Lecture Notes

2024-07-02

Functions, Objects and Visualization

```
require(tidyverse)
```

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

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── 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/ISP_Data_Science_2024/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>
```

Look at column names

```
colnames(df)
```

```
##  [1] "unitid"         "instnm"         "stabbr"         "grad_debt_mdn"
##  [5] "control"        "region"         "preddeg"        "openadmp"
##  [9] "adm_rate"       "ccbasic"        "sat_avg"        "md_earn_wne_p6"
## [13] "ugds"           "costt4_a"       "selective"      "research_u"
```

New functions: head()

```
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>
```

New functions: filter()

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

```
## # A tibble: 1 × 16
##   unitid instnm    stabbr grad_debt_mdn control region preddeg openadmp adm_rate
##    <int> <chr>     <chr>          <int> <chr>   <chr>  <chr>      <int>    <dbl>
## 1 221999 Vanderbi… TN             14962 Private South… Bachel…        2   0.0912
## # ℹ 7 more variables: ccbasic <int>, sat_avg <int>, md_earn_wne_p6 <int>,
## #   ugds <int>, costt4_a <int>, selective <dbl>, research_u <dbl>
```

New functions: select()

```
df %>%
  select(md_earn_wne_p6,costt4_a)
```

```
## # A tibble: 2,546 × 2
##    md_earn_wne_p6 costt4_a
##             <int>    <int>
##  1          25200    23053
##  2          35100    24495
##  3          30700    14800
##  4          36200    23917
##  5          22600    21866
##  6          37400    29872
##  7          23100    10493
##  8          33400       NA
##  9          30100    19849
## 10          39500    31590
## # ℹ 2,536 more rows
```

Combining functions

```
# Vanderbilt
df %>%
  filter(instnm == "Vanderbilt University") %>%
  select(md_earn_wne_p6,costt4_a)
```

```
## # A tibble: 1 × 2
##   md_earn_wne_p6 costt4_a
##            <int>    <int>
## 1          53400    70146
```

```
# Harvard
df %>%
  filter(instnm == "Harvard University") %>%
  select(md_earn_wne_p6,costt4_a)
```

```
## # A tibble: 1 × 2
##   md_earn_wne_p6 costt4_a
##            <int>    <int>
## 1          70300    73485
```

Augmented filter()

```
df %>%
  filter(str_detect(instnm, "Harvard")) %>%
  select(md_earn_wne_p6, costt4_a,sat_avg)
```

```
## # A tibble: 1 × 3
##   md_earn_wne_p6 costt4_a sat_avg
##            <int>    <int>   <int>
## 1          70300    73485    1517
```

OR logic in filter()

```
df %>%
  filter(instnm == "Vanderbilt University" | instnm == "Harvard University") %>%
  select(instnm,md_earn_wne_p6,costt4_a,sat_avg)
```

```
## # A tibble: 2 × 4
##   instnm                md_earn_wne_p6 costt4_a sat_avg
##   <chr>                          <int>    <int>   <int>
## 1 Harvard University             70300    73485    1517
## 2 Vanderbilt University          53400    70146    1515
```

```
# str_detect() version of OR
df %>%
  filter(str_detect(instnm,"Harvard|Vanderbilt")) %>%
  select(instnm,sat_avg)
```

```
## # A tibble: 2 × 2
##   instnm                sat_avg
##   <chr>                   <int>
## 1 Harvard University       1517
## 2 Vanderbilt University    1515
```

Applied test: schools with the word “of” in their name

```
df %>%
  filter(str_detect(instnm,"of")) %>%
  select(instnm,sat_avg)
```

```
## # A tibble: 521 × 2
##    instnm                                sat_avg
##    <chr>                                   <int>
##  1 University of Alabama at Birmingham    1234
##  2 University of Alabama in Huntsville    1319
##  3 The University of Alabama              1261
##  4 University of West Alabama             1041
##  5 University of Mobile                   1166
##  6 University of Montevallo               1135
##  7 University of North Alabama            1148
##  8 University of South Alabama            1166
##  9 University of Alaska Anchorage           NA
## 10 University of Alaska Fairbanks         1121
## # ℹ 511 more rows
```

```
# that are located in vermont
df %>%
  filter(str_detect(instnm,"of")) %>%
  # filter(stabbr == "VT") %>%
  filter(str_detect(instnm,"Vermont")) %>%
  select(instnm,sat_avg)
```

```
## # A tibble: 2 × 2
##   instnm                       sat_avg
##   <chr>                          <int>
## 1 Community College of Vermont      NA
## 2 University of Vermont           1287
```

```
# Combine in a single filter with commas (,)
df %>%
  filter(str_detect(instnm,"of"),
         stabbr == "VT") %>%
  select(instnm,sat_avg)
```

```
## # A tibble: 2 × 2
##   instnm                       sat_avg
##   <chr>                          <int>
## 1 Community College of Vermont      NA
## 2 University of Vermont           1287
```

RQ: What is the relationship between SAT scores and sarlary?

