PUBH6680 Lab 4.B

Exploratory Data Analysis

KOMAL BHOSLE

RStudio Link

https://posit.cloud/spaces/603138/content/10064181

Library Calls

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(openintro)
```

```
Loading required package: airports
Loading required package: cherryblossom
Loading required package: usdata
```

```
library(easystats)
```

```
# Attaching packages: easystats 0.7.2 (red = needs update)
x bayestestR  0.13.2   x correlation 0.8.4
x datawizard  0.11.0   x effectsize  0.8.8
x insight     0.20.1   x modelbased  0.8.8
x performance 0.12.0   x parameters  0.21.7
x report      0.5.8    x see         0.8.4

Restart the R-Session and update packages with `easystats::easystats_update()`.
```

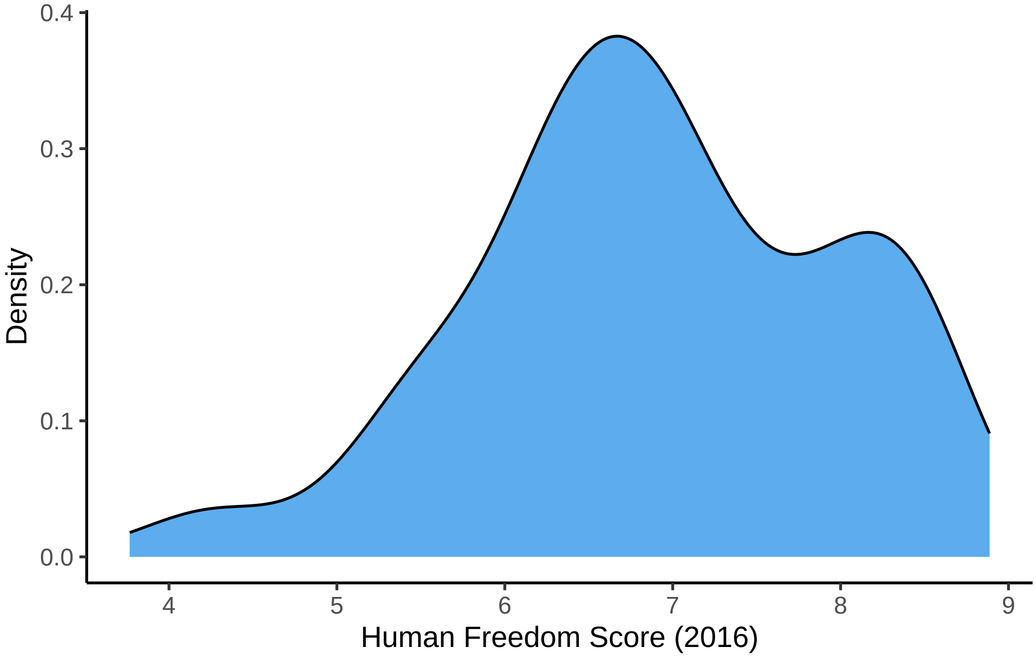
```r
library(ggpubr)
```

```

Attaching package: 'ggpubr'

The following objects are masked from 'package:datawizard':

    mean_sd, median_mad
```

Univariate Variation (One Continuous Variable)

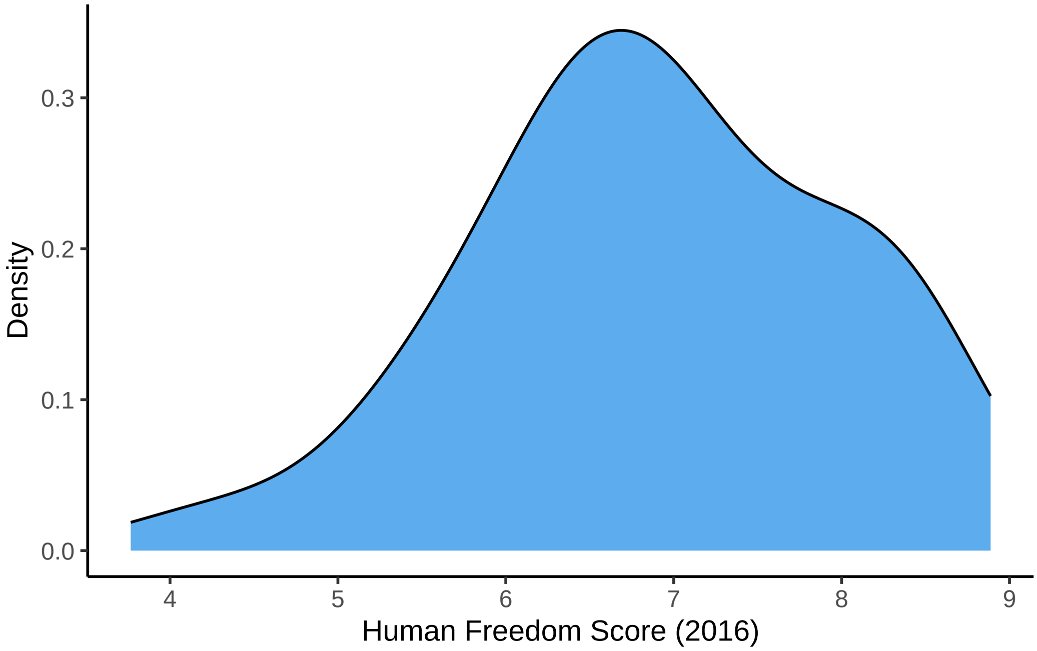
First Density Plot

```r
hfi_2016 <-
  hfi |>
  filter(
    year == 2016
  ) |>
  select(
    hf_score
  ) |>
  drop_na()

hfi_2016 |>
  ggplot(
    mapping = aes(
      x     = hf_score
    )
  ) +
    geom_density(
      fill = "steelblue2"
    ) +
    labs(
      x = "Human Freedom Score (2016)",
      y = "Density"
    ) +
    theme_classic()
```



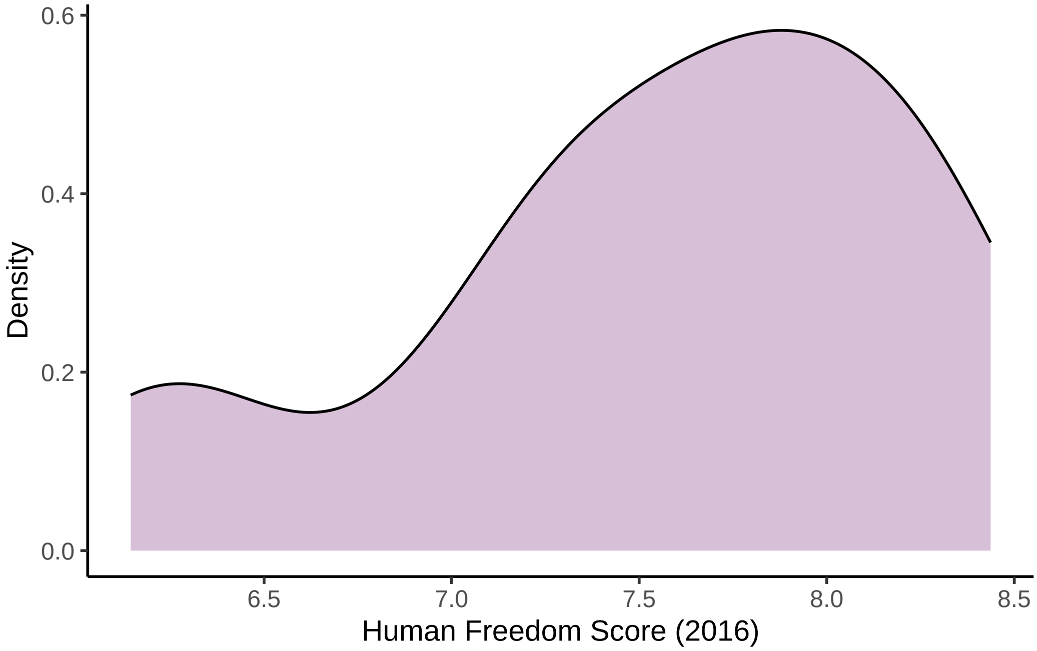
Second Density Plot

