'Descriptive Analysis of Demographic Data'

Bushra Tariq Kiyani (230204)

2022-11-09

```
library('GGally')
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library('dplyr')
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library('ggplot2')
library('gridExtra')
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library('tidyverse')
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## v readr     2.1.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x gridExtra::combine() masks dplyr::combine()
## 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 become errors
```

```
#citation(package = "gridExtra")
```

```r
#Loading the data
census_data <- read.csv('census2002_2022.csv')[,-1]
```

```r
#Changing Column Names for better readability
colnames(census_data) <- c("Country","Subregion","Region","Year",
                           "Life_exp_both","Life_exp_male","Life_exp_female",
                           "Mortality_rate_both", "Mortality_rate_male",
                           "Mortality_rate_female")
```

```r
str(census_data)
```

```
## 'data.frame':    454 obs. of  10 variables:
##  $ Country              : chr  "Afghanistan" "Afghanistan" "Albania" "Albania" ...
##  $ Subregion            : chr  "South-Central Asia" "South-Central Asia" "Southern Europe" "Southern
##  $ Region               : chr  "Asia" "Asia" "Europe" "Europe" ...
##  $ Year                 : int  2002 2022 2002 2022 2002 2022 2002 2022 2002 2022 ...
##  $ Life_exp_both        : num  46.1 53.6 75.4 79.5 73 ...
##  $ Life_exp_male        : num  45.1 52.1 72.8 76.8 72.1 ...
##  $ Life_exp_female      : num  47.1 55.3 78.4 82.3 74 ...
##  $ Mortality_rate_both  : num  213.9 154.1 24.1 12.7 40.5 ...
##  $ Mortality_rate_male  : num  220.4 161.8 25.4 13.7 42.9 ...
##  $ Mortality_rate_female: num  207 146.1 22.6 11.5 38.1 ...
```

```r
length(unique(census_data$Country))
```

```
## [1] 227
```

```r
# check for missing values in each column
colSums(is.na(census_data))
```

```
##               Country             Subregion                Region
##                     0                     4                     4
##                  Year         Life_exp_both         Life_exp_male
##                     0                     6                     6
##       Life_exp_female   Mortality_rate_both   Mortality_rate_male
##                     6                     6                     6
## Mortality_rate_female
##                     6
```

```r
# Display Region Names
unique_regions <- distinct(census_data, Region)$Region
print(unique_regions)
```

```
## [1] "Asia"     "Europe"   "Africa"   "Oceania"  "Americas" NA
```

```r
# display rows where Region and Subregion columns are missing
subset(census_data, is.na(Region) & is.na(Subregion))
```

```
##           Country Subregion Region Year Life_exp_both Life_exp_male
## 101       Curaçao      <NA>   <NA> 2002         76.20         74.10
## 102       Curaçao      <NA>   <NA> 2022         79.42         77.09
## 107 Côte d'Ivoire      <NA>   <NA> 2002         51.17         49.37
## 108 Côte d'Ivoire      <NA>   <NA> 2022         62.26         60.07
##     Life_exp_female Mortality_rate_both Mortality_rate_male
## 101           78.41               11.89               12.60
## 102           81.87                8.84                9.65
## 107           53.02              137.61              149.63
## 108           64.52               76.18               84.78
```

```
##     Mortality_rate_female
## 101                 11.15
## 102                  7.99
## 107                125.24
## 108                 67.33
```

```
# Fill missing values in the "Region" and "Subregion" columns:

country_mapping <- tibble(
  Country = c("Curaçao", "Côte d'Ivoire"),
  Region = c("Americas", "Africa"),
  Subregion = c("South America", "Western Africa")
)

census_data <- census_data %>%
  left_join(country_mapping, by = "Country", suffix = c("", "_mapped")) %>%
  mutate(
    Region = coalesce(Region, Region_mapped),
    Subregion = coalesce(Subregion, Subregion_mapped)
  ) %>%
  select(-Region_mapped, -Subregion_mapped)
```

```
# display rows where Regions and Subregions are added
 census_data %>%
  filter(Country %in% c("Curaçao", "Côte d'Ivoire")) %>%
  select(Country, Region, Subregion)
```

```
##         Country   Region      Subregion
## 1       Curaçao Americas  South America
## 2       Curaçao Americas  South America
## 3 Côte d'Ivoire   Africa Western Africa
## 4 Côte d'Ivoire   Africa Western Africa
```

```
# display rows where Life_exp_both and Under_5_Mortality_both columns are missing
subset(census_data, is.na(Life_exp_both) & is.na(Mortality_rate_both))
```

```
##           Country        Subregion   Region Year Life_exp_both Life_exp_male
## 235         Libya  Northern Africa   Africa 2002            NA            NA
## 325   Puerto Rico        Caribbean Americas 2002            NA            NA
## 379   South Sudan  Northern Africa   Africa 2002            NA            NA
## 385         Sudan  Northern Africa   Africa 2002            NA            NA
## 393         Syria     Western Asia     Asia 2002            NA            NA
## 429 United States Northern America Americas 2002            NA            NA
##     Life_exp_female Mortality_rate_both Mortality_rate_male
## 235              NA                  NA                  NA
## 325              NA                  NA                  NA
## 379              NA                  NA                  NA
## 385              NA                  NA                  NA
## 393              NA                  NA                  NA
## 429              NA                  NA                  NA
##     Mortality_rate_female
## 235                    NA
## 325                    NA
## 379                    NA
## 385                    NA
## 393                    NA
```

```
## 429                    NA
```

```
# Replace NaN values with the median
census_data <- census_data %>%
  mutate(across(c("Life_exp_both", "Life_exp_male", "Life_exp_female",
                  "Mortality_rate_both", "Mortality_rate_male",
                  "Mortality_rate_female"),
                ~ replace(., is.na(.), median(., na.rm = TRUE))))
```

```
#Ordering the Data according Region and Subregion
census_data <- census_data[order(census_data$Region,census_data$Subregion),]
head(census_data)
```

```
##      Country      Subregion Region Year Life_exp_both Life_exp_male
## 65   Burundi Eastern Africa Africa 2002         55.85         53.79
## 66   Burundi Eastern Africa Africa 2022         67.42         65.32
## 87   Comoros Eastern Africa Africa 2002         60.54         58.91
## 88   Comoros Eastern Africa Africa 2022         67.20         64.93
## 111 Djibouti Eastern Africa Africa 2002         58.33         56.08
## 112 Djibouti Eastern Africa Africa 2022         65.30         62.72
##     Life_exp_female Mortality_rate_both Mortality_rate_male
## 65            57.98              126.12               134.64
## 66            69.59               53.90                59.31
## 87            62.21              117.94               130.97
## 88            69.54               78.54                90.98
## 111           60.65               95.25               106.07
## 112           67.96               63.09                71.80
##     Mortality_rate_female
## 65                 117.34
## 66                  48.33
## 87                 104.52
## 88                  65.72
## 111                 84.09
## 112                 54.12
```

```
#Factoring with the Sub regions
census_data$Subregion <- factor(census_data$Subregion,
                                levels = unique(census_data$Subregion[order(census_data$Region)]))
```

```
#Split the Data Based on the year
census_data_2022 <- census_data %>% filter(Year == 2022)
census_data_2002 <- census_data %>% filter(Year == 2002)
```

```
#Summary of data 2022
census_data_2022   %>% summary()
```

```
##    Country                     Subregion      Region              Year
##  Length:227         Caribbean         : 24   Length:227         Min.   :2022
##  Class :character   Western Asia      : 19   Class :character   1st Qu.:2022
##  Mode  :character   Eastern Africa    : 17   Mode  :character   Median :2022
##                     Western Africa    : 17                      Mean   :2022
##                     Southern Europe   : 16                      3rd Qu.:2022
##                     South-Central Asia: 14                      Max.   :2022
##                     (Other)           :120
##  Life_exp_both   Life_exp_male   Life_exp_female Mortality_rate_both
##  Min.   :53.65   Min.   :52.10   Min.   :55.28   Min.   :  1.940
##  1st Qu.:70.05   1st Qu.:67.93   1st Qu.:72.63   1st Qu.:  7.415
```

```
##  Median :75.82   Median :73.26   Median :78.69   Median : 15.080
##  Mean   :74.58   Mean   :72.10   Mean   :77.18   Mean   : 26.677
##  3rd Qu.:79.66   3rd Qu.:77.19   3rd Qu.:82.56   3rd Qu.: 37.775
##  Max.   :89.52   Max.   :85.70   Max.   :93.49   Max.   :154.130
##
##  Mortality_rate_male Mortality_rate_female
##  Min.   :  2.03      Min.   :  1.640
##  1st Qu.:  8.32      1st Qu.:  6.345
##  Median : 17.55      Median : 13.620
##  Mean   : 29.23      Mean   : 24.012
##  3rd Qu.: 41.02      3rd Qu.: 34.470
##  Max.   :161.78      Max.   :146.090
##
```

