'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
```

```
#Loading the data
census_data <- read.csv('census2001_2021.csv')

#View the Data
head(census_data)
```

```
##   Country.Name          Subregion Region Year Life.Expectancy..Both.Sexes
## 1  Afghanistan South-Central Asia   Asia 2001                       45.81
## 2  Afghanistan South-Central Asia   Asia 2021                       53.25
## 3      Albania    Southern Europe Europe 2001                       75.14
## 4      Albania    Southern Europe Europe 2021                       79.23
## 5      Algeria    Northern Africa Africa 2001                       72.19
## 6      Algeria    Northern Africa Africa 2021                       77.79
##   Life.Expectancy..Males Life.Expectancy..Females
## 1                  44.85                    46.83
## 2                  51.73                    54.85
```

```
## 3                  72.39                    78.20
## 4                  76.55                    82.12
## 5                  71.36                    73.07
## 6                  76.32                    79.33
##   Infant.Mortality.Rate..Both.Sexes
## 1                            144.77
## 2                            106.75
## 3                             23.88
## 4                             11.10
## 5                             39.97
## 6                             20.23
```

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

```
##
## To cite package 'gridExtra' in publications use:
##
##   Auguie B (2017). _gridExtra: Miscellaneous Functions for "Grid"
##   Graphics_. R package version 2.3.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {gridExtra: Miscellaneous Functions for "Grid" Graphics},
##     author = {Baptiste Auguie},
##     year = {2017},
##     note = {R package version 2.3},
##   }
```

```
#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")
```

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

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

```
head(census_data)
```

```
##      Country      Subregion Region Year Life_exp_both Life_exp_male
## 65   Burundi Eastern Africa Africa 2001         55.06         53.02
## 66   Burundi Eastern Africa Africa 2021         67.07         64.98
## 87   Comoros Eastern Africa Africa 2001         60.17         58.58
## 88   Comoros Eastern Africa Africa 2021         66.90         64.65
## 111 Djibouti Eastern Africa Africa 2001         58.21         55.97
## 112 Djibouti Eastern Africa Africa 2021         65.00         62.40
##     Life_exp_female Mortality_rate
## 65            57.16          83.17
## 66            69.22          38.96
## 87            61.81          85.45
## 88            69.21          58.21
## 111           60.50          69.88
## 112           67.67          47.78
```

```
#Check is there any missing values in Data
colSums(is.na(census_data))
```

```
##         Country       Subregion          Region            Year   Life_exp_both
##               0               0               0               0               6
##   Life_exp_male Life_exp_female  Mortality_rate
##               6               6               6
```

```
#Check the missing value records #Mortality rate
census_data[which(is.na(census_data$Mortality_rate)),]
```

```
##           Country        Subregion   Region Year Life_exp_both Life_exp_male
## 235         Libya  Northern Africa   Africa 2001            NA            NA
## 379   South Sudan  Northern Africa   Africa 2001            NA            NA
## 385         Sudan  Northern Africa   Africa 2001            NA            NA
## 325   Puerto Rico        Caribbean Americas 2001            NA            NA
## 429 United States Northern America Americas 2001            NA            NA
## 393         Syria     Western Asia     Asia 2001            NA            NA
##     Life_exp_female Mortality_rate
## 235              NA             NA
## 379              NA             NA
## 385              NA             NA
## 325              NA             NA
## 429              NA             NA
## 393              NA             NA
```

```
#Split the Data Based on the year
census_data_2021 <- census_data %>% filter(Year == 2021)
census_data_2001 <- census_data %>% filter(Year == 2001)
```

