Assignment 3 - Data Visualization

Name: Matei Penca

Student number: s4039696

Visualization A

For this visualization we're using some data from the General Social Survey data (2016). Here you can find more information from this dataset: https://rdrr.io/github/kjhealy/socviz/man/gss_sm.html

Packages

```
library(ggplot2)
library(tidyverse)
library(socviz) # install.packages("socviz") if you haven't done so
data <- gss_sm # gss_sm is a dataset from the package socviz
```

Code to create new variable

```
data <- data %>% mutate(sibs_rec = case_when(
  sibs < 10 ~ as.character(sibs),
  sibs >= 10 ~ "10+"
))
```

Create new theme

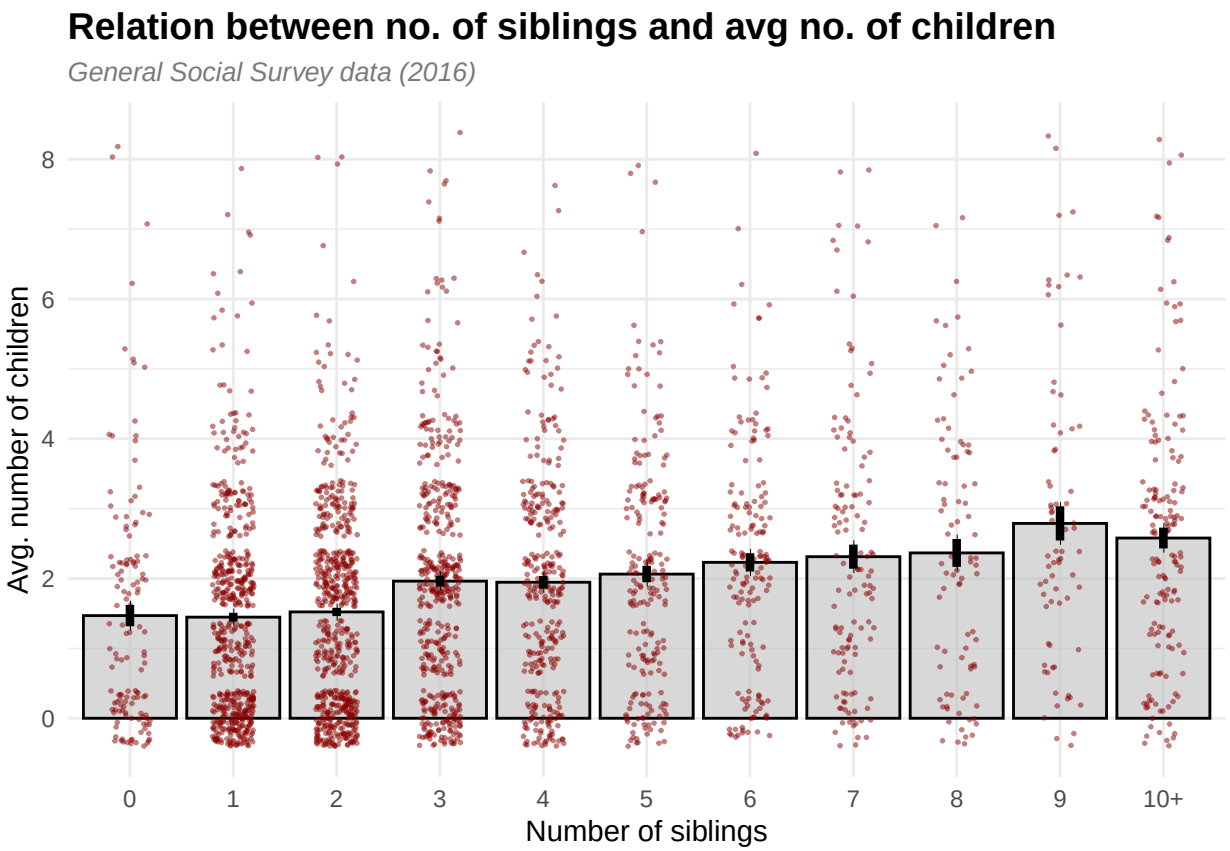
```
custom_theme <- theme(
  plot.title = element_text(size = 14, face = "bold", colour = "black"),
  plot.subtitle = element_text(size = 10, face = "italic", colour = "gray48")
)
```

