Team 2cool4school

Project Proposal

```
library(tidyverse)
```

Data 1

Introduction and data

- The dataset 'lemur_data.csv' comes from Kaggle.com and was released by Jesse Mostipak.
- The dataset was orginally collected from the 2019 data release from the Duke Lemur Center Database and further compiled by Zehr, SM, Roach RG, Haring D, Taylor J, Cameron FH, Yoder AD in their research paper.
- Animal data have been collected and entered by Duke Lemur Center staff according to standard operating procedures and USDA, AZA, and IACUC guidelines throughout the history of the center (United States Department of Agriculture, Association of Zoos and Aquariums, Institutional Animal Care and Use Committee respectively). Births, deaths, weights, enclosure moves, behaviors, and other significant events are recorded daily by animal care, veterinary, and research staff and subsequently entered into the permanent records by the DLC Registrar.
- This dataset contains information on over 3,500 observations. Each observation represent a lemur, including lemur-information such as ancestry, reproduction, longevity, and body mass (in total 54 columns).

Research question

- What are the top 3 factors that influence the lifespan of lemurs?

- Lemurs are the most endangered group of mammals. In fact, 98% of lemur species are endangered, and 31% of species are critically endangered nowadays. Therefore, it is crucial to conduct researches on the lifespan of lemurs to save the endangered species. Moreover, Duke Lemur Center has been the world leader in the study, care, and protection of lemurs since it was founded in 1966. As Duke students, we are able to utilize the resources at DLC to research on lemurs and potentially contribute to their studies.
- Through our preliminary investigation, we found that the death age of lemurs varies a lot ranging from 0 to 35. Thus, we are interested in researching on what are the determining factors of lemurs' lifespan. Our hypotheses is that taxon, sex, weight are the top 3 factors that would affect the lifespan of lemurs.
- We plan to research on all the variables within the dataset that could potentially affect lemur's lifespan. There are both categorical and quantitative variables involved in our research questions since categorical variables such as taxon and quantitative variables such as weight can all play a role in their lifespan.

