Homework 9

Ethan Chang - ehc586

This homework is due on April 11, 2023 at 11:00pm. Please submit as a pdf file on Canvas.

For all problems in this homework, we will work with the `heart_disease_data` dataset, which is a simplified and recoded version of a dataset available from kaggle. You can read about the original dataset here: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download

The `heart_disease_data` dataset contains 9 variables: `HeartDisease`(whether or not the participant has heart disease), `BMI` (body mass index), `PhysicalHealth` (how many days a month was their physical health not good), `MentalHealth` (how many days a month was their mental health not good), `ApproximateAge` (participants age), `SleepTime` (how many hours of sleep do they get in a 24-hour period), `Smoking` (1-smoker, 0-nonsmoker), `AlcoholDrinking` (1-drinks alcohol, 0-does not drink), `PhysicalActivity` (1-did physical activity or exercise during the past 30 days, 0-hardly any physical activity). Compared to the original dataset, the columns `ApproximateAge`, `Smoking`, `AlcoholDrinking`, and `PhysicalActivity` have been converted into numeric columns so they can be included in a PCA.

Note: This homework is about the contents of the plots. Don't worry about styling. It's OK to use the default theme and plot labeling.

```r
heart_data <- read_csv("https://wilkelab.org/SDS375/datasets/heart_disease_data.csv")
```

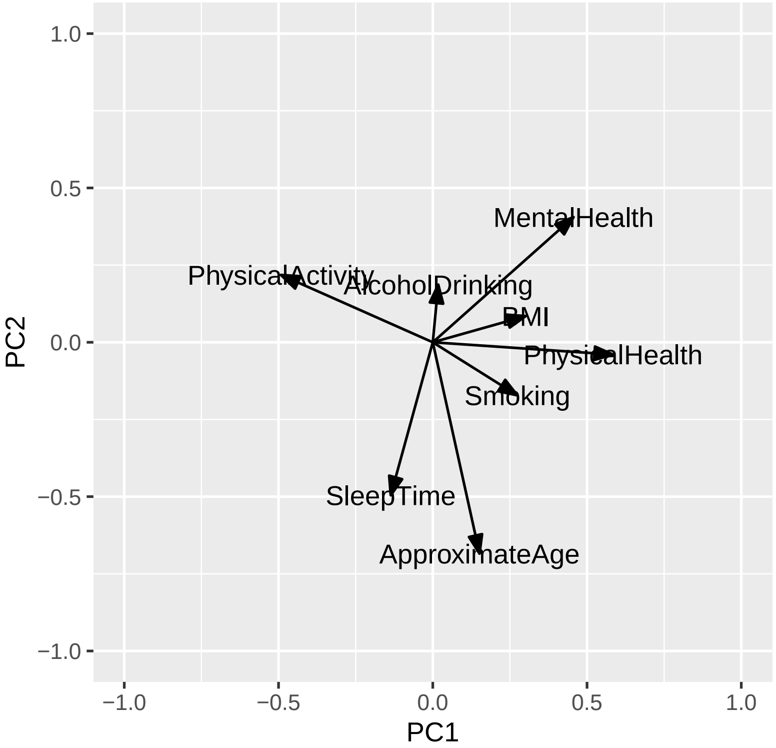
Problem 1: (5 pts)

Perform a PCA of the `heart_disease_data` dataset and make two plots, a rotation plot of components 1 and 2 and a plot of the eigenvalues, showing the amount of variance explained by the various components.

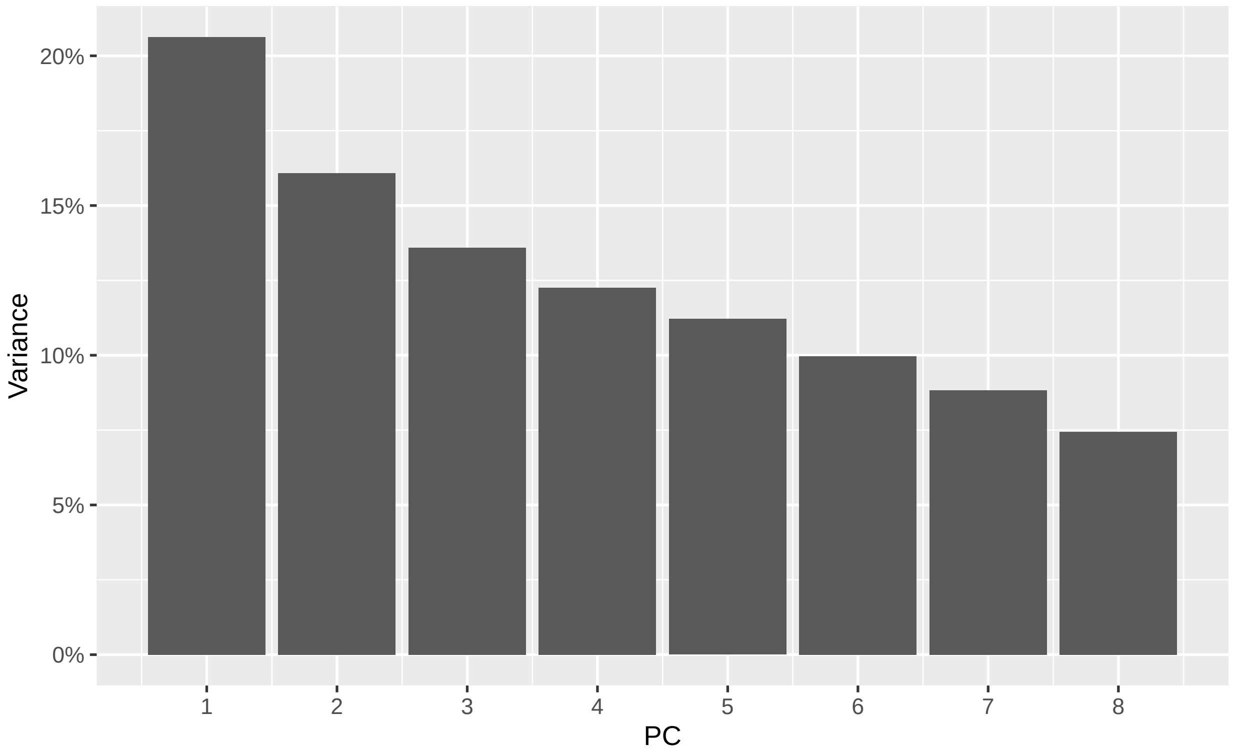
```r
pca_fit <- heart_data %>%
  select(where(is.numeric)) %>%
  scale() %>%
  prcomp()

arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)

pca_fit %>%
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text(aes(label = column)) +
  xlim(-1, 1) + ylim(-1, 1) +
  coord_fixed()
```

