Homework 10

Enter your name and EID here Connor Hanna cdh3663

This homework is due on April 25, 2022 at 11:00am. Please submit as a pdf file on Canvas.

Problem 1: (3 pts) For Problem 1, we will be using `happiness` taken from the World Happiness Report. You can read more about the dataset here: https://www.kaggle.com/datasets/unsdsn/world-happiness.

```
# data preparation
happiness <- read_csv("https://wilkelab.org/SDS375/datasets/happiness.csv")
head(happiness)
```

```
## # A tibble: 6 x 9
##   country     happiness_score   GDP family_score health_life_expe~ freedom_score
##   <chr>                 <dbl> <dbl>        <dbl>             <dbl>         <dbl>
## 1 Switzerland            7.59  1.40         1.35             0.941         0.666
## 2 Iceland                7.56  1.30         1.40             0.948         0.629
## 3 Denmark                7.53  1.33         1.36             0.875         0.649
## 4 Canada                 7.43  1.33         1.32             0.906         0.633
## 5 Finland                7.41  1.29         1.32             0.889         0.642
## 6 Netherlands            7.38  1.33         1.28             0.893         0.616
## # ... with 3 more variables: government_trust_score <dbl>,
## #   generosity_score <dbl>, dystopia_score <dbl>
```

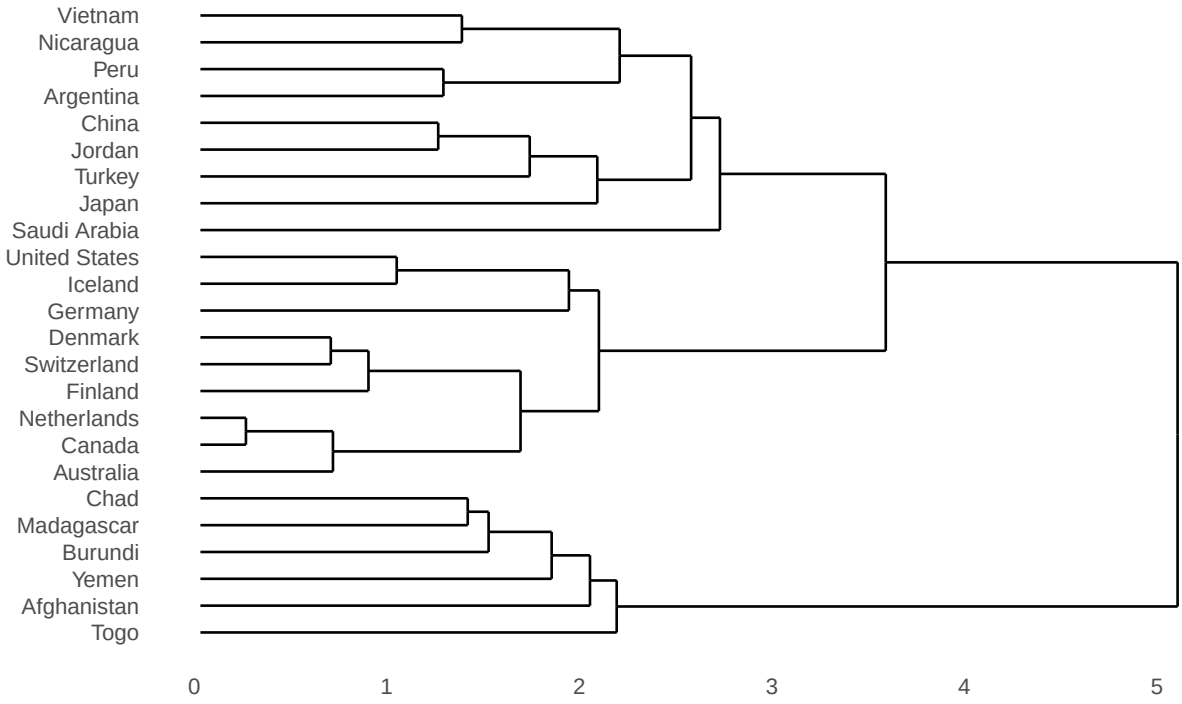
a) Perform hierarchical clustering of the countries and calculate the distance matrix. You do not need to display the distance matrix.

b) Display clustering results in a dendrogram.

```
# calculating the distance matrix
dist_out <-
  happiness |>
    column_to_rownames(var = "country") |>
    scale() |>
    dist(method = "euclidean")

# hierarchical clustering output
hc_out <- hclust(
  dist_out, method = "average"
)

# displaying the results in a dendrogram
ggdendrogram(hc_out, rotate = TRUE)
```



Problem 2: (3 pts) Use the clustering results you found in Problem 1 for Problem 2.

a) Assign clusters by cutting the dendrogram.

b) Plot a scatter plot for two numeric variables of your choice from `happy_data` and add cluster info into scatterplot.