Glimpse of data

```
lemur <- read_csv("data/lemur_data.csv")
```

```
Rows: 82609 Columns: 54
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (19): Taxon, DLC_ID, Hybrid, Sex, Name, Current_Resident, StudBook, Est...
dbl  (27): Birth_Month, Litter_Size, Expected_Gestation, Concep_Month, Dam_A...
date  (8): DOB, Estimated_Concep, Dam_DOB, Sire_DOB, DOD, Weight_Date, Conce...

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.
```

```
glimpse(lemur)
```

```
Rows: 82,609
Columns: 54
$ Taxon                     <chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG", "O~
$ DLC_ID                    <chr> "0005", "0005", "0006", "0006", "0009", "000~
$ Hybrid                    <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",~
$ Sex                       <chr> "M", "M", "F", "F", "M", "M", "M", "M", "M",~
$ Name                      <chr> "KANGA", "KANGA", "ROO", "ROO", "POOH BEAR",~
$ Current_Resident          <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",~
```

```
$ StudBook                  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ DOB                       <date> 1961-08-25, 1961-08-25, 1961-03-17, 1961-03~
$ Birth_Month               <dbl> 8, 8, 3, 3, 9, 9, 9, 5, 5, 10, 10, 6, 6, 3, ~
$ Estimated_DOB             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ Birth_Type                <chr> "CB", "CB", "CB", "CB", "CB", "CB", "CB", "C~
$ Birth_Institution         <chr> "Duke Lemur Center", "Duke Lemur Center", "D~
$ Litter_Size               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ Expected_Gestation        <dbl> 129, 129, 129, 129, 129, 129, 129, 129, 129,~
$ Estimated_Concep          <date> 1961-04-18, 1961-04-18, 1960-11-08, 1960-11~
$ Concep_Month              <dbl> 4, 4, 11, 11, 5, 5, 5, 1, 1, 6, 6, 1, 1, 11,~
$ Dam_ID                    <chr> "0001", "0001", "0001", "0001", "0001", "000~
$ Dam_Name                  <chr> "WHITE-TAIL", "WHITE-TAIL", "WHITE-TAIL", "W~
$ Dam_Taxon                 <chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG", "O~
$ Dam_DOB                   <date> 1959-01-28, 1959-01-28, 1959-01-28, 1959-01~
$ Dam_AgeAtConcep_y         <dbl> 2.22, 2.22, 1.78, 1.78, 4.32, 4.32, 4.32, 4.~
$ Sire_ID                   <chr> "0002", "0002", "0002", "0002", "0007", "000~
$ Sire_Name                 <chr> "BRUISER", "BRUISER", "BRUISER", "BRUISER", ~
$ Sire_Taxon                <chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG", "O~
$ Sire_DOB                  <date> 1959-01-28, 1959-01-28, 1959-01-28, 1959-01~
$ Sire_AgeAtConcep_y        <dbl> 2.22, 2.22, 1.78, 1.78, 4.32, 4.32, 4.32, 4.~
$ DOD                       <date> 1977-02-07, 1977-02-07, 1974-10-15, 1974-10~
$ AgeAtDeath_y              <dbl> 15.47, 15.47, 13.59, 13.59, 10.38, 10.38, 10~
$ AgeOfLiving_y             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ AgeLastVerified_y         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 14.16, 1~
$ AgeMax_LiveOrDead_y       <dbl> 15.47, 15.47, 13.59, 13.59, 10.38, 10.38, 10~
$ N_known_offspring         <dbl> 7, 7, 9, 9, 1, 1, 1, 7, 7, 5, 5, 4, 4, 1, 1,~
$ DOB_Estimated             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ Weight_g                  <dbl> 1086, 1190, 947, 1174, 899, 917, 910, 1185, ~
$ Weight_Date               <date> 1972-02-16, 1972-06-20, 1972-02-16, 1972-06~
$ MonthOfWeight             <dbl> 2, 6, 2, 6, 2, 2, 6, 2, 6, 2, 6, 2, 6, 2, 6,~
$ AgeAtWt_d                 <dbl> 3827, 3952, 3988, 4119, 3061, 3074, 3188, 28~
$ AgeAtWt_wk                <dbl> 546.71, 564.57, 569.71, 588.43, 437.29, 439.~
$ AgeAtWt_mo                <dbl> 125.82, 129.93, 131.11, 135.42, 100.64, 101.~
$ AgeAtWt_mo_NoDec          <dbl> 125, 129, 131, 135, 100, 101, 104, 92, 97, 8~
$ AgeAtWt_y                 <dbl> 10.48, 10.83, 10.93, 11.28, 8.39, 8.42, 8.73~
$ Change_Since_PrevWt_g     <dbl> NA, 104, NA, 227, NA, 18, -7, NA, 51, NA, 71~
$ Days_Since_PrevWt         <dbl> NA, 125, NA, 131, NA, 13, 114, NA, 125, NA, ~
$ Avg_Daily_WtChange_g      <dbl> NA, 0.83, NA, 1.73, NA, 1.38, -0.06, NA, 0.4~
$ DaysBeforeDeath           <dbl> 1818, 1693, 972, 841, 728, 715, 601, 2086, 1~
$ R_Min_Dam_AgeAtConcep_y   <dbl> 0.59, 0.59, 0.59, 0.59, 0.59, 0.59, 0.59, 0.~
$ Age_Category              <chr> "adult", "adult", "adult", "adult", "adult",~
$ Preg_Status               <chr> "NP", "NP", "NP", "NP", "NP", "NP", "NP", "N~
$ Expected_Gestation_d      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

```
$ ConcepDate_IfPreg         <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ InfantDOB_IfPreg          <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ DaysBeforeInfBirth_IfPreg <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ Pct_PregRemain_IfPreg     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ InfantLitSz_IfPreg        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

Data 2

Introduction and data

- The data comes from CORGIS (Collection of Really Great, Interesting, Situated Datasets) website.
- The earthquake data was originally collected on 6/7/2016 by simply collecting information from the United States Geological Survey by Ryan Whitcomb.
- Each observation contains a unique earthquake and all the information surrounding the circumstances of the earthquake, including but not limited to: coordinates of where the earthquake occurred, its magnitude, time of earthquake, and the depth of the earthquake.
- There are no ethical concerns in this data.