Task 1: Frequency Distributions of Different Variables

```
custom_theme <- theme(plot.title = element_text(hjust = 0.5, size = 13, face="bold"),
                      axis.text=element_text(size=12),
                      axis.title = element_text(size = 14),
                      axis.title.y=element_text(size=12), # Change the size of the y-axis title
                      legend.text=element_text(size=12),
                      legend.title = element_text(size=14))

binwidth <- 1
Life_exp_both  <- ggplot(census_data_2022, aes(x = Life_exp_both)) +
           geom_histogram(aes(fill = after_stat(count)), col = "black",
                          binwidth = binwidth, linewidth = 0.1) +
           scale_x_continuous(name = "Life expectancy of both sexes in years") +
           scale_y_continuous(name = "Frequency") +
           scale_fill_viridis_c() +
           ggtitle("c) Frequency Distribution of\nLife expectancy of both sexes") +
           custom_theme+
          theme(legend.position = c(0, 1), # Set legend position to upper right corner
          legend.justification = c(0, 1)) # Align legend to the upper right corner
binwidth <- 3

Mortality_rate_both <- ggplot(census_data_2022, aes(x = Mortality_rate_both)) +
  geom_histogram(aes(fill = after_stat(count)), binwidth = binwidth, color = "black",
                 linewidth = 0.1) +
  scale_x_continuous(name = "Mortality rate of both sexes in years") +
  scale_y_continuous(name = "Frequency") +
  scale_fill_viridis_c() +
  ggtitle("d) Frequency Distribution of\nMortality rate of both sexes") +
  custom_theme+
  theme(legend.position = c(1, 1), # Set legend position to upper right corner
        legend.justification = c(1, 1)) # Align legend to the upper right corner
```

```
# Create custom theme for plot
custom_theme <- theme(plot.title = element_text(hjust = 0.5, size = 13, face="bold"),
                      axis.text=element_text(size=12),
                      axis.title = element_text(size = 14),
                      axis.title.y=element_text(size=12), # Change the size of the y-axis title
                      legend.text=element_text(size=12),
```

