Project Proposal

Due November 16 at 11:59pm

Ammy Lin

Load Packages

```
library(tidyverse)
library(dplyr)
library(ggplot2)
```

Dataset

Data source: World Happiness Report (Kaggle)

Brief description: This dataset takes responses from the Gallup World Poll, which assesses subjective well-being across 155 countries. These responses are based on the Cantril ladder question, which asks respondents to think of a ladder, where the best possible life for them is a 10 and the worst possible life is a 0, and rate their own current lives on that scale. Each row in the dataset represents a country along with its happiness metrics from 2015-2019 (each year is a separate CSV file, so we will likely choose one year, say 2015, to focus on for our project), including the happiness score, GDP per capita, family support, life expectancy, freedom, and trust in government.

Research question 1: Is there a significant association between life expectancy and national happiness, after adjusting for GDP per capita?

- Outcome variable (include the name/description and type of variable): Happiness score (continuous outcome variable), which is a continuous numeric measure ranging roughly from 0 to 10 that reflects the average life satisfaction of a country's population, with higher values indicating greater happiness; it is derived from survey responses and supporting factors such as GDP per capita, social support, life expectancy, freedom, and generosity.

Research question 2: Is the likelihood of a country belonging to a region with a higher average happiness group (Low, Medium, or High) associated with greater generosity and freedom, and how do these two factors interact?

- Outcome variable (include the name/description and type of variable): Regional happiness group (ordinal categorical outcome variable), which classifies countries into three groups (low, medium, and high) based on the average happiness score of the region they belong to, with higher categories indicating regions with greater overall happiness.

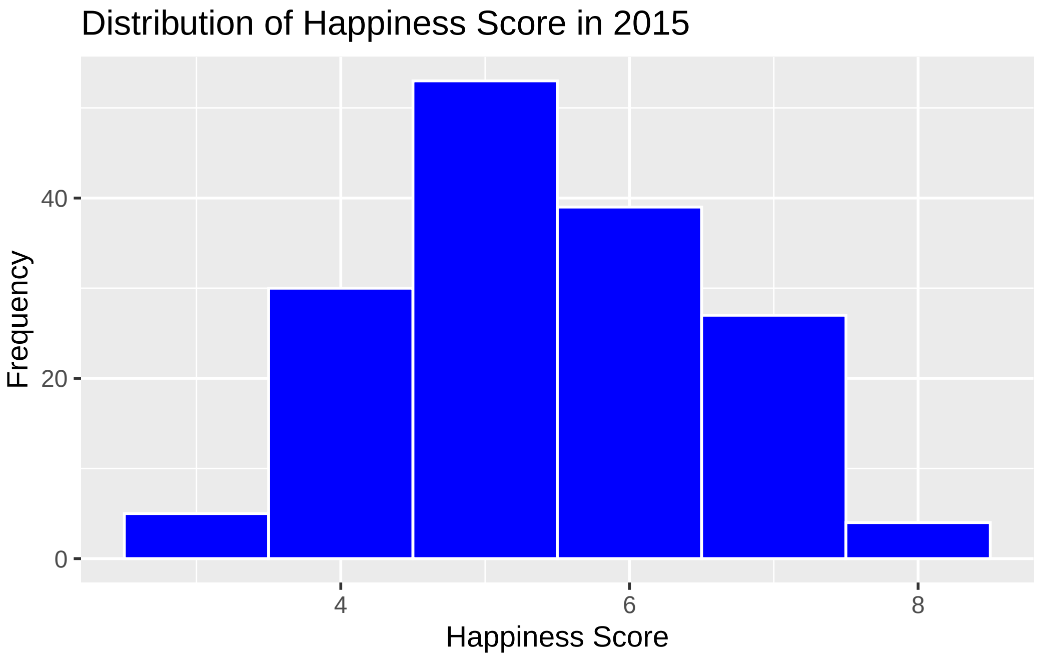
Load the data and provide a `glimpse()`:

```
happiness_2015 <- read.csv("https://github.com/lingyuehao/ids702_Team3/raw/refs/heads/main/20
glimpse(happiness_2015)
```

```
Rows: 158
Columns: 12
$ Country                       <chr> "Switzerland", "Iceland", "Denmark", "No~
$ Region                        <chr> "Western Europe", "Western Europe", "Wes~
$ Happiness.Rank                <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1~
$ Happiness.Score               <dbl> 7.587, 7.561, 7.527, 7.522, 7.427, 7.406~
$ Standard.Error                <dbl> 0.03411, 0.04884, 0.03328, 0.03880, 0.03~
$ Economy..GDP.per.Capita.      <dbl> 1.39651, 1.30232, 1.32548, 1.45900, 1.32~
$ Family                        <dbl> 1.34951, 1.40223, 1.36058, 1.33095, 1.32~
$ Health..Life.Expectancy.      <dbl> 0.94143, 0.94784, 0.87464, 0.88521, 0.90~
$ Freedom                       <dbl> 0.66557, 0.62877, 0.64938, 0.66973, 0.63~
$ Trust..Government.Corruption. <dbl> 0.41978, 0.14145, 0.48357, 0.36503, 0.32~
$ Generosity                    <dbl> 0.29678, 0.43630, 0.34139, 0.34699, 0.45~
$ Dystopia.Residual             <dbl> 2.51738, 2.70201, 2.49204, 2.46531, 2.45~
```