Research question

- Does the frequency of earthquakes at a certain location have any relationship to their magnitude or significance?
- This question is important because by determining if there are relationships between the variables, we can then consider the living conditions of people that live in those locations. There might be certain locations in which earthquakes are at their strongest and people should not be living there, or maybe locations in which we have to consider what kinds of infrastructure buildings should have to prevent them from collapsing as a result of multiple strong earthquakes.
- A rough look at the first 50 observations showed that a lot of earthquakes happened in Alaska and California. This made us wonder why this was the case and to see if those patterns are consistent throughout the rest of the data. Furthermore, because earthquakes happen more often in those states, we wonder if they are stronger than earthquakes that only occur once at a certain location and if they overall have a higher significance. Our hypothesis is that frequency and magnitude will have a positive correlation, and that the average significance of the more frequent locations will be higher than the average significance of less frequent locations.

- Our research question focuses on four variables: location, frequency, magnitude, and significance. For location, we can consider the nominal categorical variable `location.name` which gives us the state/country in which the earthquake occurred. We can also consider `location.longitude` and `location.latitude` which gives us the exact coordinates of an earthquake's location, which would be continuous numerical variables. `impact.magnitude` displays the magnitude of an earthquake which is a continuous numerical variable. We will also consider `impact.significance`, which is a discrete numerical variable and gives the significance of each earthquake based on a range of factors including estimated impact, magnitude, and felt reports. Lastly, we will calculate frequency, a discrete numerical variable, based on the location data.

Glimpse of data

```r
earthquakes <- read.csv("data/earthquakes.csv")

glimpse(earthquakes)
```

```
Rows: 8,394
Columns: 18
$ id                  <chr> "nc72666881", "us20006i0y", "nc72666891", "nc72666~
$ impact.gap          <dbl> 122.00000, 30.00000, 249.00000, 122.00000, 113.610~
$ impact.magnitude    <dbl> 1.43, 4.90, 0.06, 0.40, 0.30, 1.80, 1.00, 2.00, 1.~
$ impact.significance <int> 31, 371, 0, 2, 1, 50, 15, 62, 22, 43, 4, 12, 4, 4,~
$ location.depth      <dbl> 15.120, 97.070, 4.390, 1.090, 7.600, 1.300, 2.452,~
$ location.distance   <dbl> 0.10340000, 1.43900000, 0.02743000, 0.02699000, 0.~
$ location.full       <chr> "13km E of Livermore, California", "58km WNW of Pa~
$ location.latitude   <dbl> 37.67233, 21.51460, 37.57650, 37.59583, 39.37750, ~
$ location.longitude  <dbl> -121.6190, 94.5721, -118.8592, -118.9948, -119.845~
$ location.name       <chr> "California", "Burma", "California", "California",~
$ time.day            <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27~
$ time.epoch          <dbl> 1.469593e+12, 1.469593e+12, 1.469594e+12, 1.469594~
$ time.full           <chr> "2016-07-27 00:19:43", "2016-07-27 00:20:28", "201~
$ time.hour           <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ time.minute         <int> 19, 20, 31, 35, 41, 52, 53, 58, 3, 4, 9, 13, 17, 1~
$ time.month          <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,~
$ time.second         <int> 43, 28, 37, 44, 59, 52, 35, 45, 0, 32, 51, 31, 18,~
$ time.year           <int> 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 20~
```

Data 3

Introduction and data

To answer our question, we want to join two data sets. Both are listed below.

Data Set #1: Niche

- The first data set comes from Niche's "2023 Best Colleges in America" list

  (https://www.niche.com/colleges/search/best-colleges/)

- Niche aggregates data from a variety of sources, including the US Department of Education and reviews from students and alumni, to build their list of college rankings. The rankings list is updated monthly to reflect new data that Niche receives; although, data from the US Department o Education is only received on an annual basis. The Niche data was scraped by Maia on October 17-19 2022.

- There are 500 observations, representing the top 500 schools in the United States. Each observation has two variables: the name of the school and the rank of the school.

Data Set #2: US Department of Education

- The second data set comes from the US Department of Education's College Scorecard, which is an exhaustive summary of characteristics and statistics for all colleges and universities in the United States. There were 2,989 variables in the original data set, many of which we don't need to answer our question, and since this data set was too large to load into RStudio, we used Excel to select the variables we needed to import.

