Project Proposal

The Four-mula: Hellen Han, Lila Rogers, Amy Duan, Eva Aggarwal

```
library(tidyverse)
library(tidymodels)

data <- read_csv("data/Life-Expectancy-Data-Updated.csv")
```

Introduction - Lila

- An introduction to the subject matter you're investigating (citing any relevant literature)
- Statement of a well-developed research question.
- The motivation for your research question and why it is important
- Your team's hypotheses regarding the research question
  - This is a narrative about what you think regarding the research question, not formal statistical hypotheses.Data description

Exploratory data analysis - Eva

- The source of the data set
- A description of when and how the data were originally collected (by the original data curator, not necessarily how you found the data)
- A description of the observations and general characteristics being measured

Analysis approach

- Description of data processing you need to do to prepare for analysis, such as joining multiple data sets, handling missing data, etc.

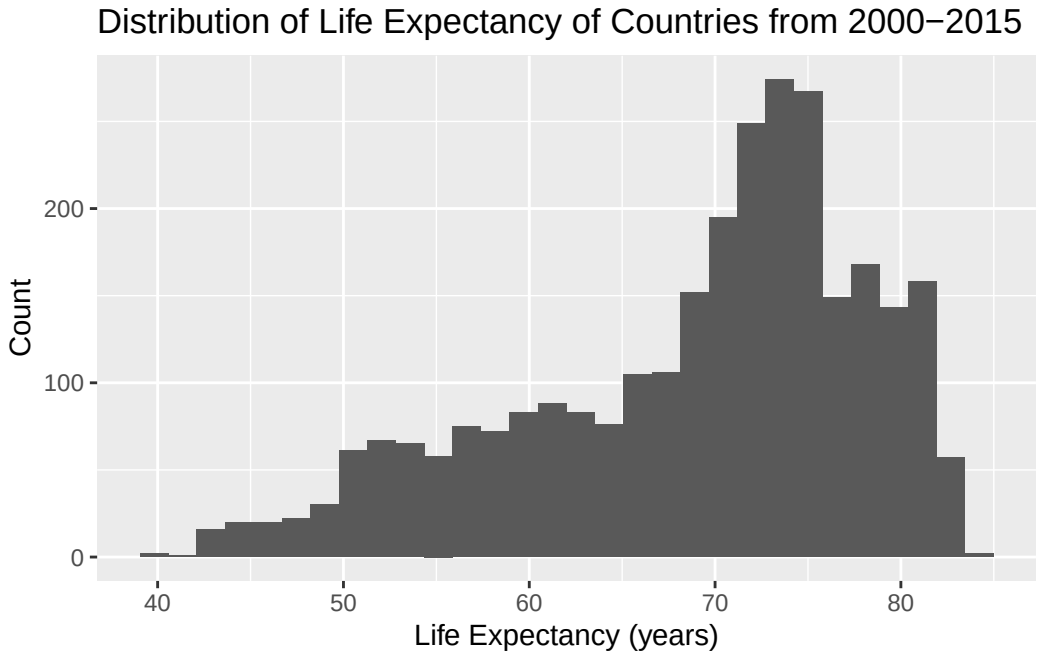
For the following life expectancy dataset, to prepare the dataset for analysis, first, we need to consider missing data values within the dataset, eliminating observations that are incomplete across our predictor/response variables. Furthermore, we need to check for data inconsistencies in the quantitative variables by analyzing the summary statistics for the predictor variables and note whether there are extreme outliers in the min/max values.

- Visualizations, summary statistics, and narrative to describe the distribution of the **Life_expectancy** variable.

```
summary(data$Life_expectancy)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.40   62.70   71.40   68.86   75.40   83.80
```

```
data %>%
  ggplot(mapping = aes(x = Life_expectancy)) +
  geom_histogram() +
  labs(title = "Distribution of Life Expectancy of Countries from 2000-2015",
       x = "Life Expectancy (years)",
       y = "Count")
```




```
data %>%
  ggplot(mapping = aes(x = Life_expectancy)) +
  geom_boxplot() +
  labs(title = "Distribution of Life Expectancy of Countries from 2000-2015",
       x = "Life Expectancy (years)",
       y = "Count")
```

The visualizations describe that the distribution of life expectancy of countries from 2000-2015 is left-skewed and unimodal, indicating that most countries have a relatively higher life expectancy. Furthermore, the distribution of life expectancy of countries have a center of approximately 71.40 years, described by the median, and a spread of approximately 12.70 years, described by the interquartile range. There seem to be a few outliers, as shown in the boxplot visualization, with a few countries with a life expectancy less than 45. The majority of countries' life expectancies tend to be between approximately 70 and 80 years.

Data dictionary - Hellen

The data dictionary can be found here [Update the link and remove this note!]