Homework 10

Sameeksha Deshatty

2025-02-18

Instructions

Answer the following questions and complete the exercises in RMarkdown. Please embed all of your code and push your final work to your repository. Your final lab report should be organized, clean, and run free from errors. Remember, you must remove the `#` for the included code chunks to run. Be sure to add your name to the author header above.

Make sure to use the formatting conventions of RMarkdown to make your report neat and clean!

Load the libraries

```r
library(tidyverse)
library(janitor)
library(naniar)
```

For this homework, we will take a departure from biological data and use data about California colleges. These data are a subset of the national college scorecard (https://collegescorecard.ed.gov/data/). Load the `ca_college_data.csv` as a new object called `colleges`.

```r
colleges <- readr::read_csv("data/ca_college_data.csv") %>% janitor::clean_names()
```

```
## Rows: 341 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): INSTNM, CITY, STABBR, ZIP
## dbl (6): ADM_RATE, SAT_AVG, PCIP26, COSTT4_A, C150_4_POOLED, PFTFTUG1_EF
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

The variables are a bit hard to decipher, here is a key:

INSTNM: Institution name
CITY: California city
STABBR: Location state
ZIP: Zip code
ADM_RATE: Admission rate
SAT_AVG: SAT average score
PCIP26: Percentage of degrees awarded in Biological And Biomedical Sciences

COSTT4_A: Annual cost of attendance
C150_4_POOLED: 4-year completion rate
PFTFTUG1_EF: Percentage of undergraduate students who are first-time, full-time degree/certificate-seeking undergraduate students

1. Use your preferred function(s) to have a look at the data and get an idea of its structure. Make sure you summarize NA's and determine whether or not the data are tidy. You may also consider dealing with any naming issues.

```
# structure
glimpse(colleges)
```

```
## Rows: 341
## Columns: 10
## $ instnm        <chr> "Grossmont College", "College of the Sequoias", "College~
## $ city          <chr> "El Cajon", "Visalia", "San Mateo", "Ventura", "Oxnard",~
## $ stabbr        <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "C~
## $ zip           <chr> "92020-1799", "93277-2214", "94402-3784", "93003-3872", ~
## $ adm_rate      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ sat_avg       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ pcip26        <dbl> 0.0016, 0.0066, 0.0038, 0.0035, 0.0085, 0.0151, 0.0000, ~
## $ costt4_a      <dbl> 7956, 8109, 8278, 8407, 8516, 8577, 8580, 9181, 9281, 93~
## $ c150_4_pooled <dbl> NA, NA, NA, NA, NA, NA, 0.2334, NA, NA, NA, NA, 0.1704, ~
## $ pftftug1_ef   <dbl> 0.3546, 0.5413, 0.3567, 0.3824, 0.2753, 0.4286, 0.2307, ~
```

```
# missing values
colleges %>% miss_var_summary()
```

```
## # A tibble: 10 x 3
##    variable      n_miss pct_miss
##    <chr>          <int>    <num>
##  1 sat_avg          276     80.9
##  2 adm_rate         240     70.4
##  3 c150_4_pooled    221     64.8
##  4 costt4_a         124     36.4
##  5 pftftug1_ef       53     15.5
##  6 pcip26            35     10.3
##  7 instnm             0      0
##  8 city               0      0
##  9 stabbr             0      0
## 10 zip                0      0
```