```
#Summary of data
census_data_2021   %>% summary()
```

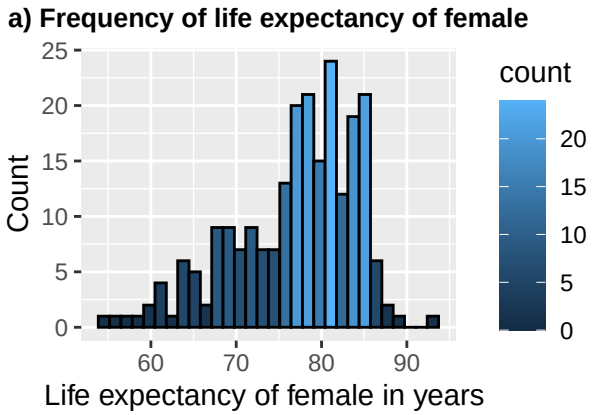
```
##    Country                       Subregion      Region              Year
##  Length:227         Caribbean         : 25   Length:227         Min.   :2021
##  Class :character   Western Asia      : 19   Class :character   1st Qu.:2021
##  Mode  :character   Eastern Africa    : 17   Mode  :character   Median :2021
##                     Western Africa    : 17                      Mean   :2021
##                     Southern Europe   : 16                      3rd Qu.:2021
##                     South-Central Asia: 14                      Max.   :2021
##                     (Other)           :119
##  Life_exp_both   Life_exp_male   Life_exp_female Mortality_rate
##  Min.   :53.25   Min.   :51.73   Min.   :54.85   Min.   :  1.53
##  1st Qu.:69.73   1st Qu.:67.58   1st Qu.:72.29   1st Qu.:  6.27
##  Median :75.56   Median :72.99   Median :78.36   Median : 12.58
##  Mean   :74.28   Mean   :71.78   Mean   :76.89   Mean   : 20.25
##  3rd Qu.:79.42   3rd Qu.:76.94   3rd Qu.:82.34   3rd Qu.: 29.48
##  Max.   :89.40   Max.   :85.55   Max.   :93.40   Max.   :106.75
##
```

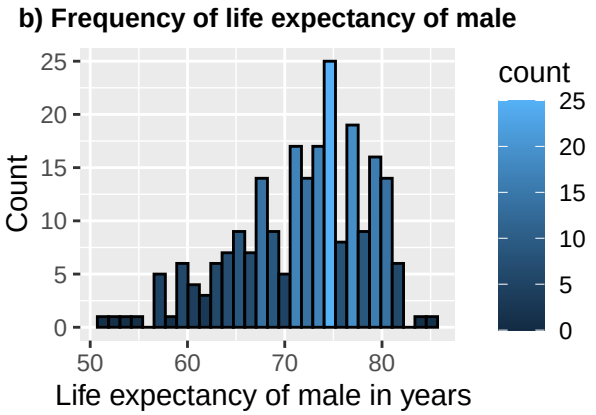
```
#Get the difference between the Life expectancy of female and male
census_data_2001$Life_exp_diff_btw_sexes <- census_data_2001$
  Life_exp_female - census_data_2001$Life_exp_male
census_data_2021$Life_exp_diff_btw_sexes <- census_data_2021$
  Life_exp_female - census_data_2021$Life_exp_male
```

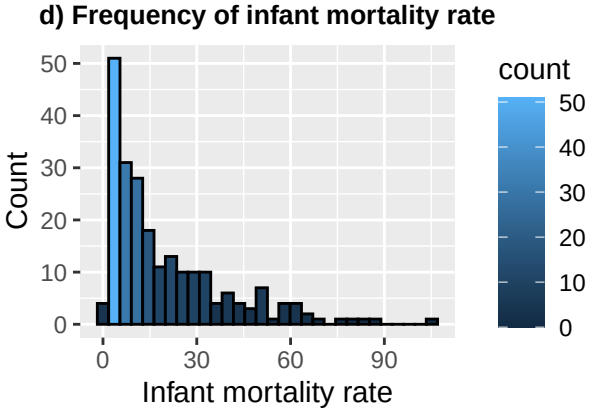
Task 1: Frequency Distributions of Different Variables

```
# Histograms
plot1  <- ggplot(census_data_2021, aes(x = Life_exp_female)) +
          geom_histogram(aes(fill = ..count..), col = "black")+
          scale_x_continuous(name = "Life expectancy of female in years ") +
          scale_y_continuous(name = "Count") +
          ggtitle("a) Frequency of life expectancy of female") +
          theme(plot.title = element_text(hjust = 0.5, size = 10, face="bold"),
            axis.text=element_text(size=9),
             axis.title=element_text(size=11))

plot2  <- ggplot(census_data_2021, aes(x = Life_exp_male)) +
           geom_histogram(aes(fill = ..count..), col = "black")+
           scale_x_continuous(name = "Life expectancy of male in years ") +
           scale_y_continuous(name = "Count") +
           ggtitle("b) Frequency of life expectancy of male") +
           theme(plot.title = element_text(hjust = 0.5, size = 10, face="bold"),
             axis.text=element_text(size=9),
             axis.title=element_text(size=11))

plot3  <- ggplot(census_data_2021, aes(x = Life_exp_both)) +
           geom_histogram(aes(fill = ..count..), col = "black")+
           scale_x_continuous(name = "Life expectancy of both sexes") +
           scale_y_continuous(name = "Count") +
           ggtitle("c) Frequency of Life expectancy of both sexes") +
           theme(plot.title = element_text(hjust = 0.5, size = 10, face="bold"),
             axis.text=element_text(size=9),
             axis.title=element_text(size=11))

plot4  <- ggplot(census_data_2021, aes(x = Mortality_rate)) +
           geom_histogram(aes(fill = ..count..), col = "black")+
           scale_x_continuous(name = "Infant mortality rate") +
           scale_y_continuous(name = "Count") +
           ggtitle("d) Frequency of infant mortality rate") +
           theme(plot.title = element_text(hjust = 0.5, size = 10, face="bold"),
             axis.text=element_text(size=9),
             axis.title=element_text(size=11))