  (https://collegescorecard.ed.gov/data/)

- The College Scorecard is updated by the Education Department as it collects new data. Data used in the scorecard comes comes from data reported by the institutions, data on federal financial aid, data from taxes, and data from other federal agencies.

- There are 6681 observations in the data set, representing all of the colleges and universities in the United States. There are 63 variables in the data set, which give information about each different school.

- There are no obvious ethical or privacy concerns with this data. The schools did not report data on individual students but rather summary data of all students.

The Department of Education data set was left joined to Niche dataset by school name to make a new data set with 500 observations and 64 variables.

Research question

- Here are three potential research questions:

  1. How do rankings compare across different types of schools (by region, by racial makeup, by school type)?
  2. Which three variable have the greatest impact on a college's ranking?
  3. Do "input variables" like ACT/SAT, GPA, and type of high school matter more to a college's ranking than "output variable" like median income and graduation rate?

- These questions are important because students applying to college trust these rankings and weigh them into their college decisions. Due to large impact that a college has on a student's life, it is important to know where these ranking come from and what they actually measure.

- In general, we want to examine how different variables affect a school's ranking on the Niche College Ranking List. We plan to look at variables that are typically thought to influence school rank such as average SAT score and acceptance rate, but we also want to look at variables that aren't typically thought of such as geographic region or endowment size.

  We hypothesize that:

  1. schools that are from the northeast, are majority-white, and are private will have higher rankings on average.
  2. ACT/SAT test score, acceptance rate, and median earnings 10 years after graduation will have the biggest impact on college rank.
  3. input variables will matter more because they are more important to students applying to college, and the ranking list is geared towards them.

- Here are the types of variables involved in our research questions:

  - Categorical (nominal):
    - city
    - state
    - zip
    - accreditor institution
    - control (public vs. private)
    - Carnegie classification

  - Numerical (continuous):
    * latitude, longitude
    * admission rate
    * Total share of enrollment by racial groups
    * average net price
    * average cost of attendance
    * average faculty salary
    * percentage of undergraduates who receive a Pell grant
    * completion rate
    * average age of entry
    * share of female students
    * share of married students
    * share of first generation students
    * median & mean family income
    * median & mean earning 10 years after entry
    * Value of school's endowment
  - Numerical (discrete)
    * 25th/50th/75th/mean ACT/SAT scores for sub-tests and composite
    * Enrollment of undergraduate certificate/degree-seeking students