```
# first few rows
head(colleges)
```

```
## # A tibble: 6 x 10
##   instnm        city  stabbr zip   adm_rate sat_avg pcip26 costt4_a c150_4_pooled
##   <chr>         <chr> <chr>  <chr>    <dbl>   <dbl>  <dbl>    <dbl>         <dbl>
## 1 Grossmont C~ El C~ CA     9202~       NA      NA 0.0016     7956            NA
## 2 College of ~ Visa~ CA     9327~       NA      NA 0.0066     8109            NA
## 3 College of ~ San ~ CA     9440~       NA      NA 0.0038     8278            NA
## 4 Ventura Col~ Vent~ CA     9300~       NA      NA 0.0035     8407            NA
## 5 Oxnard Coll~ Oxna~ CA     9303~       NA      NA 0.0085     8516            NA
## 6 Moorpark Co~ Moor~ CA     9302~       NA      NA 0.0151     8577            NA
## # i 1 more variable: pftftug1_ef <dbl>
```

2. Which cities in California have the highest number of colleges?

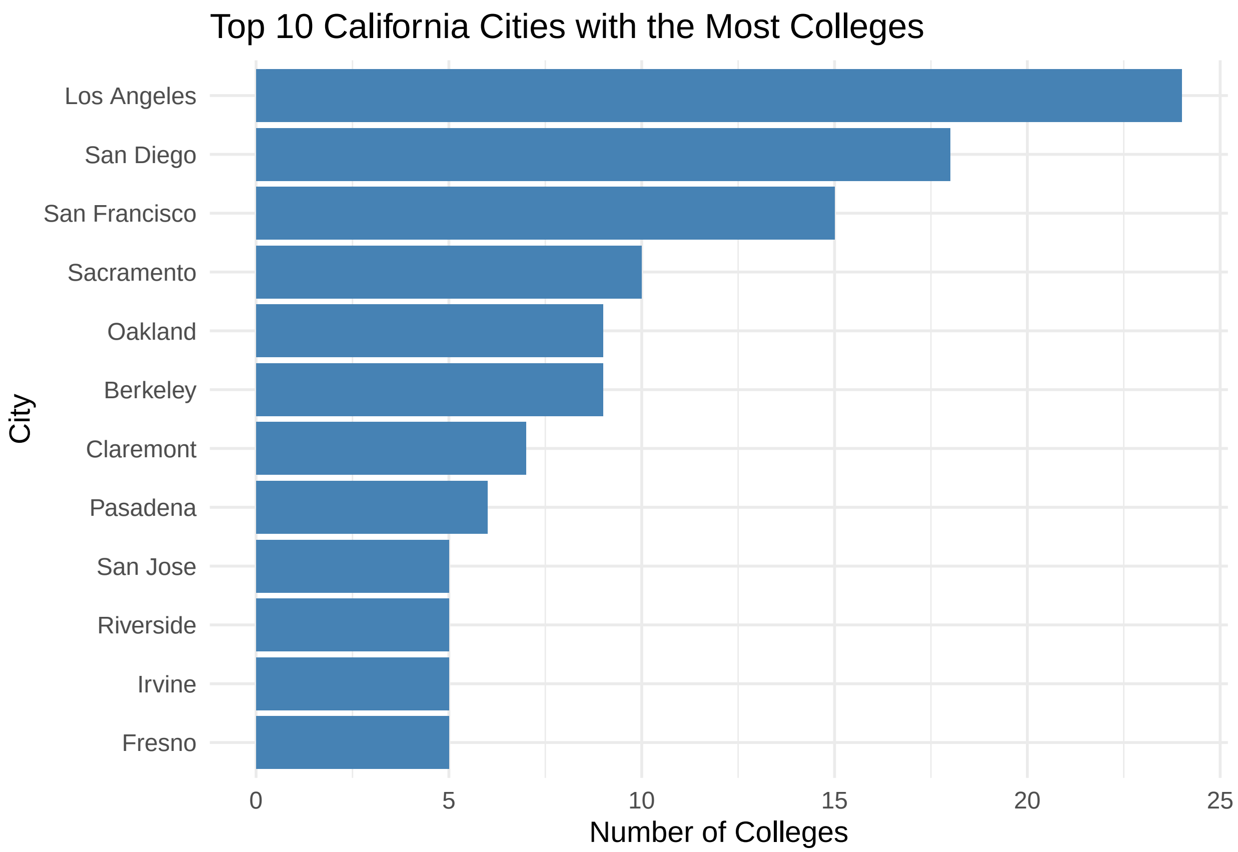
```
colleges %>%
  count(city, sort = TRUE)
```

```
## # A tibble: 161 x 2
##    city              n
##    <chr>         <int>
##  1 Los Angeles      24
##  2 San Diego        18
##  3 San Francisco    15
##  4 Sacramento       10
##  5 Berkeley          9
##  6 Oakland           9
##  7 Claremont         7
##  8 Pasadena          6
##  9 Fresno            5
## 10 Irvine            5
## # i 151 more rows
```

Los Angeles has the highest number of colleges with 24.

3. Based on your answer to #2, make a plot that shows the number of colleges in the top 10 cities.

