P8106 Midterm Project

Paula Wu

```
library(tidyverse)
library(viridis)
library(GGally)
library(caret)
library(patchwork)
```

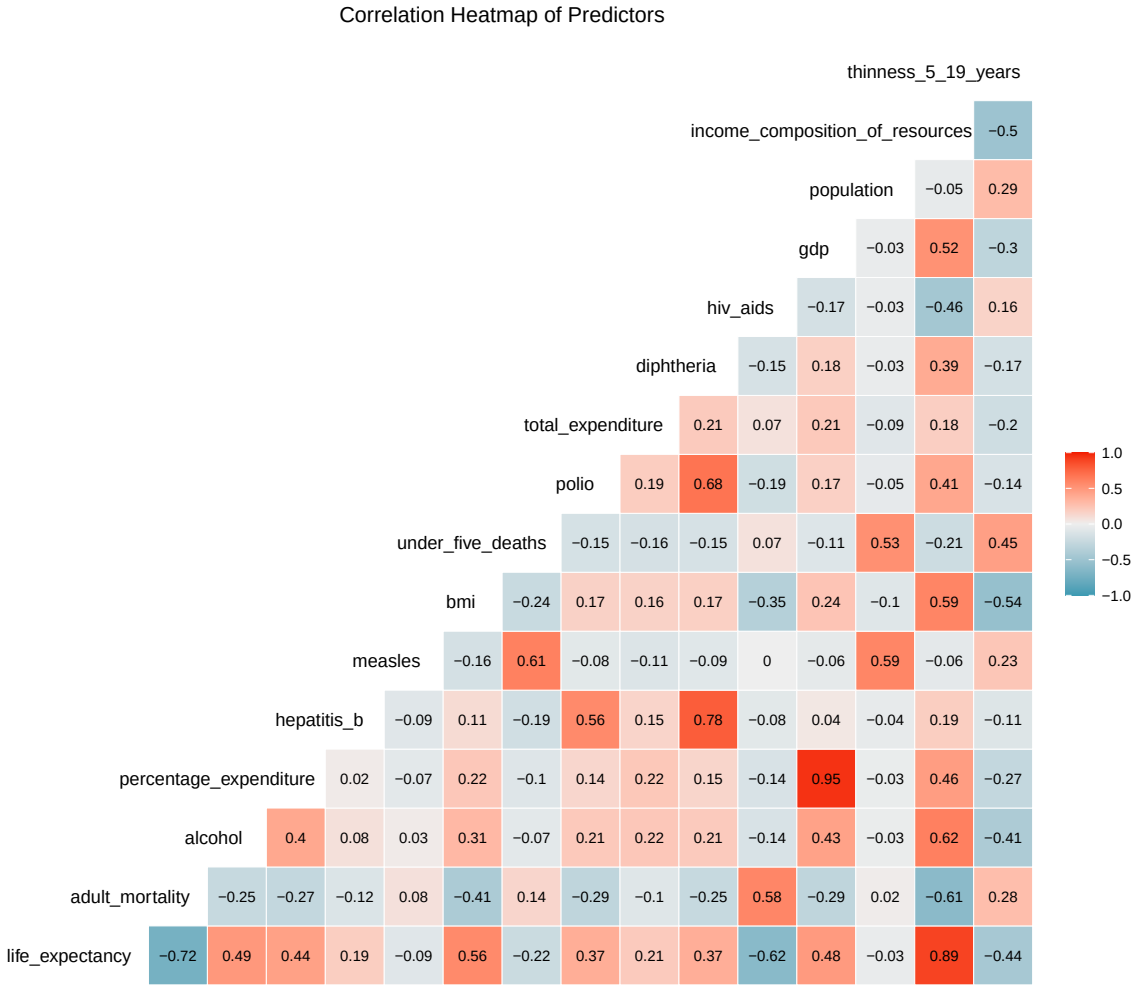
Read in data (source: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who)

```
life_exp = read_csv("./data/Life Expectancy Data.csv") %>%
  janitor::clean_names() %>%
  drop_na() %>%
  filter(year %in% c(2011, 2012, 2013, 2014, 2015)) %>%
  mutate(status = factor(status, levels = c("Developing", "Developed")),
         thinness_5_19_years = thinness_1_19_years + thinness_5_9_years) %>%
  select(-infant_deaths, -country, -thinness_1_19_years, -thinness_5_9_years, -schooling, -year)

life_exp
```

```
## # A tibble: 522 x 17
##    status  life_expectancy adult_mortality alcohol percentage_expen~ hepatitis_b
##    <fct>             <dbl>           <dbl>   <dbl>             <dbl>       <dbl>
##  1 Develo~            65               263    0.01             71.3           65
##  2 Develo~            59.9             271    0.01             73.5           62
##  3 Develo~            59.9             268    0.01             73.2           64
##  4 Develo~            59.5             272    0.01             78.2           67
##  5 Develo~            59.2             275    0.01              7.10          68
##  6 Develo~            77.8              74    4.6             365.            99
##  7 Develo~            77.5               8    4.51            429.            98
##  8 Develo~            77.2              84    4.76            431.            99
##  9 Develo~            76.9              86    5.14            412.            99
## 10 Develo~            76.6              88    5.37            437.            99
## # ... with 512 more rows, and 11 more variables: measles <dbl>, bmi <dbl>,
## #   under_five_deaths <dbl>, polio <dbl>, total_expenditure <dbl>,
## #   diphtheria <dbl>, hiv_aids <dbl>, gdp <dbl>, population <dbl>,
## #   income_composition_of_resources <dbl>, thinness_5_19_years <dbl>
```

```
life_exp %>%
  ggcorr(label=TRUE, hjust = 0.9, layout.exp = 2, label_size = 3, label_round = 2) +
  ggtitle("Correlation Heatmap of Predictors") +
  theme(plot.title = element_text(hjust = 0.5))
```



test train split (70:30)

```
# partition the dataset
set.seed(123)
indexTrain = createDataPartition(y = life_exp$life_expectancy, p = 0.7, list = FALSE)
trainData = life_exp[indexTrain, ]
testData = life_exp[-indexTrain, ]

# matrix
x = model.matrix(life_expectancy ~., trainData)[, -1]
y = trainData$life_expectancy
```

Linear models

Least square