Glimpse of data

```
niche_data_500 <- read_csv("data/niche_data_500.csv")
```

```
Rows: 500 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): college
dbl (1): rank

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.
```

```r
glimpse(niche_data_500)
```

```
Rows: 500
Columns: 2
$ college <chr> "Massachusetts Institute of Technology", "Stanford University"~
$ rank    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,~
```

```r
us_dep_of_ed <- read_csv("data/us_dep_of_ed.csv")
```

```
Rows: 6681 Columns: 63
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (60): OPEID, OPEID6, INSTNM, CITY, STABBR, ZIP, ACCREDAGENCY, LATITUDE, ...
dbl  (3): UNITID, REGION, CONTROL

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.
```

```r
glimpse(us_dep_of_ed)
```

```
Rows: 6,681
Columns: 63
$ UNITID          <dbl> 100654, 100663, 100690, 100706, 100724, 100751, 100760~
$ OPEID           <chr> "100200", "105200", "2503400", "105500", "100500", "10~
$ OPEID6          <chr> "1002", "1052", "25034", "1055", "1005", "1051", "1007~
$ INSTNM          <chr> "Alabama A & M University", "University of Alabama at ~
$ CITY            <chr> "Normal", "Birmingham", "Montgomery", "Huntsville", "M~
$ STABBR          <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", ~
$ ZIP             <chr> "35762", "35294-0110", "36117-3553", "35899", "36104-0~
$ ACCREDAGENCY    <chr> "Southern Association of Colleges and Schools Commissi~
$ LATITUDE        <chr> "34.783368", "33.505697", "32.362609", "34.724557", "3~
$ LONGITUDE       <chr> "-86.568502", "-86.799345", "-86.17401", "-86.640449",~
$ REGION          <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
$ CCBASIC         <chr> "18", "15", "20", "16", "19", "15", "2", "22", "18", "~
$ CONTROL         <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 3, 1, 1, 2, 1, 3, ~
$ ADM_RATE        <chr> "0.8965", "0.806", "NULL", "0.7711", "0.9888", "0.8039~
$ SATVR25         <chr> "430", "560", "NULL", "590", "438", "540", "NULL", "NU~
$ SATVR75         <chr> "520", "668", "NULL", "700", "531", "660", "NULL", "NU~
$ SATMT25         <chr> "410", "530", "NULL", "580", "406", "530", "NULL", "NU~
```

```
$ SATMT75         <chr> "500", "660", "NULL", "730", "518", "670", "NULL", "NU~
$ SATWR25         <chr> "370", "NULL", "NULL", "NULL", "NULL", "480", "NULL", ~
$ SATWR75         <chr> "457", "NULL", "NULL", "NULL", "NULL", "600", "NULL", ~
$ SATVRMID        <chr> "475", "614", "NULL", "645", "485", "600", "NULL", "NU~
$ SATMTMID        <chr> "455", "595", "NULL", "655", "462", "600", "NULL", "NU~
$ SATWRMID        <chr> "414", "NULL", "NULL", "NULL", "NULL", "540", "NULL", ~
$ ACTCM25         <chr> "15", "22", "NULL", "24", "14", "23", "NULL", "NULL", ~
$ ACTCM75         <chr> "20", "30", "NULL", "31", "20", "31", "NULL", "NULL", ~
$ ACTEN25         <chr> "14", "22", "NULL", "24", "14", "23", "NULL", "NULL", ~
$ ACTEN75         <chr> "20", "33", "NULL", "33", "20", "33", "NULL", "NULL", ~
$ ACTMT25         <chr> "15", "20", "NULL", "23", "14", "21", "NULL", "NULL", ~
$ ACTMT75         <chr> "18", "27", "NULL", "29", "20", "29", "NULL", "NULL", ~
$ ACTWR25         <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "7", "NULL", "~
$ ACTWR75         <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "8", "NULL", "~
$ ACTCMMID        <chr> "18", "26", "NULL", "28", "17", "27", "NULL", "NULL", ~
$ ACTENMID        <chr> "17", "28", "NULL", "29", "17", "28", "NULL", "NULL", ~
$ ACTMTMID        <chr> "17", "24", "NULL", "26", "17", "25", "NULL", "NULL", ~
$ ACTWRMID        <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "8", "NULL", "~
$ SAT_AVG         <chr> "959", "1245", "NULL", "1300", "938", "1262", "NULL", ~
