Project 3

Introduction:

The project dataset I used is about college admissions that tends to study the level of economic segregation in different universities. The dataset contains universities names, parental income in percentiles, attend, and interaction variables like attend_sat, rel_apply_sat, and etc. I want to explore the question of:

What different characteristics do students from each UC school have?

The explanatory variables I decide to include are:

1. par_income_bin which is the percentile of parent's household income;
2. attend which stands for the fraction of students attending this university among all test_takers with similar parental income.
3. rel_apply which shows given the students' test scores, how much more or less likely are they to apply to this school compared to average college.
4. rel_apply_sat: relative application rate for a specific test score band.
5. attend_instate: attendance rate for in state students.
6. attend_instate_sat: absolute estimates on a specific test score for in-state students. Only available for public schools.
7. attend_oostate: attendance rate for out of state students.
8. attend_oostate_sat: absolute estimates on a specific test score for out-of-state students. Only available for public schools.
9. attend_sat: absolute attendance rate for specific test score band.

Question: What different characteristics do students from each UC school have?

Approach:

In order to study what different characteristics do students from each UC school have, I ran data wrangling to filter out schools in the UC system. Then, I selected some explanatory variables that are significant to school enrollment: par_income_bin, attend, rel_apply, rel_apply_sat, attend_instate, attend_instate_sat, attend_oostate, attend_oostate_sat, attend_sat. After that, I chose to do PCA analysis becuase by using the loading as the element of eigenvector, I am able to see how much each variable contributes to each principle component.

I ran PCA analysis after filtering out numeric columns and standardize each column. After naming the PCA analysis pca_fit, I created a rotation matrix of loadings.

The next step is creating a scatter plot of fitted PC2 versus fitted PC1 because scatterplot allows me to get more information on how each observations in each school are at with PC1 being the x axis and PC2 at the y axis. It allows me to deeper connect with the rotation matrix. I colored the students from different UC schools: with dark blue being the best school in UC system, which

is UCLA, through the lightest version of blue representing the 5th best school in UC system, UC Irvine. After the 5th place, UC Santa Babara is colored light pink, and the worst school is colored darker pink.

Lastly, I align the rotation matrix and scatter plot side by side for easier comparison and analysis.

Analysis:

```
college_admissions <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/2024/2024-09-10/college_admissions.csv')
```

```
Rows: 1946 Columns: 80
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (4): name, par_income_lab, tier, test_band_tier
dbl (74): super_opeid, par_income_bin, attend, stderr_attend, attend_level, ...
lgl  (2): public, flagship

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.
```

```
cal_college<- college_admissions |>
  na.omit() |>
  filter(str_detect(name, "University Of California", ))|>
  select(2,3,5,8, 11,16,21,24,27,30)

pca_fit <- cal_college |>
  select(where(is.numeric)) |>
  scale() |>
  prcomp()

#Rotation Plot
pca_fit |>
  tidy(matrix="loadings")
```