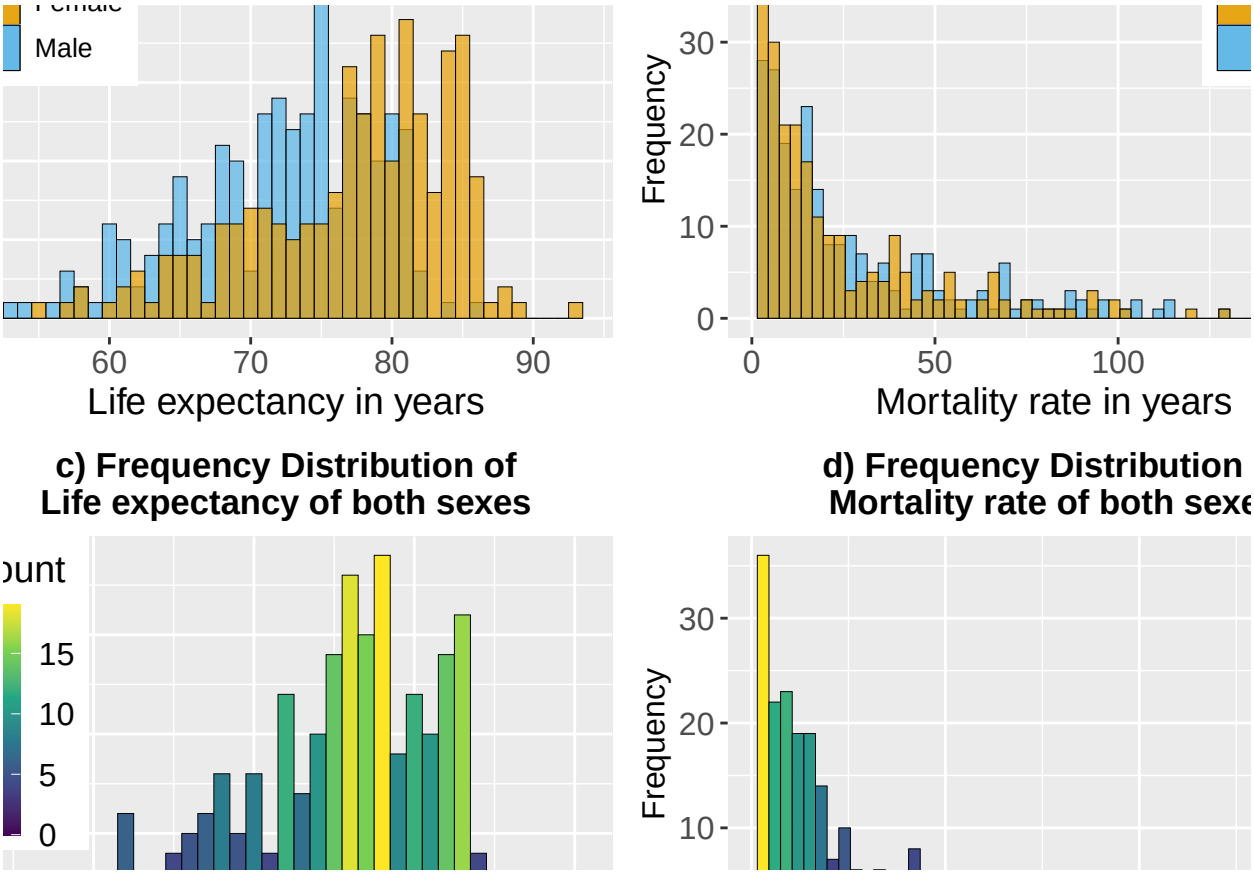
```
                      legend.title = element_text(size=14))
# Define custom color palette for male and female
custom_palette <- c("#E69F00", "#56B4E9")
# Create ggplot object
combined_life_plot <- ggplot() +
  geom_histogram(aes(x=census_data_2022$Life_exp_male, fill="Male"),
                 data=data.frame(census_data_2022$Life_exp_male),col = "black",
                 size = 0.1, alpha=0.7, binwidth=1) +
  geom_histogram(aes(x=census_data_2022$Life_exp_female, fill="Female"),
                 data=data.frame(census_data_2022$Life_exp_female), col = "black",
                 size = 0.1, alpha=0.7, binwidth=1) +
  labs(title="a) Frequency Distribution\nof life expectancy in female and male",
       x="Life expectancy in years", y="Frequency") +
  scale_fill_manual(values = custom_palette) +
  labs(fill = "Gender") +
  custom_theme +
  theme(legend.position = c(0, 1), # Set legend position to upper right corner
        legend.justification = c(0, 1), # Align legend to the upper right corner
        legend.title = element_text(size = 11), # Reduce legend title size
        legend.text = element_text(size = 10)) # Reduce legend text size
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Create custom theme for plot
custom_theme <- theme(plot.title = element_text(hjust = 0.5, size = 13, face="bold"),
                      axis.text=element_text(size=12),
                      axis.title = element_text(size = 14),
                      axis.title.y=element_text(size=12), # Change the size of the y-axis title
                      legend.text=element_text(size=12),
                      legend.title = element_text(size=14))
# Define custom color palette for male and female
custom_palette <- c("#E69F00", "#56B4E9","darkgreen")
# Create ggplot object
combined_mortality_plot <- ggplot() +
  geom_histogram(aes(x=census_data_2022$Mortality_rate_male, fill="Male"),
                 data=data.frame(census_data_2022$Mortality_rate_male),col = "black",
                 size = 0.1, alpha=0.7, binwidth=3) +
  geom_histogram(aes(x=census_data_2022$Mortality_rate_female, fill="Female"),
                 data=data.frame(census_data_2022$Mortality_rate_female), col = "black",
                 size = 0.1, alpha=0.7, binwidth=3) +
  labs(title="b) Frequency Distribution\nof Mortality rate in female and male",
       x="Mortality rate in years", y="Frequency") +
  scale_fill_manual(values = custom_palette) +
  labs(fill = "Gender") +
  custom_theme +
  theme(legend.position = c(1, 1), # Set legend position to upper right corner
        legend.justification = c(1, 1), # Align legend to the upper right corner
        legend.title = element_text(size = 11), # Reduce legend title size
        legend.text = element_text(size = 10)) # Reduce legend text size
```

```
# Plot the ggplot object
#print(combined_mortality_plot)

# Combine the plots into a grid with adjusted heights
final_plot_all_n <- grid.arrange(combined_life_plot, combined_mortality_plot,
                                 Life_exp_both, Mortality_rate_both, ncol = 2, nrow = 2,
                                 widths = unit(c(4, 4), "in"), heights = unit(c(3.2,3.2), "in"))
```




```
# Save the final plot to a file with adjusted dimensions
ggsave("histograms_all_n.pdf", plot = final_plot_all_n, width = 8.5, height = 6.5, units = "in")
final_plot_all_n
```

```
## TableGrob (2 x 2) "arrange": 4 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## 3 3 (2-2,1-1) arrange gtable[layout]
## 4 4 (2-2,2-2) arrange gtable[layout]
```

```
# life expectancy by 'Region'
boxplot_life_expectancy <- ggplot(census_data_2022, aes(x = Region, y = Life_exp_both,
                                                        fill = Region)) +
  geom_boxplot() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face="bold"),
        axis.text=element_text(size=12),
        axis.title=element_text(size=14)) +
  xlab("Region") +
```

```
  ylab("Life Expectancy (Years)") +
  ggtitle("Life Expectancy by Region") +
  scale_fill_brewer(palette="Dark2")

ggsave("boxplots_life_expectancy_by_region.pdf", plot = boxplot_life_expectancy)
```

```
## Saving 6.5 x 4.5 in image
```

```
# Mortality rate by 'Region'
boxplot_mortality_rate <- ggplot(census_data_2022, aes(x = Region, y = Mortality_rate_both,
                                                       fill = Region)) +
  geom_boxplot() +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face="bold"),
        axis.text=element_text(size=12),
        axis.title=element_text(size=14)) +
  xlab("Region") +
  ylab("Under age 5 Mortality Rate") +
  ggtitle("Under age 5 Mortality Rate by Region") +
  scale_fill_brewer(palette="Dark2")
ggsave("boxplots_mortality_by_region.pdf", plot = boxplot_mortality_rate)
```

```
## Saving 6.5 x 4.5 in image
```

```
install.packages("patchwork")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

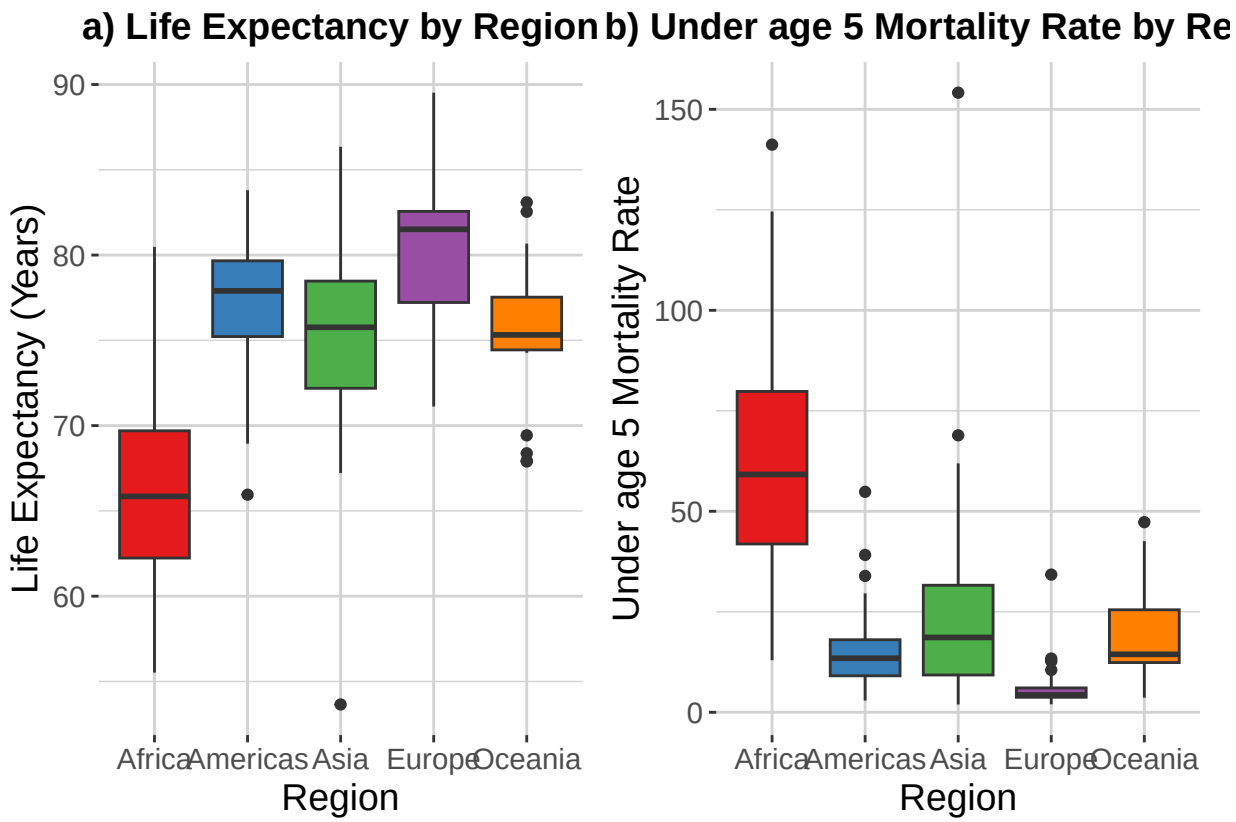
```
library(patchwork)
boxplot_life_expectancy <- ggplot(census_data_2022, aes(x = Region, y = Life_exp_both)) +
  geom_boxplot(aes(fill = Region), show.legend = FALSE) +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face="bold"),
        axis.text=element_text(size=11),
        axis.title=element_text(size=14),
        panel.background = element_rect(fill = "white"),
        panel.grid = element_line(colour = "lightgrey"))  +
  xlab("Region") +
  ylab("Life Expectancy (Years)") +
  ggtitle("a) Life Expectancy by Region") +
  scale_fill_brewer(palette="Set1", guide = FALSE)

