Lecture 3 Notes

2024-01-18

Lecture 3: Working with `tidyverse()`

- This is bullet 1
- This is bullet 2

```
2+2
```

```
## [1] 4
```

```
mean(c(1,50,120,-3))
```

```
## [1] 42
```

Downloading data with `tidyverse()`

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

```
df <- read_rds("https://github.com/jbisbee1/DS1000_S2024/raw/main/data/sc_debt.Rds")

df
```

```
## # A tibble: 2,546 × 16
##    unitid instnm   stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##     <int> <chr>    <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
##  1 100654 Alabama… AL             33375 Public  South… Bachel…        2    0.918
##  2 100663 Univers… AL             22500 Public  South… Bachel…        2    0.737
##  3 100690 Amridge… AL             27334 Private South… Associ…        1   NA    
##  4 100706 Univers… AL             21607 Public  South… Bachel…        2    0.826
##  5 100724 Alabama… AL             32000 Public  South… Bachel…        2    0.969
##  6 100751 The Uni… AL             23250 Public  South… Bachel…        2    0.827
##  7 100760 Central… AL             12500 Public  South… Associ…        1   NA    
##  8 100812 Athens … AL             19500 Public  South… Bachel…       NA   NA    
##  9 100830 Auburn … AL             24826 Public  South… Bachel…        2    0.904
## 10 100858 Auburn … AL             21281 Public  South… Bachel…        2    0.807
## # ℹ 2,536 more rows
## # ℹ 7 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>
```

```
df %>%
  head()
```

```
## # A tibble: 6 × 16
##   unitid instnm    stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##    <int> <chr>     <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
## 1 100654 Alabama … AL             33375 Public  South… Bachel…        2    0.918
## 2 100663 Universi… AL             22500 Public  South… Bachel…        2    0.737
## 3 100690 Amridge … AL             27334 Private South… Associ…        1   NA    
## 4 100706 Universi… AL             21607 Public  South… Bachel…        2    0.826
## 5 100724 Alabama … AL             32000 Public  South… Bachel…        2    0.969
## 6 100751 The Univ… AL             23250 Public  South… Bachel…        2    0.827
## # ℹ 7 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>
```

```
df %>%
  tail()
```

```
## # A tibble: 6 × 16
##   unitid instnm    stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##    <int> <chr>     <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
## 1 493716 Yeshiva … NJ                NA Private North… Associ…        2    0.477
## 2 493725 Universi… AR                NA Public  South… Bachel…        1   NA    
## 3 493822 College … RI                NA Private New E… Bachel…        1   NA    
## 4 494630 Christ M… TX                NA Private South… Bachel…        1   NA    
## 5 494685 Urshan C… MO                NA Private Plains Bachel…        2    0.836
## 6 494737 Yeshiva … NY                NA Private North… Bachel…        1   NA    
## # ℹ 7 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>
```

```
tail(df)
```

```
## # A tibble: 6 × 16
##   unitid instnm    stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##    <int> <chr>     <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
## 1 493716 Yeshiva … NJ                NA Private North… Associ…        2    0.477
## 2 493725 Universi… AR                NA Public  South… Bachel…        1   NA
## 3 493822 College … RI                NA Private New E… Bachel…        1   NA
## 4 494630 Christ M… TX                NA Private South… Bachel…        1   NA
## 5 494685 Urshan C… MO                NA Private Plains Bachel…        2    0.836
## 6 494737 Yeshiva … NY                NA Private North… Bachel…        1   NA
## # ℹ 7 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>
```