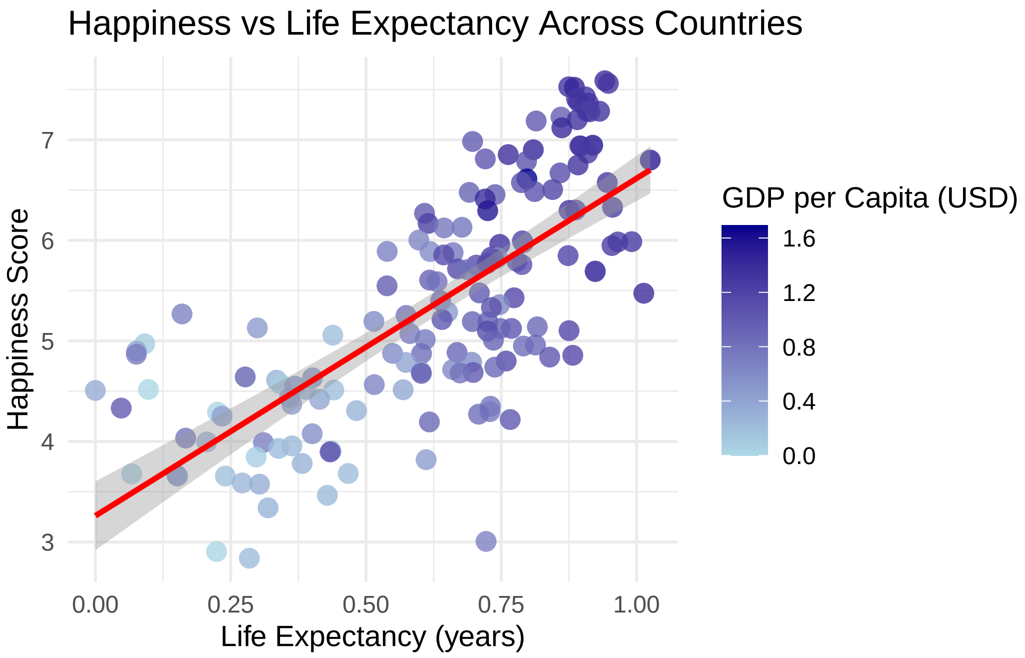
Exploratory Plots:

```
# Research Question 1
# Outcome variable: happiness score (continuous)

# 1. Plot the outcome variable
ggplot(data=happiness_2015, aes(x = Happiness.Score)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "white") +
  labs(title = "Distribution of Happiness Score in 2015",
       x = "Happiness Score",
       y = "Frequency")
```



```r
# 2. Exploratory plot
# Dependent variable: happiness score (continuous)
# Primary independent variable: life expectancy (continuous)
# Control variable: GDP per capita (continuous)

# Scatter plot for happiness vs. life expectancy
ggplot(happiness_2015, aes(x = Health..Life.Expectancy., y = Happiness.Score)) +
  geom_point(aes(color = Economy..GDP.per.Capita.), size = 3, alpha = 0.8) +
  scale_color_gradient(low = "lightblue", high = "darkblue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    title = "Happiness vs Life Expectancy Across Countries",
    x = "Life Expectancy (years)",
    y = "Happiness Score",
    color = "GDP per Capita (USD)"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

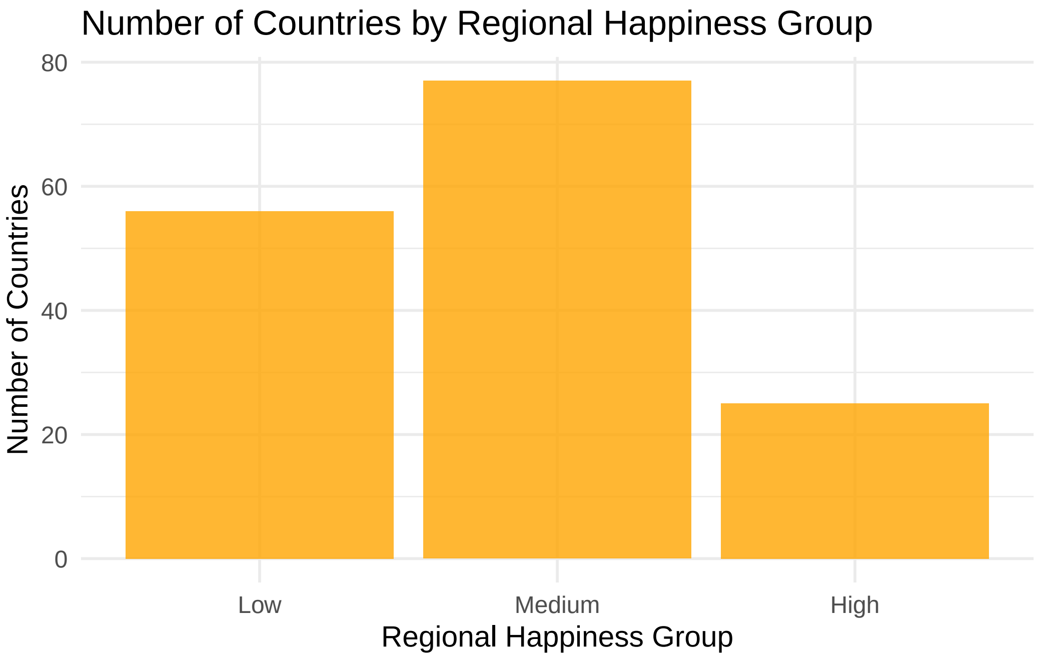