boxplot_mortality_rate <- ggplot(census_data_2022, aes(x = Region, y = Mortality_rate_both)) +
  geom_boxplot(aes(fill = Region), show.legend = FALSE) +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face="bold"),
        axis.text=element_text(size=11),
        axis.title=element_text(size=14),
        panel.background = element_rect(fill = "white"),
        panel.grid = element_line(colour = "lightgrey")) +
  xlab("Region") +
  ylab("Under age 5 Mortality Rate") +
  ggtitle("b) Under age 5 Mortality Rate by Region") +
  scale_fill_brewer(palette="Set1")
combined_plot <- boxplot_life_expectancy + boxplot_mortality_rate
ggsave(filename = "combined_plot.pdf", plot = combined_plot, width = 8.5, height = 4, units = "in")
```

```
## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in
```

```
## ggplot2 3.3.4.
## i Please use "none" instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
print(combined_plot)
```




```
# Females Life expectancy above 90 (Highest)
females_above90 = census_data_2022[census_data_2022$
                                   Life_exp_female > 90,][c("Country","Region","Life_exp_female")]
females_above90
```

```
##     Country Region Life_exp_female
## 204  Monaco Europe           93.49
```

```
# Females Life expectancy below 55 (lowest)
females_below55 = census_data_2022[census_data_2022$
                                   Life_exp_female < 60,][c("Country","Life_exp_female")]
females_below55
```

```
##                       Country Life_exp_female
## 10                 Mozambique           58.49
## 13                    Somalia           58.12
## 20  Central African Republic           56.88
## 114               Afghanistan           55.28
```

```
# Highest female frequency interval range
females_interval = census_data_2022[census_data_2022$Life_exp_female > 76 &
                                    census_data_2022$Life_exp_female <= 86,
```

```
                                    c("Country", "Region", "Life_exp_female")]
count(females_interval)
```

```
##     n
## 1 139
```

```
# Males Life expectancy above 85 (Highest)
males_above85 = census_data_2022[census_data_2022$
                                   Life_exp_male > 82,][c("Country","Region","Life_exp_male")]
males_above85
```

```
##       Country Region Life_exp_male
## 111     Macau   Asia         82.09
## 135 Singapore   Asia         83.65
## 204    Monaco Europe         85.70
```

```
# Males Life expectancy below 55 (Lowest)
males_below55 = census_data_2022[census_data_2022$
                                   Life_exp_male < 55,][c("Country","Life_exp_male")]
males_below55
```

```
##                      Country Life_exp_male
## 13                   Somalia         53.39
## 20  Central African Republic         54.19
## 114              Afghanistan         52.10
```

```
# Highest male frequency interval range
males_interval = census_data_2022[census_data_2022$Life_exp_male > 71
                                  & census_data_2022$Life_exp_male <= 81,
                                  c("Country", "Region", "Life_exp_male")]
count(males_interval)
```

```
##     n
## 1 137
```

```
# Highest Mortality rate females
mortality_F = census_data_2022[census_data_2022$
                                 Mortality_rate_female >100,][c("Country","Mortality_rate_female")]
mortality_F
```

```
##                      Country Mortality_rate_female
## 13                   Somalia                128.81
## 20  Central African Republic                119.95
## 21                     Chad                101.80
## 114              Afghanistan                146.09
```

```
# lowest Mortality rate females
mortality_F = census_data_2022[census_data_2022$Mortality_rate_female < 2,][c("Country","Mortality_rate
mortality_F
```

```
##       Country Mortality_rate_female
## 135 Singapore                 1.85
## 196  Slovenia                 1.68
## 204    Monaco                 1.64
```

```
# Lowest rate interval females
mortality_F = census_data_2022[census_data_2022$
                                 Mortality_rate_female >1 & census_data_2022$
```

```
                                Mortality_rate_female <= 6,][c("Country","Mortality_rate_female")]
count(mortality_F)
```

```
##    n
## 1 55
```

```
# Lowest rate interval males
mortality_M = census_data_2022[census_data_2022$
                                Mortality_rate_male >1 & census_data_2022$
                                Mortality_rate_male <= 6,][c("Country","Mortality_rate_male")]
count(mortality_M)
```

```
##    n
## 1 46
```

```
# Highest Mortality rate males
mortality_M = census_data_2022[census_data_2022$
                                Mortality_rate_male >100,][c("Country","Mortality_rate_male")]
mortality_M
```

```
##                      Country Mortality_rate_male
## 13                   Somalia              153.23
## 20  Central African Republic              129.08
## 21                      Chad              114.55
## 24         Equatorial Guinea              113.76
## 31               South Sudan              104.67
## 48                      Mali              101.28
## 50                     Niger              109.77
## 54              Sierra Leone              104.06
## 114              Afghanistan              161.78
```

```
# lowest Mortality rate females
mortality_m = census_data_2022[census_data_2022$Mortality_rate_male < 3,][c("Country","Mortality_rate_ma
mortality_m
```

```
##       Country Mortality_rate_male
## 108     Japan                2.67
## 135 Singapore                2.03
## 171   Finland                2.74
## 173   Iceland                2.91
## 196  Slovenia                2.29
## 204    Monaco                2.31
```

```
# Highest rate interval Life Expectancy both
Life_B = census_data_2022[census_data_2022$
                                Life_exp_both >85 & census_data_2022$
                                 Life_exp_both <= 90,][c("Country","Life_exp_both")]
Life_B
```

