Lab 1

First I am creating a data fram reading in other data sets and adding there data to the data fram below

```
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(countrycode)
```

```
## Warning: package 'countrycode' was built under R version 4.3.3
```

```
library(tinytex)

pop <- read.csv("pop.csv")
country = pop$country
p2010 = pop$X2010
# colnames(pop)
life_exp <- read.csv("lex.csv")

Life2010 <- life_exp$X2010
life_cont <- life_exp[, c("country", "X2010")]
#print(life_cont)
# do country names to then merge it with the df

df <- data.frame(
  country = country,
  population = rep(0, length(country)),
  life_expectancy = rep(0, length(country)),
  daily_income = rep(0, length(country)),
  babies_per_woman = rep(0, length(country)),
  child_mortality = rep(0, length(country)),
  co2_emissions_per_person = rep(0, length(country)),
  gdp_per_capita = rep(0, length(country)),
  health_spending_per_person = rep(0, length(country)),
  population_density = rep(0, length(country)),
```

```
  basic_water_source_percentage = rep(0, length(country)),
  murders = rep(0, length(country)),
  continent = rep(0, length(country)),
  stringsAsFactors = FALSE
)
# convert population M and B
text.to.num <- function(x){
xx = as.character(x)
res=rep(0, length(x))
for(i in 1:length(x)){
if(is.na(xx[i])){
res[i]=NA
next
}
if (grepl("M", xx[i], ignore.case = TRUE)) {
res[i]=as.numeric(gsub("M", "", xx[i], ignore.case = TRUE)) * 1e6
} else if (grepl("k", xx[i], ignore.case = TRUE)) {
res[i]=as.numeric(gsub("k", "", xx[i], ignore.case = TRUE)) * 1e3
} else {
res[i]=as.numeric(xx[i])
}
}
res
}

#country code
df$continent <- countrycode(sourcevar = df[, "country"],
origin = "country.name",
destination = "continent")


# ading columns
df$population <- p2010
df$population <- text.to.num(df$population)
```

```
## Warning in text.to.num(df$population): NAs introduced by coercion

## Warning in text.to.num(df$population): NAs introduced by coercion
```

```