```
colleges %>%
  count(city, sort = TRUE) %>%
  slice_max(n, n = 10) %>%
  ggplot(aes(x = reorder(city, n), y = n)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 10 California Cities with the Most Colleges",
       x = "City", y = "Number of Colleges") +
  theme_minimal()
```



4. The column `COSTT4_A` is the annual cost of each institution. Which city has the highest average cost? Where is it located?

```
colleges %>%
  group_by(city) %>%
  summarise(avg_cost = mean(costt4_a, na.rm = TRUE)) %>%
  arrange(desc(avg_cost))
```

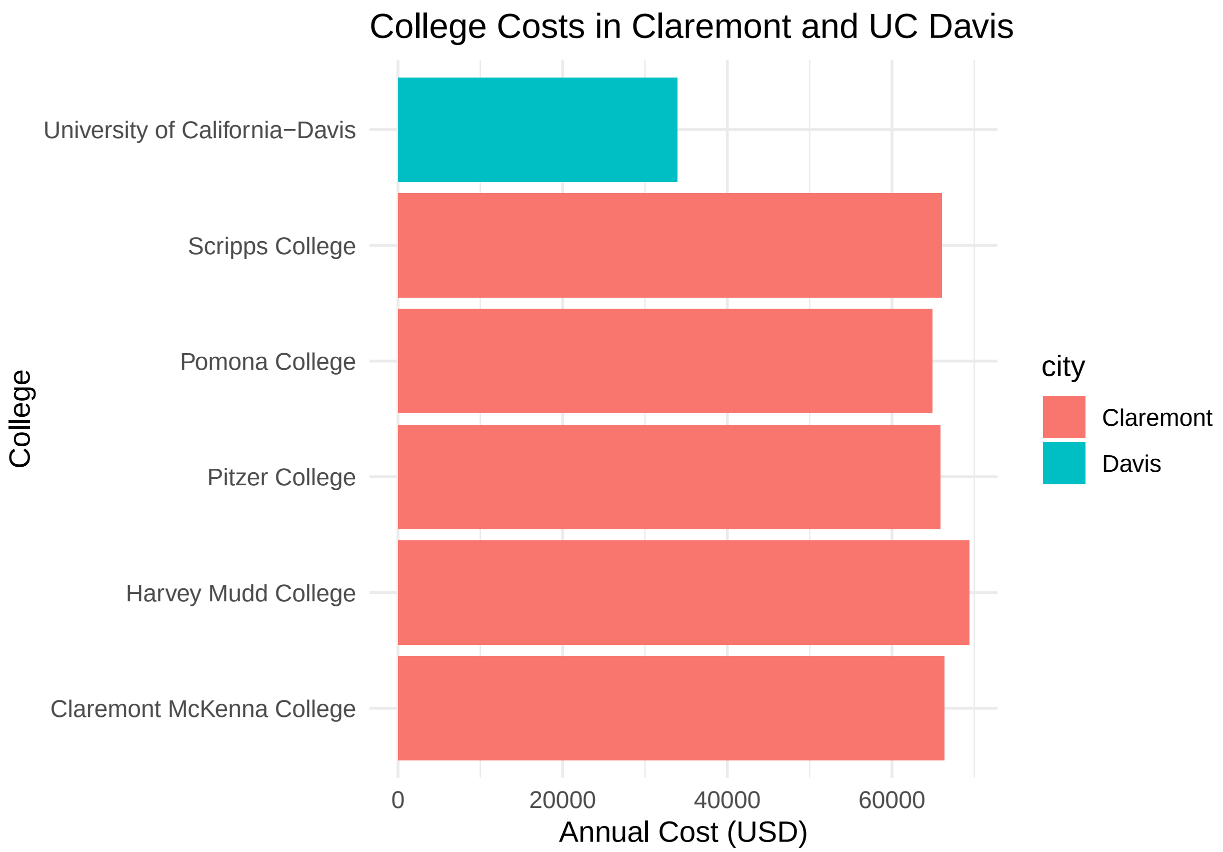
```
## # A tibble: 161 x 2
##    city                avg_cost
##    <chr>                  <dbl>
##  1 Claremont              66498
##  2 Malibu                 66152
##  3 Valencia               64686
##  4 Orange                 64501
##  5 Redlands               61542
##  6 Moraga                 61095
##  7 Atherton               56035
##  8 Thousand Oaks          54373
##  9 Rancho Palos Verdes    50758
## 10 La Verne               50603
## # i 151 more rows
```

The highest annual cost is $66,489 per year at Claremont college in Claremont.

5. Based on your answer to #4, make a plot that compares the cost of the individual colleges in the most expensive city. Bonus! Add UC Davis here to see how it compares :>).

