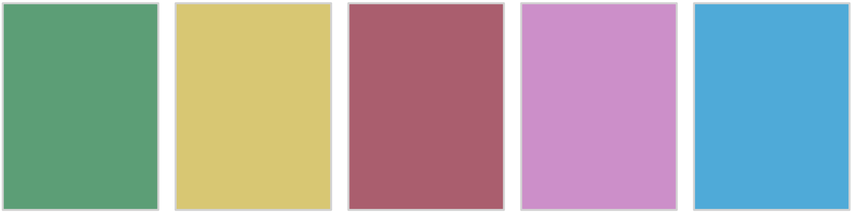
PCA and Rotation plots of components 1 & 2

Problem 1: Use the color picker app from the **colorspace** package (`colorspace::choose_color()`) to create a qualitative color scale containing five colors. One of the five colors should be `#5C9E76`, so you need to find four additional colors that go with this one.

```
# colors from colorspace
colors <- c('#5C9E76', '#D8C773', '#AA5E6E', '#CC8FC9', '#4FAAD8')
swatchplot(colors)
```



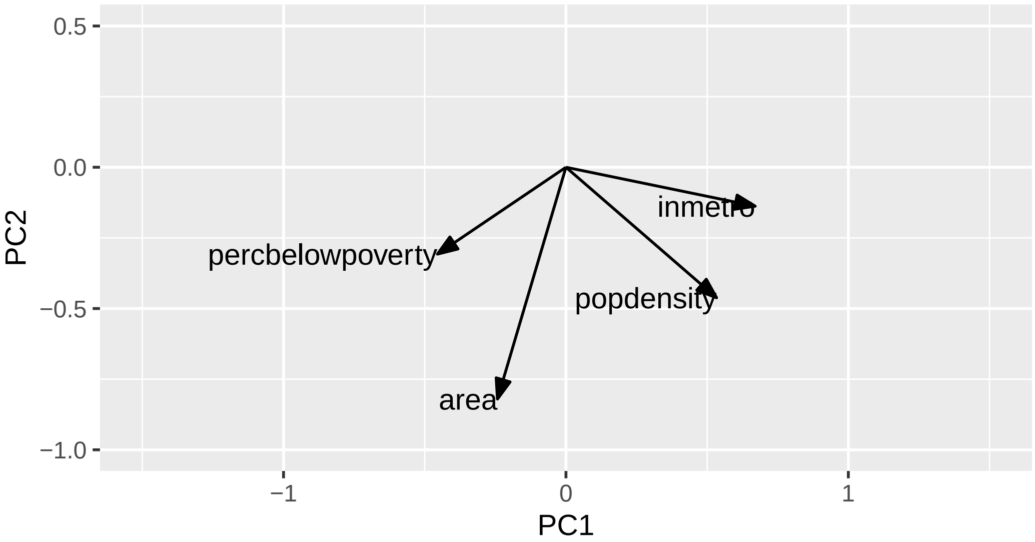
For the rest of this homework, we will be working with the `midwest_clean` dataset, which is a cleaned up version of the **ggplot2** `midwest` dataset.

```
midwest_clean <- midwest %>%
  select(
    state, county, area, popdensity, percbelowpoverty, inmetro
  ) %>%        # keep only a subset of data
  na.omit()    # remove any rows with missing data
```

Problem 2: Perform a PCA of the `midwest_clean` dataset and make a rotation plot of components 1 and 2.

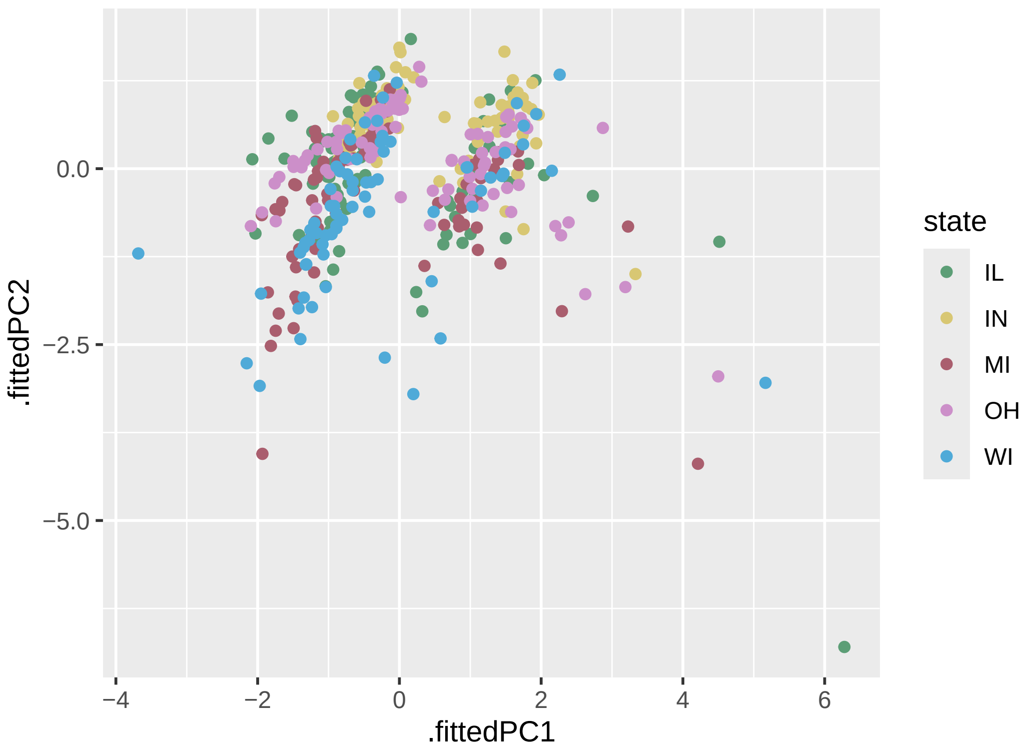
```
# running the PCA and storing the result
pca_fit <- midwest_clean %>%
  select(where(is.numeric)) %>% # retain only numeric columns
  scale() %>%                   # scale to zero mean and unit variance
  prcomp()

# plotting the rotation matrix
arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)
```

```
pca_fit %>%
  # extract rotation matrix
  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), hjust = 1) +
  xlim(-1.5, 1.5) + ylim(-1, 0.5) +
  coord_fixed()
```



Problem 3: Make a scatter plot of PC 2 versus PC 1 and color by state. You should use the custom colorscale you created in Problem 1. Then use the rotation plot from Problem 2 to describe where Chicago, Illinois can be found on the scatter plot. Provide any additional evidence used to support your answer.

```
# plotting PC 2 versus PC 1
pca_fit %>%
  # add PCs to the original dataset
  augment(midwest_clean) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = state)) +
  scale_color_manual(values = colors)
```



Answer: *Based on the rotation plot we generated, the population density (popdensity) variable is pointing to bottom right hand side in the projected space of the rotated coordinate system. This dataset does not contain city information. We will rely on county information to help us decide where Chicago is located in the plot. As we know, Chicago is the most populous city in Illinois and it is located in Cook County. There is only one data point in the plot that is at the very bottom right hand side and we can safely say that that data point belongs to Cook county. We may further prove it by adding a line in the above code* **(geom_text(aes(label = county)))** *and print the county name on the plot.*