df$population <- text.to.num(df$population)

#Life
df <- merge(df, life_cont, by = "country", all.x = TRUE)
df$life_expectancy <- df$X2010
df$X2010 <- NULL
#child
child <- read.csv("child.csv")
cd2010 <- child[, c("country", "X2010")]
df <- merge(df, cd2010, by = "country", all.x = TRUE)
df$child_mortality <- df$X2010
df$X2010 <- NULL

#income
```

```r
income <- read.csv("income.csv")
income_2010 <- income[, c("country", "X2010")]
df <- merge(df, income_2010, by = "country", all.x = TRUE)
df$daily_income <- df$X2010
df$X2010 <- NULL

#babies per women
babies <- read.csv("CPW.csv")
b2010 <- babies[, c("country", "X2010")]
df <- merge(df, b2010, by = "country", all.x = TRUE)
df$babies_per_woman <- df$X2010
df$X2010 <- NULL

# co2 emissions
co2 <- read.csv("co2.csv")
co2010 <- co2[, c("country", "X2010")]
df<- merge(df, co2010, by = "country", all.x = TRUE)
df$co2_emissions_per_person <- df$X2010
df$X2010 <- NULL

# gdp
gdp <- read.csv("gdp.csv")
gd2010 <- gdp[, c("country","X2010")]
df <-  merge(df, gd2010, by = "country", all.x = TRUE)
df$gdp_per_capita <- df$X2010
df$X2010 <- NULL
df$gdp_per_capita <- text.to.num(df$gdp_per_capita)

# spending per person
hm <- read.csv("health_money.csv")
spending <- hm[, c("country", "X2010")]
df <- merge(df , spending, by = "country", all.x = TRUE)
df$health_spending_per_person <- df$X2010
df$X2010 <- NULL

#population density
popd <- read.csv("popd.csv")
popd2010 <- popd[, c("country", "X2010")]
df <- merge(df, popd2010, by = "country", all.x= TRUE)
df$population_density <- df$X2010
df$X2010 <- NULL

#water supply
water <- read.csv("water.csv")
water2010<- water[, c("country", "X2010")]
df <-  merge(df, water2010, by = "country", all.x = TRUE)
df$basic_water_source_percentage <- df$X2010
df$X2010 <- NULL

# murder
mur <- read.csv("murder.csv")
dead2010 <- mur[, c("country", "X2010")]
df<- merge(df, dead2010, by = "country", all.x = TRUE)
```

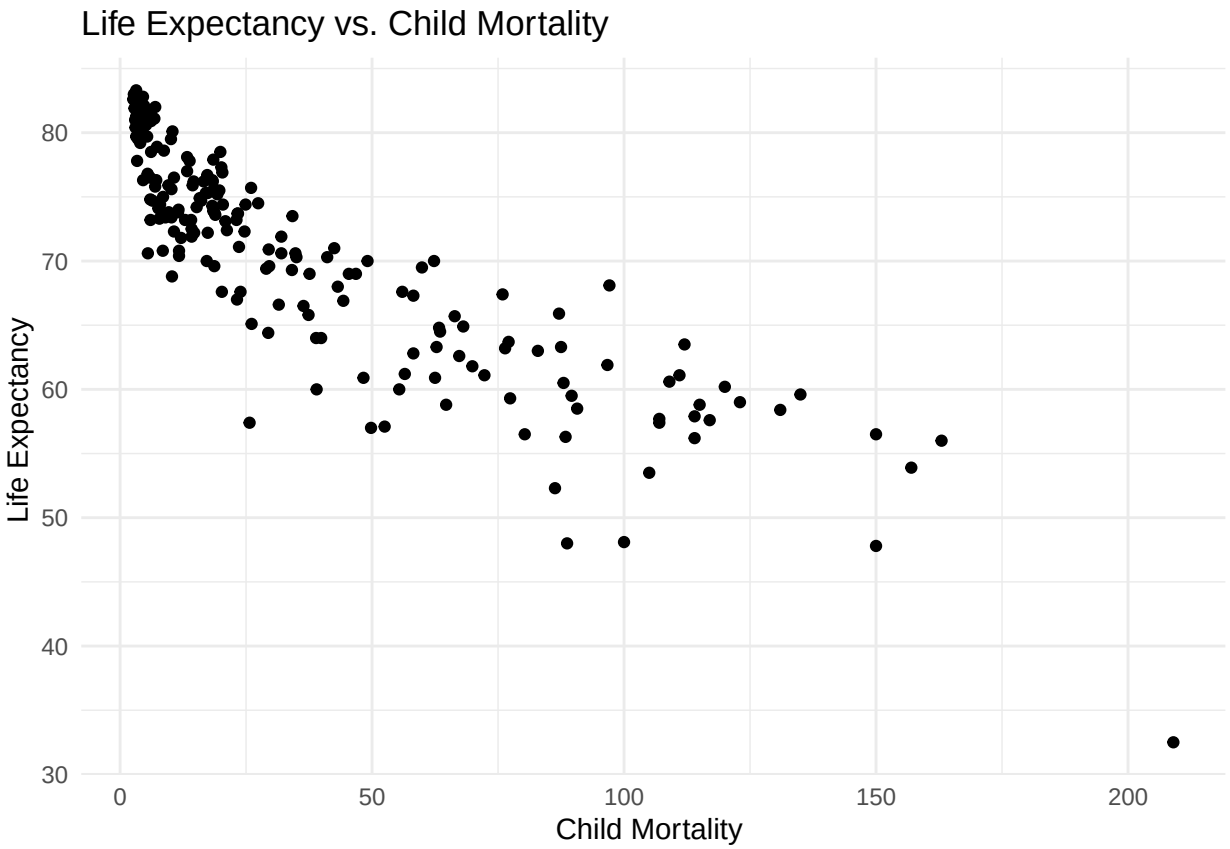
```r
df$murders <- df$X2010
df$X2010 <- NULL

glimpse(df)