```
hfi_2016 |>
  ggplot(
    mapping = aes(
      x     = hf_score
    )
  ) +
    geom_density(
      fill = "steelblue2",
      bw   =  0.5
    ) +
    labs(
      x = "Human Freedom Score (2016)",
      y = "Density"
    ) +
    theme_classic()
```



Third Density Plot

```
hfi_Europe_2016 <-
  hfi |>
  filter(
    year == 2016,
    region == "Eastern Europe"
  ) |>
  select(
    hf_score
  ) |>
  drop_na()

hfi_Europe_2016 |>
  ggplot(
    mapping = aes(
      x      = hf_score
    )
  ) +
    geom_density(
      fill = "thistle"
    ) +
    labs(
      x = "Human Freedom Score (2016)",
      y = "Density"
    ) +
    theme_classic()
```



Coefficient of Variation

```
hfi_co_var <-
  hfi |>
  filter(
    year   == 2016,
    region == "Eastern Europe"
  ) |>
  select(-ends_with("Score"),
         -ends_with("Rank"),
         -contains("sex"),
         -contains("ss")
  ) |>
  pivot_longer(
    starts_with("pf"),
    names_to  = "hfi_measure",
    values_to = "value"
  ) |>
  select(
    hfi_measure,
    value
  ) |>
  drop_na(
    value
  ) |>
  group_by(
    hfi_measure
  ) |>
  summarise(
    co_var = sd(value, na.rm = TRUE) / mean(value, na.rm = TRUE)
  )
```

```
hfi_co_var_desc <-
  hfi_co_var |>
  arrange(desc(co_var))
```

```
hfi_co_var_asc <-
  hfi_co_var |>
  arrange(co_var)
```

Skewness and Kurtosis

```
hfi_east_euro_2016 <-
  hfi |>
  filter(
    year   == 2016,
    region == "Eastern Europe"
  ) |>
  select(
    ef_score,
    hf_score
  ) |>
  drop_na()
```

```
skewness(hfi_east_euro_2016)
```

```
Parameter | Skewness |    SE
----------------------------
ef_score  |   -0.504 | 0.457
hf_score  |   -0.786 | 0.457
```

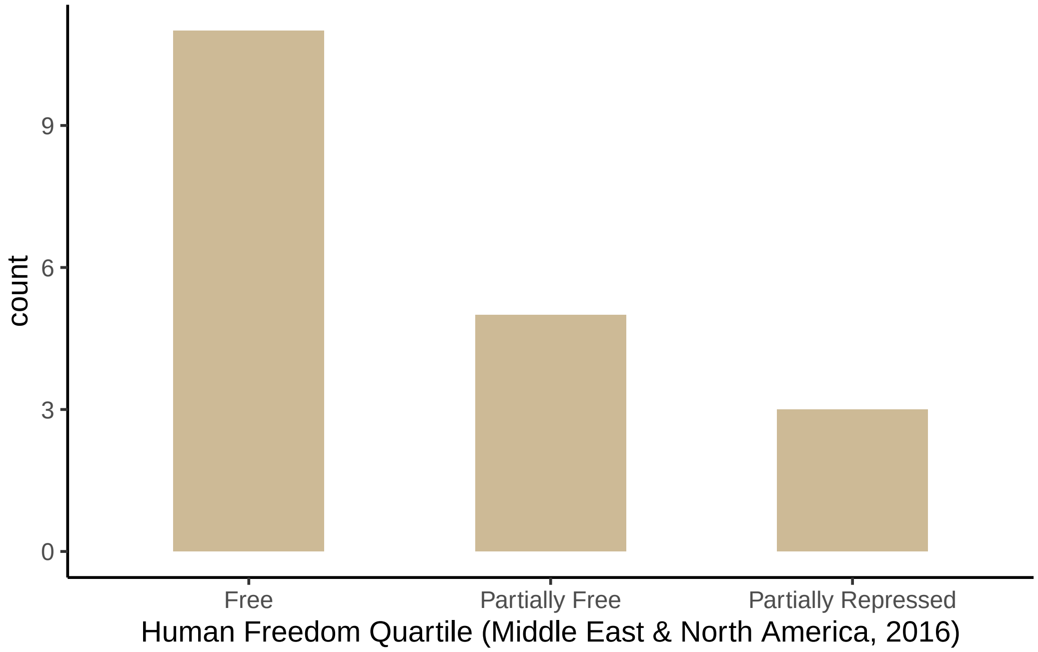
```
kurtosis(hfi_east_euro_2016)
```

```
Parameter | Kurtosis |    SE
----------------------------
ef_score  |   -0.410 | 0.750
hf_score  |   -0.164 | 0.750
```

Univariate Variation (One Discrete Variable)

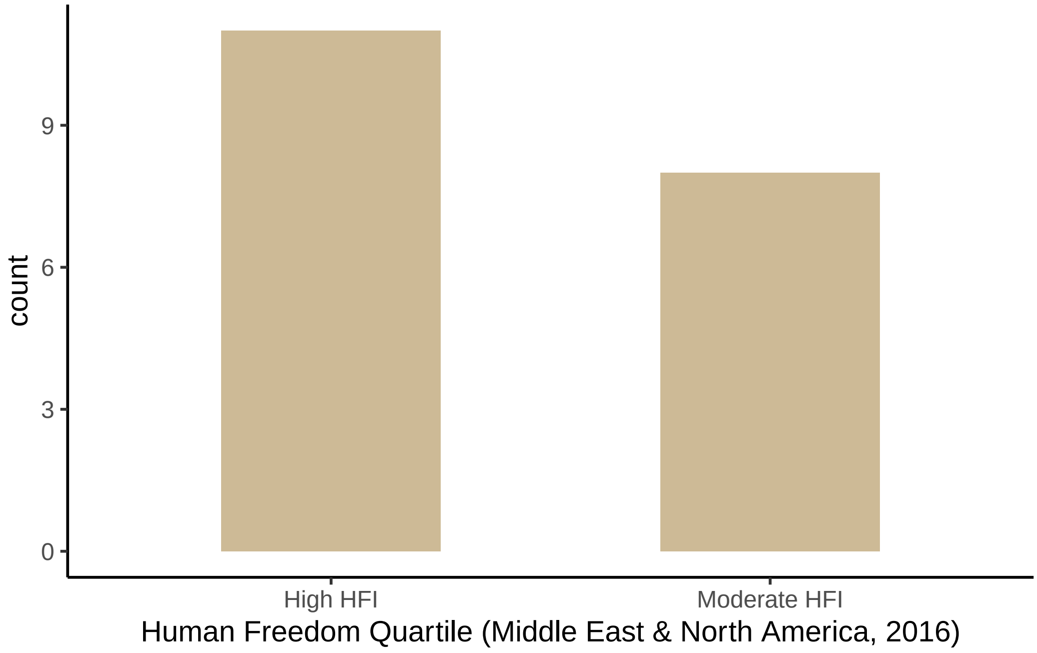
First Bar Plot