final_plot1 <- grid.arrange(plot1, plot2, plot3, plot4, ncol=2, nrow = 2)
```

```
## `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`.
```








```
ggsave("histograms.pdf",plot = final_plot1)
```

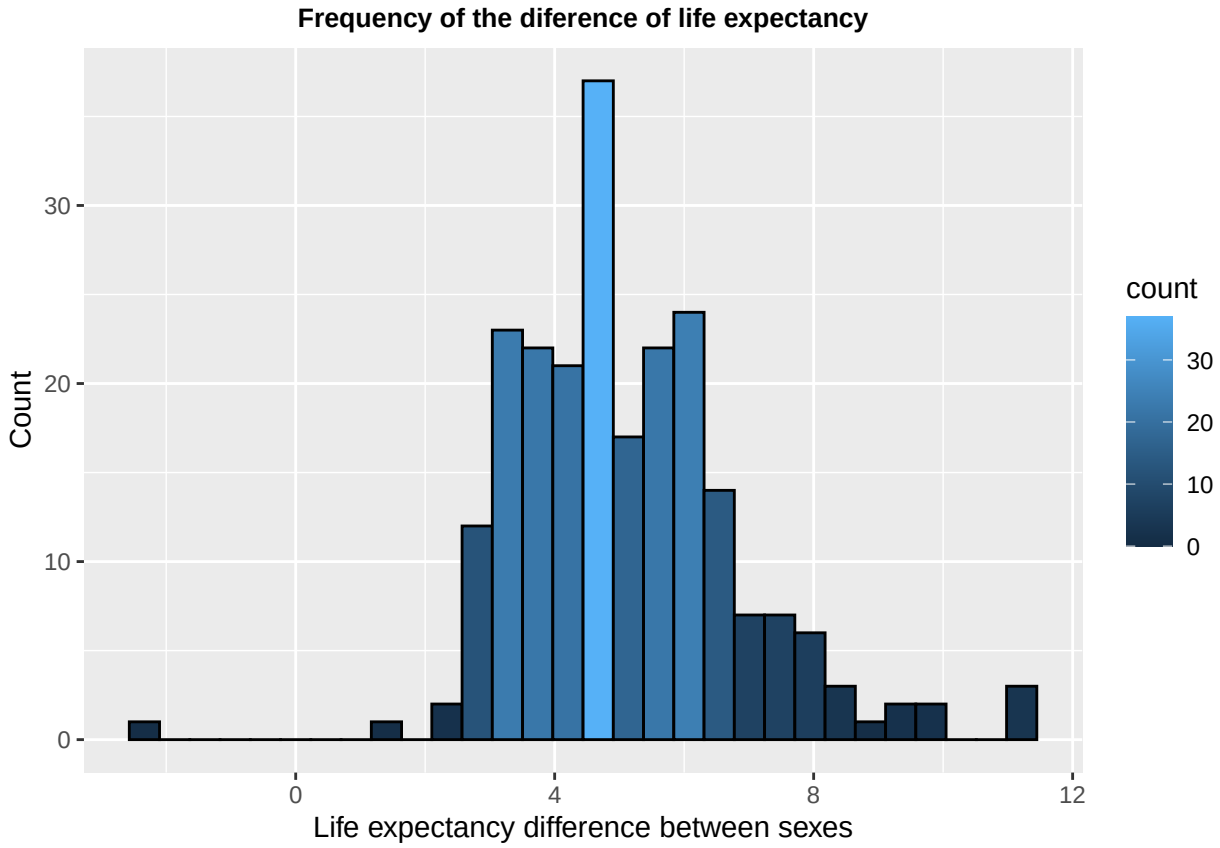
```
## Saving 6.5 x 4.5 in image
```

```
final_plot1
```

```
## 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]
```

```
plot5  <- ggplot(census_data_2021, aes(x = Life_exp_diff_btw_sexes)) +
           geom_histogram(aes(fill = ..count..), col = "black")+
           scale_x_continuous(name = "Life expectancy difference between sexes") +
           scale_y_continuous(name = "Count") +
           ggtitle("Frequency of the diference of life expectancy") +
           theme(plot.title = element_text(hjust = 0.5, size = 10, face="bold"),
             axis.text=element_text(size=9),
             axis.title=element_text(size=11))

final_plot2 <- grid.arrange(plot5, ncol=1, nrow = 1)
```

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




```
ggsave("histogram-d.pdf",plot = final_plot2)
```

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

```
final_plot2
```

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

```
# Females Life expectancy above 90 (Highest)
females_above90 = census_data_2021[census_data_2021$
                                   Life_exp_female > 90,][c("Country","Region","Life_exp_female")]
females_above90
```