```
##       Country Life_exp_both
## 135 Singapore         86.35
## 204    Monaco         89.52
```

```
# Highest rate interval Life Expectancy both
Life_B = census_data_2022[census_data_2022$
                                Life_exp_both >80 & census_data_2022$
                                 Life_exp_both <= 81,][c("Country","Life_exp_both","Region")]
```

```
Life_B
```

```
##                                         Country Life_exp_both   Region
## 52  Saint Helena, Ascension, and Tristan da Cunha         80.48   Africa
## 70                                Saint Barthelemy         80.58 Americas
## 73                                    Saint Martin         80.58 Americas
## 77                      Turks and Caicos Islands         80.82 Americas
## 79                          Virgin Islands, U.S.         80.27 Americas
## 92                                 United States         80.59 Americas
## 186                                    Gibraltar         80.42   Europe
## 227                              Wallis and Futuna         80.67  Oceania
```

```
# Highest Mortality rate
mortality_h = census_data_2022[census_data_2022$
                               Mortality_rate_both > 150,][c("Country","Mortality_rate_both")]
mortality_h
```

```
##         Country Mortality_rate_both
## 114 Afghanistan              154.13
```

```
# Lowest Mortality rate
mortality_l = census_data_2022[census_data_2022$
                               Mortality_rate_both< 2,][c("Country","Region","Mortality_rate_both")]
mortality_l
```

```
##       Country Region Mortality_rate_both
## 135 Singapore   Asia                1.94
## 196  Slovenia Europe                1.99
## 204    Monaco Europe                1.98
```

```
# Lowest Mortality rate interval
mortality_l = census_data_2022[census_data_2022$
                               Mortality_rate_both > 100 & census_data_2022$
                               Mortality_rate_both <= 160,][c("Country","Region","Mortality_rate_both
mortality_l
```

```
##                      Country Region Mortality_rate_both
## 13                   Somalia Africa              141.20
## 20  Central African Republic Africa              124.58
## 21                      Chad Africa              108.30
## 24         Equatorial Guinea Africa              106.43
## 50                     Niger Africa              104.72
## 114             Afghanistan   Asia              154.13
```

```
census_data_2022 %>%
  group_by(Region) %>%
  summarise(Num_Countries = n())
```

```
## # A tibble: 5 x 2
##   Region   Num_Countries
##   <chr>            <int>
## 1 Africa              55
## 2 Americas            50
## 3 Asia                52
## 4 Europe              49
## 5 Oceania             21
```

```r
census_data_2022 %>%
  filter(Subregion == "Southern Europe") %>%
  arrange(desc(Mortality_rate_female)) %>%
  select(Country, Mortality_rate_female) %>%
  slice(1:4)
```

```
##           Country Mortality_rate_female
## 1          Kosovo                 31.58
## 2         Albania                 11.53
## 3         Croatia                 10.90
## 4 North Macedonia                  7.42
```

```r
americas_median_life_exp <- census_data_2022 %>%
  filter(Region == "Oceania") %>%
  group_by(Region) %>%
  summarize(medianLifeExp = median(Life_exp_both))

# To print the result
print(americas_median_life_exp)
```

```
## # A tibble: 1 x 2
##   Region  medianLifeExp
##   <chr>           <dbl>
## 1 Oceania          75.3
```

```r
americas_median_Mor <- census_data_2022 %>%
  filter(Region == "Europe") %>%
  group_by(Region) %>%
  summarize(medianmor = median(Mortality_rate_both))

# To print the result
print(americas_median_Mor)
```

```
## # A tibble: 1 x 2
##   Region medianmor
##   <chr>      <dbl>
## 1 Europe      4.32
```

```r
top_two_countries_oceania <- census_data_2022 %>%
  filter(Region == "Europe") %>%
  arrange(desc(Mortality_rate_female)) %>%
  slice(1:3)

# To print the result
print(top_two_countries_oceania)
```

```
##   Country       Subregion Region Year Life_exp_both Life_exp_male
## 1  Kosovo Southern Europe Europe 2022         71.12         68.83
## 2 Albania Southern Europe Europe 2022         79.47         76.80
## 3 Moldova  Eastern Europe Europe 2022         72.44         68.60
##   Life_exp_female Mortality_rate_both Mortality_rate_male Mortality_rate_female
## 1           73.58               34.25               36.73                 31.58
## 2           82.33               12.66               13.71                 11.53
## 3           76.52               13.36               15.48                 11.12
```

Task 2: Analysis of variability within and between subregions.

```
#Summary for box plot for life expectancy of both sexes in regions and subregions
census_data_2022 %>%                                    # Summary by group using dplyr
  group_by(Region) %>%
  dplyr::summarize(min = min(Life_exp_both),
            q1 = quantile(Life_exp_both, 0.25),
            median = median(Life_exp_both),
            q3 = quantile(Life_exp_both, 0.75),
            max = max(Life_exp_both))
```

```
## # A tibble: 5 x 6
##   Region     min    q1 median    q3   max
##   <chr>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Africa    55.5  62.2   65.8  69.7  80.5
## 2 Americas  66.0  75.2   77.9  79.7  83.8
## 3 Asia      53.6  72.2   75.8  78.5  86.4
## 4 Europe    71.1  77.2   81.5  82.6  89.5
## 5 Oceania   67.9  74.4   75.3  77.5  83.1
```

```
census_data_2022 %>%                                    # Summary by group using dplyr
  group_by(Region) %>%
  dplyr::summarize(min = min(Mortality_rate_both),
            q1 = quantile(Mortality_rate_both, 0.25),
            median = median(Mortality_rate_both),
            q3 = quantile(Mortality_rate_both, 0.75),
            max = max(Mortality_rate_both))
```

```
## # A tibble: 5 x 6
##   Region     min    q1 median    q3   max
##   <chr>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Africa   13.0  41.9   59.2  79.8  141.
## 2 Americas  2.91  9.07  13.4  18.0   54.8
## 3 Asia      1.94  9.28  18.6  31.6  154.
## 4 Europe    1.98  3.74   4.32  6.05  34.2
## 5 Oceania   3.63 12.4   14.4  25.5   47.3
```

```
census_data_2022 %>%                                    # Summary by group using dplyr
  group_by(Subregion) %>%
  dplyr::summarize(min = min(Life_exp_both),
            q1 = quantile(Life_exp_both, 0.25),
            median = median(Life_exp_both),
            q3 = quantile(Life_exp_both, 0.75),
            max = max(Life_exp_both))
```

```
## # A tibble: 21 x 6
##    Subregion          min    q1 median    q3   max
##    <fct>            <dbl> <dbl>  <dbl> <dbl> <dbl>
##  1 Eastern Africa    55.7  65.8   67.4  69.7  76.1
##  2 Middle Africa     55.5  61.8   62.1  63.7  69.7
##  3 Northern Africa   59.2  70.4   74.4  77    78.0
##  4 Southern Africa   59.6  59.7   65.3  65.6  66.5
##  5 Western Africa    58.8  62.3   63.9  69.4  80.5
##  6 Caribbean         66.0  76.1   78.4  80.3  82.2
##  7 Central America   72.3  74.3   75.3  76.3  79.6
```

