

Why code?

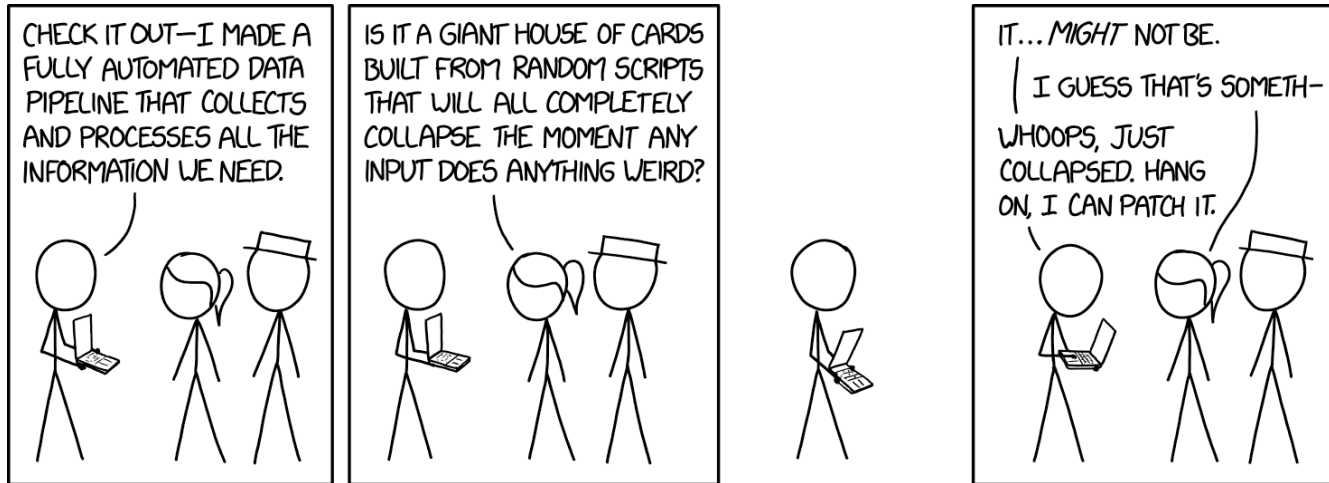
Joseph Ciesielski

12/10/2019

Grounding

1. Data analysis offers the power to critically examine and improve our organizations and advance their missions
2. The way we do data analysis isn't (always) conducive to this
3. Using code can help

A story



Four tables

Ran 30+ reports and organized into four tables

1. Characteristics
2. Activity
3. Test scores
4. Outcomes

Allows me to do any analysis I need and not rely on specific reports

Characteristics

- ID
- name
- race
- gender
- program
- program start date

Activity

```
## # A tibble: 1,500 x 3
##       id activity      date
##   <int> <chr>      <date>
## 1     54 employment 2017-02-18
## 2     40 case management 2017-02-02
## 3     32 education 2018-06-05
## 4     62 education 2016-12-08
## 5     31 education 2018-08-11
## 6    100 education 2015-12-18
## 7     33 employment 2016-07-04
## 8     13 education 2017-01-17
## 9     54 case management 2015-04-29
## 10    10 employment 2017-11-25
## # . with 1,490 more rows
```

Test scores

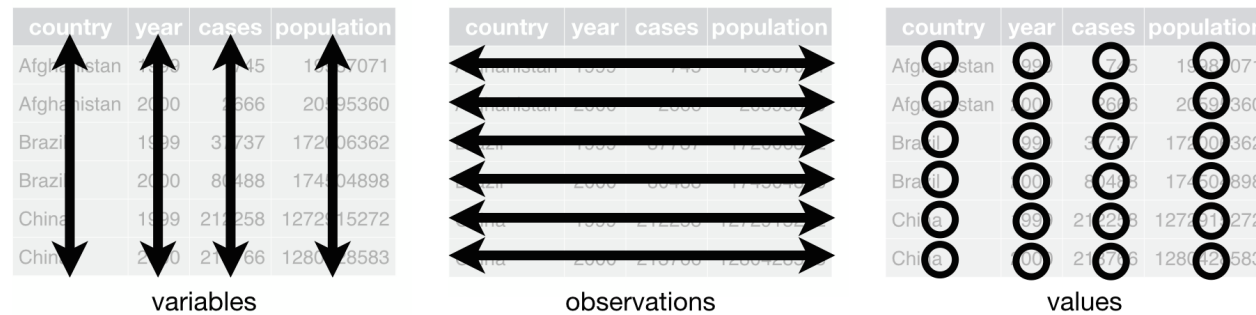
```
## # A tibble: 200 x 4
##       id test  test_date  score
##   <int> <chr> <date>    <dbl>
##  1     23 GED   2017-11-14  16.6
##  2     21 TABE  2019-09-05  70.1
##  3     32 GED   2016-08-30  84.9
##  4     11 GED   2015-10-29  87.3
##  5     50 GED   2016-02-25  57.6
##  6     61 TABE  2017-10-24  50.0
##  7     37 TABE  2016-01-18  90.0
##  8     69 GED   2017-02-18  39.2
##  9      2 GED   2017-10-25  23.0
## 10    71 TABE  2017-11-22  81.5
## # . with 190 more rows
```