```
hfi_me_na_2016 <-
  hfi |>
  filter(
    year   == 2016,
    region == "Middle East & North Africa"
  ) |>
  select(
    hf_quartile
  ) |>
  drop_na() |>
  mutate(
    hf_quartile = case_when(
      hf_quartile == 1 ~ "Repressed",
      hf_quartile == 2 ~ "Partially Repressed",
      hf_quartile == 3 ~ "Partially Free",
      hf_quartile == 4 ~ "Free"
    )
  ) |>
  mutate(
    factor(hf_quartile,
           levels  = c("Repressed",
                       "Partially Repressed",
                       "Partially Free",
                       "Free"),
           ordered = TRUE)
  )

hfi_me_na_2016 |>
  ggplot(
    mapping = aes(
      x     = hf_quartile
    )
  ) +
  geom_bar(
    fill  = "wheat3",
    width = 0.5
  ) +
  labs(
    x = "Human Freedom Quartile (Middle East & North America, 2016)",
    y = "count"
  ) +
  theme_classic()
```



Second Bar Plot

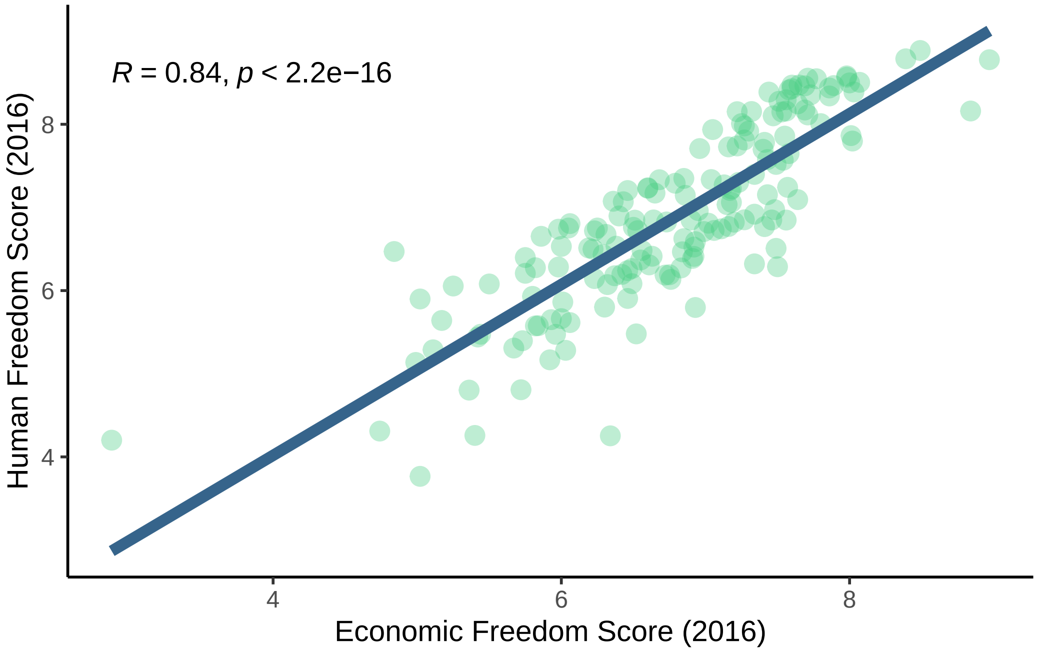
```
hfi_me_na_2016 <-
  hfi |>
  filter(
    year   == 2016,
    region == "Middle East & North Africa"
  ) |>
  select(
    hf_quartile
  ) |>
  drop_na() |>
  mutate(
    hf_quartile   = case_when(
      hf_quartile == 1 ~ "Low HFI",
      hf_quartile %in% c(2,3) ~ "Moderate HFI",
      hf_quartile == 4 ~ "High HFI"
    )
  ) |>
  mutate(
    factor(hf_quartile,
           levels  = c("Low HFI",
                       "Moderate HFI",
                       "High HFI"),
           ordered = TRUE)
  )

hfi_me_na_2016 |>
  ggplot(
    mapping = aes(
      x     = hf_quartile
    )
  ) +
  geom_bar(
    fill  = "wheat3",
    width = 0.5
  ) +
  labs(
    x = "Human Freedom Quartile (Middle East & North America, 2016)",
    y = "count"
  ) +
  theme_classic()
```



Bivariate Covariance (Two Continuous Variables)

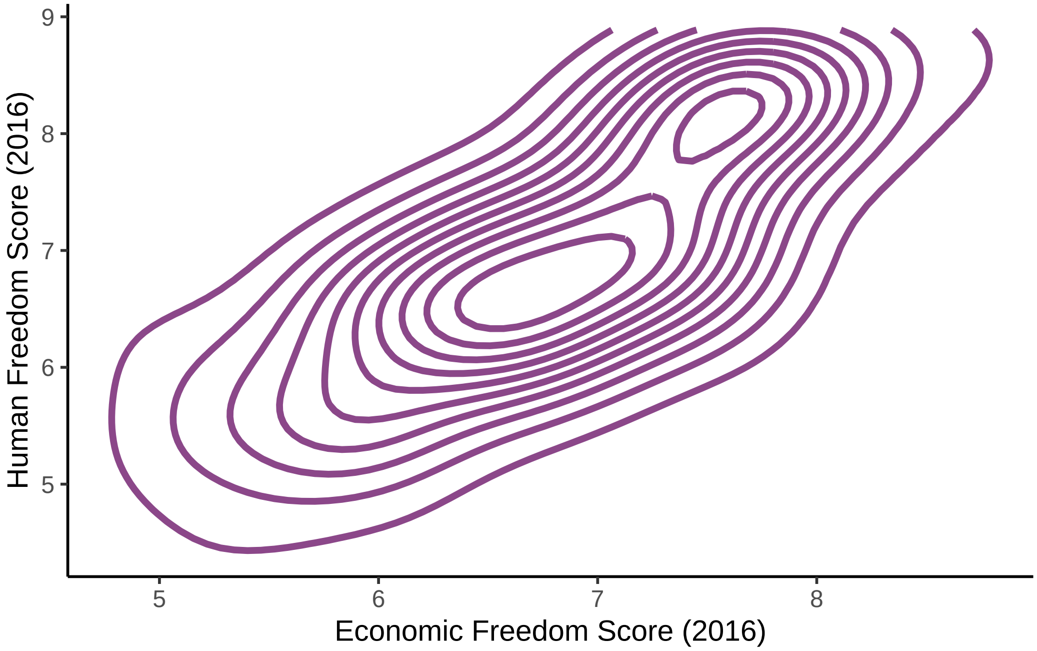
First Scatterplot