```

```
## Rows: 197
## Columns: 13
## $ country                        <chr> "Afghanistan", "Albania", "Algeria", "An~
## $ population                     <dbl> 2.83e+07, 2.93e+06, 3.62e+07, 8.07e+04, ~
## $ life_expectancy                <dbl> 60.5, 78.1, 74.5, 81.8, 60.2, 75.9, 75.9~
## $ daily_income                   <dbl> 4.50, 9.77, 9.08, 66.50, 6.29, 18.40, 23~
## $ babies_per_woman               <dbl> 6.20, 1.65, 2.88, 1.28, 6.19, 1.79, 2.41~
## $ child_mortality                <dbl> 88.00, 13.30, 27.40, 4.18, 120.00, 9.59,~
## $ co2_emissions_per_person       <dbl> 0.290, 2.260, 3.280, 6.120, 1.240, 5.960~
## $ gdp_per_capita                 <dbl> 2030, 10700, 11000, 60500, 7690, 20100, ~
## $ health_spending_per_person     <dbl> 37.7, 241.0, 178.0, 3100.0, 123.0, 690.0~
## $ population_density             <chr> "43.6", "107", "15.2", "172", "18.7", "1~
## $ basic_water_source_percentage <dbl> 48.8, 91.4, 92.3, 100.0, 50.4, 98.4, 98.~
## $ murders                        <chr> "4130", "65.9", "530", "0.5", "824", "5.~
## $ continent                      <chr> "Asia", "Europe", "Africa", "Europe", "A~
```

```r
# Scatter plot of life expectancy vs childmortality
library(ggplot2)
ggplot(data = df, aes(x = child_mortality, y = life_expectancy)) +
  geom_point() +
  labs(title = "Life Expectancy vs. Child Mortality",
       x = "Child Mortality ",
       y = "Life Expectancy ") +
  theme_minimal()
```

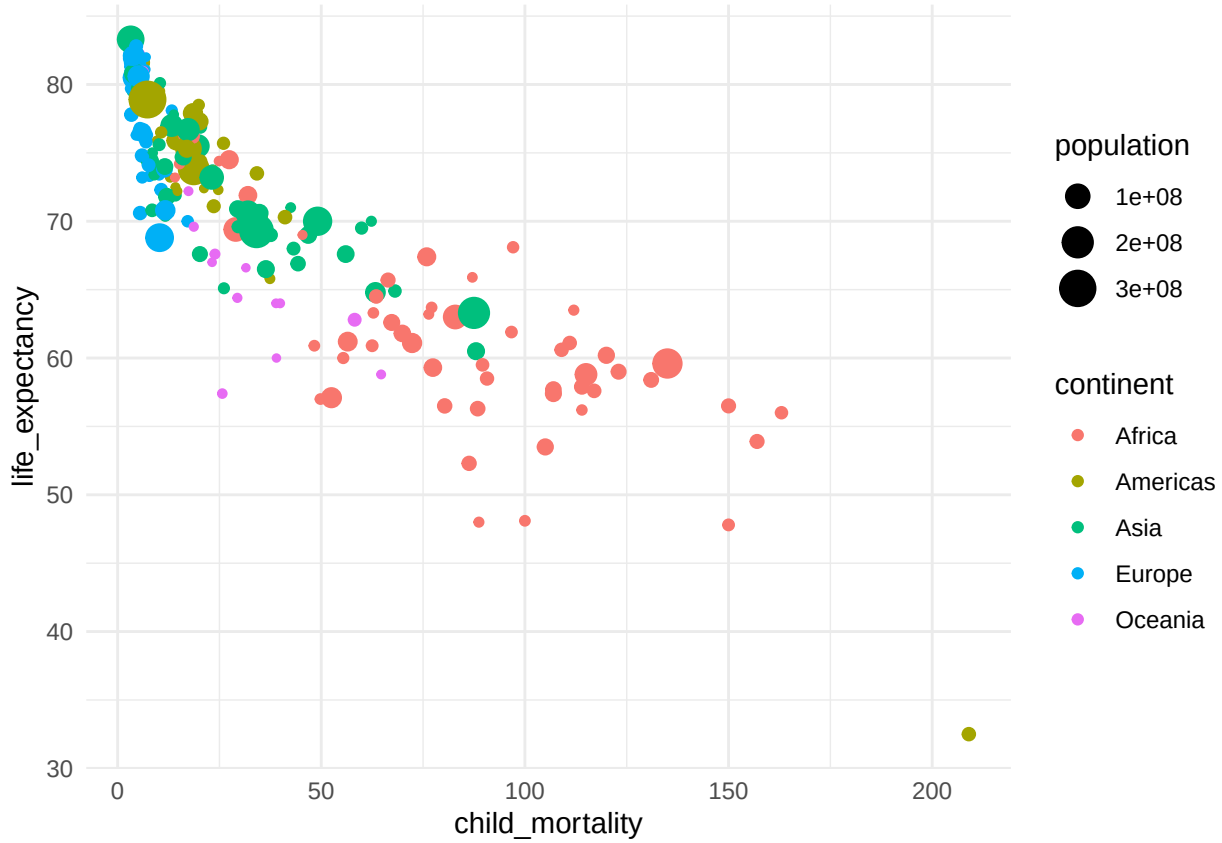
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



For the above graph the correlation between the graphs shows the lower the child mortality the higher the life expectancy is. Same with the opposite, the higher the child mortality the lower the life expectancy. So we can conclude there is a relatinship between child mortality and life expectancy where the lower the child mortality the higher the life expectancy. This scatter plot therefore has a negative relationship and looks to be approximately linear.