```
# Research Question 2
# Outcome variable: Regional Happiness Group (ordinal)

# 1. Create the outcome variable
# Calculate average happiness score per region and rank
region_levels <- happiness_2015 %>%
  group_by(Region) %>%
  summarize(avg_happiness = mean(Happiness.Score, na.rm = TRUE)) %>%
  arrange(avg_happiness) %>%
  mutate(region_happiness_level = dense_rank(avg_happiness))

# Add Low / Medium / High grouping directly to region_levels
region_levels <- region_levels %>%
  mutate(region_happiness_group = cut(region_happiness_level,
                                      breaks = c(0, 3, 7, 10),
                                      labels = c("Low", "Medium", "High"),
                                      ordered_result = TRUE))

# Check the region rankings directly (now includes Low/Medium/High)
region_levels %>%
  arrange(region_happiness_level) %>%
  select(region_happiness_level, region_happiness_group, Region, avg_happiness)
```

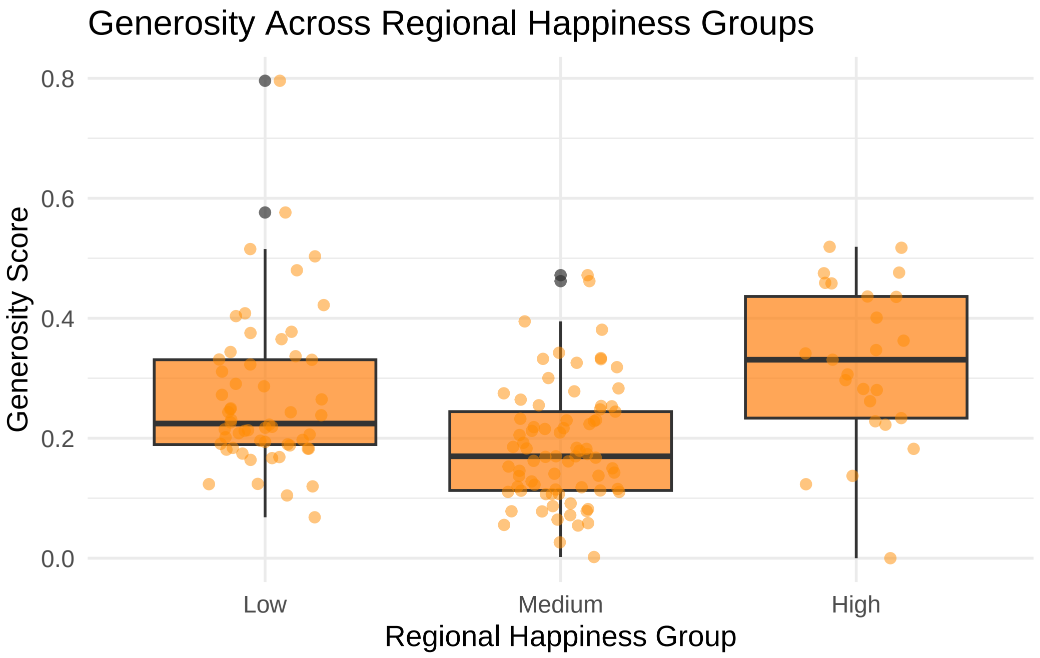
```
# A tibble: 10 x 4
```