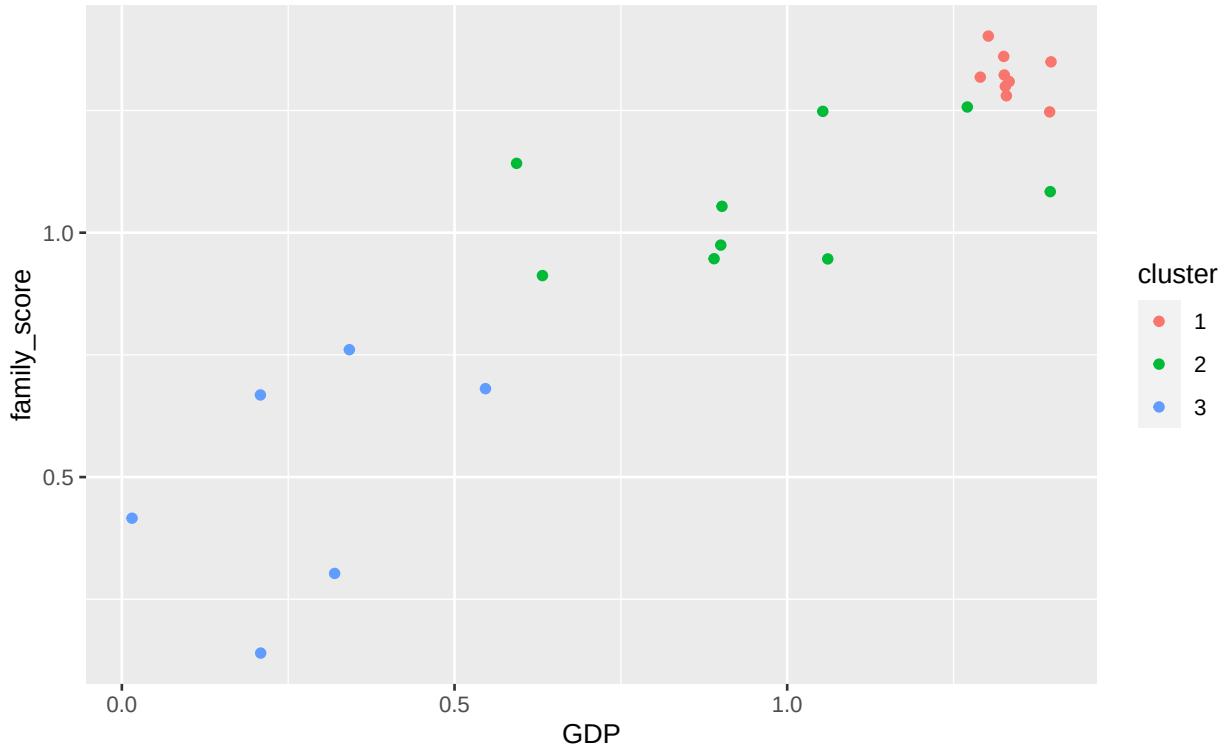
c) Interpret the plot.

```
# cutting clusters using cutree
cluster <- cutree(hc_out, k = 3)
cluster
```

```
##    Switzerland       Iceland       Denmark        Canada       Finland
##              1             1             1             1             1
##    Netherlands     Australia United States       Germany     Argentina
##              1             1             1             1             2
##   Saudi Arabia         Japan     Nicaragua          Peru       Vietnam
##              2             2             2             2             2
##         Turkey        Jordan         China         Yemen    Madagascar
##              2             2             2             3             3
##           Chad   Afghanistan       Burundi          Togo
##              3             3             3             3
```

```
happiness |>
  left_join(
    tibble(
```

```
    country = names(cluster),
    cluster = factor(cluster)
  )
) |>
ggplot(aes(GDP, family_score)) +
geom_point(aes(color = cluster))
```



Looking at the chart, it looks like GDP and family score are closely aligned. Rich countries in the red cluster tend to have high scores in both variables. At the risk of drawing a causal connection from a simple correlation, it would seem that higher productivity enables workers to spend more time at home.