```
##  8 Northern America  74.0  80.6   81.4  82.0  83.8
##  9 South America     68.9  72.5   75.9  78.4  79.8
## 10 Eastern Asia      71.4  76.2   82.1  83.9  85.0
## # i 11 more rows
```

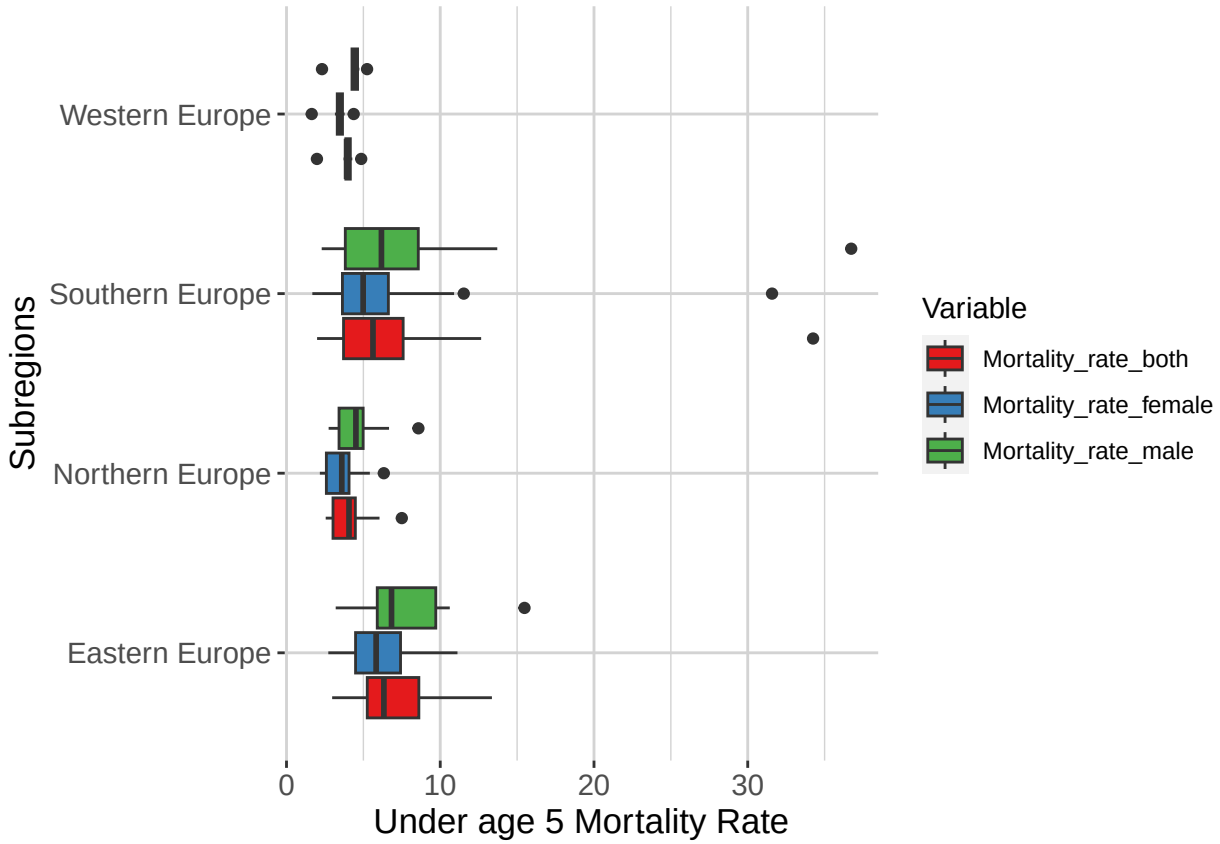
```
census_data_2022 %>%                                # Summary by group using dplyr
  group_by(Subregion) %>%
  dplyr::summarize(min = min(Mortality_rate_both),
            q1 = quantile(Mortality_rate_both, 0.25),
            median = median(Mortality_rate_both),
            q3 = quantile(Mortality_rate_both, 0.75),
            max = max(Mortality_rate_both))
```

```
## # A tibble: 21 x 6
##    Subregion          min    q1 median     q3   max
##    <fct>            <dbl> <dbl>  <dbl>  <dbl> <dbl>
##  1 Eastern Africa   14.3  42.0   53.4   59.8  141.
##  2 Middle Africa    41.7  71.1   81.1  106.   125.
##  3 Northern Africa  13.0  17.2   22.8   42.5   98.3
##  4 Southern Africa  31.6  32.2   43.1   50.7   63.6
##  5 Western Africa   22.1  59.0   73.9   84.3  105.
##  6 Caribbean         4.45  8.80  13.0   15.3   54.8
##  7 Central America   9.61 13.4   15.5   21.7   33.9
##  8 Northern America  2.91  5.07   6.27   9.41  11.0
##  9 South America     7.61 11     16.8   26.4   39.1
## 10 Eastern Asia      2.54  3.45   5.01  12.0   27.6
## # i 11 more rows
```

```
# Filter data to only include the region named Europe
europe_data <- census_data_2022 %>%
  filter(Region == "Europe") %>%
# Reshape data to have a single column for mortality rates and a new column to indicate gender
  gather(Variable, value, Mortality_rate_male, Mortality_rate_female, Mortality_rate_both)

# Create box plots for mortality_rate_male and mortality_rate_female by subregion
mortality_europe <- europe_data %>%
  ggplot(aes(x = Subregion, y = value, fill = Variable)) +
  geom_boxplot() +
  #scale_fill_brewer(palette = "Set1") +
  scale_fill_brewer(palette = "Set1") +
  coord_flip() +
theme(plot.title = element_text(hjust = 0.5, size = 13, face="bold"),
        axis.text=element_text(size=11),
        axis.title=element_text(size=13),
        panel.background = element_rect(fill = "white"),
        panel.grid = element_line(colour = "lightgrey"))+
  xlab("Subregions") + ylab("Under age 5 Mortality Rate")