```
##     Country Region Life_exp_female
## 204  Monaco Europe            93.4
```

```
# Females Life expectancy below 55 (lowest)
females_below55 = census_data_2021[census_data_2021$
                                   Life_exp_female < 55,][c("Country","Life_exp_female")]
females_below55
```

```
##         Country Life_exp_female
## 114 Afghanistan           54.85
```

```
# Males Life expectancy above 85 (Highest)
males_above85 = census_data_2021[census_data_2021$
                                 Life_exp_male > 82,][c("Country","Region","Life_exp_male")]
males_above85
```

```
##       Country Region Life_exp_male
## 135 Singapore   Asia         83.48
## 204    Monaco Europe         85.55
```

```
# Males Life expectancy below 55 (Lowest)
males_below55 = census_data_2021[census_data_2021$
                                   Life_exp_male < 55,][c("Country","Life_exp_male")]
males_below55
```

```
##                      Country Life_exp_male
## 13                   Somalia         53.02
## 20  Central African Republic         53.74
## 114              Afghanistan         51.73
```

```
# Countries with higher life expectancy of males than females
males_h_Ex = census_data_2021[census_data_2021$
                                Life_exp_diff_btw_sexes < 0,][c("Country","Life_exp_diff_btw_sexes")]
males_h_Ex
```

```
##       Country Life_exp_diff_btw_sexes
## 69 Montserrat                   -2.11
```

```
# Highest Mortality rate
mortality_h = census_data_2021[census_data_2021$
                                 Mortality_rate > 85,][c("Country","Mortality_rate")]
mortality_h
```

```
##         Country Mortality_rate
## 13      Somalia          88.03
## 114 Afghanistan         106.75
```

```
# Lowest Mortality rate
mortality_l = census_data_2021[census_data_2021$
                                 Mortality_rate < 2,][c("Country","Region","Mortality_rate")]
mortality_l
```

```
##       Country Region Mortality_rate
## 108     Japan   Asia           1.92
## 135 Singapore   Asia           1.56
## 173   Iceland Europe           1.66
## 196  Slovenia Europe           1.53
## 204    Monaco Europe           1.78
```

Task 2: Bivariate correlations between the variables

```
#Pairplot
scat_plot  <- ggpairs(census_data_2021, columns = 5:8,
                      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)
```

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

```
## `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`.
```

Task 3: Analysis of variability within and between subregions.

```
#Summary for box plot for life expectancy of both sexes in regions and subregions
census_data_2021 %>%                                    # 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    55.1  61.8   65.5  69.3  80.2
## 2 Americas  65.6  75.0   77.7  79.4  83.6
## 3 Asia      53.2  71.6   75.5  78.2  86.2
## 4 Europe    70.1  77.0   81.3  82.4  89.4
## 5 Oceania   67.6  74.2   75.1  77.3  82.9
```

```
census_data_2021 %>%                                    # Summary by group using dplyr
  group_by(Region) %>%
  dplyr::summarize(min = min(Mortality_rate),
            q1 = quantile(Mortality_rate, 0.25),
            median = median(Mortality_rate),
            q3 = quantile(Mortality_rate, 0.75),
            max = max(Mortality_rate))
```

```
## # A tibble: 5 x 6
##   Region     min    q1 median    q3   max
##   <chr>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Africa   10.8  29.4   42.4  57.3   88.0
## 2 Americas  2.21  8.21  11.5  15.2   41.3
## 3 Asia      1.56  7.5   15.6  26.3  107.
## 4 Europe    1.53  3.24   3.64  5.32  29.5
## 5 Oceania   3.05  7.96  12.7  20.5   34.4
```