Code to create visualization

```r
# Calculate average per age and standard error
childs_mean <- data %>%
  filter(!is.na(sibs_rec)) %>%
  group_by(sibs_rec) %>%
  summarise(
    mean_childs = mean(childs, na.rm = TRUE),
    n = n(),
    sd = sd(childs, na.rm = TRUE),
    se = sd / sqrt(n),
    lower = mean_childs - se,
    upper = mean_childs + se
  )

# Set the order to be increasing
order <- c("0", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10+")

ggplot(filter(data, !is.na(sibs_rec)), aes(x = sibs_rec)) +
  geom_bar(
    data = childs_mean, aes(y = mean_childs), stat = "identity",
    color = "black", fill = "grey", alpha = 0.6
  ) +
  geom_jitter(aes(y = childs),
    colour = "darkred", alpha = 0.5, size = 0.3,
    width = 0.2
  ) +
  geom_errorbar(
    data = childs_mean, aes(ymin = lower, ymax = upper),
    width = 0, size = 1.5
  ) +
  scale_x_discrete(limits = order) +
  labs(
    x = "Number of siblings", y = "Avg. number of children",
    title = "Relation between no. of siblings and avg no. of children",
    subtitle = "General Social Survey data (2016)"
  ) +
  theme_minimal() +
  custom_theme
```



Description of visualization

In the graph above we can observe a relation between the number of siblings a person might have and the number of children that they will have. People with no siblings or at most 2, will have on average less than 2 children. Starting from people with 3 siblings, we can see the average number of children is around 2 and keeps increasing, almost reaching 3 for people with 9 siblings.

Looking at the raw data, we can see that there are substantially more people with 0, 1, 2, or 3 siblings in our data set than there are with 4 or more which makes sense as nowadays people tend to have fewer children than in the past. For the 10+ category, it is hard to take out any exact information, but it is clear from the graph that even with a lower combined sample size, their average number of children is on the higher spectrum.

There are outliers in all categories, but because the number of data points is quite large they are not enough to skew the averages. The same can be said about the error bars which are small for the categories where the sample size is large. We can notice that for 8 and 9 siblings category the data points seem to be a lot more uniformly distributed, having no clusters. The outliters might influence these categories more as the (longer) error bars might also indicate.

Visualization B

For this visualization we're using the `gapminder` dataset again.

