dundies               8.7        3213 2005-09-20
##  8      2       2 Sexual Harassment         8.2        2736 2005-09-27
##  9      2       3 Office Olympics           8.4        2742 2005-10-04
## 10      2       4 The Fire                  8.4        2713 2005-10-11
## # … with 178 more rows
```

There's a package for that!

```r
install.packages("tidyquant")
```

```r
my_ratings_with_s_e <- my_ratings %>%
  mutate(s_e = paste("s", season, "e", episode, sep = ""))

office_ratings_with_s_e <- office_ratings %>%
  mutate(s_e = paste("s", season, "e", episode, sep = ""))
```

```r
my_ratings_with_more_info <- my_ratings_with_s_e %>%
  mutate(
    title = tidyquant::VLOOKUP(s_e, office_ratings_with_s_e, s_e, title),
    air_date = tidyquant::VLOOKUP(s_e, office_ratings_with_s_e, s_e, air_date)
         )
```

```
## # A tibble: 3 x 7
##   date       season episode my_rating s_e   title            air_date  
##   <chr>       <dbl>   <dbl>     <dbl> <chr> <chr>            <date>    
## 1 2021-06-25      1       1         6 s1e1  Pilot            2005-03-24
## 2 2021-08-01      3       4         8 s3e4  Grief Counseling 2006-10-12
## 3 2021-10-15      2       4         9 s2e4  The Fire         2005-10-11
```

But what is VLOOKUP, really?

date	season	episode	my_rating
2021-06-25	1	1	6
2021-08-01	3	4	8
2021-10-15	2	4	9

season	episode	title	air_date
1	1	Pilot	2005-03-24
1	2	Diversity Day	2005-03-29
1	3	Health Care	2005-04-05
1	4	The Alliance	2005-04-12
1	5	Basketball	2005-04-19

This is a join!

VLOOKUP with the tidyverse: dplyr joins

```
left_join(my_ratings, office_ratings)
```

```
## Joining, by = c("season", "episode")
```

```
## # A tibble: 3 x 8
##   date     season episode my_rating title     imdb_rating total_votes air_date
##   <chr>     <dbl>   <dbl>     <dbl> <chr>           <dbl>       <dbl> <date>
## 1 2021-06…      1       1         6 Pilot             7.6        3706 2005-03-24
## 2 2021-08…      3       4         8 Grief Co…         8          2311 2006-10-12
## 3 2021-10…      2       4         9 The Fire          8.4        2713 2005-10-11
```

```
left_join(my_ratings, office_ratings) %>% select(-imdb_rating, -total_votes)
```

```
## Joining, by = c("season", "episode")
```

```
## # A tibble: 3 x 6
##   date       season episode my_rating title            air_date
##   <chr>       <dbl>   <dbl>     <dbl> <chr>            <date>
## 1 2021-06-25      1       1         6 Pilot            2005-03-24
## 2 2021-08-01      3       4         8 Grief Counseling 2006-10-12
## 3 2021-10-15      2       4         9 The Fire         2005-10-11
```

Why use joins?

- You can join by multiple columns
    - In this example: No need to create a "s1e1" column to use as an id
- You can add multiple columns at once
    - In this example: no need for repetitious code in the mutate code
- You can use different join functions to suit the specific needs of your use case

Obstacle: What if the data is messier?

```
schrute_data
```

```
## # A tibble: 55,130 x 12
##   index season episode episode_name director  writer     character text
##   <int>  <int>   <int> <chr>        <chr>     <chr>      <chr>     <chr>
## 1     1      1       1 Pilot        Ken Kwap… Ricky Ger… Michael   All right Ji…
## 2     2      1       1 Pilot        Ken Kwap… Ricky Ger… Jim       Oh, I told y…
## 3     3      1       1 Pilot        Ken Kwap… Ricky Ger… Michael   So you've co…
## 4     4      1       1 Pilot        Ken Kwap… Ricky Ger… Jim       Actually, yo…
## 5     5      1       1 Pilot        Ken Kwap… Ricky Ger… Michael   All right. W…
## # … with 55,125 more rows, and 4 more variables: text_w_direction <chr>,
## #   imdb_rating <dbl>, total_votes <int>, air_date <fct>
```

```
left_join(my_ratings, schrute_data)
```

```
## Joining, by = c("season", "episode")

## # A tibble: 784 x 14
##   date    season episode my_rating index episode_name director writer   character
##   <chr>    <dbl>   <dbl>     <dbl> <int> <chr>        <chr>    <chr>    <chr>
## 1 2021-…       1       1         6     1 Pilot        Ken Kwa… Ricky G… Michael
## 2 2021-…       1       1         6     2 Pilot        Ken Kwa… Ricky G… Jim
## 3 2021-…       1       1         6     3 Pilot        Ken Kwa… Ricky G… Michael
## 4 2021-…       1       1         6     4 Pilot        Ken Kwa… Ricky G… Jim
## 5 2021-…       1       1         6     5 Pilot        Ken Kwa… Ricky G… Michael
## # … with 779 more rows, and 5 more variables: text <chr>,
## #   text_w_direction <chr>, imdb_rating <dbl>, total_votes <int>,
```

Solution: Create the tibble you'll need

```
schrute_data_for_join <- schrute_data %>%
  select(season, episode, title = episode_name, air_date) %>%
  distinct()

left_join(my_ratings, schrute_data_for_join)
```

```
## Joining, by = c("season", "episode")

## # A tibble: 3 x 6
##   date       season episode my_rating title             air_date
##   <chr>       <dbl>   <dbl>     <dbl> <chr>             <fct>
## 1 2021-06-25      1       1         6 Pilot             2005-03-24
## 2 2021-08-01      3       4         8 Grief Counseling 2006-10-12
## 3 2021-10-15      2       4         9 The Fire          2005-10-11
```

Conclusion: How to do VLOOKUP in R

- If you need to replicate VLOOKUP as close to the excel version as possible, you can use the `VLOOKUP()` function from the `tidyquant` package.
- Consider using dplyr joins instead.
- If the dataframe that has the information you want has more columns than you want to add or it has duplicates in your lookup column(s), create the tibble you'll need before doing the join.

Thank you!!

Alejandra Gerosa - @alejagerosa - hello@alegerosa.com