```
census_data_2021 %>%                                    # 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))
```

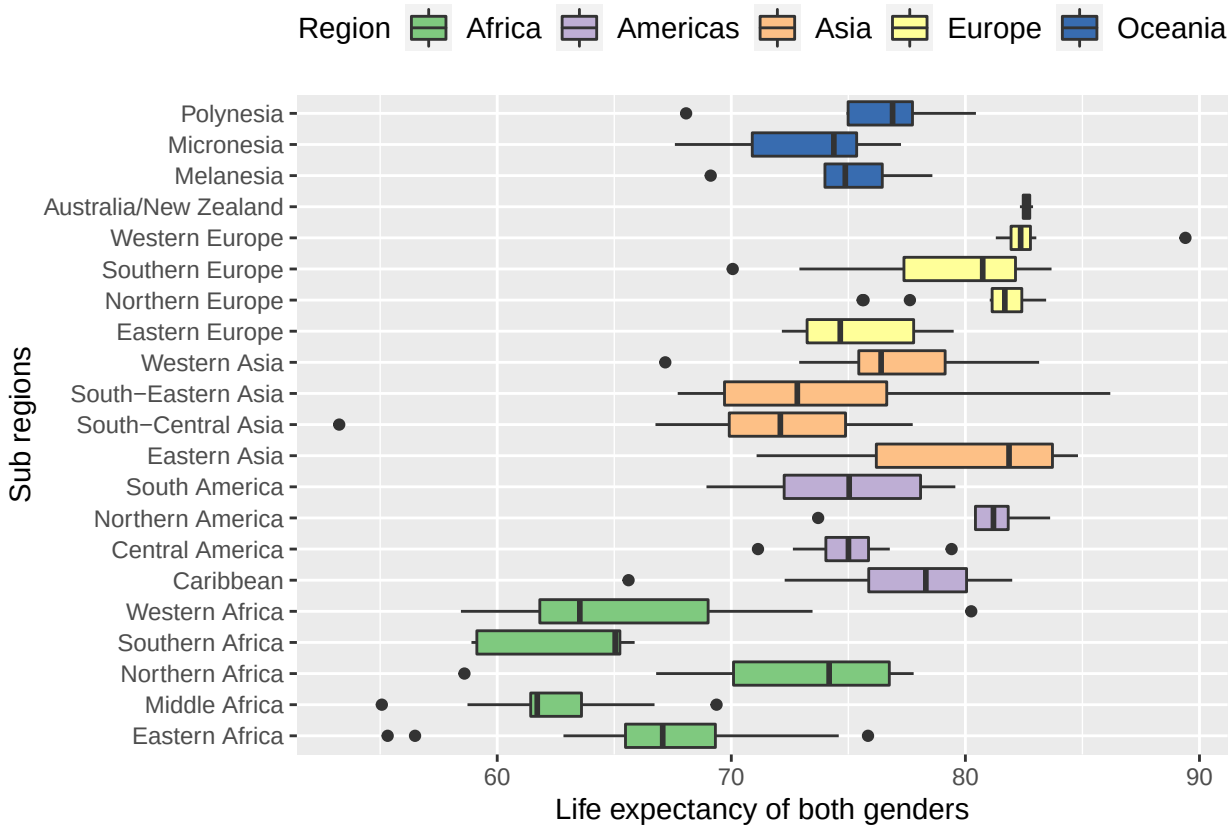
```
## # A tibble: 21 x 6
##    Subregion          min    q1 median    q3   max
##    <fct>            <dbl> <dbl>  <dbl> <dbl> <dbl>
##  1 Eastern Africa    55.3  65.5   67.1  69.3  75.8
##  2 Middle Africa     55.1  61.4   61.7  63.6  69.4
##  3 Northern Africa   58.6  70.1   74.2  76.8  77.8
##  4 Southern Africa   58.9  59.1   65.0  65.2  65.9
##  5 Western Africa    58.4  61.8   63.5  69.0  80.2
##  6 Caribbean         65.6  75.9   78.3  80.0  82
##  7 Central America   71.1  74.0   75.0  75.9  79.4
##  8 Northern America  73.7  80.4   81.2  81.8  83.6
##  9 South America     68.9  72.3   75.0  78.1  79.6
## 10 Eastern Asia      71.1  76.2   81.9  83.7  84.8
## # ... with 11 more rows
```

```
census_data_2021 %>%                                    # Summary by group using dplyr
  group_by(Subregion) %>%
  dplyr::summarize(min = min(Mortality_rate),
            q1 = quantile(Mortality_rate, 0.25),
            median = median(Mortality_rate),
            q3 = quantile(Mortality_rate, 0.75),
            max = max(Mortality_rate))
```

```
## # A tibble: 21 x 6
##    Subregion          min    q1 median    q3   max
```