Packages

```
library(ggplot2)
library(gapminder)
```

Code to create visualization

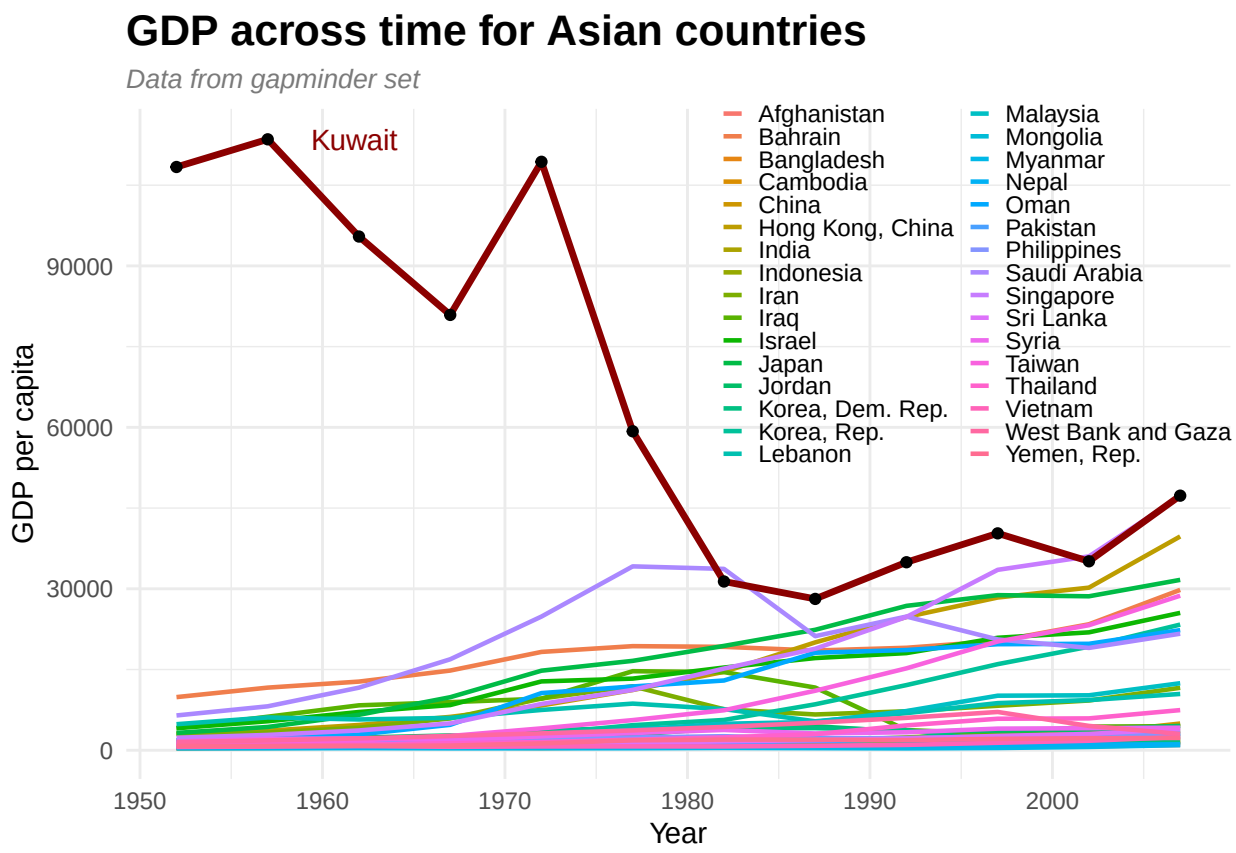
```
# Filter out all the other continents
asian_countries <- gapminder %>% filter(continent == "Asia")

# Find the country with the highest GDP per cap and get it's name(Kuwait)
declining_country <- gapminder %>% filter(gdpPercap == max(gdpPercap))
country_name <- as.character(declining_country$country)

# Take out Kuwait from first set and make it its own
asian_countries <- asian_countries %>% filter(country != country_name)
declinig_GDP <- gapminder %>% filter(country == country_name)

# Create custom theme for labels and legend position
custom_theme <- theme(
  plot.title = element_text(size = 16, face = "bold", colour = "black"),
  plot.subtitle = element_text(size = 10, face = "italic", colour = "gray48"),
  legend.position = c(.77, .75),
  legend.key.size = unit(0.3, "cm"),
  legend.title = element_blank()
)

ggplot(asian_countries, aes(x = year, y = gdpPercap)) +
  # Plot all other Asian countries
  geom_line(aes(group = country, color = country), size = 0.8) +
  # Plot the interesting one(Kuwait) with a hand-picked color
  geom_line(data = declinig_GDP, color = "darkred", size = 1.2) +
  geom_point(data = declinig_GDP, color = "black") +
  geom_text(
    data = declining_country, aes(label = country), size = 4,
    color = "darkred", hjust = -0.5
  ) +
  labs(
    x = "Year", y = "GDP per capita",
    title = "GDP across time for Asian countries",
    subtitle = "Data from gapminder set"
  ) +
  theme_minimal() +
  custom_theme
```


GDP across time for Asian countries
Data from gapminder set
Kuwait
Afghanistan
Bahrain
Bangladesh
Cambodia
China
Hong Kong, China
India
Indonesia
Iran
Iraq
Israel
Japan
Jordan
Korea, Dem. Rep.
Korea, Rep.
Lebanon
Malaysia
Mongolia
Myanmar
Nepal
Oman
Pakistan
Philippines
Saudi Arabia
Singapore
Sri Lanka
Syria
Taiwan
Thailand
Vietnam
West Bank and Gaza
Yemen, Rep.
GDP per capita
90000
60000
30000
0
1950
1960
1970
1980
1990
2000
Year