```
#2 life expectancy and child mortality in color
ggplot(data = df, aes(x= child_mortality, y= life_expectancy, color= continent, size=population)) + geo
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



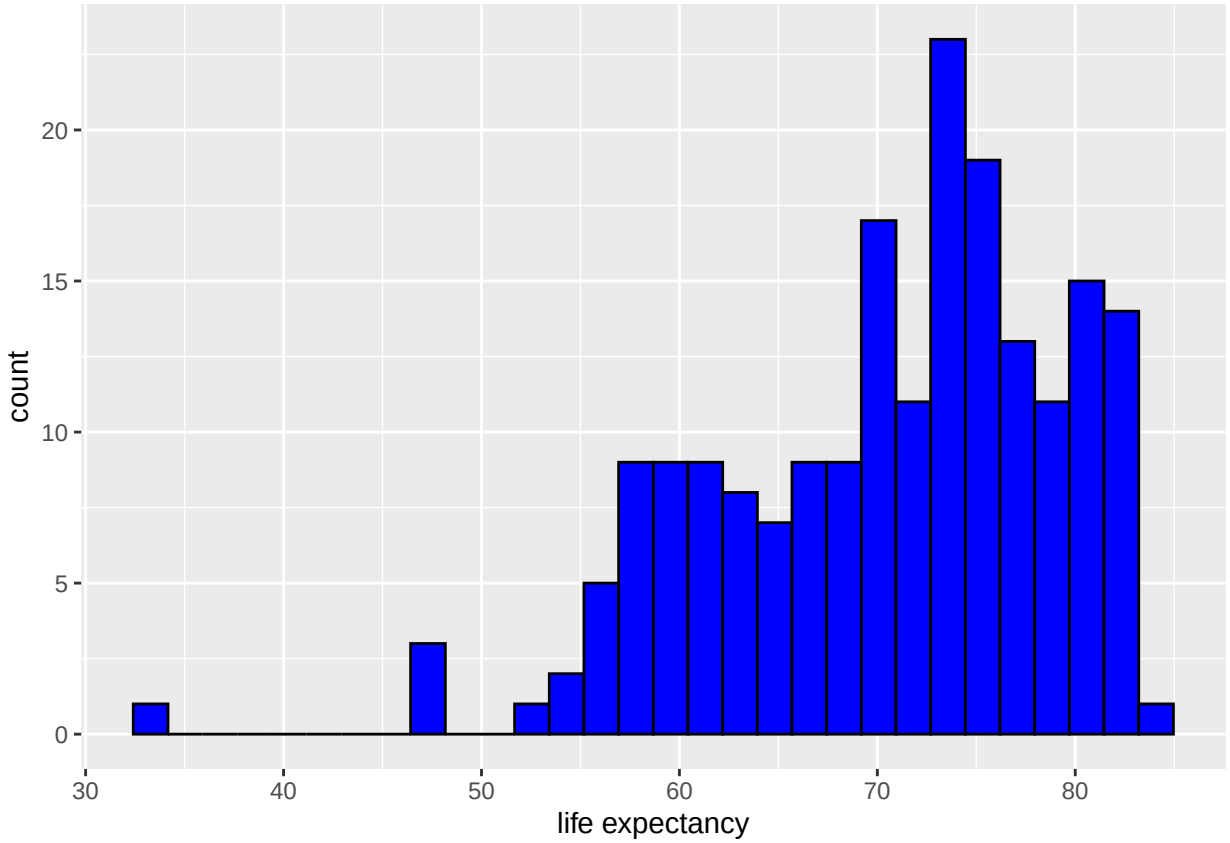
Similar to the relationship discussed above the graph shows that Africa has a higher child mortality rate and therefore a lower life expectancy. Where as continents like the Americas and Europe have a higher life expectancy and a lower child mortality rate. This graph can help break the relaitonship by countries so we can see other than the one outlier in the Americas the trend by Countries.