```r
hfi_ef_hf_2016 <-
  hfi |>
  filter(
    year == 2016
  ) |>
  select (
    ef_score,
    hf_score
  ) |>
  drop_na()

hfi_ef_hf_2016 |>
  ggplot(
    mapping = aes(
    x       = ef_score,
    y       = hf_score
    )
  ) +
  geom_point(
    size  = 3,
    alpha = 0.35,
    color = "seagreen3"
  ) +
  geom_smooth(
    method    = "lm",
    formula   = y ~ x,
    se        = FALSE,
    linewidth = 2,
    color     = "steelblue4",
    alpha     = 0.7
  ) +
  stat_cor() +
  labs(
    x = "Economic Freedom Score (2016)",
    y = "Human Freedom Score (2016)"
  ) +
  theme_classic()
```



2D Density Plot

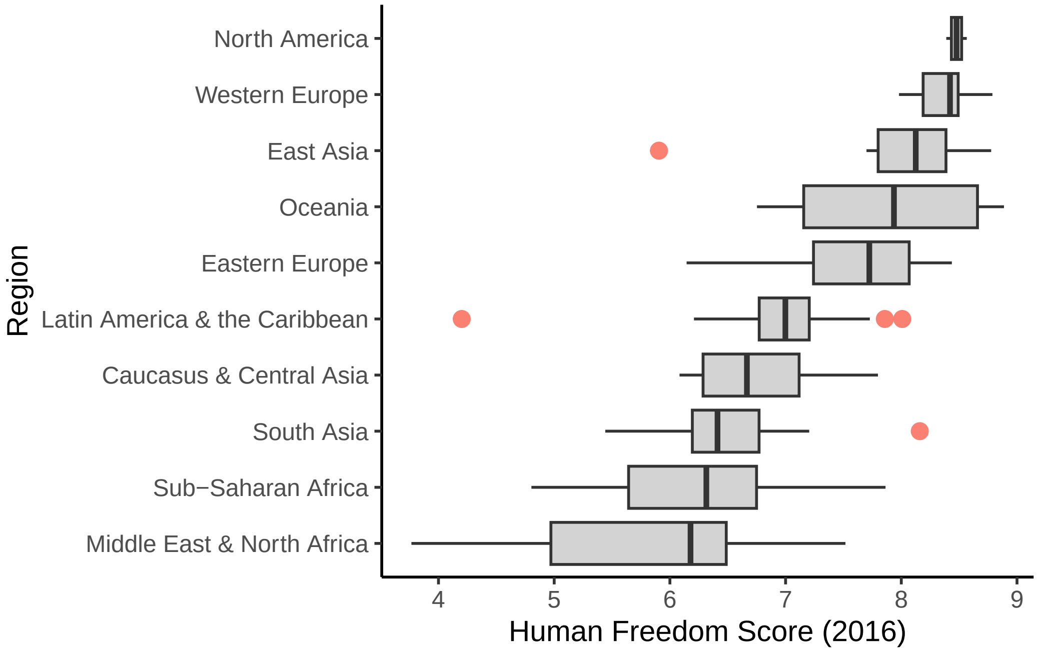
```
hfi_ef_hf_2016 <-
  hfi |>
  filter(
    year == 2016
  ) |>
  select (
    ef_score,
    hf_score
  ) |>
  drop_na()

hfi_ef_hf_2016 |>
  ggplot(
    mapping = aes(
    x       = ef_score,
    y       = hf_score
    )
  ) +
  geom_density_2d(
    linewidth = 1.2,
    color      = "orchid4"
  ) +
  labs(
    x = "Economic Freedom Score (2016)",
    y = "Human Freedom Score (2016)"
  ) +
  theme_classic()
```



Bivariate Covariance (Between One Continuous and One Discrete Variables)

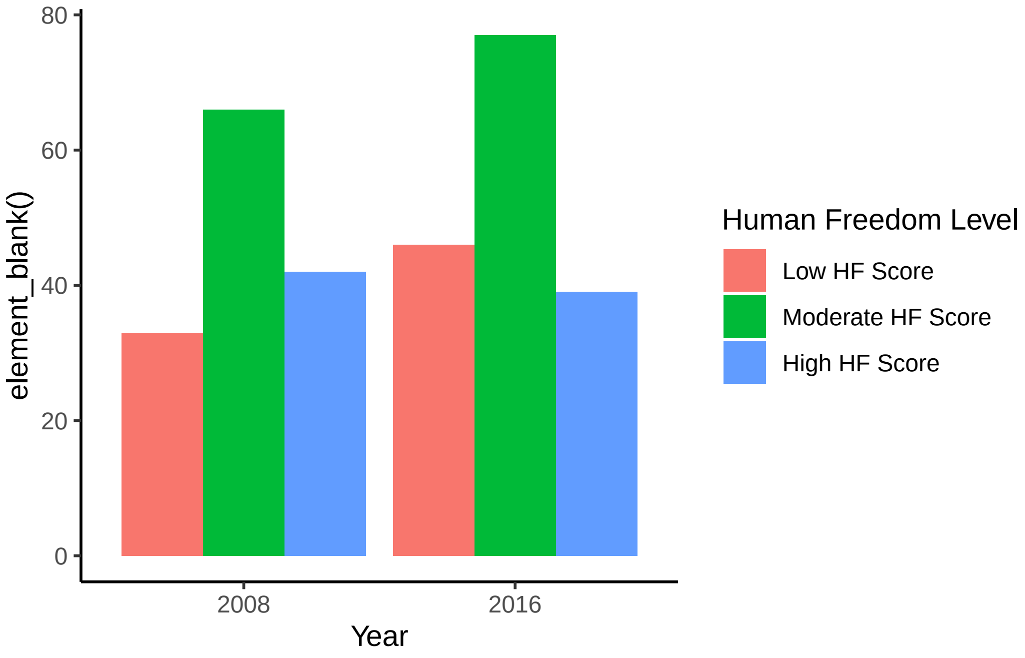
First Box Plot

```
hfi_re_hf_2016 <-
  hfi |>
  filter(
    year == 2016
  ) |>
  select(
    region,
    hf_score
  ) |>
  drop_na(
    hf_score
  ) |>
  mutate(
    region  = fct_reorder(
      region,
      hf_score,
      .fun  = median,
      na.rm = TRUE
    )
  )

hfi_re_hf_2016 |>
  ggplot(
    mapping = aes(
    x       = region,
    y       = hf_score
    )
  ) +
  geom_boxplot(
    fill          = "lightgray",
    outlier.color = "salmon",
    outlier.size  = 2.5
  ) +
  coord_flip(
  ) +
  labs(
    x = "Region",
    y = "Human Freedom Score (2016)"
  ) +
  theme_classic()
```



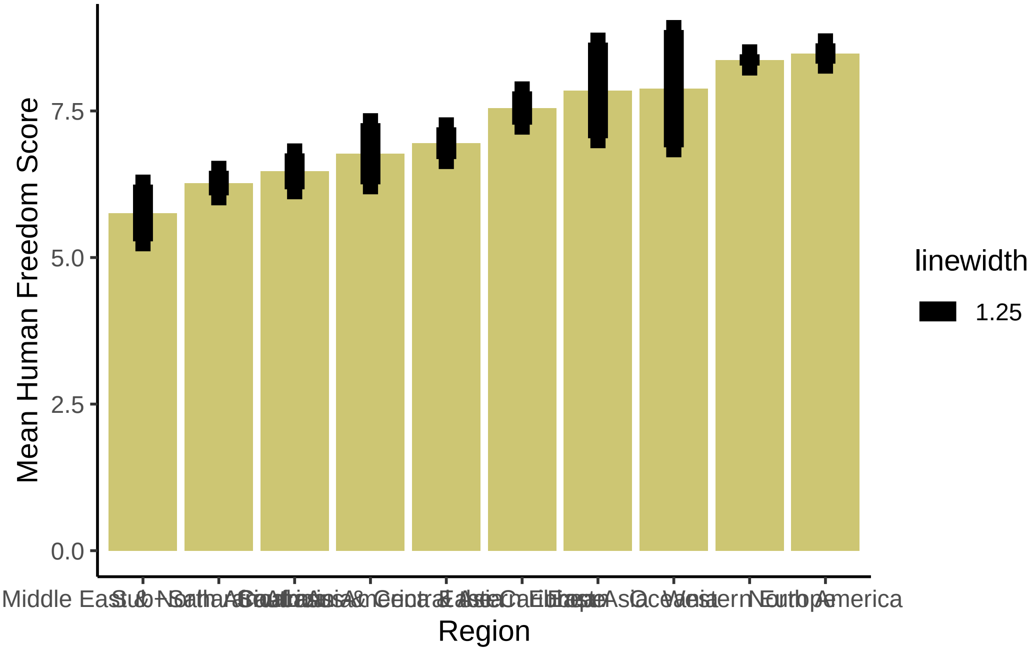
Third Bar Plot