Outcomes

```
## # A tibble: 50 x 3
##       id outcome_date outcome
##   <int> <date>      <chr>
## 1     3 2018-02-13  education
## 2    38 2015-10-04  employment
## 3    58 2018-07-28  employment
## 4     7 2017-10-08  education
## 5    21 2018-08-26  employment
## 6    35 2017-11-17  employment
## 7    86 2017-09-27  education
## 8    86 2016-07-31  employment
## 9    60 2019-02-08  employment
## 10   60 2018-10-07  employment
## # . with 40 more rows
```


Tidy data

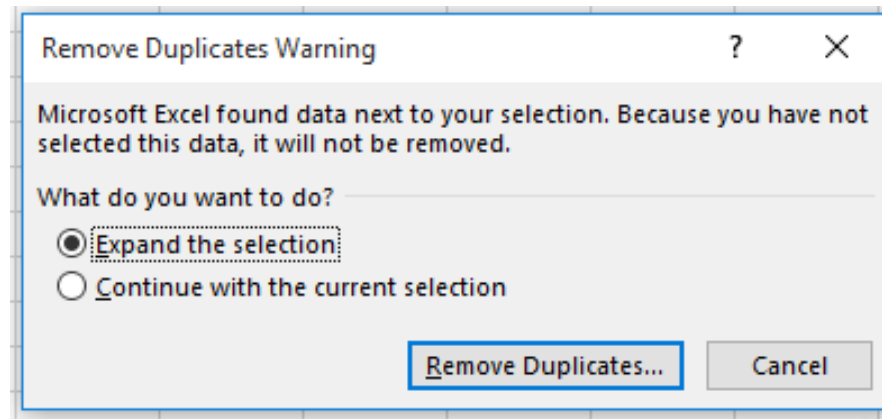
<https://vita.had.co.nz/papers/tidy-data.pdf>



- Recreated relational databases that worked for me
- If you can't use code, can still get data in this format
 - load into Tableau, Power BI, etc. for reproducible analysis

Why is code helpful?

Provenance



<https://www.youtube.com/watch?v=cpbtcsGE0OA>

Reproducible

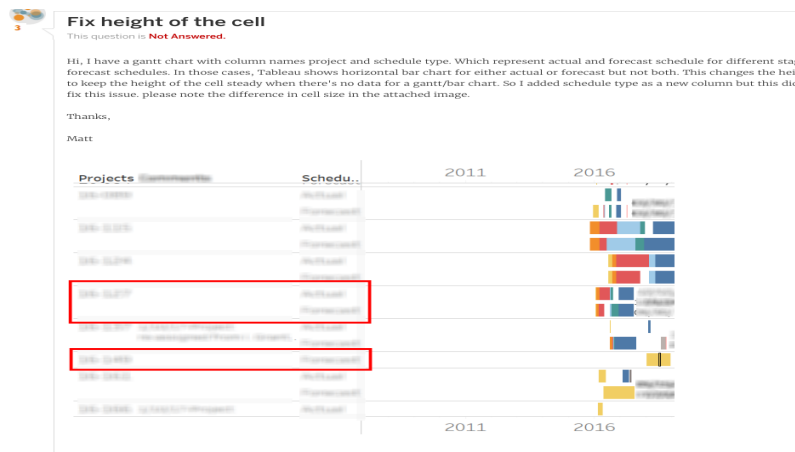


- Shareable
- Modifiable
- Updatable

Plain Text

Copy and paste

non-pasteable



pasteable

By using the `merge` function and its optional parameters:

Inner join: `merge(df1, df2)` will work for these examples because R automatically joins the frames by common variable names, but you would most likely want to specify `merge(df1, df2, by = "CustomerId")` to make sure that you were matching on only the fields you desired. You can also use the `by.x` and `by.y` parameters if the matching variables have different names in the different data frames.