ggsave('mortality_EU.pdf', plot = mortality_europe, width = 8.4, height = 3, units = "in")
mortality_europe
```

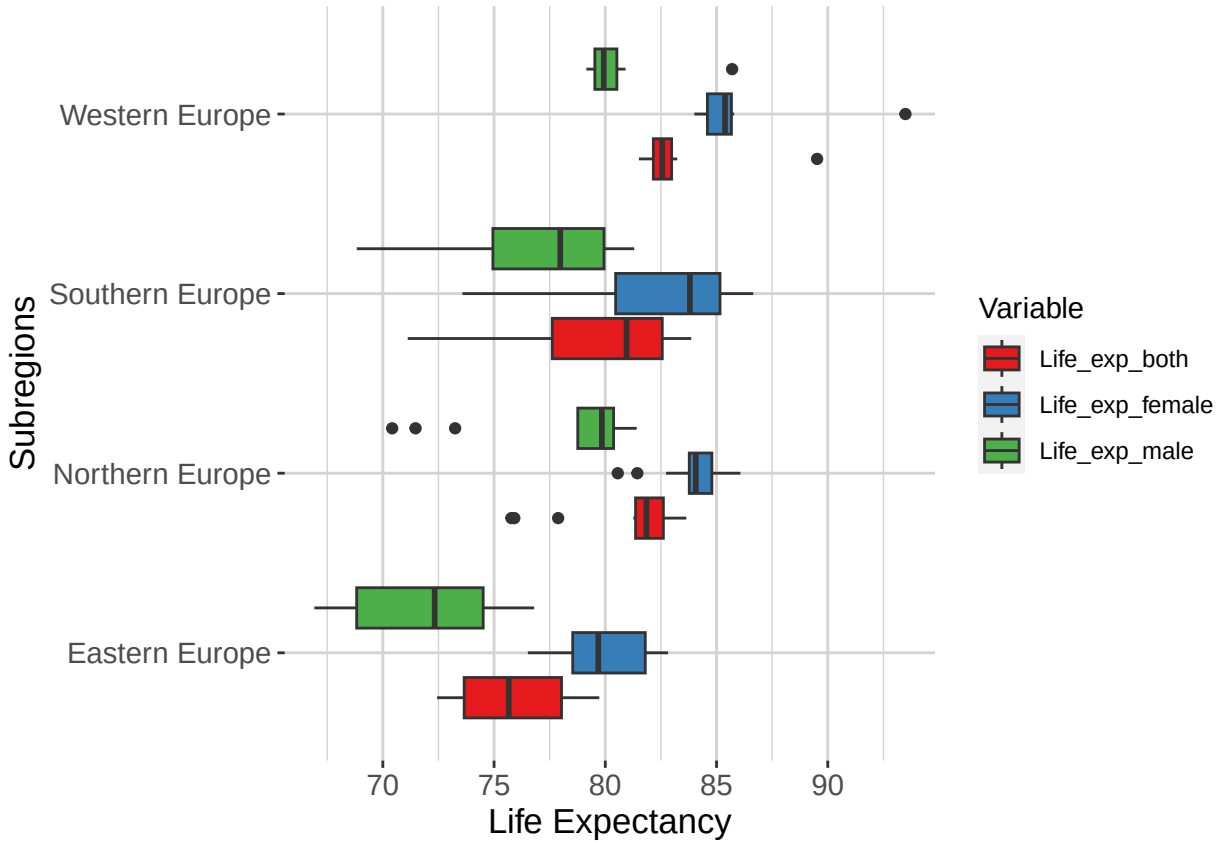



```
europe_data <- census_data_2022 %>%
  filter(Region == "Europe") %>%
# Reshape data to have a single column for mortality rates and a new column to indicate gender
  gather(Variable, value, Life_exp_male, Life_exp_female, Life_exp_both)

# Create box plots for Life expectancy_male and Life expectancy female by subregion

life_europe <- europe_data %>%
  ggplot(aes(x = Subregion, y = value, fill = Variable)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Set1") +
  coord_flip() +
  theme(plot.title = element_text(hjust = 0.5, size = 13, face="bold"),
        axis.text=element_text(size=11),
        axis.title=element_text(size=13),
        panel.background = element_rect(fill = "white"),
        panel.grid = element_line(colour = "lightgrey")) +
  xlab("Subregions") + ylab("Life Expectancy")

ggsave('life_EU.pdf', plot = life_europe, width = 8.4 , height = 3, units = "in")
life_europe
```



Task 3: Bivariate correlations between the variables

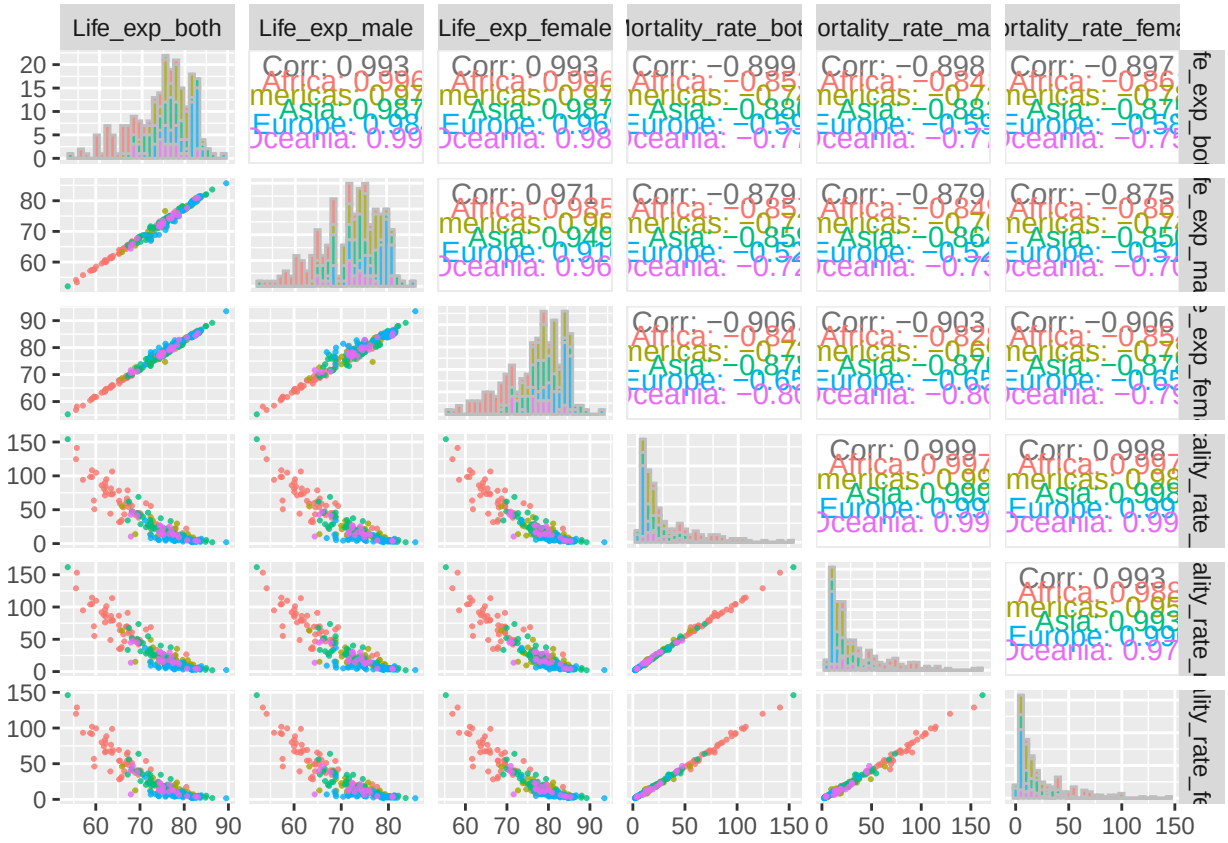
```
#Pairplot
scat_plot  <- ggpairs(census_data_2022, columns = 5:10,
                      upper = list(continuous = GGally::wrap(ggally_cor, stars = F)),
                      diag = list(continuous = wrap("barDiag", alpha = 0.8, color="grey")),
                      lower = list(continuous = wrap("points", alpha = 0.8,size=0.4),
                      combo = wrap("dot", alpha = 0.8,size=0.2) ),
                      mapping=ggplot2::aes(colour = Region)) +
                      theme(axis.text=element_text(size=9),
                      axis.title=element_text(size=11))
ggsave("corr_plot.pdf",plot = scat_plot, width = 8.5,height = 8.5, units = "in")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
scat_plot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Task 4: comparison of 2002 with 2022

```
countries  <- census_data[which(is.na(census_data$Mortality_rate_both)),]$Country
countries
```

```
## character(0)
```

```
census_data_2002  <- census_data_2002  %>% filter(!Country %in% countries)
census_data_2022  <- census_data_2022  %>% filter(!Country %in% countries)
```

```
scat_plot1 <- ggplot(data = NULL, aes(x = census_data_2002$Life_exp_both,
                        y = census_data_2022$Life_exp_both,
                        color = census_data_2002$Region)) +
      geom_point(size = 2.5) +
   guides(colour = guide_legend(title = "Subregion", size = 16, override.aes = list(shape = 15))) +
      geom_abline(intercept =0 , slope = 1)+ xlim(40,90)+ylim(40,90) +
      xlab("Life expectancy of both sexes in 2002") + ylab("Life expectancy of both sexes in 2022") +
      theme(plot.title = element_text(hjust = 0.5, size = 12, face="bold"),
          legend.position = c(0.15, 0.83),legend.background = element_rect(fill = "transparent"),
          legend.text = element_text(size = 10),
          axis.text=element_text(size=12),
          axis.title=element_text(size=14))+
  ggtitle("a) Life expectancy 2002 vs 2022") +
  theme(plot.title = element_text(size = 14))
```

```
scat_plot2  <- ggplot(data = NULL, aes(x = census_data_2002$Mortality_rate_both,
                        y = census_data_2022$Mortality_rate_both,
                        color = census_data_2002$Region)) + geom_abline(intercept =0 , slope = 1) +
      xlim(0,150)+ylim(0,150) +
      geom_point(size = 2.5) +
   guides(colour = guide_legend(title = "Subregion", size = 16, override.aes = list(shape = 15))) +
      xlab("Under age 5 mortality rate in 2002") + ylab("Under age 5 mortality rate in 2022") +
      theme(plot.title = element_text(hjust = 0.5, size = 24, face="bold"),
          legend.position = c(0.15, 0.83),legend.background = element_rect(fill = "transparent"),
          legend.text = element_text(size = 10),
          axis.text=element_text(size=12),
          axis.title=element_text(size=14))+
  ggtitle("b) Mortality rate 2002 Vs 2022") +
  theme(plot.title = element_text(size = 14))
```

