Project 3

Ethan Chang - ehc586

Introduction: For this project, we work with the Weather Forecast Accuracy dataset, a dataset consisting of information from the USA National Weather Service with 16 months of forecasts and observations from 167 cities. More specifically though, we look at the `cities` subset of this dataset which contains information about the cities. Each row in the dataset corresponds to a city with multiple columns denoting that city's geography and climate. More specifically, the dataset contains a city's name, residing state, longitude, latitude, Koppen climate classification, elevation, its distance to a coast, average wind speed, greatest elevation change of the four and eight closest points, and average annual precipitation.

To answer the provided question, we work with 7 main variables from the dataset, the `lon`gitude, `lat`itude, `koppen` climate classification, `elevation`, `distance_to_coast`, `wind` speed, and `avg_annual_precip`itation. The longitude, latitude, elevation (meters), distance to coast (miles), wind speed, and precipitation (inches) are all given as numeric values, more specifically doubles. Koppen was given as a character code that described the type of climate in the area.

```
# read in cities subset
cities <- readr::read_csv(
  'https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2022/2022-12-20/cities.csv
  show_col_types = FALSE
  )

# wrangle data for PCA fit
new_cities <- cities %>%
  # extract main Koppen climate group (first character from koppen)
  mutate(koppen_climate = substring(koppen, 1, 1)) %>%
  select(where(is.numeric),-elevation_change_four, -elevation_change_eight, koppen_climate) %>%
  drop_na()
```

Question: Which variables best differentiate between the main Koppen climate groups?

Approach: To answer this question, we would first wrangle the data by extracting the main koppen climate groups using `mutate()` and `substring()` to create the variable `koppen_climate` containing the first character of `koppen`. We would then select the newly created variable `koppen_climate` and all numeric variables for analysis, excluding `elevation_change_four` and `elevation_change_eight` as they do not provide any useful or interesting input for climate classification. Finally, we remove any observations with `NA` using `drop_na()` to clean up the data and avoid potential errors.

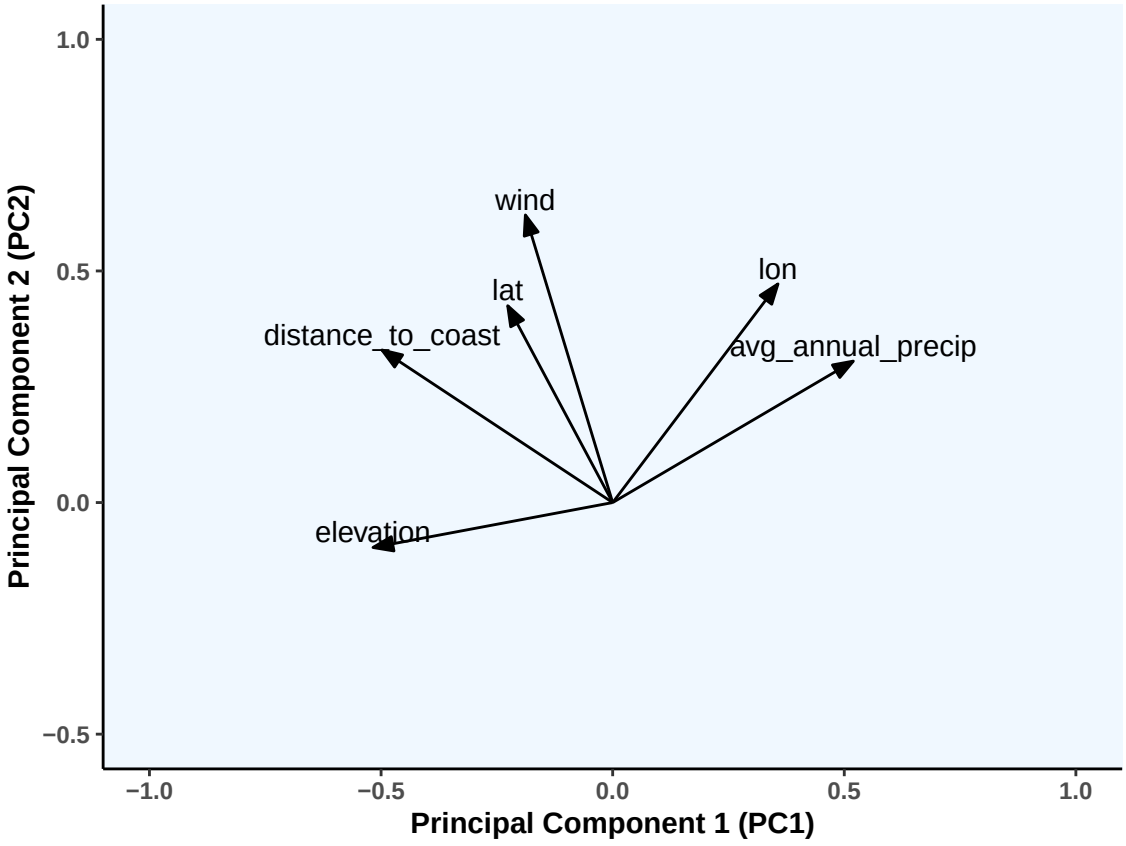
With our wrangled data, we could then answer the question using Principal Component Analysis (PCA), creating a rotation matrix to visualize the effects of each variable and a scatter plot to visualize the spread of the Koppen climate groups. Using a rotation matrix is best for identifying the effects of each variable, which is what we want to analyze with respect to Koppen climate groups. Using a scatter plot (`geom_point()`) is best for showing the spread of points to then analyze using the respective rotation matrix, allowing for easier comparison between the two plots. We color the points by the main Koppen climate group as that is what we want to compare between using the numerical variables. To connect the two plots, we plot the two graphs using principal components 1 and 2 (PC1 and PC2).