Outer join: `merge(x = df1, y = df2, by = "CustomerId", all = TRUE)`

Left outer: `merge(x = df1, y = df2, by = "CustomerId", all.x = TRUE)`

Right outer: `merge(x = df1, y = df2, by = "CustomerId", all.y = TRUE)`

Cross join: `merge(x = df1, y = df2, by = NULL)`

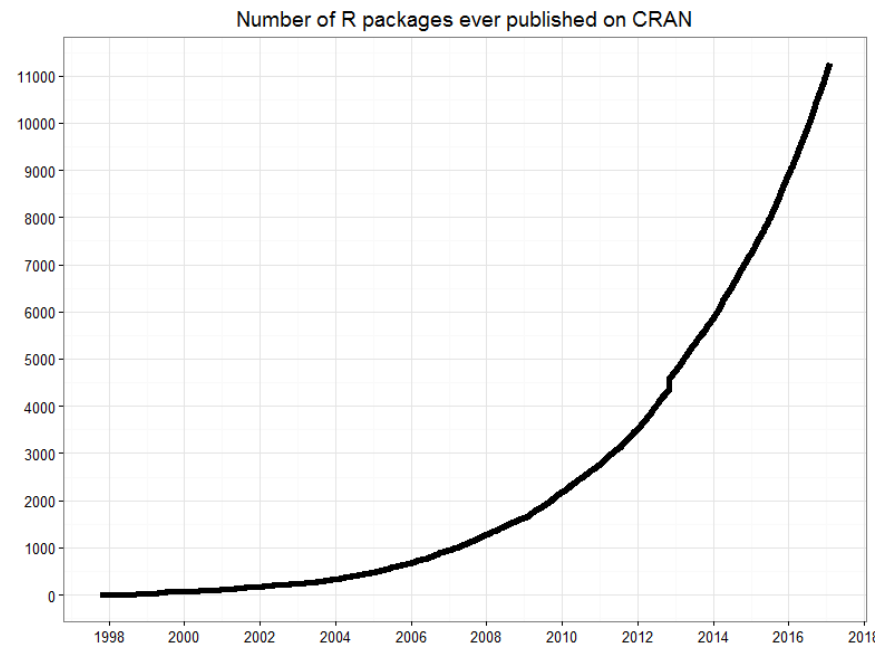
Just as with the inner join, you would probably want to explicitly pass "CustomerId" to R as the matching variable. I think it's almost always best to explicitly state the identifiers on which you want to merge; it's safer if the input data frames change unexpectedly and easier to read later on.

You can merge on multiple columns by giving `by` a vector, e.g., `by = c("CustomerId", "OrderId")`.

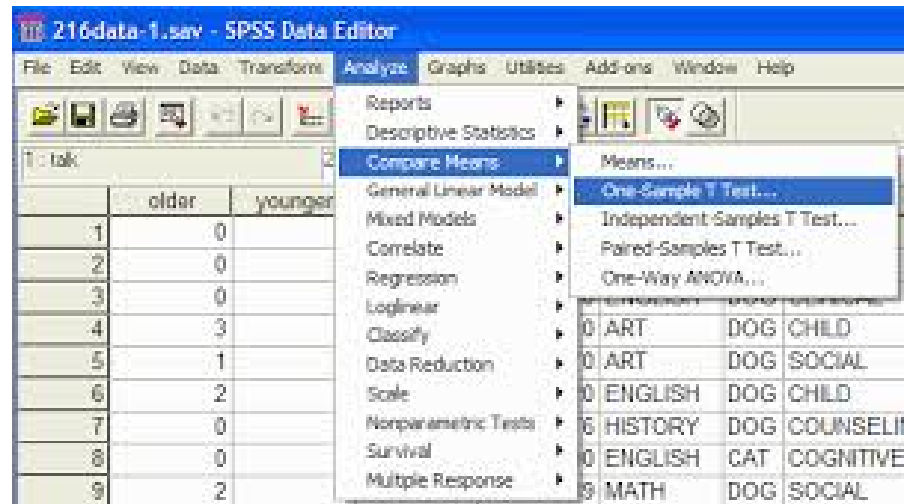
If the column names to merge on are not the same, you can specify, e.g., `by.x = "CustomerId_in_df1", by.y = "CustomerId_in_df2"` where `CustomerId_in_df1` is the name of the column in the first data frame and `CustomerId_in_df2` is the name of the column in the second data frame. (These can also be vectors if you need to merge on multiple columns.)

