Homework 5

This homework is due on the deadline posted on edX. Please submit a .pdf file of your output and upload a .zip file with your .Rmd file.

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.

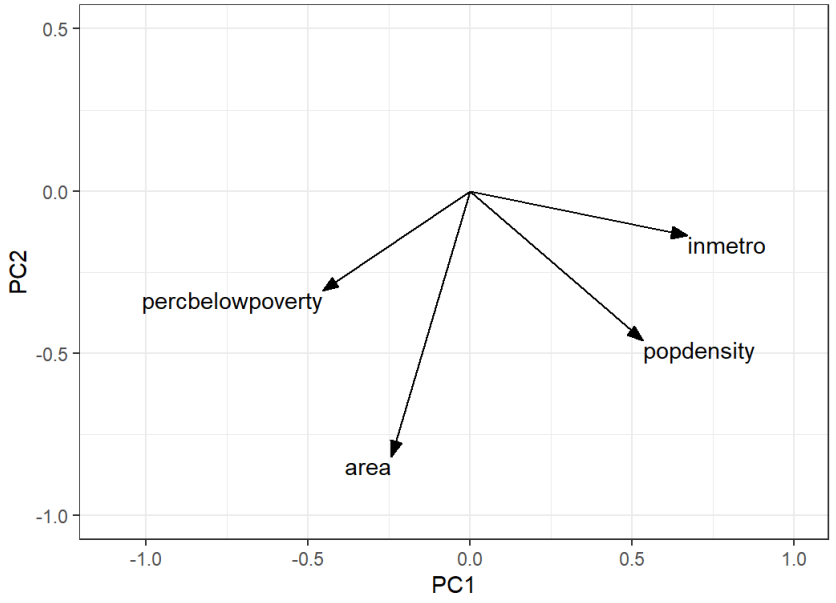
```
# replace "#FFFFFF" with your own colors
colors <- c("#5C9E76", "#5499B4", "#A185BE", "#C17BA7", "#C59A4A")
swatchplot(colors)
```

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
```

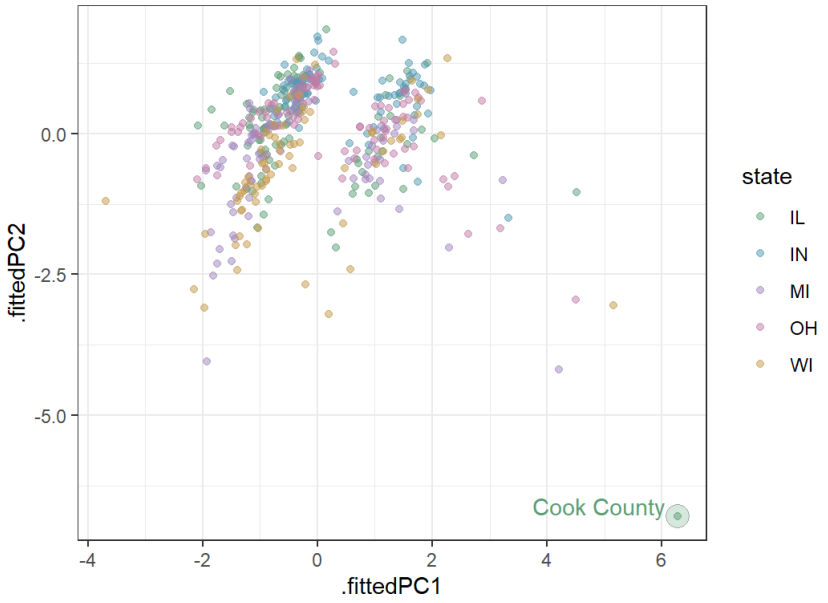
Problem 2: Perform a PCA of the `midwest_clean` dataset and make a rotation plot of components 1 and 2.

```
# Rotate Coordinates
pca_fit <- midwest_clean %>%
  select(where(is.numeric)) %>% #select only numeric columns (skip if all numeric)
  scale() %>% #scale to zero mean and unit variance
  prcomp() #do PC

# Plot the Rotation Matrix
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
  ) +
  xlim(c(-1.1, 1)) +
  ylim(c(-1, .5)) +
  geom_text(aes(label = column), hjust = c(1,0,1,0),vjust = 1) +
  coord_fixed() +
  theme_bw()
```



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.

```
# Isolate the county containing Chicago for highlighting
chicago <- pca_fit %>%
  augment(midwest_clean) %>%
  filter(county=='COOK')

pca_fit %>%
  augment(midwest_clean) %>%
  ggplot(aes(.fittedPC1, .fittedPC2, color = state)) +
  geom_point(alpha = 0.5) +
  scale_color_manual(values = colors) +
  geom_point( #highlight point representing Chicago (Cook County)
    data = chicago,
    aes(.fittedPC1, .fittedPC2),
    color='black',
    fill = colors[1],
    size = 5,
    alpha = 0.25,
    shape = 21) +
  geom_text(
    data = chicago,
    aes(label = 'Cook County'),
    hjust = 1.1,
    vjust = -.1,
    show.legend = FALSE) +
theme_bw()
```




```
chicago %>%
  select(-.rownames)
```

```
## # A tibble: 1 x 10
##   state county  area popdensity percbelowpoverty inmetro .fittedPC1 .fittedPC2
##   <chr> <chr>  <dbl>      <dbl>            <dbl>   <int>      <dbl>      <dbl>
## 1 IL    COOK   0.058     88018.             14.2       1       6.28      -6.80
## # ... with 2 more variables: .fittedPC3 <dbl>, .fittedPC4 <dbl>
```

```
summary(midwest_clean)
```

```
##     state              county              area           popdensity
##  Length:437         Length:437         Min.   :0.00500   Min.   :   85.05
##  Class :character   Class :character   1st Qu.:0.02400   1st Qu.:  622.41
##  Mode  :character   Mode  :character   Median :0.03000   Median : 1156.21
##                                        Mean   :0.03317   Mean   : 3097.74
##                                        3rd Qu.:0.03800   3rd Qu.: 2330.00
##                                        Max.   :0.11000   Max.   :88018.40
##  percbelowpoverty    inmetro
##  Min.   : 2.180   Min.   :0.0000
##  1st Qu.: 9.199   1st Qu.:0.0000
##  Median :11.822   Median :0.0000
##  Mean   :12.511   Mean   :0.3432
##  3rd Qu.:15.133   3rd Qu.:1.0000
##  Max.   :48.691   Max.   :1.0000
```

Given the Chicago is in Cook county and filtering by that value, we see that the county has a population density of 88,018 residents. Looking at the graph, we see that this is the rightmost point located about halfway up and shown in green. This most likely also corresponds with the outlier on the bottom right of the above rotation plot from Problem 2.