```
Q1 <- 6.3

Q3 <- 7.8

hfi_bar <-
  hfi |>
  filter(
    year %in% c(2008, 2016)
  ) |>
  select(
    year,
    hf_score
  ) |>
  drop_na(hf_score) |>
  mutate(
    hf_score = case_when(
      hf_score < Q1 ~ "Low HF Score",
      hf_score <= Q3 ~ "Moderate HF Score",
      hf_score > Q3 ~ "High HF Score"
    ),
    hf_score = factor(
    hf_score,
    levels   = c("Low HF Score",
                 "Moderate HF Score",
                 "High HF Score")),
    year     = as_factor(year)
  )

hfi_bar |>
  ggplot(
    mapping = aes(
      x     = year,
      fill  = hf_score
    )
  ) +
  geom_bar(position = "dodge") +
  labs(
    x    = "Year",
    y    = "element_blank()",
    fill = "Human Freedom Level"
  ) +
  theme_classic()
```



Fourth Bar Plot

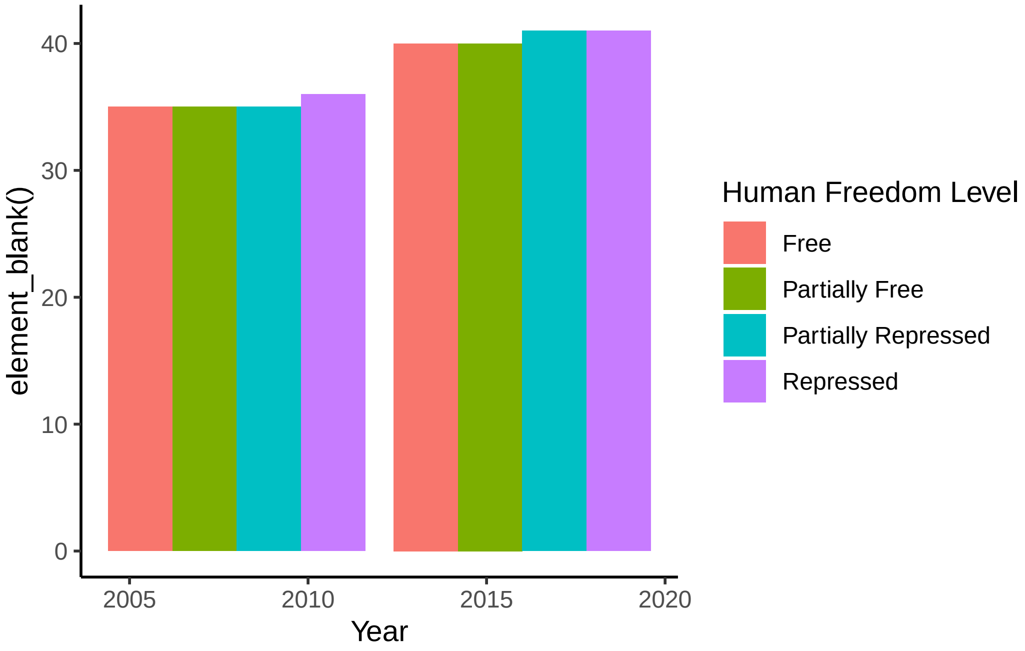
```
hfi_error_bar_2016 <-
  hfi |>
  filter(
    year == 2016
  ) |>
  select(
    region,
    hf_score
  ) |>
  drop_na(
    hf_score
  ) |>
  group_by(
    region
  ) |>
  summarise(
  mean_hf_score = round(mean(hf_score), 2),
  se      = sd(hf_score) / sqrt(n()),
  se_ymin = mean_hf_score - 1.96 * se,
  se_ymax = mean_hf_score + 1.96 * se
  ) |>
  mutate(
    region = fct_reorder(
      region,
      mean_hf_score
    )
  )

hfi_error_bar_2016 |>
  ggplot(
    mapping = aes(
      x     = region,
      y     = mean_hf_score
    )
  ) +
  geom_col(
    fill = "khaki3"
  ) +
  geom_errorbar( aes(
    ymin      = se_ymin,
    ymax       = se_ymax,
    width      = 0.2,
    linewidth = 1.25
    )
  ) +
  labs(
    x = "Region",
    y = "Mean Human Freedom Score"
  ) +
  theme_classic()
```



Bivariate Covariance (Two Discrete Variables)

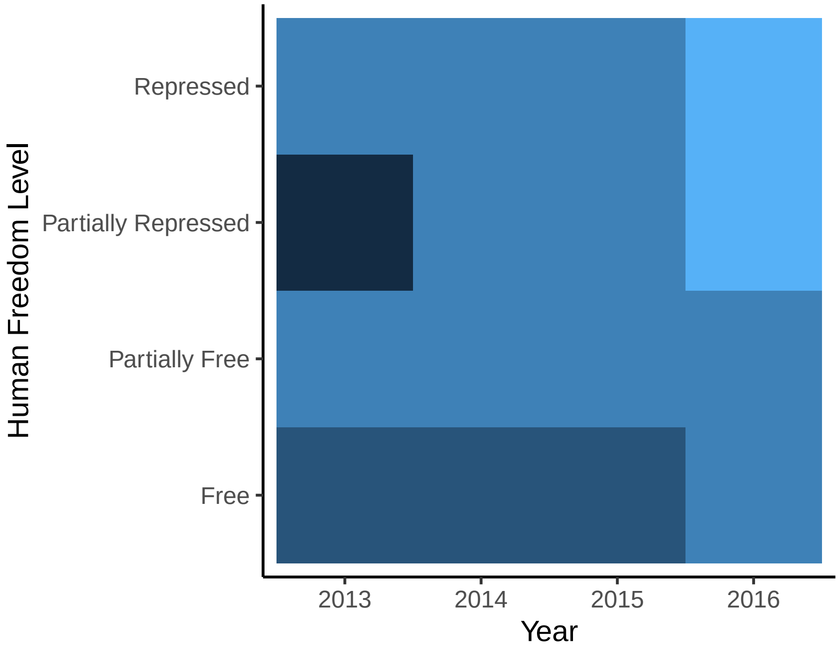
Fifth Bar Plot

```
hfi_2008_2016 <-
  hfi |>
  filter(
    year %in% c(2008, 2016)
  ) |>
  select(
    year,
    hf_quartile
  ) |>
  drop_na(hf_quartile) |>
  mutate(
    hf_quartile = case_when(
      hf_quartile == 1 ~ "Repressed",
      hf_quartile == 2 ~ "Partially Repressed",
      hf_quartile == 3 ~ "Partially Free",
      hf_quartile == 4 ~ "Free"
    )
  ) |>
  mutate(
    factor(hf_quartile,
           levels  = c("Repressed",
                       "Partially Repressed",
                       "Partially Free",
                       "Free"),
           ordered = TRUE)
  )

hfi_2008_2016 |>
  ggplot(
    mapping = aes(
      x     = year,
      fill  = hf_quartile
    )
  ) +
  geom_bar(position = "dodge") +
  labs(
    x    = "Year",
    y    = "element_blank()",
    fill = "Human Freedom Level"
  ) +
  theme_classic()
```



Heatmap