```
   region_happiness_level region_happiness_group Region            avg_happiness
                    <int> <ord>                  <chr>                     <dbl>
 1                      1 Low                    Sub-Saharan Afri~          4.20
 2                      2 Low                    Southern Asia              4.58
 3                      3 Low                    Southeastern Asia          5.32
 4                      4 Medium                 Central and East~          5.33
 5                      5 Medium                 Middle East and ~          5.41
 6                      6 Medium                 Eastern Asia               5.63
 7                      7 Medium                 Latin America an~          6.14
 8                      8 High                   Western Europe             6.69
 9                      9 High                   North America              7.27
10                     10 High                   Australia and Ne~          7.28
```

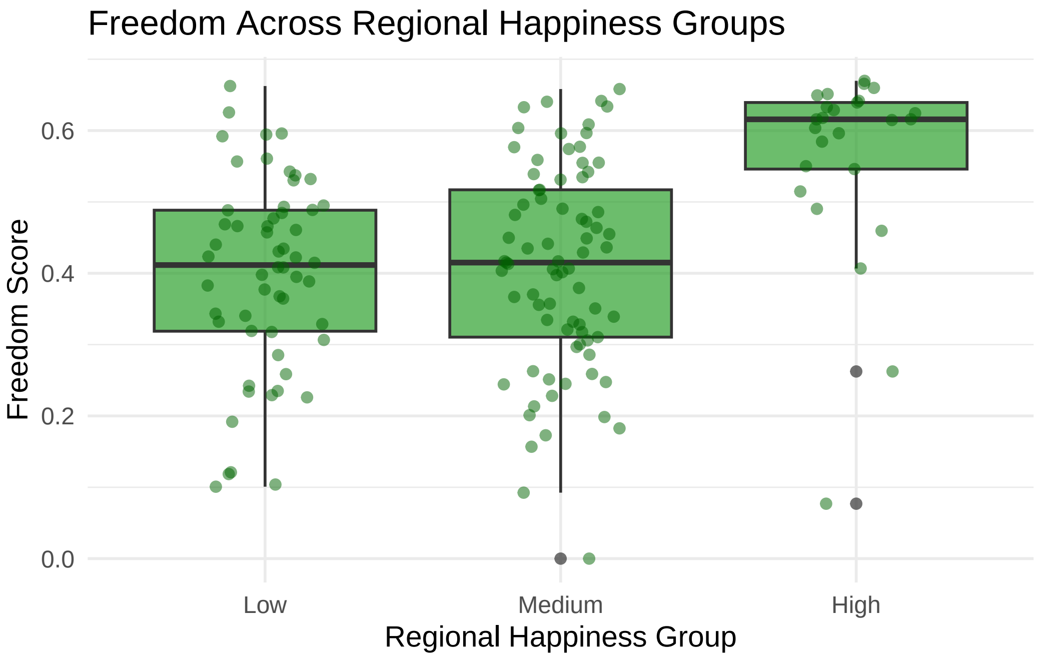
```
# Merge ordinal variable and grouping back into main dataset
happiness_2015 <- happiness_2015 %>%
  left_join(region_levels %>% select(Region,
                                     region_happiness_level,
                                     region_happiness_group),
            by = "Region")

# Convert to ordered factors
happiness_2015 <- happiness_2015 %>%
  mutate(region_happiness_level = factor(region_happiness_level,
                                         levels = sort(unique(region_happiness_level)),
                                         ordered = TRUE),
         region_happiness_group = factor(region_happiness_group,
                                         levels = c("Low", "Medium", "High"),
                                         ordered = TRUE))

