Data Wrangling with dplyr and tidyr

...

By Shiloh Bradley

Outline

What is Data Wrangling?

Analysis Techniques with dplyr() and tidyr() functions

Issues and Limitations

What is Data Wrangling?

Form of data preparation and preprocessing

Numerous functions in R

Used to prepare data for further analytics

Majority of the time is spent on this rather than analytics

Analysis Technique

Diabetes data set

dplyr techniques - Subset Observations

- distinct()
- sample_frac()
- group_by()

tidyr techniques

Results

```
> summary(d)
  Pregnancies      PG.Concentration  Diastolic.BP     Tri.Fold.Thick    Serum.Ins
 Min.   : 0.000   Min.   :  0.0    Min.   :  0.00   Min.   : 0.00   Min.   :  0.0
 1st Qu.: 1.000   1st Qu.: 99.0    1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0
 Median : 3.000   Median :117.0    Median : 72.00   Median :23.00   Median : 30.5
 Mean   : 3.845   Mean   :120.9    Mean   : 69.11   Mean   :20.54   Mean   : 79.8
 3rd Qu.: 6.000   3rd Qu.:140.2    3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2
 Max.   :17.000   Max.   :199.0    Max.   :122.00   Max.   :99.00   Max.   :846.0
 NA's   :3        NA's   :3        NA's   :3        NA's   :3       NA's   :3
      BMI         DP.Function          Age             Diabetes
 Min.   : 0.00   Min.   :0.0780   Min.   :21.00          :  3
 1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   Healthy:500
 Median :32.00   Median :0.3725   Median :29.00   Sick   :268
 Mean   :31.99   Mean   :0.4719   Mean   :33.24
 3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :67.10   Max.   :2.4200   Max.   :81.00
 NA's   :3       NA's   :3        NA's   :3
```

Statistical summary (character) of the data set before any tests

Results - dplyr

distinct()

Remove duplicate rows

Did not change the character of the data

sample_frac()

Randomly select a fraction of the rows

group_by()

Results

```
> summary(sample_frac(d, 0.5, replace = TRUE))
  Pregnancies      PG.Concentration  Diastolic.BP     Tri.Fold.Thick    Serum.Ins          BMI          DP.Function          Age             Diabetes
 Min.   : 0.000   Min.   :  0.0    Min.   :  0.00   Min.   : 0.0   Min.   :  0.00   Min.   : 0.0   Min.   :0.0850   Min.   :21.00   Healthy:257
 1st Qu.: 1.000   1st Qu.: 97.0    1st Qu.: 64.00   1st Qu.: 0.0   1st Qu.:  0.00   1st Qu.:26.5   1st Qu.:0.2350   1st Qu.:24.00   Sick   :128
 Median : 3.000   Median :114.0    Median : 70.00   Median :22.0   Median : 23.00   Median :31.2   Median :0.3830   Median :29.00
 Mean   : 3.948   Mean   :117.5    Mean   : 68.29   Mean   :19.7   Mean   : 77.92   Mean   :31.4   Mean   :0.4415   Mean   :33.51
 3rd Qu.: 6.000   3rd Qu.:135.0    3rd Qu.: 80.00   3rd Qu.:32.0   3rd Qu.:126.00   3rd Qu.:36.5   3rd Qu.:0.6010   3rd Qu.:41.00
 Max.   :14.000   Max.   :199.0    Max.   :110.00   Max.   :99.0   Max.   :543.00   Max.   :57.3   Max.   :1.6980   Max.   :72.00
 NA's   :1        NA's   :1        NA's   :1        NA's   :1      NA's   :1        NA's   :1      NA's   :1        NA's   :1
>
> summary(sample_frac(d, 0.5, replace = TRUE))
  Pregnancies      PG.Concentration  Diastolic.BP     Tri.Fold.Thick    Serum.Ins          BMI          DP.Function          Age             Diabetes
 Min.   : 0.000   Min.   :  0.0    Min.   :  0.00   Min.   : 0.00   Min.   :  0.00   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   Healthy:231
 1st Qu.: 1.000   1st Qu.: 97.0    1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.:26.40   1st Qu.:0.2370   1st Qu.:24.00   Sick   :153
 Median : 3.000   Median :115.5    Median : 70.00   Median :22.00   Median :  0.00   Median :31.20   Median :0.3450   Median :29.00
 Mean   : 3.826   Mean   :120.2    Mean   : 68.06   Mean   :19.59   Mean   : 79.17   Mean   :31.54   Mean   :0.4709   Mean   :33.49
 3rd Qu.: 6.000   3rd Qu.:141.0    3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:125.25   3rd Qu.:36.30   3rd Qu.:0.6400   3rd Qu.:41.00
 Max.   :15.000   Max.   :199.0    Max.   :122.00   Max.   :52.00   Max.   :600.00   Max.   :57.30   Max.   :2.1370   Max.   :81.00
 NA's   :2        NA's   :2        NA's   :2        NA's   :2       NA's   :2        NA's   :2       NA's   :2        NA's   :2
>
> summary(sample_frac(d, 0.5, replace = TRUE))
  Pregnancies      PG.Concentration  Diastolic.BP     Tri.Fold.Thick    Serum.Ins          BMI          DP.Function          Age             Diabetes
 Min.   : 0.00    Min.   :  0.0    Min.   :  0.00   Min.   : 0.00   Min.   :  0.00   Min.   : 0.00   Min.   :0.0840   Min.   :21.00   Healthy:262
 1st Qu.: 1.00    1st Qu.:100.0    1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.:26.20   1st Qu.:0.2510   1st Qu.:24.00   Sick   :123
 Median : 3.00    Median :116.0    Median : 72.00   Median :22.00   Median : 45.00   Median :31.60   Median :0.3650   Median :29.00
 Mean   : 3.73    Mean   :119.5    Mean   : 67.59   Mean   :20.34   Mean   : 85.07   Mean   :31.55   Mean   :0.4891   Mean   :32.92
 3rd Qu.: 6.00    3rd Qu.:137.0    3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:135.00   3rd Qu.:36.10   3rd Qu.:0.6470   3rd Qu.:39.00
 Max.   :15.00    Max.   :198.0    Max.   :110.00   Max.   :60.00   Max.   :846.00   Max.   :53.20   Max.   :2.1370   Max.   :69.00
 NA's   :1        NA's   :1        NA's   :1        NA's   :1       NA's   :1        NA's   :1       NA's   :1        NA's   :1
>
```

sample_frac summary statistics

Results - tidyr

gather()

Collapses information down to two columns

Technique does not apply

spread()

Spread rows into columns

Technique does not apply

Issues and Limitations

Cost analysis

About 50-80% of the time could be spent on Data Wrangling

Software/ technology discrepancies

Review

What is Data Wrangling?

Analysis Techniques with dplyr() and tidyr() functions

Issues and Limitations

References

Boehmke, B. (2014). *Data Processing with dplyr & tidyr.* Retrieved from RPub.com.

Dasu, T., Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning.* Wiley.

Data Wrangling with dplyr and tidyr. Retrieved from RStudio.com

Grolemund, G. (2015). *Data Wrangling with R: How to work with the structures of your data.* Retrieved from RStudio.com.

Kahn, M. (1994). Diabetes data set. Retrieved from University of California, Irvine - Machine Learning Repository.

Vaidyanathan, R. *Data Wrangling I.* Retrieved from ramnathv.github.io.

Questions?