$ UGDS            <chr> "5090", "13549", "298", "7825", "3603", "30610", "994"~
$ UG              <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "NULL", "NULL"~
$ UGDS_WHITE      <chr> "0.0159", "0.5496", "0.255", "0.7173", "0.0167", "0.76~
$ UGDS_BLACK      <chr> "0.9022", "0.2401", "0.6913", "0.0907", "0.9265", "0.1~
$ UGDS_HISP       <chr> "0.0116", "0.061", "0.0268", "0.0599", "0.013", "0.051~
$ UGDS_ASIAN      <chr> "0.0012", "0.0704", "0.0034", "0.0354", "0.0019", "0.0~
$ UGDS_AIAN       <chr> "0.0028", "0.0024", "0", "0.0083", "0.0017", "0.0033",~
$ UGDS_NHPI       <chr> "0.0008", "0.0004", "0", "0.001", "0.0017", "0.0008", ~
$ UGDS_2MOR       <chr> "0.0143", "0.0469", "0", "0.0431", "0.0119", "0.0359",~
$ UGDS_NRA        <chr> "0.0073", "0.0232", "0", "0.019", "0.0155", "0.0187", ~
$ UGDS_UNKN       <chr> "0.044", "0.0059", "0.0235", "0.0252", "0.0111", "0.00~
$ NPT4_PUB        <chr> "15529", "16530", "NULL", "17208", "19534", "20917", "~
$ NPT4_PRIV       <chr> "NULL", "NULL", "17618", "NULL", "NULL", "NULL", "NULL~
$ COSTT4_A        <chr> "23445", "25542", "20100", "24861", "21892", "30016", ~
$ COSTT4_P        <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "NULL", "NULL"~
$ AVGFACSAL       <chr> "7599", "11380", "4545", "9697", "7194", "10349", "658~
$ PCTPELL         <chr> "0.7095", "0.3397", "0.7452", "0.2403", "0.7368", "0.1~
$ C150_4          <chr> "0.2866", "0.6117", "0.25", "0.5714", "0.3177", "0.721~
$ AGE_ENTRY       <chr> "20.28374137", "23.60797466", "33.6722973", "22.727919~
$ FEMALE          <chr> "0.564030132", "0.63909074", "0.648648649", "0.4763499~
$ MARRIED         <chr> "0.009102323", "0.105086641", "0.236486487", "0.100460~
$ FIRST_GEN       <chr> "0.365828092", "0.341223671", "0.5125", "0.310132159",~
$ FAMINC          <chr> "32362.82611", "51306.67431", "21079.47297", "61096.58~
$ MD_FAMINC       <chr> "23553", "34489", "15033.5", "44787", "22080.5", "6673~
```

```
$ MN_EARN_WNE_P10 <chr> "35500", "48400", "47600", "52000", "30600", "51600", ~
$ MD_EARN_WNE_P10 <chr> "36339", "46990", "37895", "54361", "32084", "52751", ~
$ ENDOWBEGIN      <chr> "NULL", "537349307", "174805", "77250279", "94536751",~
```

```
colleges <- niche_data_500 |>
  left_join(us_dep_of_ed, by = c("college" ="INSTNM"))

glimpse(colleges)
```

```
Rows: 500
Columns: 64
$ college         <chr> "Massachusetts Institute of Technology", "Stanford Uni~
$ rank            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
$ UNITID          <dbl> 166683, 243744, 166027, 130794, 186131, 227757, 110404~
$ OPEID           <chr> "217800", "130500", "215500", "142600", "262700", "360~
$ OPEID6          <chr> "2178", "1305", "2155", "1426", "2627", "3604", "1131"~
$ CITY            <chr> "Cambridge", "Stanford", "Cambridge", "New Haven", "Pr~
$ STABBR          <chr> "MA", "CA", "MA", "CT", "NJ", "TX", "CA", "NC", "RI", ~
$ ZIP             <chr> "02139-4307", "94305", "2138", "6520", "08544-0070", "~
$ ACCREDAGENCY    <chr> "New England Commission on Higher Education", "Western~
$ LATITUDE        <chr> "42.359243", "37.429434", "42.374471", "41.311158", "4~
$ LONGITUDE       <chr> "-71.093226", "-122.167359", "-71.118313", "-72.926688~
$ REGION          <dbl> 1, 8, 1, 1, 2, 6, 8, 5, 1, 1, 2, NA, 5, 3, NA, 3, 2, 8~
$ CCBASIC         <chr> "15", "15", "15", "15", "15", "15", "15", "15", "15", ~
$ CONTROL         <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, NA, 2, 2, NA, 2, 2, 2~
$ ADM_RATE        <chr> "0.0726", "0.0519", "0.0501", "0.0653", "0.0563", "0.1~
$ SATVR25         <chr> "730", "700", "720", "720", "710", "710", "740", "720"~
$ SATVR75         <chr> "780", "770", "780", "780", "770", "770", "780", "770"~
$ SATMT25         <chr> "780", "720", "740", "740", "740", "750", "790", "750"~
