

Dataframes in R: Takeaways

by Dataquest Labs, Inc. - All rights reserved © 2020

Concepts

- Tabular data is organized into rows, where one row represents a single entity and columns represent different characteristics of this row.
- Microsoft Excel, Google Sheets, and CSV files are common ways that we see tabular data.
- Tibbles are a data structure that implements tabular data in R and the `tidyverse`.
- Piping enables us to create pipelines with all of the functions we learned, allowing us to convert raw data in tibbles to more refined datasets.

Syntax

- Import a dataset:

```
library(readr)
data <- read_csv("name_of_file_with_data.csv")
```

- Learn about a tibble's columns, types and dimensions:

```
> glimpse(recent_grads)
#> # A tibble: 173 x 18
#>   Rank Major_code Major Major_code Major
#>   <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1     1 2419 2419 2419 2419
#> 2     2 2416 2416 2416 2416
#> 3     3 2416 2416 2416 2416
#> 4     4 2416 2416 2416 2416
#> 5     5 2416 2416 2416 2416
#> 6     6 2416 2416 2416 2416
#> 7     7 2416 2416 2416 2416
#> 8     8 2416 2416 2416 2416
#> 9     9 2416 2416 2416 2416
#> 10    10 2416 2416 2416 2416
#> 11    11 2416 2416 2416 2416
#> 12    12 2416 2416 2416 2416
#> 13    13 2416 2416 2416 2416
#> 14    14 2416 2416 2416 2416
#> 15    15 2416 2416 2416 2416
#> 16    16 2416 2416 2416 2416
#> 17    17 2416 2416 2416 2416
#> 18    18 2416 2416 2416 2416
#> 19    19 2416 2416 2416 2416
#> 20    20 2416 2416 2416 2416
#> 21    21 2416 2416 2416 2416
#> 22    22 2416 2416 2416 2416
#> 23    23 2416 2416 2416 2416
#> 24    24 2416 2416 2416 2416
#> 25    25 2416 2416 2416 2416
#> 26    26 2416 2416 2416 2416
#> 27    27 2416 2416 2416 2416
#> 28    28 2416 2416 2416 2416
#> 29    29 2416 2416 2416 2416
#> 30    30 2416 2416 2416 2416
#> 31    31 2416 2416 2416 2416
#> 32    32 2416 2416 2416 2416
#> 33    33 2416 2416 2416 2416
#> 34    34 2416 2416 2416 2416
#> 35    35 2416 2416 2416 2416
#> 36    36 2416 2416 2416 2416
#> 37    37 2416 2416 2416 2416
#> 38    38 2416 2416 2416 2416
#> 39    39 2416 2416 2416 2416
#> 40    40 2416 2416 2416 2416
#> 41    41 2416 2416 2416 2416
#> 42    42 2416 2416 2416 2416
#> 43    43 2416 2416 2416 2416
#> 44    44 2416 2416 2416 2416
#> 45    45 2416 2416 2416 2416
#> 46    46 2416 2416 2416 2416
#> 47    47 2416 2416 2416 2416
#> 48    48 2416 2416 2416 2416
#> 49    49 2416 2416 2416 2416
#> 50    50 2416 2416 2416 2416
#> 51    51 2416 2416 2416 2416
#> 52    52 2416 2416 2416 2416
#> 53    53 2416 2416 2416 2416
#> 54    54 2416 2416 2416 2416
#> 55    55 2416 2416 2416 2416
#> 56    56 2416 2416 2416 2416
#> 57    57 2416 2416 2416 2416
#> 58    58 2416 2416 2416 2416
#> 59    59 2416 2416 2416 2416
#> 60    60 2416 2416 2416 2416
#> 61    61 2416 2416 2416 2416
#> 62    62 2416 2416 2416 2416
#> 63    63 2416 2416 2416 2416
#> 64    64 2416 2416 2416 2416
#> 65    65 2416 2416 2416 2416
#> 66    66 2416 2416 2416 2416
#> 67    67 2416 2416 2416 2416
#> 68    68 2416 2416 2416 2416
#> 69    69 2416 2416 2416 2416
#> 70    70 2416 2416 2416 2416
#> 71    71 2416 2416 2416 2416
#> 72    72 2416 2416 2416 2416
#> 73    73 2416 2416 2416 2416
#> 74    74 2416 2416 2416 2416
#> 75    75 2416 2416 2416 2416
#> 76    76 2416 2416 2416 2416
#> 77    77 2416 2416 2416 2416
#> 78    78 2416 2416 2416 2416
#> 79    79 2416 2416 2416 2416
#> 80    80 2416 2416 2416 2416
#> 81    81 2416 2416 2416 2416
#> 82    82 2416 2416 2416 2416
#> 83    83 2416 2416 2416 2416
#> 84    84 2416 2416 2416 2416
#> 85    85 2416 2416 2416 2416
#> 86    86 2416 2416 2416 2416
#> 87    87 2416 2416 2416 2416
#> 88    88 2416 2416 2416 2416
#> 89    89 2416 2416 2416 2416
#> 90    90 2416 2416 2416 2416
#> 91    91 2416 2416 2416 2416
#> 92    92 2416 2416 2416 2416
#> 93    93 2416 2416 2416 2416
#> 94    94 2416 2416 2416 2416
#> 95    95 2416 2416 2416 2416
#> 96    96 2416 2416 2416 2416
#> 97    97 2416 2416 2416 2416
#> 98    98 2416 2416 2416 2416
#> 99    99 2416 2416 2416 2416
#> 100   100 2416 2416 2416 2416
#> 101   101 2416 2416 2416 2416
#> 102   102 2416 2416 2416 2416
#> 103   103 2416 2416 2416 2416
#> 104   104 2416 2416 2416 2416
#> 105   105 2416 2416 2416 2416
#> 106   106 2416 2416 2416 2416
#> 107   107 2416 2416 2416 2416
#> 108   108 2416 2416 2416 2416
#> 109   109 2416 2416 2416 2416
#> 110   110 2416 2416 2416 2416
#> 111   111 2416 2416 2416 2416
#> 112   112 2416 2416 2416 2416
#> 113   113 2416 2416 2416 2416
#> 114   114 2416 2416 2416 2416
#> 115   115 2416 2416 2416 2416
#> 116   116 2416 2416 2416 2416
#> 117   117 2416 2416 2416 2416
#> 118   118 2416 2416 2416 2416
#> 119   119 2416 2416 2416 2416
#> 120   120 2416 2416 2416 2416
#> 121   121 2416 2416 2416 2416
#> 122   122 2416 2416 2416 2416
#> 123   123 2416 2416 2416 2416
#> 124   124 2416 2416 2416 2416
#> 125   125 2416 2416 2416 2416
#> 126   126 2416 2416 2416 2416
#> 127   127 2416 2416 2416 2416
#> 128   128 2416 2416 2416 2416
#> 129   129 2416 2416 2416 2416
#> 130   130 2416 2416 2416 2416
#> 131   131 2416 2416 2416 2416
#> 132   132 2416 2416 2416 2416
#> 133   133 2416 2416 2416 2416
#> 134   134 2416 2416 2416 2416
#> 135   135 2416 2416 2416 2416
#> 136   136 2416 2416 2416 2416
#> 137   137 2416 2416 2416 2416
#> 138   138 2416 2416 2416 2416
#> 139   139 2416 2416 2416 2416
#> 140   140 2416 2416 2416 2416
#> 141   141 2416 2416 2416 2416
#> 142   142 2416 2416 2416 2416
#> 143   143 2416 2416 2416 2416
#> 144   144 2416 2416 2416 2416
#> 145   145 2416 2416 2416 2416
#> 146   146 2416 2416 2416 2416
#> 147   147 2416 2416 2416 2416
#> 148   148 2416 2416 2416 2416
#> 149   149 2416 2416 2416 2416
#> 150   150 2416 2416 2416 2416
#> 151   151 2416 2416 2416 2416
#> 152   152 2416 2416 2416 2416
#> 153   153 2416 2416 2416 2416
#> 154   154 2416 2416 2416 2416
#> 155   155 2416 2416 2416 2416
#> 156   156 2416 2416 2416 2416
#> 157   157 2416 2416 2416 2416
#> 158   158 2416 2416 2416 2416
#> 159   159 2416 2416 2416 2416
#> 160   160 2416 2416 2416 2416
#> 161   161 2416 2416 2416 2416
#> 162   162 2416 2416 2416 2416
#> 163   163 2416 2416 2416 2416
#> 164   164 2416 2416 2416 2416
#> 165   165 2416 2416 2416 2416
#> 166   166 2416 2416 2416 2416
#> 167   167 2416 2416 2416 2416
#> 168   168 2416 2416 2416 2416
#> 169   169 2416 2416 2416 2416
#> 170   170 2416 2416 2416 2416
#> 171   171 2416 2416 2416 2416
#> 172   172 2416 2416 2416 2416
#> 173   173 2416 2416 2416 2416
```