```r
most_expensive_city <- colleges %>%
  group_by(city) %>%
  summarise(avg_cost = mean(costt4_a, na.rm = TRUE)) %>%
  arrange(desc(avg_cost)) %>%
  slice(1) %>%
  pull(city)

colleges %>%
  filter(city == most_expensive_city | instnm == "University of California-Davis") %>%
  filter(!is.na(costt4_a)) %>%
  ggplot(aes(x = instnm, y = costt4_a, fill = city)) +
  geom_col() +
  coord_flip() +
  labs(title = paste("College Costs in", most_expensive_city, "and UC Davis"),
       x = "College", y = "Annual Cost (USD)") +
  theme_minimal()
```



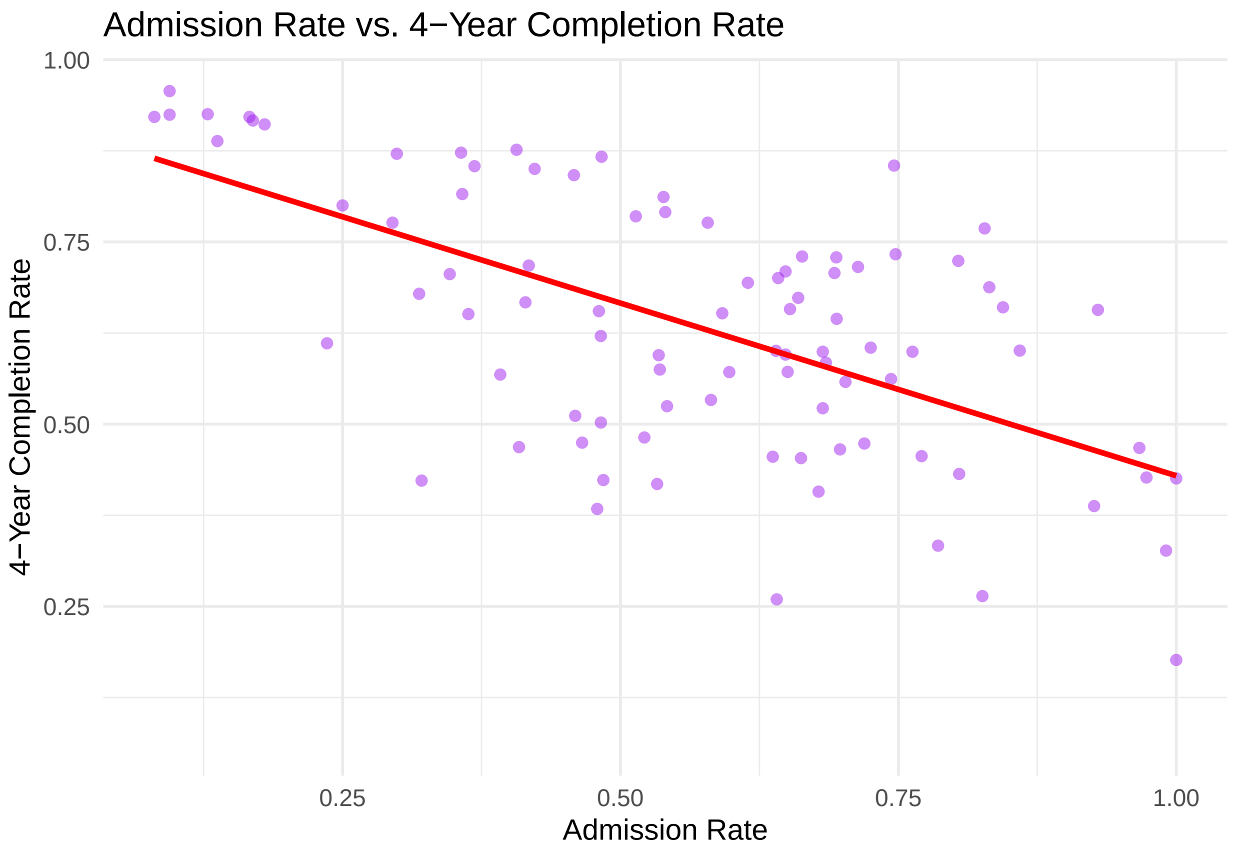
6. The column `ADM_RATE` is the admissions rate by college and `C150_4_POOLED` is the four-year completion rate. Use a scatterplot to show the relationship between these two variables. What do you think this means?