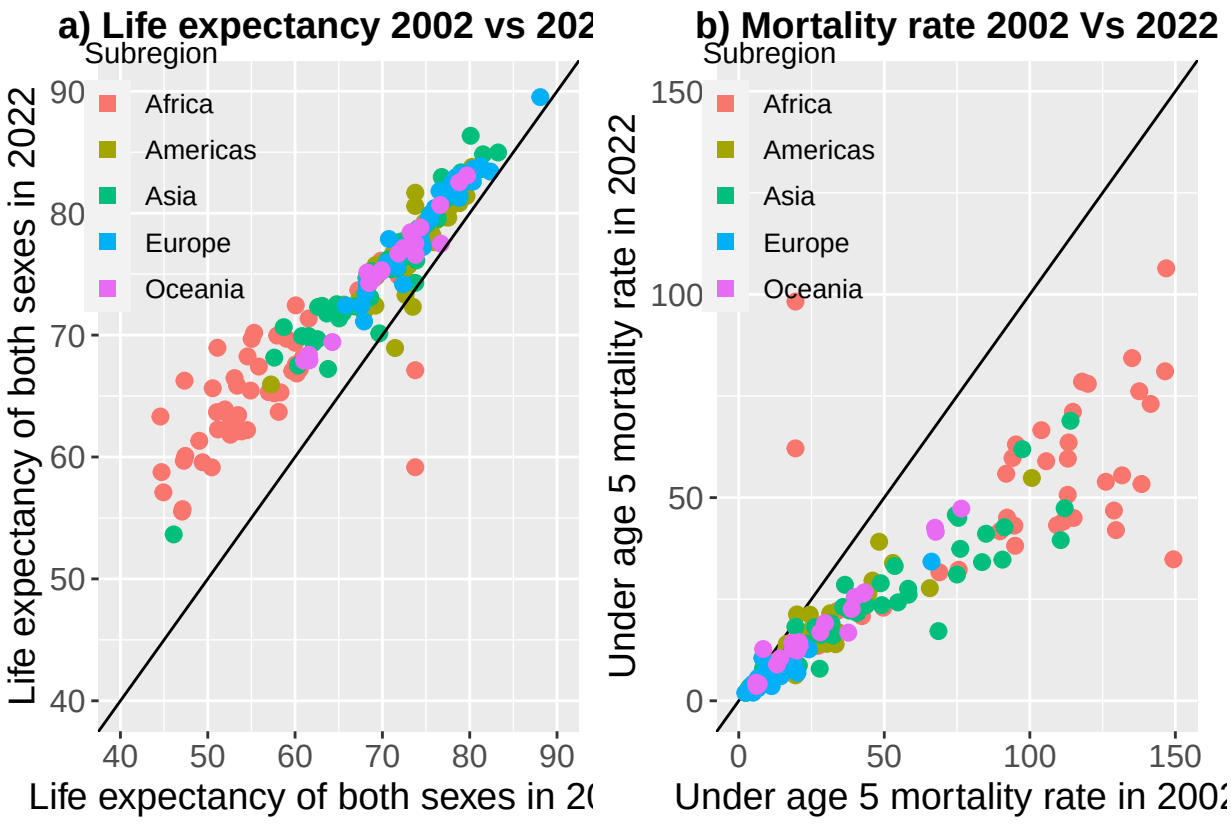
```
combined_plot2 <- scat_plot1 + scat_plot2 + plot_layout(ncol = 2)

ggsave("combined_plot2.pdf", combined_plot2, width = 8.4, height = 5, units = "in")
```

```
## Warning: Removed 14 rows containing missing values (`geom_point()`).
```

```
combined_plot2
```

```
## Warning: Removed 14 rows containing missing values (`geom_point()`).
```




```
# Countries for which mortality rate increased in 2022 as compared to 2002
mortality_change = census_data_2022[census_data_2022$
                                    Mortality_rate_both >
                  census_data_2002$Mortality_rate_both,][c("Country","Region","Mortality_rate_both")]
```

```
mortality_change
```

```
##         Country   Region Mortality_rate_both
## 31  South Sudan   Africa               98.26
## 32        Sudan   Africa               62.12
## 87       Panama Americas               21.26
## 185     Croatia   Europe               10.53
## 214        Guam  Oceania               12.74
```

```
# Countries for which Life expectancy decreased in 2022 as compared to 2002
lifeEx_change = census_data_2022[census_data_2022$
                                 Life_exp_both <
                                 census_data_2002$Life_exp_both,][c("Country","Region","Life_exp_both"
lifeEx_change
```

```
##         Country   Region Life_exp_both
## 31  South Sudan   Africa         59.16
## 32        Sudan   Africa         67.12
## 85       Mexico Americas         72.32
## 102        Peru Americas         68.94
```