- Return the number of rows or columns from a tibble:

```
nrow(data) # returns the number of rows in `data`  
ncol(data) # returns the number of rows in `data`
```

- Pick columns to keep or remove from your data:

```
# Keeping data  
filtered_data <- select(recent_grads, Rank, Major)  
# Removing data  
filtered_data <- select(recent_grads, - College_jobs)
```

- Filter rows based on conditions:

```
top_100_majors <- filter(recent_grads, Rank < 100)
```

- Chain together tidyverse functions into a pipeline:

```
library(dplyr)  
low_total_ranked_majors <- recent_grads %>%  
  select(., Rank, Major, Total) %>%  
  filter(., ranked_majors, Total < 2000)
```

- Create new columns:

```
new_recent_grads <- recent_grads %>%  
  mutate(  
    prop_male = Men / Total  
  )
```

- Sort data by a particular or multiple columns:

```
new_recent_grads <- recent_grads %>%  
  mutate(  
    prop_male = Men / Total  
  ) %>%  
  arrange( -prop_male)
```

- Use `head()` to return just the first few rows of a tibble

```
> head(new_recent_grads)  
# A tibble: 6 x 3  
  Total    Men prop_male  
1   124   124         1  
2  4790  4419       0.923  
3 18498 16820       0.909  
4   756   679       0.898  
5  1258  1123       0.893  
6 91227 80320       0.880
```

- Use `summarize()` to calculate some summary values based on entire columns:

```
summary_table <- recent_grads %>%  
  summarize(  
    avg_unemp = mean(Unemployment_rate),  
    min_unemp = min(Unemployment_rate),  
    max_unemp = max(Unemployment_rate)  
  )
```



Takeaways by Dataquest Labs, Inc. - All rights reserved © 2020