```
colleges %>%
  ggplot(aes(x = adm_rate, y = c150_4_pooled)) +
  geom_point(color = "purple", alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Admission Rate vs. 4-Year Completion Rate",
       x = "Admission Rate", y = "4-Year Completion Rate") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 251 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 251 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



The scatterplot is showing that a lower admission rate has a correlation with a higher 4-year completion rate which makes sense as more prestigious colleges have lower admit rates and also higher completion rates.

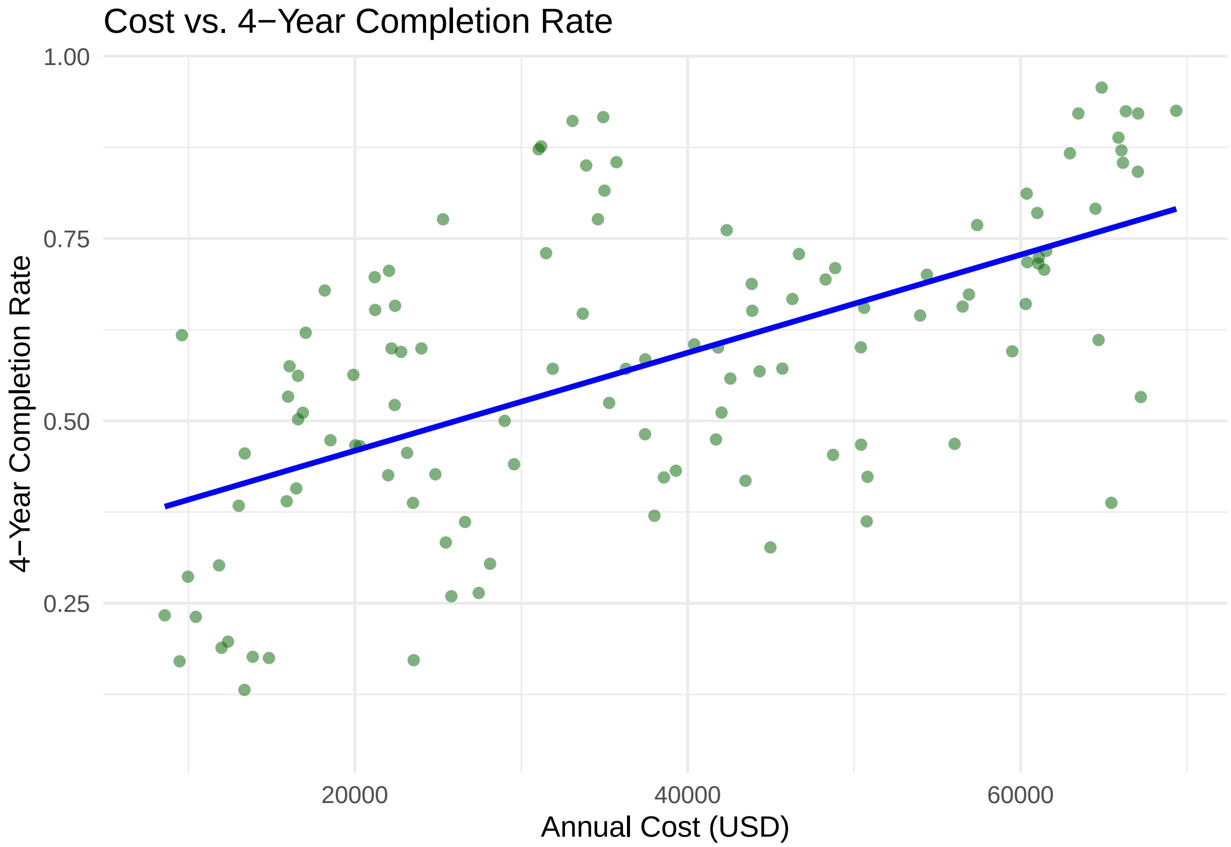
7. Is there a relationship between cost and four-year completion rate? (You don't need to do the stats, just produce a plot). What do you think this means?

```
colleges %>%
  ggplot(aes(x = costt4_a, y = c150_4_pooled)) +
  geom_point(color = "darkgreen", alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Cost vs. 4-Year Completion Rate",
       x = "Annual Cost (USD)", y = "4-Year Completion Rate") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 225 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 225 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



This scatterplot is showing that a higher annual cost correlates with a higher 4-year completions, again this makes sense as more prestigious colleges have higher annual attendance costs.

8. The column titled `INSTNM` is the institution name. We are only interested in the University of California colleges. Make a new data frame that is restricted to UC institutions. You can remove `Hastings College of Law` and `UC San Francisco` as we are only interested in undergraduate institutions.