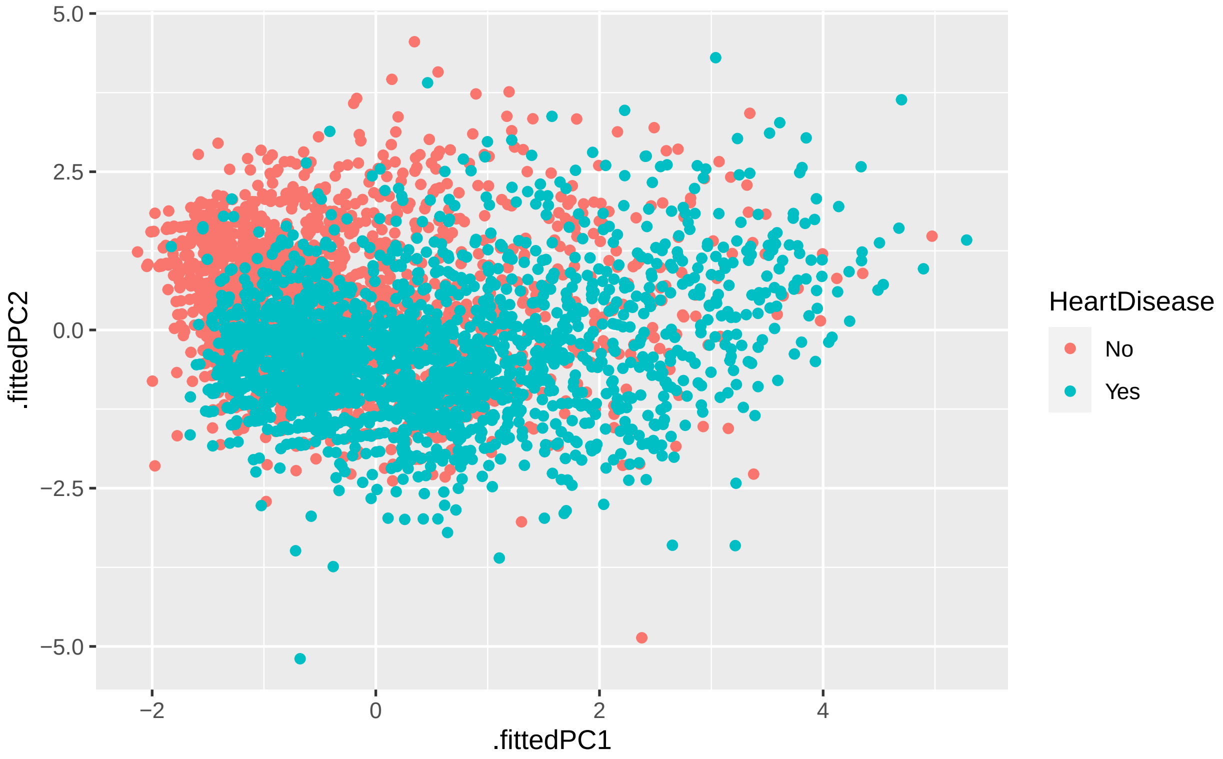


```
pca_fit %>%
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(PC, percent)) +
  geom_col() +
  scale_x_continuous(
    breaks = 1:8
  ) +
  scale_y_continuous(
    name = "Variance",
    label = scales::label_percent(accuracy = 1)
  )
```



Problem 2: (5 pts) Make a scatter plot of PC 2 versus PC 1 and color by heart disease status. Then use the rotation plot from Problem 1 to describe the variables/factors by which we can separate the study participants with heart disease from the study participants without heart disease.

```
pca_fit %>%
  augment(heart_data) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = HeartDisease))
```



When looking at the given plot and comparing it to the rotation plot from problem 1, it can be seen that variables/factors that could be used to separate the study participants with heart disease from those without heart disease are physical activity and smoking. The majority of patients without heart disease tend to reside in the upper left portion whereas those with heart disease tend to reside in the bottom right portion. As such, the variables with the greatest impact are the ones that reside in those directions: physical activity and smoking. Patients with more physical activity and less smoking appear to be less likely to suffer from heart disease while those with less physical activity and more smoking are more likely to have heart disease.