```
hfi_heatmap <-
  hfi |>
  filter(
    year %in% c(2013, 2014, 2015, 2016)
  ) |>
  select(
    year,
    hf_quartile
  ) |>
  drop_na(hf_quartile) |>
  mutate(
    year = factor(year, levels = c(2013, 2014, 2015, 2016)),
    hf_quartile = case_when(
      hf_quartile == 1 ~ "Repressed",
      hf_quartile == 2 ~ "Partially Repressed",
      hf_quartile == 3 ~ "Partially Free",
      hf_quartile == 4 ~ "Free"
    )
  ) |>
  mutate(
    factor(hf_quartile,
           levels = c("Repressed",
                      "Partially Repressed",
                      "Partially Free",
                      "Free"),
           ordered = TRUE)
  ) |>
  group_by(
    year,
    hf_quartile
  ) |>
  summarise(
    n       = n(),
    .groups = "drop")

hfi_heatmap |>
  ggplot(
    mapping = aes(
      x     = year,
      y     = hf_quartile,
      fill  = n
    )
  ) +
  geom_tile() +
  coord_fixed() +
  labs(
    x    = "Year",
    y    = "Human Freedom Level",
    fill = "Element Blank ()"
  ) +
  theme_classic() +
  theme(legend.position = "None")
```

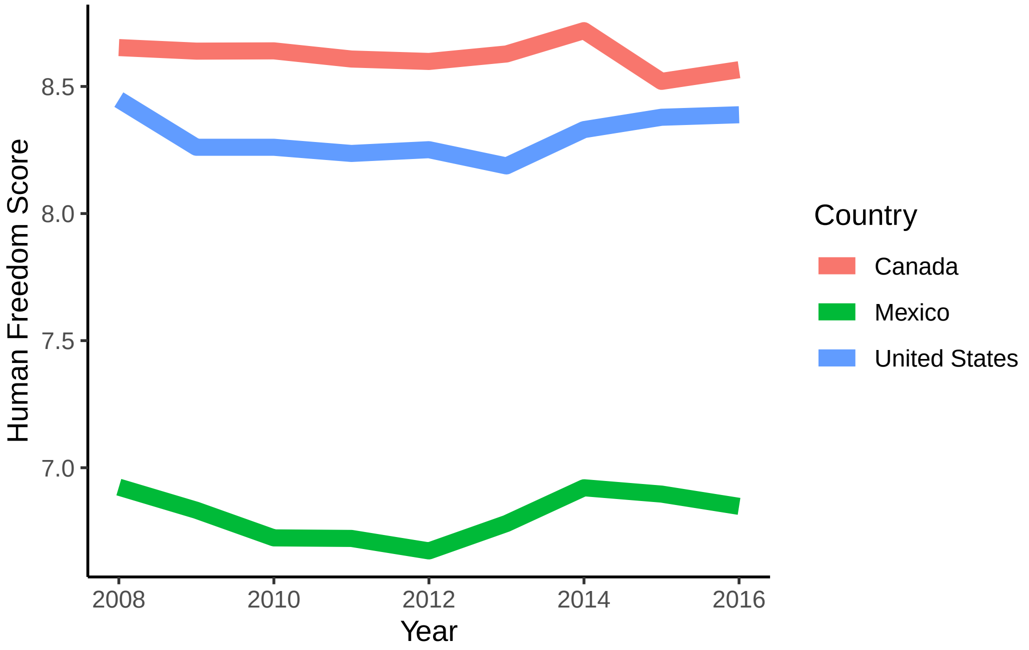


Multivariate Covariance

Line Plot

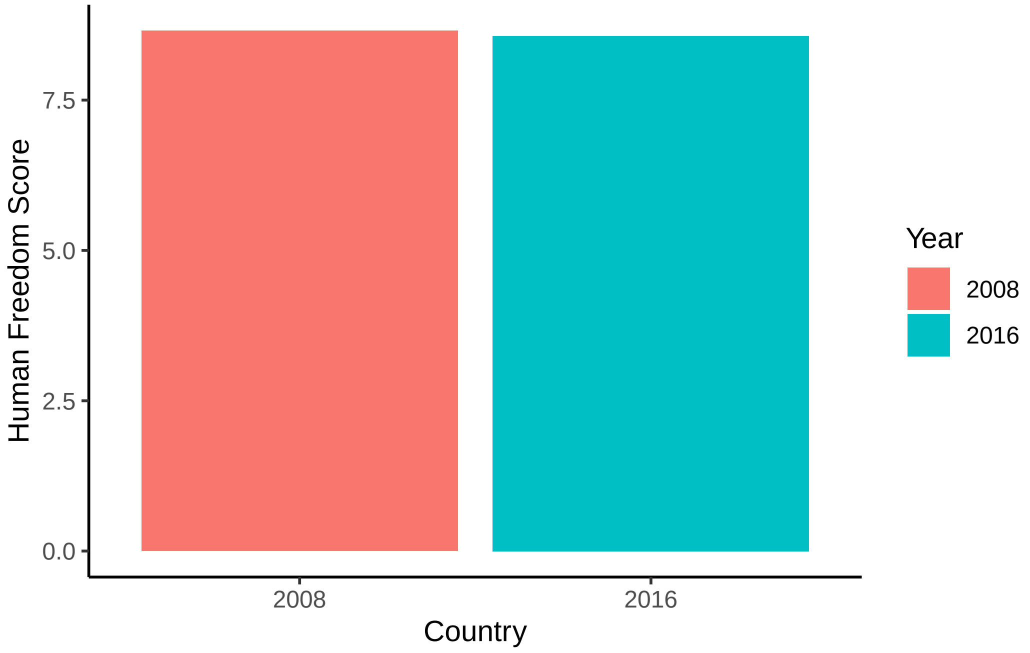
```
hfi_line_plot <-
  hfi |>
  filter(countries %in% c("Canada",
                          "United States",
                          "Mexico")
  ) |>
  select(
    year,
    countries,
    hf_score
  ) |>
  drop_na() |>
  mutate(
    year      = as.integer(year),
    countries = as.factor(countries)
  )