```
##    <fct>            <dbl> <dbl>  <dbl> <dbl> <dbl>
##  1 Eastern Africa   10.8  29.4   34.6  42.4   88.0
##  2 Middle Africa    29.4  49.3   60.6  67.0   84.2
##  3 Northern Africa  11.5  15.2   19.7  31.7   64.8
##  4 Southern Africa  26.0  26.8   30.4  41.2   50.2
##  5 Western Africa   19.8  38.0   50.7  57.4   74.6
##  6 Caribbean         3.11  7.84  10.7  13.0   41.3
##  7 Central America   8.59 11.6   13.9  18.2   26.8
##  8 Northern America  2.21  4.44   5.22  8.35   8.9
##  9 South America     6.68 10.5   16.3  22.4   30.6
## 10 Eastern Asia      1.92  2.83   4.36 10.2   22.4
## # ... with 11 more rows
```

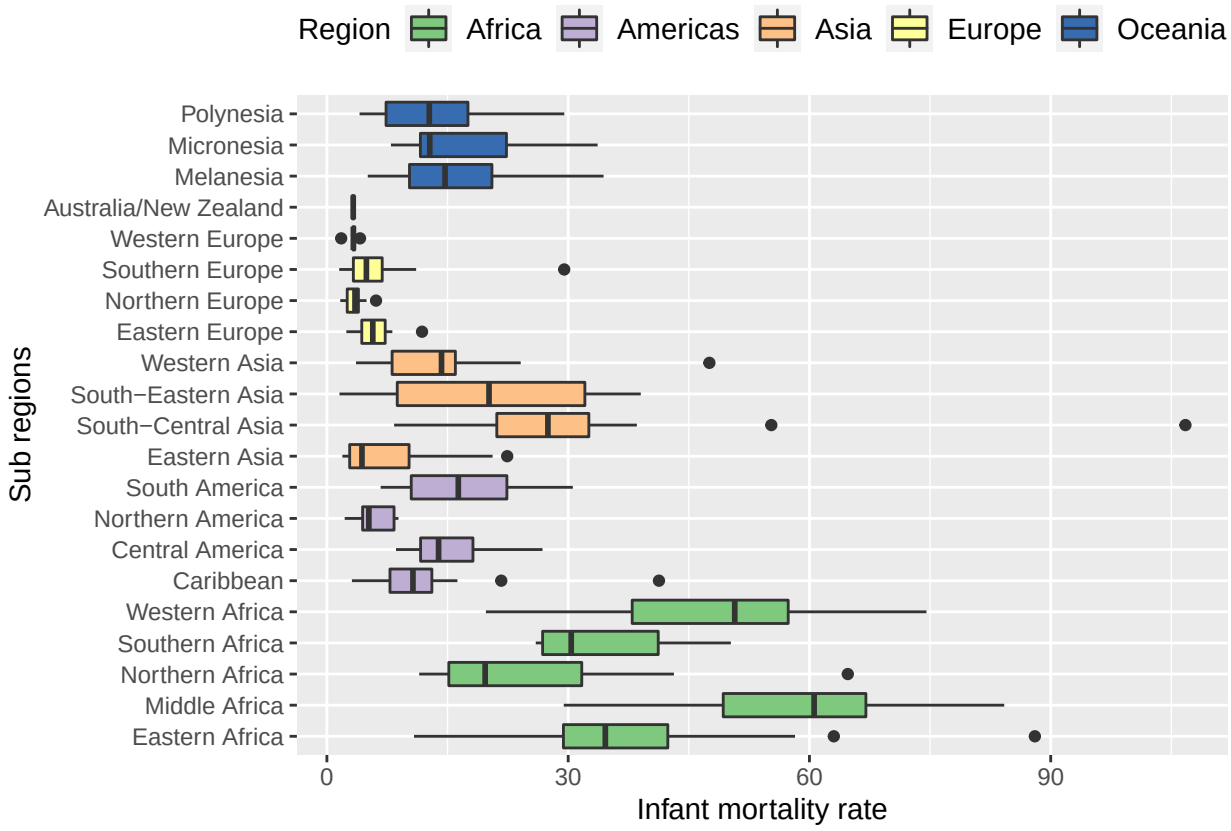
```
#Comparing the Life Expectancies of Male in Sub Regions
box_plot1 <- census_data_2021 %>%
  ggplot(aes(x=Subregion, y=Life_exp_both, fill=Region)) +
  geom_boxplot() +
  coord_flip()+ scale_fill_brewer(palette="Accent") +
  theme(legend.position="top",
        axis.text = element_text(vjust = 0.5, size = 9),
        legend.text = element_text(size = 11),
        axis.title=element_text(size=11)
       )+
  xlab("Sub regions") + ylab("Life expectancy of both genders")
box_plot1
```



```
ggsave('Boxplot1.pdf', plot = box_plot1)
```

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

```
#Comparing the infant mortality rate in sub regions
box_plot2 <-  census_data_2021 %>%
  ggplot(aes(x=Subregion, y=Mortality_rate, fill=Region)) +
  geom_boxplot() + scale_fill_brewer(palette="Accent") +
  coord_flip()+
  theme(legend.position="top",
        axis.text = element_text(vjust = 0.5, size = 9),
        legend.text = element_text(size = 11),
        axis.title=element_text(size=11)
       )+
  xlab("Sub regions") + ylab("Infant mortality rate")
box_plot2
```




```
ggsave('Boxplot2.pdf', plot = box_plot2)
```

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

Task 4: comparison of 2001 with 2021

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