Analysis:

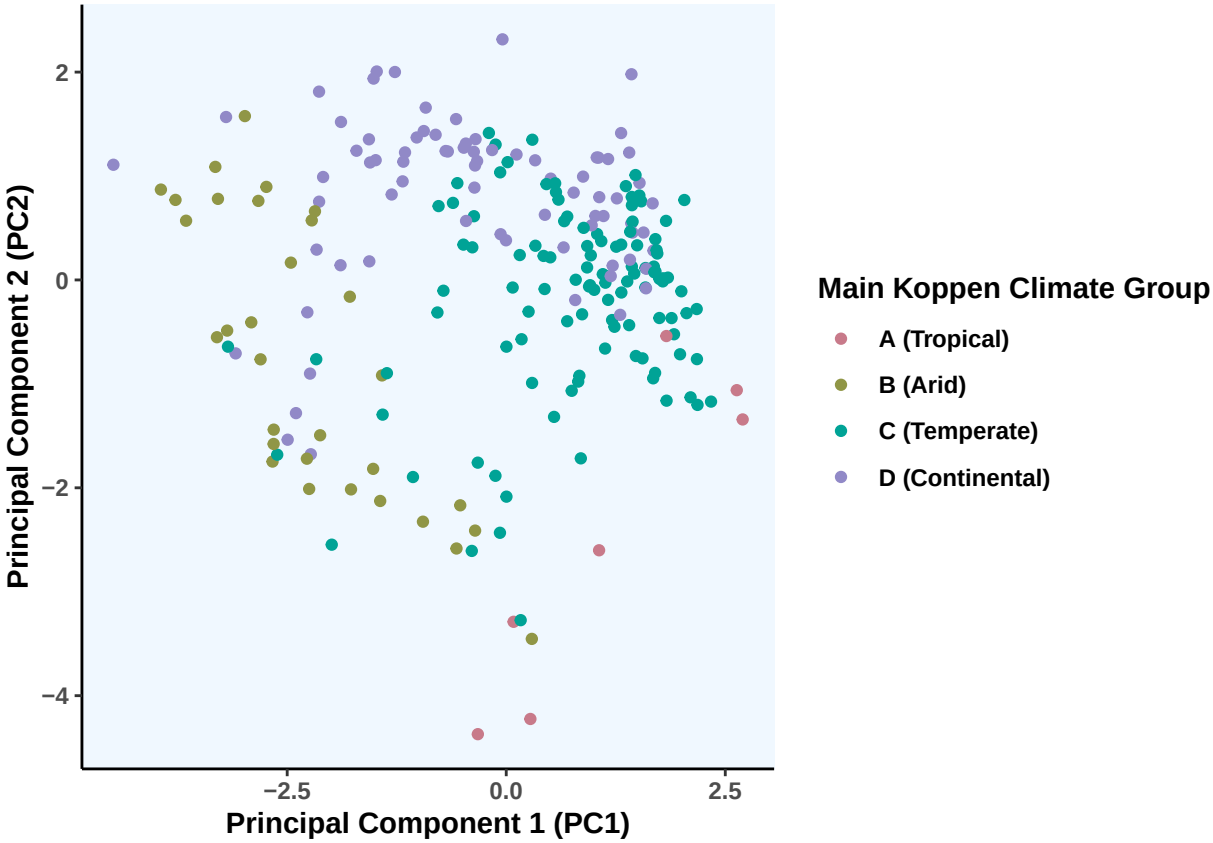
```r
# perform PCA
pca_fit <- new_cities %>%
  select(where(is.numeric)) %>%
  scale() %>%
  prcomp()

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

# plot rotation matrix
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), vjust = -0.25) +
  coord_fixed() +
  scale_x_continuous(name = "Principal Component 1 (PC1)", limits = c(-1, 1)) +
  scale_y_continuous(name = "Principal Component 2 (PC2)", limits = c(-0.5, 1)) +
  theme_classic() +
  theme(text = element_text(face = "bold"),
        plot.title = element_text(hjust = 0.5),
        panel.background = element_rect(fill = "aliceblue"))
```




```
# plot scatter plot
pca_fit %>%
  augment(new_cities) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = koppen_climate)) +
  scale_x_continuous(name = "Principal Component 1 (PC1)") +
  scale_y_continuous(name = "Principal Component 2 (PC2)") +
  scale_color_discrete_qualitative(
    name = "Main Koppen Climate Group",
    labels = c(A = "A (Tropical)", B = "B (Arid)", C = "C (Temperate)", D = "D (Continental)"),
    palette = "Dark 2"
  ) +
  theme_classic() +
  theme(
    text = element_text(face = "bold"),
    plot.title = element_text(hjust = 0.5),
    panel.background = element_rect(fill = "aliceblue")
  )
```



Discussion: From the rotation matrix created in figure 1, it can be seen that the wind speed, latitude, and longitude are best categorized in the PC2 group as they are more responsible for changes vertically whereas distance to coast, elevation, and annual precipitation are best categorized in the PC1 group as they are more responsible for changes horizontally. Additionally, distance to coast, latitude, and wind speed are more responsible for changes along the top left and bottom right areas of the plot while longitude, elevation, and annual precipitation are more responsible for changes along th top right and bottom left areas of the plot. When Looking at the scatter plot created in figure 2, it can be seen that the primary locations in the plot for each main Koppen climate group are A in the bottom-right, B in the left to top-left, C in the right to top-right, and D in the top. These groups can be categorized using the variables organized in the rotation matrix, with A seeming to have more annual precipitation, but lower wind speed, latitude, distance to coast, and elevation; B seeming to have less annual precipitation, but greater distance to coast and elevation; C seeming to have more annual precipitation and longitude, but lower elevation; and D seeming to have more annual precipitation, longitude, latitude, wind speed, and distance to coast, but lower elevation.

These variable categorizations make sense when considering the types of climates each climate group (A, B, C, D) is related to. Tropical climates tend to have more rain, are closer to the coast, leveled with the ground, and located in lower latitudes. Arid climates tend to have less rain, are farther from the coast, and higher up above sea level. Temperate climates tend to have more rain and are level with the ground. Finally, continental climates tend to have more rain and are farther from the coast with higher latitudes and longitudes, and are more level with the ground.