Google / www.stackoverflow.com

Open Source



Powerful



So what?

Go beyond counting

Ask different kinds of questions

1. What did ... ?
2. Why did ... ?
3. What will ... ?

Look at relationships

```
library(tidyverse)

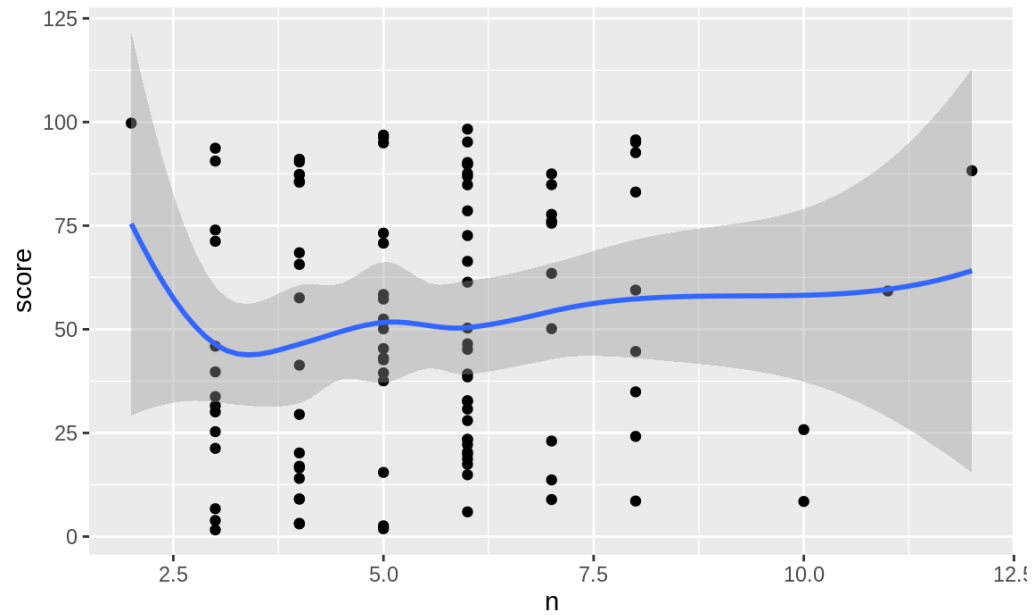
test_cm <- act %>%
  filter(activity == "case management") %>%
  count(id) %>%
  inner_join(test, by = "id") %>%
  filter(test == "GED")
```

```
test_cm
```

```
## # A tibble: 105 x 5
##       id     n test  test_date  score
##   <int> <int> <chr> <date>    <dbl>
## 1     1    11 GED   2019-01-30 59.2
## 2     2     7 GED   2017-10-25 23.0
## 3     2     7 GED   2017-01-15  8.93
## 4     3     6 GED   2017-11-28 50.3
## 5     4     5 GED   2015-09-24 15.5
## 6     5     6 GED   2015-09-19 66.4
## 7     5     6 GED   2018-10-21 14.9
## 8     6     4 GED   2017-11-21 90.4
## 9     6     4 GED   2015-03-02 91.0
```

Look at relationships

```
test_cm %>%  
  ggplot(aes(n, score)) +  
    geom_point() +  
    geom_smooth()
```



Regression

```
cm_model <- lm(score ~ n, data = test_cm)
```

```
summary(cm_model)
```

```
##
## Call:
## lm(formula = score ~ n, data = test_cm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -49.107 -28.894  -4.095   27.748   55.153
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.332      9.594   4.308 3.77e-05 ***
## n              1.624      1.663   0.977  0.331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.94 on 103 degrees of freedom
## Multiple R-squared:  0.00918,    Adjusted R-squared:  -0.0004401
## F-statistic: 0.9543 on 1 and 103 DF,  p-value: 0.3309
```

Prediction

```
library(randomForest)
```

```
randomForest(outcome ~ ., data = outcomes)
```

Where to start with coding?

Languages

- SQL and (R or Python)

Resources

- [R for Data Science](#) and [learning community](#)
- Google
- Twitter
- SQL is easiest to learn at work

Focus on questions that matter

- Incentives are not designed to critically examine our work
- Using code and good data organization principles allows us to focus energy on asking deeper questions and investigating relationships

Contact Info

Joe Ciesielski

jtcies@gmail.com