```
census_data_2001  <- census_data_2001  %>% filter(!Country %in% countries)
census_data_2021  <- census_data_2021  %>% filter(!Country %in% countries)
```

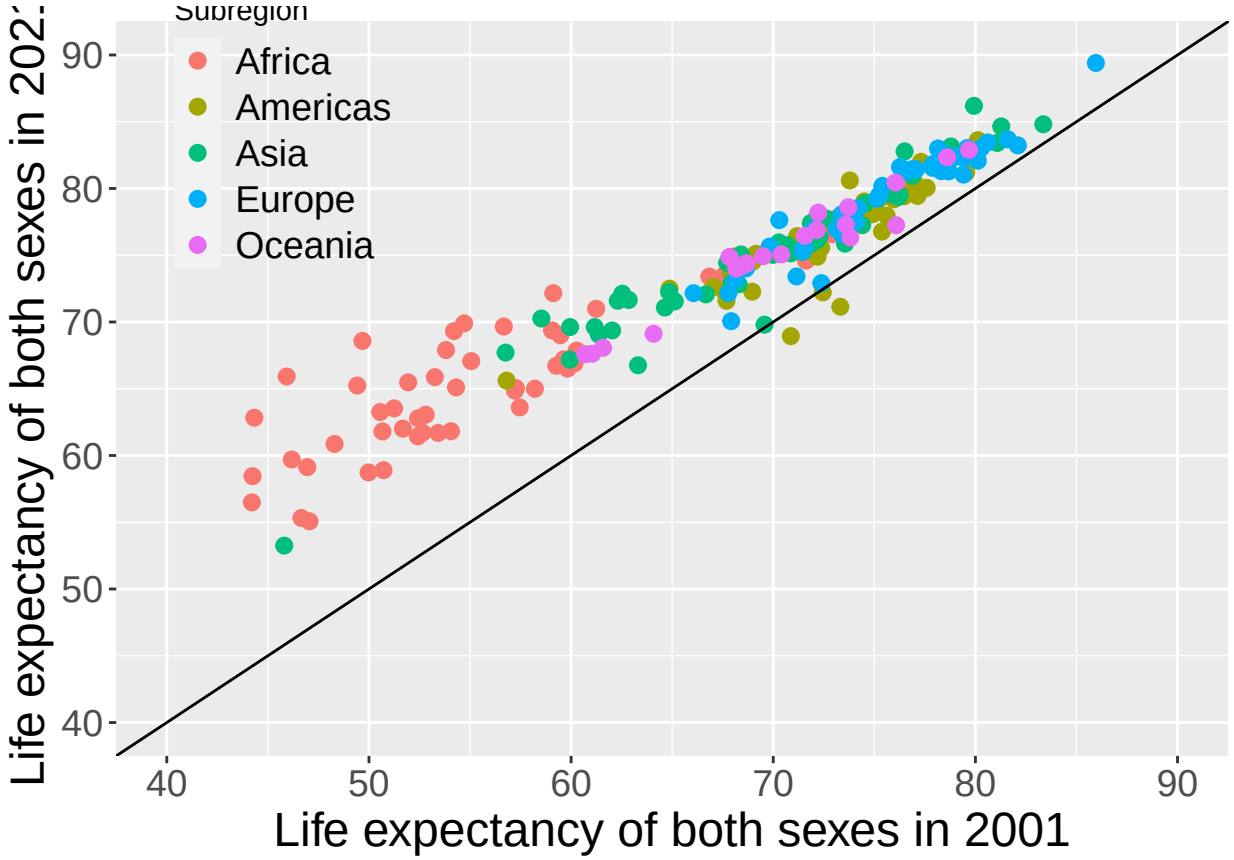
```
scat_plot1 <- ggplot(data = NULL, aes(x = census_data_2001$Life_exp_both,
                        y = census_data_2021$Life_exp_both,
                        color = census_data_2001$Region)) +
      geom_point(size = 2.5) + guides(colour = guide_legend(title = "Subregion", size = 16)) +
      geom_abline(intercept =0 , slope = 1)+ xlim(40,90)+ylim(40,90) +
      xlab("Life expectancy of both sexes in 2001") + ylab("Life expectancy of both sexes in 2021") +
      theme(plot.title = element_text(hjust = 0.5, size = 12, face="bold"),
          legend.position = c(0.15, 0.85),legend.background = element_rect(fill = "transparent"),
          legend.text = element_text(size = 14),
          axis.text=element_text(size=14),
          axis.title=element_text(size=18))

scat_plot2  <- ggplot(data = NULL, aes(x = census_data_2001$Mortality_rate,
                        y = census_data_2021$Mortality_rate,
                        color = census_data_2001$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)) +
      xlab("Infant mortality rate in 2001") + ylab("Infant mortality rate in 2021") +
      theme(plot.title = element_text(hjust = 0.5, size = 24, face="bold"),
          legend.position = c(0.15, 0.85),legend.background = element_rect(fill = "transparent"),
          legend.text = element_text(size = 14),
          axis.text=element_text(size=14),
          axis.title=element_text(size=18))