hfi_line_plot |>
```

```
ggplot(
  mapping = aes(
    x     = year,
    y     = hf_score,
    color = countries
  )
) +
geom_line(
  linewidth = 3
) +
labs(
  x     = "Year",
  y     = "Human Freedom Score",
  color = "Country"
) +
theme_classic()
```



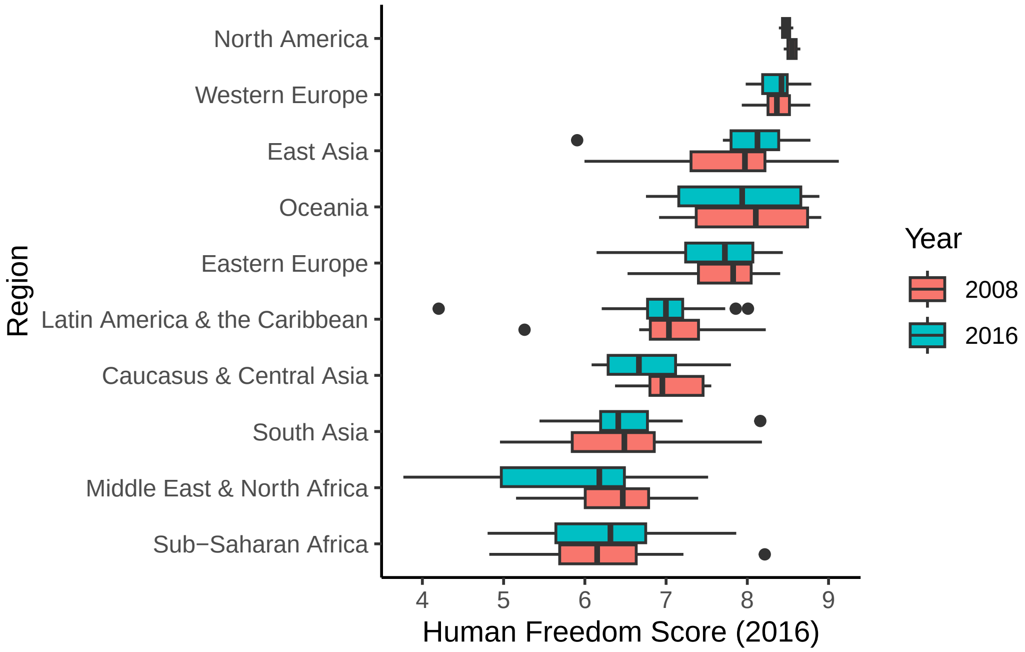
Sixth Bar Plot

```
hfi_bar_plot <-
  hfi |>
  filter(
    year %in% c(2008, 2016),
    countries %in% c("Canada",
                     "United States",
                     "Mexico")
  ) |>
  select(
    year,
    countries,
    hf_score
  ) |>
  drop_na() |>
  mutate(
    year      = factor(
      year,
      levels  = c(2008, 2016)),
    countries = as.factor(countries)
  )

hfi_bar_plot |>
  ggplot(
    mapping = aes(
      x     = year,
      y     = hf_score,
      fill  = year
    )
  ) +
  geom_bar(
    stat     = "identity",
    position = "dodge"
  ) +
  labs(
    x    = "Country",
    y    = "Human Freedom Score",
    fill = "Year"
  ) +
  theme_classic()
```



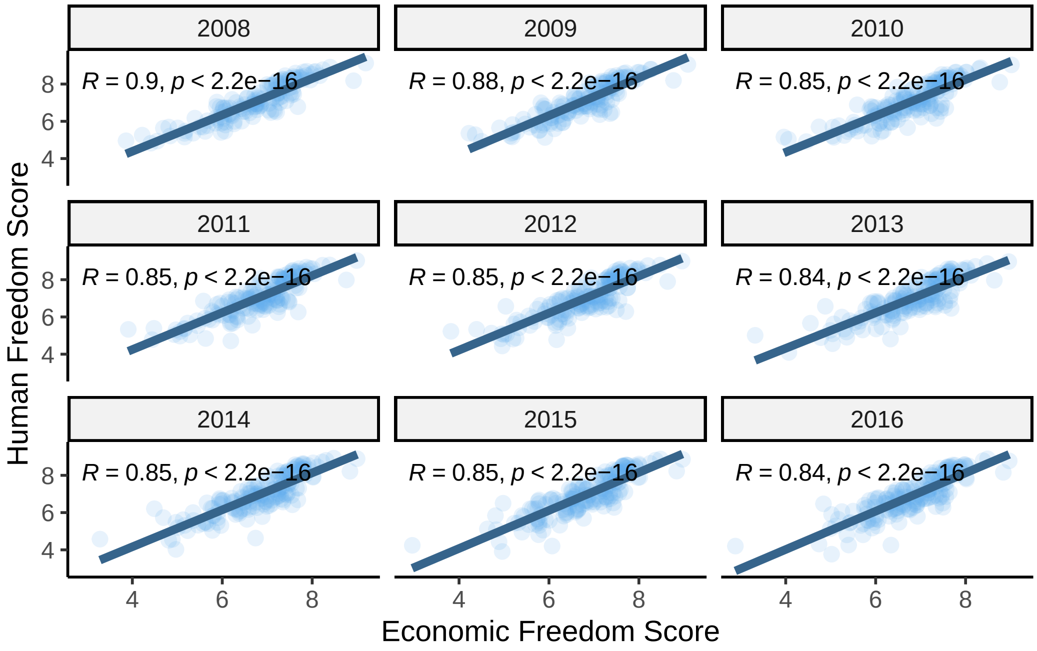
Second Box Plot

```
hfi_re_hf_2016 <-
  hfi |>
  filter(
    year %in% c(2008, 2016),
  ) |>
  select(
    year,
    region,
    hf_score
  ) |>
  drop_na(
    hf_score
  ) |>
  mutate(
    region = fct_reorder(
      region,
      hf_score,
      .fun = median
    ),
    year = factor(
      year, levels = c(2008, 2016)
    )
  )

hfi_re_hf_2016 |>
  ggplot(
    mapping = aes(
    x       = region,
    y       = hf_score,
    fill    = year,
    group   = interaction(
      year,
      region)
    )
  ) +
  geom_boxplot(
  ) +
  coord_flip(
  ) +
  labs(
    x    = "Region",
    y    = "Human Freedom Score (2016)",
    fill = "Year"
  ) +
  theme_classic()
```



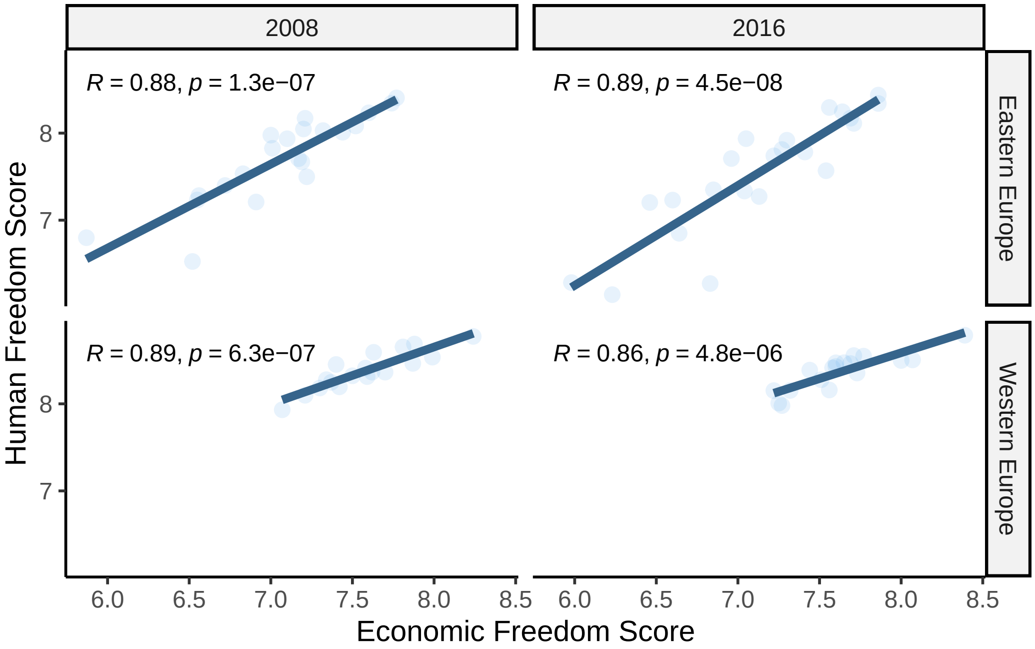
Second Scatterplot