```
univ_calif <- colleges %>%
  filter(str_detect(instnm, "University of California")) %>%
  filter(!instnm %in% c("University of California-Hastings College of the Law",
                                "University of California-San Francisco"))

glimpse(univ_calif)
```

```
## Rows: 9
## Columns: 10
## $ instnm        <chr> "University of California-San Diego", "University of Cal~
## $ city          <chr> "La Jolla", "Irvine", "Riverside", "Los Angeles", "Davis~
## $ stabbr        <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA"
## $ zip           <chr> "92093", "92697", "92521", "90095-1405", "95616-8678", "~
## $ adm_rate      <dbl> 0.3566, 0.4065, 0.6634, 0.1799, 0.4228, 0.5785, 0.1693, ~
## $ sat_avg       <dbl> 1324, 1206, 1078, 1334, 1218, 1201, 1422, 1281, NA
## $ pcip26        <dbl> 0.2165, 0.1073, 0.1491, 0.1548, 0.1975, 0.1927, 0.1053, ~
## $ costt4_a      <dbl> 31043, 31198, 31494, 33078, 33904, 34608, 34924, 34998, ~
## $ c150_4_pooled <dbl> 0.8724, 0.8764, 0.7300, 0.9112, 0.8502, 0.7764, 0.9165, ~
## $ pftftug1_ef   <dbl> 0.6622, 0.7254, 0.8111, 0.6607, 0.6049, 0.7856, 0.7087, ~
```

Remove `Hastings College of Law` and `UC San Francisco` and store the final data frame as a new object `univ_calif_final`. WAS COMPLETED ABOVE

Use `separate()` to separate institution name into two new columns "UNIV" and "CAMPUS".

```
univ_calif_final <- univ_calif %>%
  mutate(
    univ = "University of California",
    campus = str_extract(instnm, "(?<=-).*")
  )