```
#3 this is a Histogram of life expectancy

ggplot(data = df, aes(x = life_expectancy))+ geom_histogram(fill="blue", color="black")+ labs(x = "life
```

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

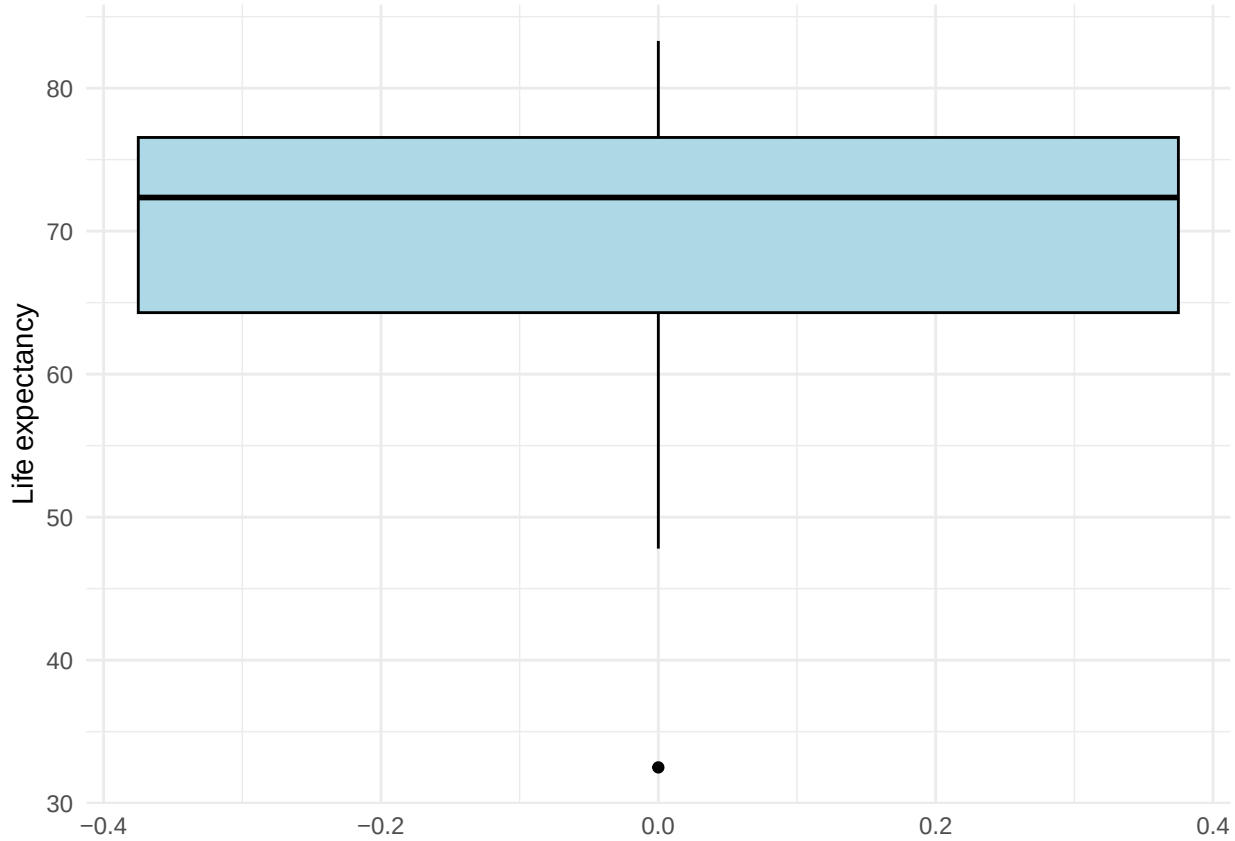
```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```



explanation : This data is skewed to the left meaning most of the life expectancy are higher. When I say higher I mean 60 and above with a lone outlier being around 35.