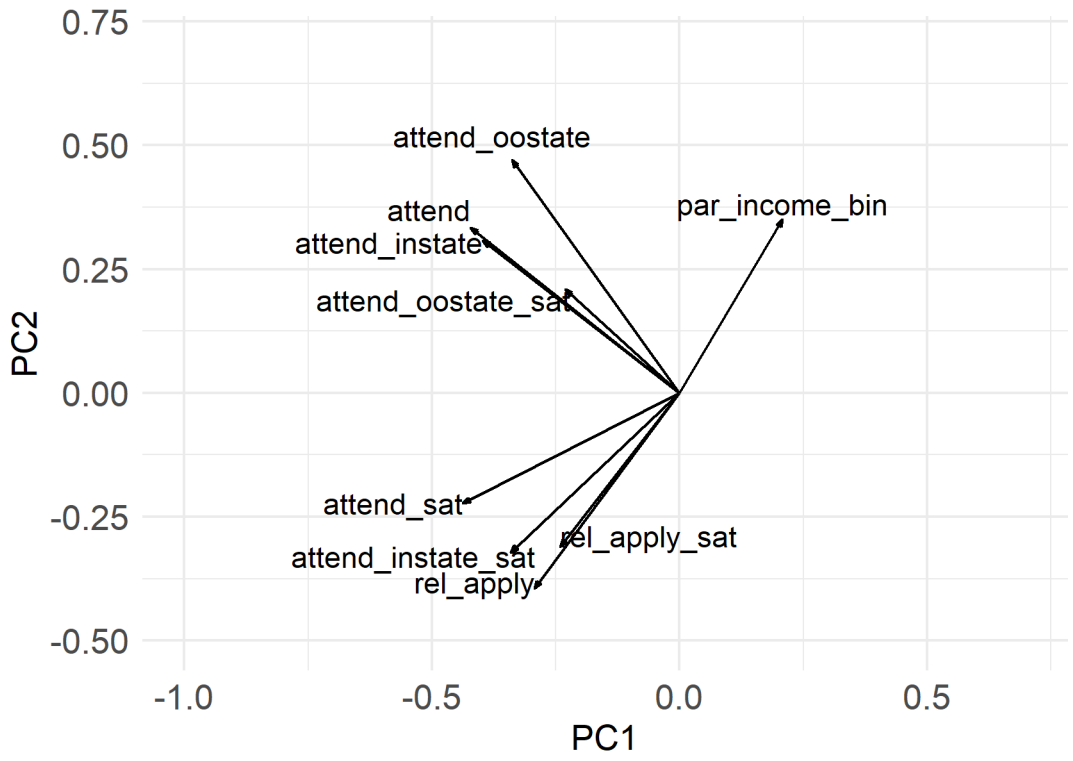
```
# A tibble: 81 × 3
   column            PC   value
   <chr>          <dbl>   <dbl>
 1 par_income_bin     1  0.208
 2 par_income_bin     2  0.351
 3 par_income_bin     3 -0.485
 4 par_income_bin     4  0.0200
 5 par_income_bin     5  0.755
 6 par_income_bin     6 -0.0795
 7 par_income_bin     7  0.0851
```

```
 8 par_income_bin     8 -0.0941
 9 par_income_bin     9 -0.0719
10 attend             1 -0.422
# ℹ 71 more rows
```

```
arrow_style <- arrow(
  angle=20, length= grid::unit(3, "pt"),
  ends="first", type="closed"
)

p1<- pca_fit |>
  tidy(matrix = "rotation") |>  # extract rotation matrix
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) |>
  mutate(
     hjust = c(0.5, 1, 1, 1, 0, 1, 0.9,0.6,0.98),
    vjust = c(-0.2, -0.3, 0.5, 0.2, 0,0.6, 0.7, -0.6,1)
  ) |>
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text(aes(
    label = column,
    hjust = hjust,
    vjust= vjust)) +
  coord_fixed(
    xlim = c(-1, .7),
    ylim = c(-0.5, 0.7)
  )+
  theme_minimal()+
  theme(
    axis.text=element_text(size=13),
    axis.title=element_text(size=13)
  )

p1
```