$ SATMT75         <chr> "800", "800", "800", "800", "800", "800", "800", "800"~
$ SATWR25         <chr> "690", "690", "710", "710", "710", "680", "730", "690"~
$ SATWR75         <chr> "780", "780", "790", "790", "790", "770", "800", "780"~
$ SATVRMID        <chr> "755", "735", "750", "750", "740", "740", "760", "745"~
$ SATMTMID        <chr> "790", "760", "770", "770", "770", "775", "795", "775"~
$ SATWRMID        <chr> "735", "735", "750", "750", "750", "725", "765", "735"~
$ ACTCM25         <chr> "34", "31", "33", "33", "32", "34", "35", "34", "33", ~
$ ACTCM75         <chr> "36", "35", "35", "35", "35", "36", "36", "35", "35", ~
$ ACTEN25         <chr> "35", "33", "35", "34", "34", "34", "35", "35", "34", ~
$ ACTEN75         <chr> "36", "36", "36", "36", "36", "36", "36", "36", "36", ~
$ ACTMT25         <chr> "34", "30", "31", "31", "31", "32", "35", "32", "30", ~
$ ACTMT75         <chr> "36", "35", "35", "35", "35", "35", "36", "35", "35", ~
```

```
$ ACTWR25         <chr> "8", "NULL", "8", "NULL", "8", "8", "8", "8", "NULL", ~
$ ACTWR75         <chr> "10", "NULL", "10", "NULL", "10", "10", "10", "10", "N~
$ ACTCMMID        <chr> "35", "33", "34", "34", "34", "35", "36", "35", "34", ~
$ ACTENMID        <chr> "36", "35", "36", "35", "35", "35", "36", "36", "35", ~
$ ACTMTMID        <chr> "35", "33", "33", "33", "33", "34", "36", "34", "33", ~
$ ACTWRMID        <chr> "9", "NULL", "9", "NULL", "9", "9", "9", "9", "NULL", ~
$ SAT_AVG         <chr> "1550", "1491", "1520", "1520", "1506", "1533", "1566"~
$ UGDS            <chr> "4360", "6366", "6099", "4701", "4688", "4052", "901",~
$ UG              <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "NULL", "NULL"~
$ UGDS_WHITE      <chr> "0.2562", "0.29", "0.3596", "0.3493", "0.3631", "0.307~
$ UGDS_BLACK      <chr> "0.067", "0.0743", "0.1058", "0.091", "0.0892", "0.076~
$ UGDS_HISP       <chr> "0.1557", "0.17", "0.123", "0.1504", "0.1064", "0.1604~
$ UGDS_ASIAN      <chr> "0.3236", "0.248", "0.2066", "0.2389", "0.2515", "0.27~
$ UGDS_AIAN       <chr> "0.0018", "0.0097", "0.0025", "0.0028", "0.0019", "0.0~
$ UGDS_NHPI       <chr> "0.0007", "0.0035", "0.0016", "0.0011", "0.0011", "0.0~
$ UGDS_2MOR       <chr> "0.0764", "0.0957", "0.0689", "0.0623", "0.0591", "0.0~
$ UGDS_NRA        <chr> "0.1028", "0.1067", "0.1123", "0.1004", "0.1165", "0.1~
$ UGDS_UNKN       <chr> "0.0158", "0.0022", "0.0198", "0.0038", "0.0113", "0.0~
$ NPT4_PUB        <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "NULL", "NULL"~
$ NPT4_PRIV       <chr> "16407", "12894", "13872", "15296", "9836", "16076", "~
$ COSTT4_A        <chr> "73160", "74570", "75914", "76645", "74150", "67102", ~
$ COSTT4_P        <chr> "NULL", "NULL", "NULL", "NULL", "NULL", "NULL", "NULL"~
$ AVGFACSAL       <chr> "19640", "20971", "21143", "19490", "20835", "16378", ~
$ PCTPELL         <chr> "0.1896", "0.1762", "0.1133", "0.184", "0.2136", "0.15~
$ C150_4          <chr> "0.9558", "0.9523", "0.9763", "0.9639", "0.9802", "0.9~
$ AGE_ENTRY       <chr> "19.34846267", "19.70933014", "22.46358974", "19.53214~
$ FEMALE          <chr> "0.371888726", "0.494019139", "0.501538462", "0.500683~
$ MARRIED         <chr> "PrivacySuppressed", "PrivacySuppressed", "0.063589744~
$ FIRST_GEN       <chr> "0.258513932", "0.303385417", "0.25708061", "0.25", "0~
$ FAMINC          <chr> "86738.81698", "80447.9067", "62458.19385", "80258.132~
$ MD_FAMINC       <chr> "53870", "44842", "33066", "44004", "37036", "65041", ~
$ MN_EARN_WNE_P10 <chr> "153600", "141300", "139100", "124400", "116300", "916~
$ MD_EARN_WNE_P10 <chr> "111222", "97798", "84918", "88655", "95689", "77683",~
$ ENDOWBEGIN      <chr> "17443750000", "27699834000", "40929700000", "30314816~
```