```
# 4 box plot of life expectancy
ggplot(data = df, aes(y = life_expectancy)) + geom_boxplot( fill = "lightblue", color = "black" ) + labs
```

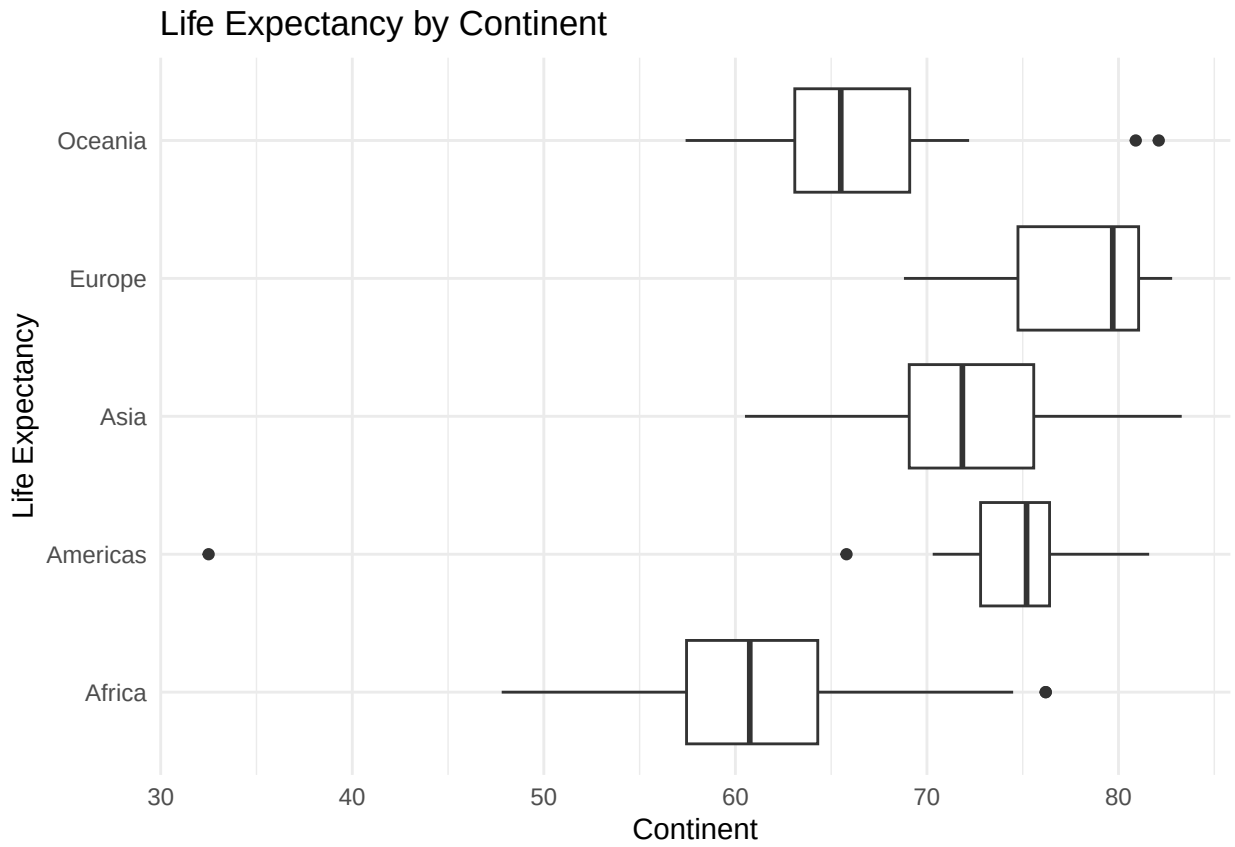
```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



There is a standalone dot thats not included in the box or on the line, the dot is around 30 and is far to the left because that country has a much shorter life expectancy than the others. But the box plot also shows the average life expectancy is around 70 years of age.