# 2. Plot the outcome variable
ggplot(happiness_2015, aes(x = region_happiness_group)) +
  geom_bar(fill = "orange", alpha = 0.8) +
  labs(
    title = "Number of Countries by Regional Happiness Group",
    x = "Regional Happiness Group",
    y = "Number of Countries"
  ) +
  theme_minimal()
```



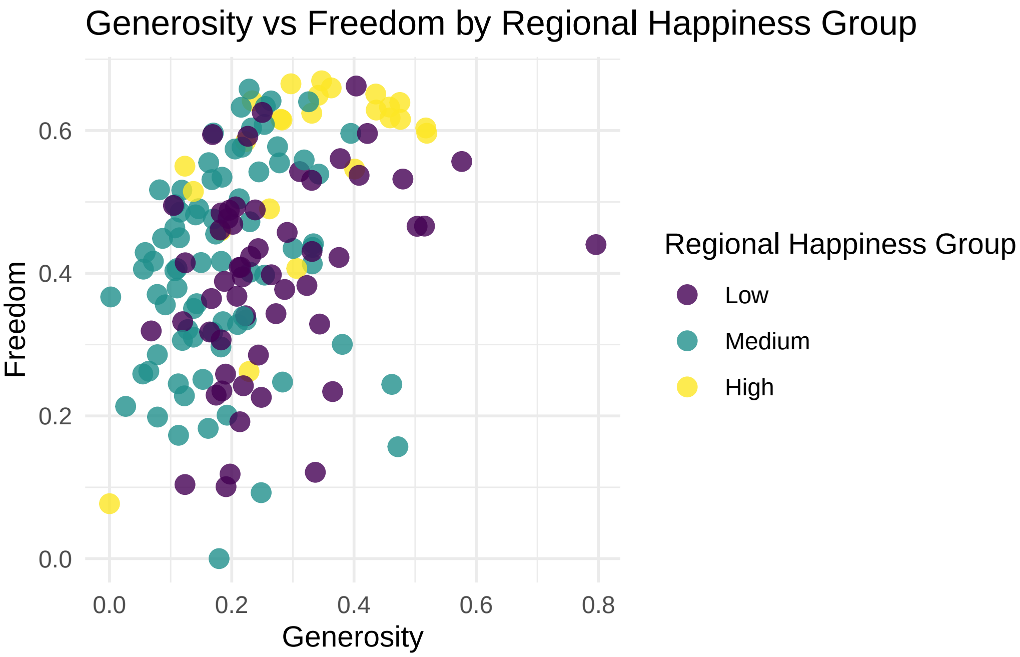
```
# 3. Exploratory plots
# Dependent variable: regional happiness group (ordinal; low/medium/high)
# Primary independent variables: generosity (continuous), freedom (continuous)
# Interaction term: generosity * freedom

# Boxplot of generosity by regional happiness group
ggplot(happiness_2015, aes(x = region_happiness_group, y = Generosity)) +
  geom_boxplot(fill = "#ff7f0e", alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.5, color = "darkorange") +
  labs(
    title = "Generosity Across Regional Happiness Groups",
    x = "Regional Happiness Group",
    y = "Generosity Score"
  ) +
  theme_minimal()
```



```
# Boxplot of freedom by regional happiness group
ggplot(happiness_2015, aes(x = region_happiness_group, y = Freedom)) +
  geom_boxplot(fill = "#2ca02c", alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.5, color = "darkgreen") +
  labs(
    title = "Freedom Across Regional Happiness Groups",
    x = "Regional Happiness Group",
    y = "Freedom Score"
  ) +
  theme_minimal()
```



```
# Interaction plot
ggplot(happiness_2015, aes(x = Generosity, y = Freedom, color = region_happiness_group)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(
    title = "Generosity vs Freedom by Regional Happiness Group",
    x = "Generosity",
    y = "Freedom",
    color = "Regional Happiness Group"
  ) +
  theme_minimal()
```



Generosity vs Freedom by Regional Happiness Group
Freedom
0.0
0.2
0.4
0.6
Generosity
0.0
0.2
0.4
0.6
0.8
Regional Happiness Group
Low
Medium
High

Dataset 2

Data source:

Brief description:

Research question 1:

- Outcome variable (include the name/description and type of variable):

Research question 2:

- Outcome variable (include the name/description and type of variable):

Load the data and provide a `glimpse()`:

Exploratory Plots:

Dataset 3 (optional)

Data source:

Brief description:

Research question 1:

- Outcome variable (include the name/description and type of variable):

Research question 2:

- Outcome variable (include the name/description and type of variable):

Load the data and provide a `glimpse()`:

Exploratory Plots:

Team Charter

When will you meet as a team to work on the project components? Will these meetings be held in person or virtually?

What is your group policy on missing team meetings (e.g., how much advance notice should be provided)?

How will your team communicate (email, Slack, text messages)? What is your policy on appropriate response time (within a certain number of hours? Nights/weekends?)?