```
df %>%
  select(instnm,sat_avg,md_earn_wne_p6)
```

```
## # A tibble: 2,546 × 3
##    instnm                              sat_avg md_earn_wne_p6
##    <chr>                                 <int>          <int>
##  1 Alabama A & M University                939          25200
##  2 University of Alabama at Birmingham    1234          35100
##  3 Amridge University                        NA          30700
##  4 University of Alabama in Huntsville    1319          36200
##  5 Alabama State University                946          22600
##  6 The University of Alabama              1261          37400
##  7 Central Alabama Community College         NA          23100
##  8 Athens State University                   NA          33400
##  9 Auburn University at Montgomery        1082          30100
## 10 Auburn University                      1300          39500
## # ℹ 2,536 more rows
```

New function: summarise()

```
df %>%
  summarise(overall_avg_sat = mean(sat_avg,na.rm = TRUE))
```

```
## # A tibble: 1 × 1
##   overall_avg_sat
##             <dbl>
## 1           1141.
```

```
df %>%
  summarise(overall_avg_earnings = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 1 × 1
##   overall_avg_earnings
##                  <dbl>
## 1                33028.
```

Creating relationships with filter() %>% summarise()

```
df %>%
  filter(sat_avg < 1141) %>%
  summarise(low_sat_salary = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 1 × 1
##   low_sat_salary
##            <dbl>
## 1         33327.
```

```
df %>%
  filter(sat_avg > 1141) %>%
  summarise(high_sat_salary = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 1 × 1
##   high_sat_salary
##             <dbl>
## 1          41052.
```

New function: mutate()

```
df %>%
  mutate(new_column = 1) %>%
  select(instnm,new_column)
```

```
## # A tibble: 2,546 × 2
##    instnm                              new_column
##    <chr>                                    <dbl>
##  1 Alabama A & M University                     1
##  2 University of Alabama at Birmingham          1
##  3 Amridge University                           1
##  4 University of Alabama in Huntsville          1
##  5 Alabama State University                     1
##  6 The University of Alabama                    1
##  7 Central Alabama Community College            1
##  8 Athens State University                      1
##  9 Auburn University at Montgomery              1
## 10 Auburn University                            1
## # ℹ 2,536 more rows
```

```
# NOTE: this will produce an error because we didn't add the new column to the df
# df %>%
#   select(new_column,sat_avg,md_earn_wne_p6)

# Save a new column
df <- df %>%
  mutate(new_column = 1)

df %>%
  select(instnm,new_column)
```

```
## # A tibble: 2,546 × 2
##    instnm                                new_column
##    <chr>                                      <dbl>
##  1 Alabama A & M University                       1
##  2 University of Alabama at Birmingham            1
##  3 Amridge University                             1
##  4 University of Alabama in Huntsville            1
##  5 Alabama State University                       1
##  6 The University of Alabama                      1
##  7 Central Alabama Community College              1
##  8 Athens State University                        1
##  9 Auburn University at Montgomery                1
## 10 Auburn University                              1
## # ℹ 2,536 more rows
```

```
df <- df %>%
  filter(instnm == "Vanderbilt") %>%
  select(instnm)

df
```

```
## # A tibble: 0 × 1
## # ℹ 1 variable: instnm <chr>
```

```
df <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/sc_debt.
Rds")
```

Augmented mutate()

```
df %>%
  mutate(sat_quality = ifelse(sat_avg > 1141,
                              "high SAT school",
                              "low SAT school")) %>%
  select(instnm,sat_avg,sat_quality)
```

```
## # A tibble: 2,546 × 3
##    instnm                              sat_avg sat_quality
##    <chr>                                 <int> <chr>
##  1 Alabama A & M University                939 low SAT school
##  2 University of Alabama at Birmingham    1234 high SAT school
##  3 Amridge University                       NA <NA>
##  4 University of Alabama in Huntsville    1319 high SAT school
##  5 Alabama State University                946 low SAT school
##  6 The University of Alabama              1261 high SAT school
##  7 Central Alabama Community College        NA <NA>
##  8 Athens State University                  NA <NA>
##  9 Auburn University at Montgomery        1082 low SAT school
## 10 Auburn University                      1300 high SAT school
## # ℹ 2,536 more rows
```

```
df <- df %>%
  mutate(sat_quality = ifelse(sat_avg > 1141,
                              "high SAT school",
                              "low SAT school"))
```

New function: group_by()

```
df %>%
  group_by(sat_quality) %>%
  summarise(avg_salary = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 3 × 2
##   sat_quality     avg_salary
##   <chr>                <dbl>
## 1 high SAT school     41052.
## 2 low SAT school      33321.
## 3 <NA>                29250.
```

```
df %>%
  group_by(stabbr) %>%
  summarise(avg_sat = mean(sat_avg,na.rm=T))
```

```
## # A tibble: 51 × 2
##    stabbr avg_sat
##    <chr>    <dbl>
##  1 AK       1121
##  2 AL       1123.
##  3 AR       1141.
##  4 AZ       1147.
##  5 CA       1183.
##  6 CO       1132.
##  7 CT       1194.
##  8 DC       1262
##  9 DE       1043
## 10 FL       1142.
## # ℹ 41 more rows
```

LAST new function: arrange()

```
df %>%
  group_by(stabbr) %>%
  summarise(avg_sat = mean(sat_avg,na.rm=T)) %>%
  arrange(-avg_sat)
```

```
## # A tibble: 51 × 2
##    stabbr avg_sat
##    <chr>    <dbl>
##  1 NH       1335
##  2 DC       1262
##  3 VT       1250.
##  4 MA       1226.
##  5 RI       1226.
##  6 UT       1215
##  7 WY       1203
##  8 NY       1195.
##  9 CT       1194.
## 10 CA       1183.
## # ℹ 41 more rows
```

```
df %>%
  group_by(stabbr) %>%
  summarise(avg_sat = mean(sat_avg,na.rm=T)) %>%
  arrange(desc(avg_sat))
```

```
## # A tibble: 51 × 2
##    stabbr avg_sat
##    <chr>    <dbl>
##  1 NH       1335
##  2 DC       1262
##  3 VT       1250.
##  4 MA       1226.
##  5 RI       1226.
##  6 UT       1215
##  7 WY       1203
##  8 NY       1195.
##  9 CT       1194.
## 10 CA       1183.
## # ℹ 41 more rows
```