glimpse(univ_calif_final)
```

```
## Rows: 9
## Columns: 12
## $ instnm        <chr> "University of California-San Diego", "University of Cal~
## $ city          <chr> "La Jolla", "Irvine", "Riverside", "Los Angeles", "Davis~
## $ stabbr        <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA"
## $ zip           <chr> "92093", "92697", "92521", "90095-1405", "95616-8678", "~
## $ adm_rate      <dbl> 0.3566, 0.4065, 0.6634, 0.1799, 0.4228, 0.5785, 0.1693, ~
## $ sat_avg       <dbl> 1324, 1206, 1078, 1334, 1218, 1201, 1422, 1281, NA
## $ pcip26        <dbl> 0.2165, 0.1073, 0.1491, 0.1548, 0.1975, 0.1927, 0.1053, ~
## $ costt4_a      <dbl> 31043, 31198, 31494, 33078, 33904, 34608, 34924, 34998, ~
## $ c150_4_pooled <dbl> 0.8724, 0.8764, 0.7300, 0.9112, 0.8502, 0.7764, 0.9165, ~
## $ pftftug1_ef   <dbl> 0.6622, 0.7254, 0.8111, 0.6607, 0.6049, 0.7856, 0.7087, ~
## $ univ          <chr> "University of California", "University of California", ~
## $ campus        <chr> "San Diego", "Irvine", "Riverside", "Los Angeles", "Davi~
```

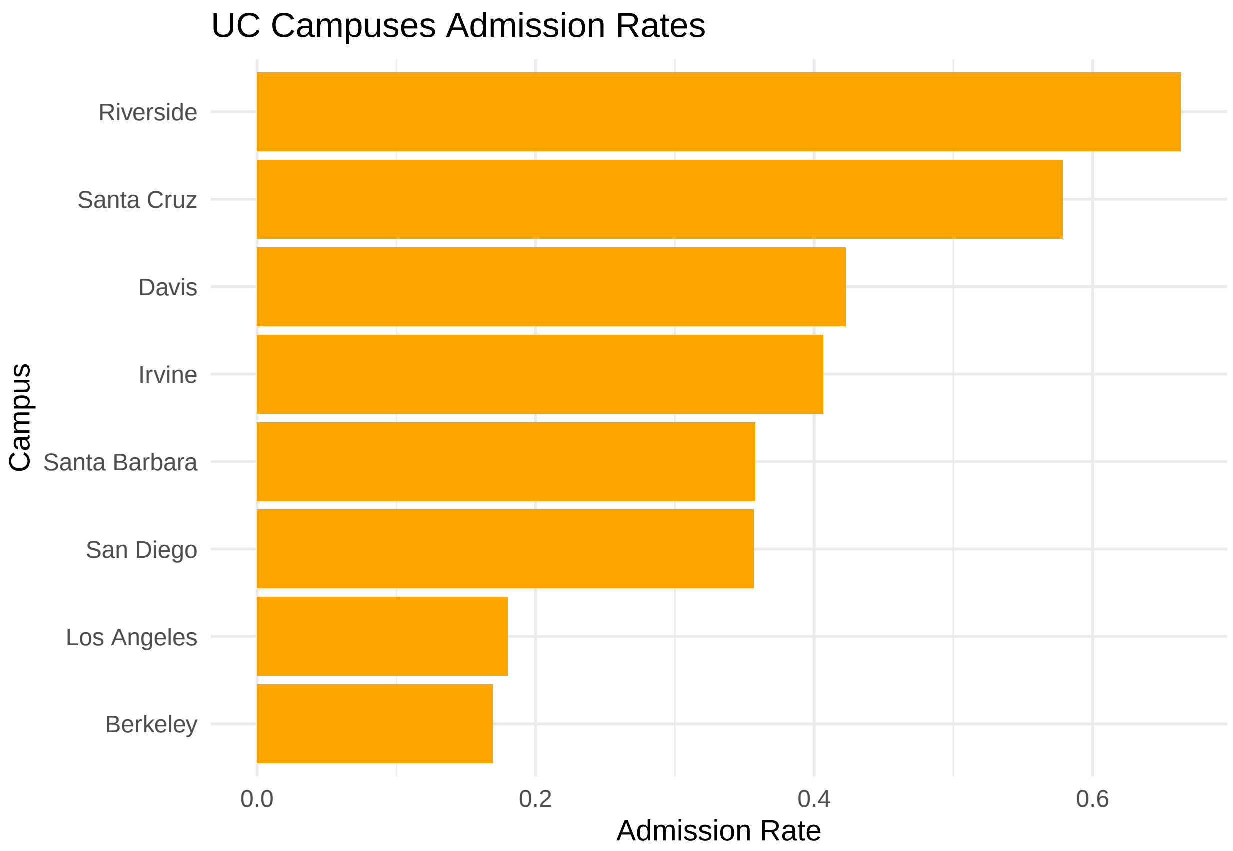
9. The column `ADM_RATE` is the admissions rate by campus. Which UC has the lowest and highest admissions rates? Produce a numerical summary and an appropriate plot.

```
univ_calif_final <- univ_calif_final %>%
  filter(campus != "Hastings College of Law")

univ_calif_final %>%
  summarise(max_admission = max(adm_rate, na.rm = TRUE),
            min_admission = min(adm_rate, na.rm = TRUE))
```

```
## # A tibble: 1 x 2
##   max_admission min_admission
##           <dbl>         <dbl>
## 1         0.663         0.169
```