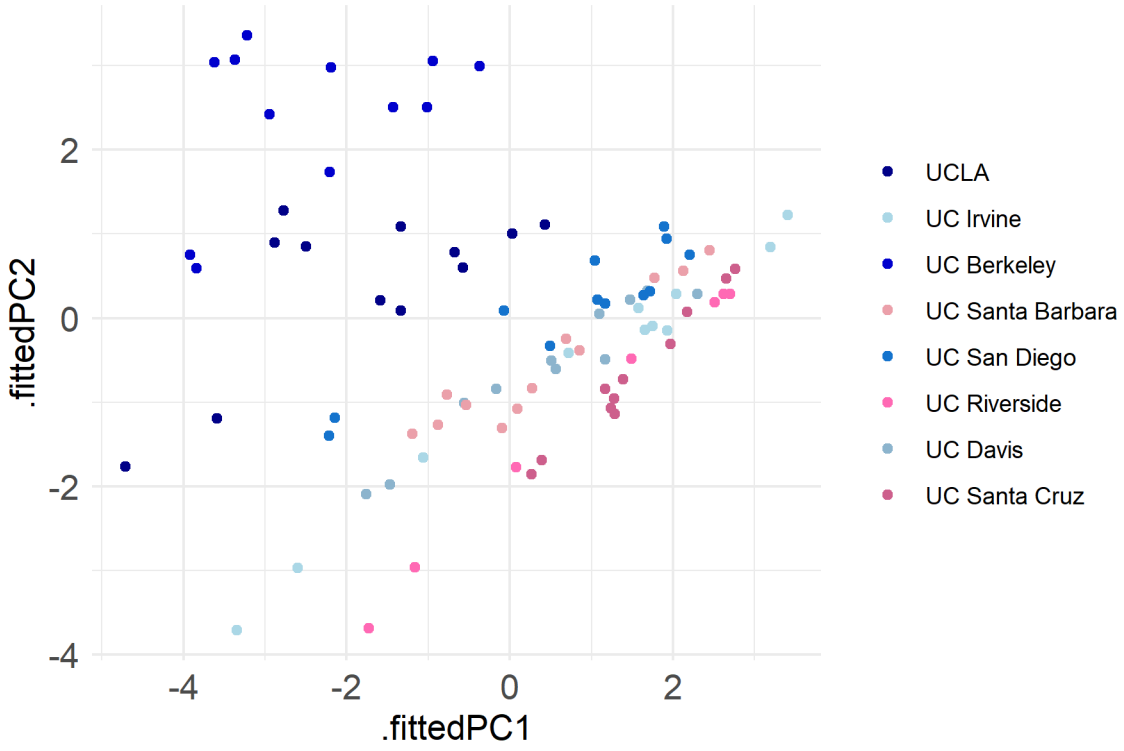



```
#Scatterplot
p2<- pca_fit |>
  augment(cal_college) |>
  ggplot(aes(.fittedPC1, .fittedPC2))+
  geom_point(aes(color=factor(name)))+
  scale_color_manual(
    name=NULL,
    values=c(
      "University Of California, Los Angeles"="darkblue",
      "University Of California, Berkeley"="blue3",
      "University Of California, San Diego"="dodgerblue3",
      "University Of California, Davis"="lightskyblue3",
      "University Of California, Irvine"="lightblue",
      "University Of California, Santa Barbara"="lightpink2",
      "University Of California, Riverside"="hotpink",
      "University Of California, Santa Cruz"="hotpink3"
      ),
    labels=c(
      "University Of California, Los Angeles"="UCLA",
      "University Of California, Berkeley"="UC Berkeley",
      "University Of California, San Diego"="UC San Diego",
      "University Of California, Davis"="UC Davis",
      "University Of California, Irvine"="UC Irvine",
      "University Of California, Santa Barbara"="UC Santa Barbara",
```

```
      "University Of California, Riverside"="UC Riverside",
      "University Of California, Santa Cruz"="UC Santa Cruz"
      ),
    breaks=c(
      "University Of California, Los Angeles",
      "University Of California, Irvine",
      "University Of California, Berkeley",
      "University Of California, Santa Barbara",

      "University Of California, San Diego",


      "University Of California, Riverside",
      "University Of California, Davis",
      "University Of California, Santa Cruz")
  )+
   theme_minimal()+
  theme(legend.position="right",
        axis.text=element_text(size=13),
        axis.title=element_text(size=13)
        )
p2
```



Discussion:

Observing the scatter plot, we can notice that all the better schools are aligned at the upper left of the scatter plot, and the schools that do not perform as well are located at the middle right corner of the scatterplot, forming a big cluster of observations. According to the rotation matrix, the attend, attendance rate from out of state and in state, and attend_oostate_sat all point to the upper left corner, and only parental income variable points to the upper right direction.

Combining findings from both plots, I find that among all test takers with similar parental income, students are likely to attend UCLA and UC Berkeley. The same applies to both in state and out of state students: they tend to choose to attend UCLA and UC Berkeley. However, there is something interesting: the variable attend_level_oostate_sat, which is the estimate of attendance rate on a specific test score band for out-of-state students, points to the upper left corver as well. But attend_instate_sat does not point to the same direction. This implies that the top 2 schools in the UC system emphasize more heavily on out of state students' SAT score compared to in state students.

On the other hand, parental income seems to play a huge role in UC San Diego, UC Davis, UC Irvine, UC Santa Barbara, UC Riverside, and UC Santa Cruz students' characteristics, and it is the only factor the PCA shows. To put it in another words, the students from these schools tend to have better parental incomes, and these schools do not emphasize on test scores that musch, and they do not have a outstanding attendance rate.