Problem 3: (4 pts) For Problem 3, we will work with the dataset `texas_income`.

a) Bin the `median_income` column into 3 bins (20K-40K, 40K-60K, 60K-90K). Hint: use `case_when()`.

b) Make a choropleth map of Texas counties colored by median income bin (3 total colors).

c) Use an appropriate color scale and use a theme that shows longitude and latitude (nearly any theme other than `theme_void()` will work).

```
# data preparation
texas_income <- readRDS(url("https://wilkelab.org/SDS375/datasets/Texas_income.rds"))
head(texas_income)
```

```
## Simple feature collection with 6 features and 4 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
```

```
## Bounding box:  xmin: -103.0647 ymin: 27.83954 xmax: -94.12963 ymax: 35.18324
## Geodetic CRS:  NAD83
##    FIPS    county median_income  moe                       geometry
## 1 48001  Anderson         41327 1842 MULTIPOLYGON (((-96.0648 31...
## 2 48003   Andrews         70423 6038 MULTIPOLYGON (((-103.0647 3...
## 3 48005  Angelina         44223 1611 MULTIPOLYGON (((-95.00488 3...
## 4 48007   Aransas         41690 3678 MULTIPOLYGON (((-96.8229 28...
## 5 48009    Archer         60275 5182 MULTIPOLYGON (((-98.95382 3...
## 6 48011 Armstrong         59737 4968 MULTIPOLYGON (((-101.6294 3...
```

```
# binning median_income values
texas_income <-
  texas_income |>
    mutate(
      income_bin = case_when(
        median_income < 40000 ~ "low",
        median_income < 60000 ~ "medium",
        median_income < 90000 ~ "high",
        TRUE ~ "NA"
        )
           )

# making a chloropleth map of Texas using the income bins
library(viridis)
```

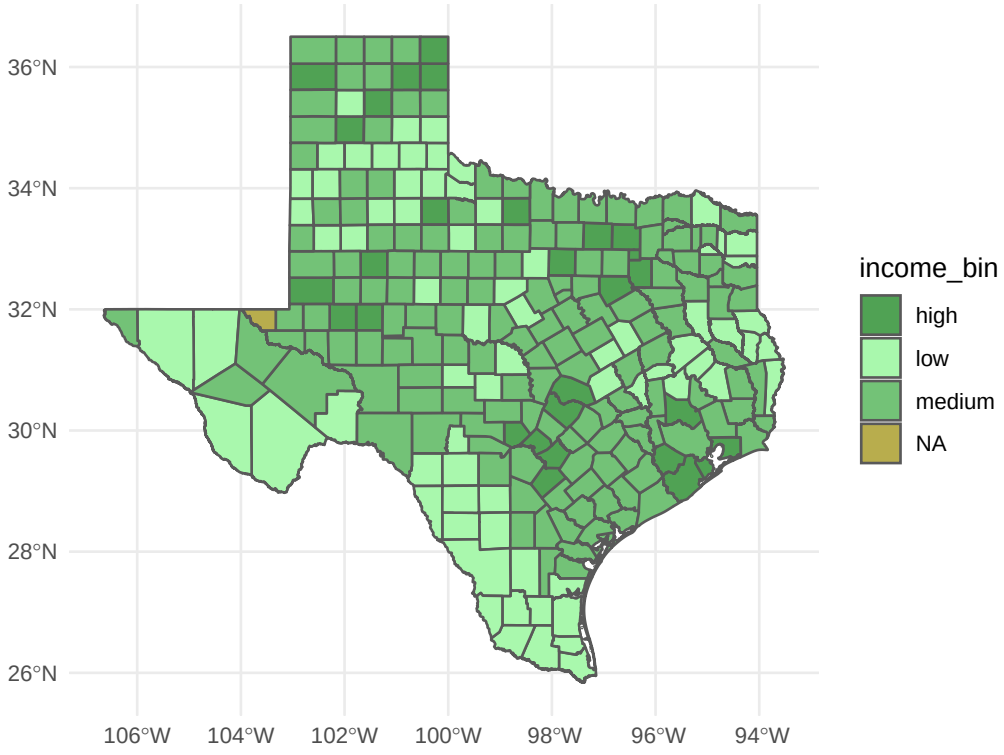
```
## Warning: package 'viridis' was built under R version 4.1.3
```

```
## Loading required package: viridisLite
```

```
ggplot(texas_income, aes(fill = income_bin)) +
  geom_sf() +
  scale_fill_manual(values = c("#50A254", "#A9F8AD", "#73C277", "#B8AD4D")) +
  theme_minimal()
```

```
## old-style crs object detected; please recreate object with a recent sf::st_crs()
```

```
## old-style crs object detected; please recreate object with a recent sf::st_crs()
## old-style crs object detected; please recreate object with a recent sf::st_crs()
```


36°N
34°N
32°N
30°N
28°N
26°N
106°W
104°W
102°W
100°W
98°W
96°W
94°W
income_bin
high
low
medium
NA