ggsave("final_plot1.pdf",plot = scat_plot1)
```

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

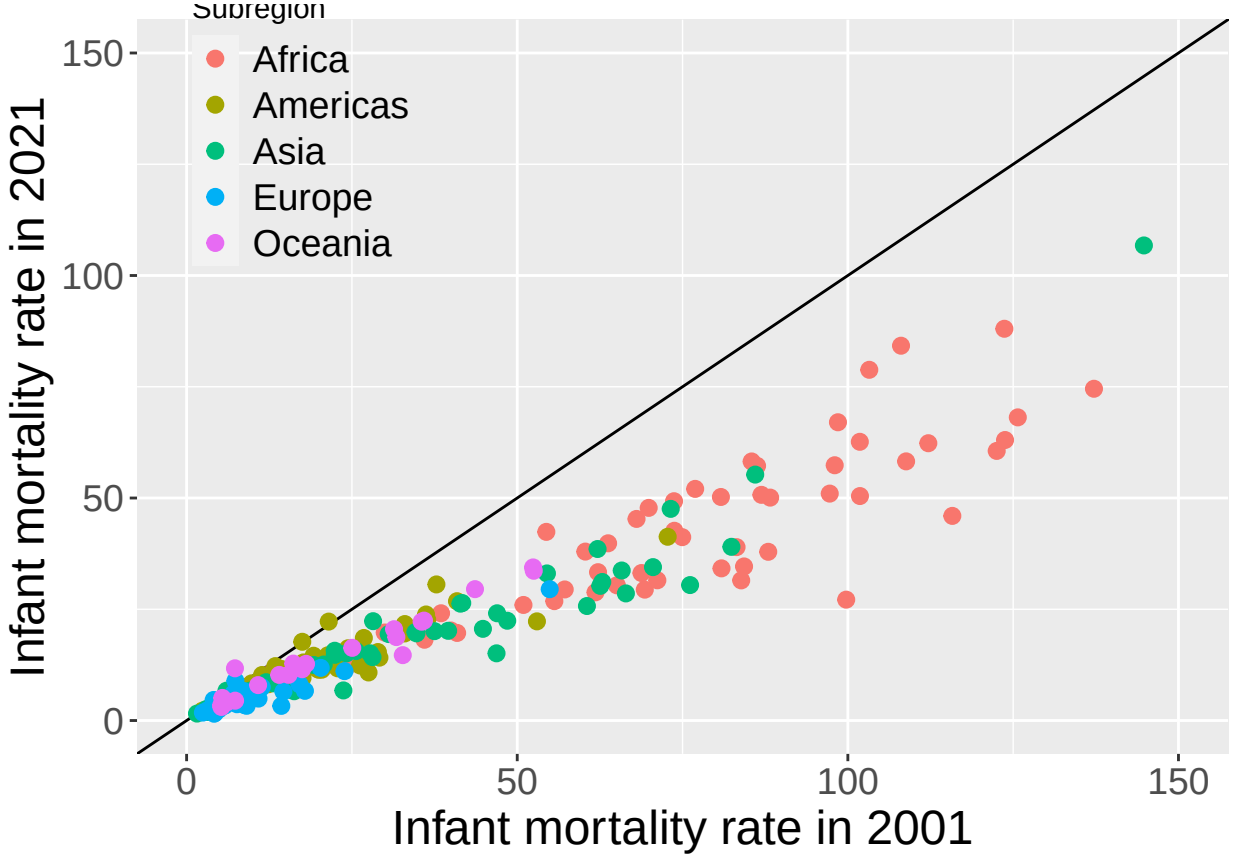
```
ggsave("final_plot2.pdf",plot = scat_plot2)
```

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

```
scat_plot1
```


Subregion
Africa
Americas
Asia
Europe
Oceania
Life expectancy of both sexes in 202
90
80
70
60
50
40
40
50
60
70
80
90
Life expectancy of both sexes in 2001

```
scat_plot2
```




```
# Countries for which mortality rate increased in 2021 as compared to 2001
mortality_change = census_data_2021[census_data_2021$
                                      Mortality_rate >
                   census_data_2001$Mortality_rate,][c("Country","Region","Mortality_rate")]
mortality_change
```

```
##       Country   Region Mortality_rate
## 84     Panama Americas          17.69
## 100 Venezuela Americas          22.23
## 128  Malaysia     Asia           6.70
## 179   Croatia   Europe           8.91
## 184     Malta   Europe           4.62
## 208      Guam  Oceania          11.73
```

```
# Countries for which Life expectancy decreased in 2021 as compared to 2001
lifeEx_change = census_data_2021[census_data_2021$
                                   Life_exp_both <
                                   census_data_2001$Life_exp_both,][c("Country","Region","Life_exp_both
lifeEx_change
```

```
##       Country   Region Life_exp_both
## 82     Mexico Americas         71.14
## 97       Peru Americas         68.94
## 100 Venezuela Americas         72.22
```