Looking for Vandy

```
df %>%
  filter(instnm == "Vanderbilt University") %>%
  select(instnm,md_earn_wne_p6)
```

```
## # A tibble: 1 × 2
##   instnm                md_earn_wne_p6
##   <chr>                          <int>
## 1 Vanderbilt University          53400
```

```
# arrange()
df %>%
  select(instnm,md_earn_wne_p6) %>%
  arrange(md_earn_wne_p6)
```

```
## # A tibble: 2,546 × 2
##    instnm                                   md_earn_wne_p6
##    <chr>                                             <int>
##  1 Stone Child College                               10600
##  2 White Earth Tribal and Community College          11000
##  3 United Tribes Technical College                   11800
##  4 Blackfeet Community College                       11900
##  5 Chief Dull Knife College                          12200
##  6 Rabbinical College of Ohr Shimon Yisroel          12200
##  7 Yeshivath Viznitz                                 12800
##  8 Yeshiva Gedolah Imrei Yosef D'spinka              12900
##  9 Yeshiva of Machzikai Hadas                        12900
## 10 United Talmudical Seminary                        13000
## # ℹ 2,536 more rows
```

```
df %>%
  select(instnm,md_earn_wne_p6) %>%
  arrange(desc(md_earn_wne_p6))
```

```
## # A tibble: 2,546 × 2
##    instnm                                                       md_earn_wne_p6
##    <chr>                                                                 <int>
##  1 University of Health Sciences and Pharmacy in St. Louis              120400
##  2 Albany College of Pharmacy and Health Sciences                       112100
##  3 Samuel Merritt University                                            100100
##  4 Massachusetts Institute of Technology                                 82200
##  5 Oregon Health & Science University                                    80000
##  6 Louisiana State University Health Sciences Center-Shreveport          78200
##  7 Cochran School of Nursing                                             77300
##  8 Duke University                                                       76300
##  9 MCPHS University                                                      75700
## 10 Los Angeles County College of Nursing and Allied Health               75300
## # ℹ 2,536 more rows
```

```
df %>%
  filter(instnm == 'Vanderbilt')
```

```
## # A tibble: 0 × 16
## # ℹ 16 variables: unitid <int>, instnm <chr>, stabbr <chr>,
## #   grad_debt_mdn <int>, control <chr>, region <chr>, preddeg <chr>,
## #   openadmp <int>, adm_rate <dbl>, ccbasic <int>, sat_avg <int>,
## #   md_earn_wne_p6 <int>, ugds <int>, costt4_a <int>, selective <dbl>,
## #   research_u <dbl>
```

```
df %>%
  filter(str_detect(instnm,"Vand")) %>%
  select(instnm,stabbr)
```

```
## # A tibble: 2 × 2
##   instnm                      stabbr
##   <chr>                       <chr>
## 1 VanderCook College of Music IL
## 2 Vanderbilt University       TN
```

```
df %>%
  filter(grepl("Vand",instnm)) %>%
  select(instnm,stabbr,md_earn_wne_p6,grad_debt_mdn)
```

```
## # A tibble: 2 × 4
##   instnm                      stabbr md_earn_wne_p6 grad_debt_mdn
##   <chr>                       <chr>           <int>         <int>
## 1 VanderCook College of Music IL                 NA         27000
## 2 Vanderbilt University       TN              53400         14962
```

```
# Negation: !
df %>%
  filter(instnm != "Vanderbilt University")
```

```
## # A tibble: 2,545 × 16
##    unitid instnm   stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##     <int> <chr>    <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
##  1 100654 Alabama… AL             33375 Public  South… Bachel…        2    0.918
##  2 100663 Univers… AL             22500 Public  South… Bachel…        2    0.737
##  3 100690 Amridge… AL             27334 Private South… Associ…        1   NA
##  4 100706 Univers… AL             21607 Public  South… Bachel…        2    0.826
##  5 100724 Alabama… AL             32000 Public  South… Bachel…        2    0.969
##  6 100751 The Uni… AL             23250 Public  South… Bachel…        2    0.827
##  7 100760 Central… AL             12500 Public  South… Associ…        1   NA
##  8 100812 Athens … AL             19500 Public  South… Bachel…       NA   NA
##  9 100830 Auburn … AL             24826 Public  South… Bachel…        2    0.904
## 10 100858 Auburn … AL             21281 Public  South… Bachel…        2    0.807
## # ℹ 2,535 more rows
## # ℹ 7 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>
```

Summarizing data with `summarise()`

```
df %>%
  summarise(avg_earn = mean(md_earn_wne_p6,na.rm=T),
            avg_debt = mean(grad_debt_mdn,na.rm=T))
```

```
## # A tibble: 1 × 2
##   avg_earn avg_debt
##      <dbl>    <dbl>
## 1   33028.   19646.
```

```
# Comparing with filter
df %>%
  filter(sat_avg > 1200) %>%
  summarise(avg_earn = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 1 × 1
##   avg_earn
##      <dbl>
## 1   43703.
```

```
df %>%
  filter(sat_avg < 1200) %>%
  summarise(avg_earn = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 1 × 1
##   avg_earn
##      <dbl>
## 1   33968.
```

Introducing `group_by()`

```
df %>%
  mutate(sat_HL = ifelse(sat_avg > 1200,
                        "High SAT",
                        "Low SAT")) %>%
  group_by(sat_HL) %>%
  summarise(avg_earn = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 3 × 2
##   sat_HL   avg_earn
##   <chr>       <dbl>
## 1 High SAT   43703.
## 2 Low SAT    33960.
## 3 <NA>       29250.
```