```
univ_calif_final %>%
  ggplot(aes(x = reorder(campus, adm_rate), y = adm_rate)) +
  geom_col(fill = "orange") +
  coord_flip() +
  labs(title = "UC Campuses Admission Rates",
       x = "Campus", y = "Admission Rate") +
  theme_minimal()
```

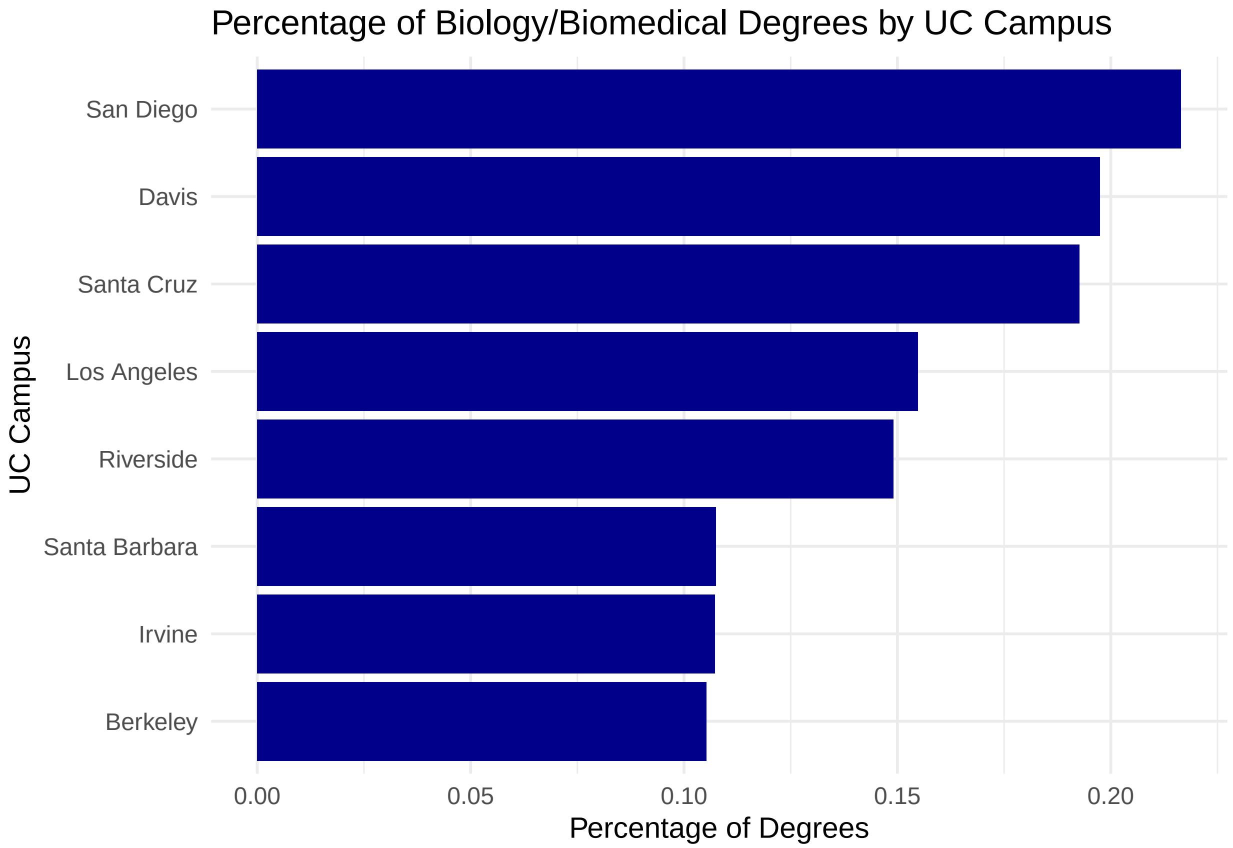


10. If you wanted to get a degree in biological or biomedical sciences, which campus confers the majority of these degrees? Produce a numerical summary and an appropriate plot.

```
univ_calif_final %>%
  arrange(desc(pcip26)) %>%
  select(campus, pcip26)
```

```
## # A tibble: 8 x 2
##   campus        pcip26
##   <chr>          <dbl>
## 1 San Diego      0.216
## 2 Davis           0.198
## 3 Santa Cruz      0.193
## 4 Los Angeles     0.155
## 5 Riverside       0.149
## 6 Santa Barbara   0.108
## 7 Irvine          0.107
## 8 Berkeley        0.105
```

```
univ_calif_final %>%
  ggplot(aes(x = reorder(campus, pcip26), y = pcip26)) +
  geom_col(fill = "darkblue") +
  coord_flip() +
  labs(title = "Percentage of Biology/Biomedical Degrees by UC Campus",
       x = "UC Campus", y = "Percentage of Degrees") +
  theme_minimal()
```



San Diego has the highest percentage of these degrees.

Knit Your Output and Post to [GitHub]