```
hfi_scatterplot_2 <-
  hfi |>
  select (
    year,
    ef_score,
    hf_score
  ) |>
  mutate(
    year = as.integer(year)
  ) |>
  drop_na()

hfi_scatterplot_2 |>
  ggplot(
    mapping = aes(
    x       = ef_score,
    y       = hf_score
    )
  ) +
  geom_point(
    size  = 2.25,
    alpha = 0.15,
    color = "steelblue2"
  ) +
  geom_smooth(
    method    = "lm",
    formula   = y ~ x,
    se        = FALSE,
    linewidth = 1.5,
    color     = "steelblue4",
    alpha     = 0.7
  ) +
  stat_cor(
    size = 3
  ) +
  facet_wrap(
    ~year,
    ncol = 3
  ) +
  labs(
    x = "Economic Freedom Score",
    y = "Human Freedom Score"
  ) +
  theme_classic() +
  theme(
    strip.background = element_rect(
      fill = "#f2f2f2"
    )
  )
```



Third Scatterplot

```
hfi_scatterplot_3 <-
  hfi |>
  filter(
    year %in% c(2008, 2016),
    region %in% c("Western Europe",
                  "Eastern Europe")
  ) |>
  select (
    year,
    region,
    ef_score,
    hf_score
  ) |>
  mutate(
    year = as.integer(year)
  ) |>
  drop_na()

hfi_scatterplot_3 |>
  ggplot(
    mapping = aes(
    x       = ef_score,
    y       = hf_score
    )
  ) +
  geom_point(
    size  = 2.25,
    alpha = 0.15,
    color = "steelblue2"
  ) +
  geom_smooth(
    method    = "lm",
    formula   = y ~ x,
    se        = FALSE,
    linewidth = 1.5,
    color     = "steelblue4",
    alpha     = 0.7
  ) +
  stat_cor(
    method = "pearson",
    size   = 3
  ) +
  facet_grid(
    rows = vars(region),
    cols = vars(year)
  ) +
  labs(
    x = "Economic Freedom Score",
    y = "Human Freedom Score"
  ) +
  theme_classic() +
  theme(
    strip.background = element_rect(
      fill = "#f2f2f2"
    )
  )
```



Questions on Data Insights

Questions on Univariate Variation (One Continuous Variable)

1. Does the first density plot indicate multiple modes, possible outliers, or both? Explain your rationale.

First density plot interprets the presence of both. As we can see a tail in the beggining of the plot indicates possible outliers, while the two peaks in the curve represnt multiple modes.

2. What does the bandwidth adjustment in the second density plot suggest? Explain your rationale.

The bandwidth adjustment has smoothened the plot, but this might indicate that details must have lost which could cause more prominent waved plot.

3. In the third density plot, a categorical level within a factor was isolated. Does this plot indicate multiple modes, possible outliers, or both? Explain your rationale.

The third density plot shows two modes which could be bimodal, and a steep dip in curve might be due to possible outliers.

4. Using the coefficients of variation, identify an Human Freedom Index measure that has high consistency among countries and an HFI measure that has low

consistency among countries. Do more than merely listing the encoded version of the measurement. Explain the HFI measure through its defining characteristics.

The highest consistency among countries is identified to be Identity, which says that majority of the countries have similar policies for personal identity, where as lowest consistency if found to be about foreign movement, which indicates that different countries have majorly different policies for the foreign travel and transportation.

5. Describe the EF and HF scores in terms of skewness and kurtosis?

The EF skewness is -0.5 which indicates a slightly left skew. The HF skewness is -0.786 which also indicates a moderate left skew.

The EF kurtosis is -0.41 and HF kurtosis is -0.16, both these values are close to zero which indicate that the distribution is majorly flat with very slight peak.

Questions on Univariate Variation (One Discrete Variable)

6. What insights can be logically deduced from the first and second bar plots?

From first and second bar plots, The Human Freedom Index was highest in Middle East and North America during the year 2016.

Questions on Bivariate Covariation (Two Continuous Variables)

7. Describe the relationship between the EF and HF scores based upon the first scatterplot, correlation coefficient, and coefficient of determination.

The Scatterplot shows a positive correlation between EF and HF as we see an upward trend.This indicates that EF is positively related to HF.

8. How many density clusters appear in the 2D density plot?

10

Questions on Bivariate Covariation (Between One Continuous and One Discrete Variables)

9. Using the boxplot, which regions have outliers within the distribution of their HI scores?

East Asia, Latin America & the Carribean, South Asia.

10. What other insights can you find within the first box plot?

Middle East and North Africa had the highest Human Freedom score in the year 2016, followed by Oceanina. While north america had the lowest score.

11. Compare the HF levels between the years 2008 and 2016 using the third bar chart. Summarize the key insights.

During the year 2008, the human freedom index score was higher compared to Human freedom index score during the year 2016.

12. Imagine you are a decision-maker with the World Bank. Using the fourth bar plot, which region would you fund for projects expanding human freedom? Which region's mean HF score would you trust the least? Explain your rationale.

Western europe has the least confidence interval with high mean HF score so I would fund Eastern Europe for projects, where as Oceania has the highest confidence interval because of which I would trust Oceania the least.

Questions on Bivariate Covariation (Two Discrete Variables)

13. What insights might be deduced from the fifth bar plot? What types of supplemental analyses would you conduct, if any?Explain your rationale—-including why you would (or would not) conduct additional analyses.

The fifth bar plot doesnot provide any characteristic insights on Human Freedom Index. An additional analysis would help provide specific insights.

14. Explain the proper setup of a heatmap. What is a heatmap? Which variables should be placed on the x and y axes? What types of variables can you visualize with a heatmap? What should be anchored in the "lower left" corner? What does the color of each "tile" represent?

Heat map uses colors to differentiate and visualize the data matrix. The X axis has continous variable where as y axis has categorical variable. The categorical and numerical variable can be best visualized. Lower left corner often anchors the smallest values. The colors show intensity and differentiates from other vairables by colors.

Questions on Multivariate Covariation

15. What trends can you obtain from the line plot? What additional analysis would you suggest?

This plot interprets that Canada followed by America have high human freedom index with mexico has comparatively less Human freedom index score.

16. Review the second box plot. Which two regions had significant improvements in their HF scores between 2008 and 2016? Which two regions had significant deterioration in their HF scores between 2008 and 2016? Explain your rationale.

East Asia and Oceania had highest HF score in 2008, where as North America has the least in 2008. In year 2016 sub - saharan africa and oceania had high hf score with north america having less score.

17. How should workflows be built before adding faceting (small multiples)? Which functions map variables?

Use function aes to map the variables, with facet_wrap function to add faceting for single variables.

18. What are some differences between facet wraps and facet grids?

Facet_wrap is used for single variable where as facet_grid is used for two variable faceting.

19. Review the facet wrap. How does the relationship between EF and HF scores change over time? Explain your rationale.

Facet wraps gives analysis of data over time by providing different plots according to time variable. This helps in understanding the change over time and also their correlation.

20. Review the facet grid. How does the relationship between EF and HF scores compare between Eastern Europe and Western Europe during the years 2008 and 2016?

Facet grid allows us to assess the information in two variables. These scatterplots show similarity between hf scores and ef score in 2008 and 2016 in east europe and west europe.