```
# #5box plot by continent
ggplot(data= df, aes(x = life_expectancy, y = continent))+geom_boxplot()+labs(title = "Life Expectancy
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```

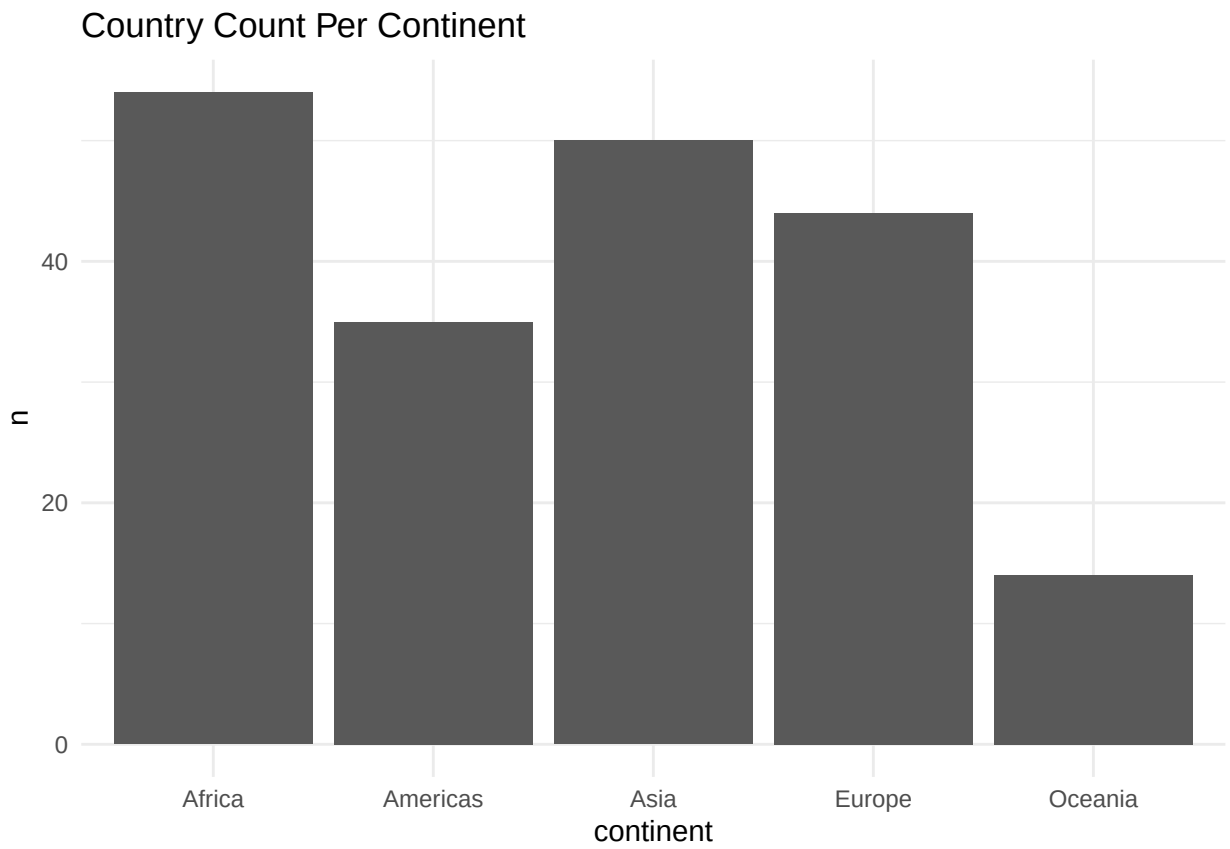


the plots remain similar, but African has a significantly lower average life expectancy than the other continents. Europe has the highest life expectancy well other continents like Asia and the Americas have similar means and ranges. We can also see there are 2 negative outliers for the Americas and 2 positive outliers for Oceania.

```
library(dplyr)
c_counts <- df %>%
count(continent)
print(c_counts)
```

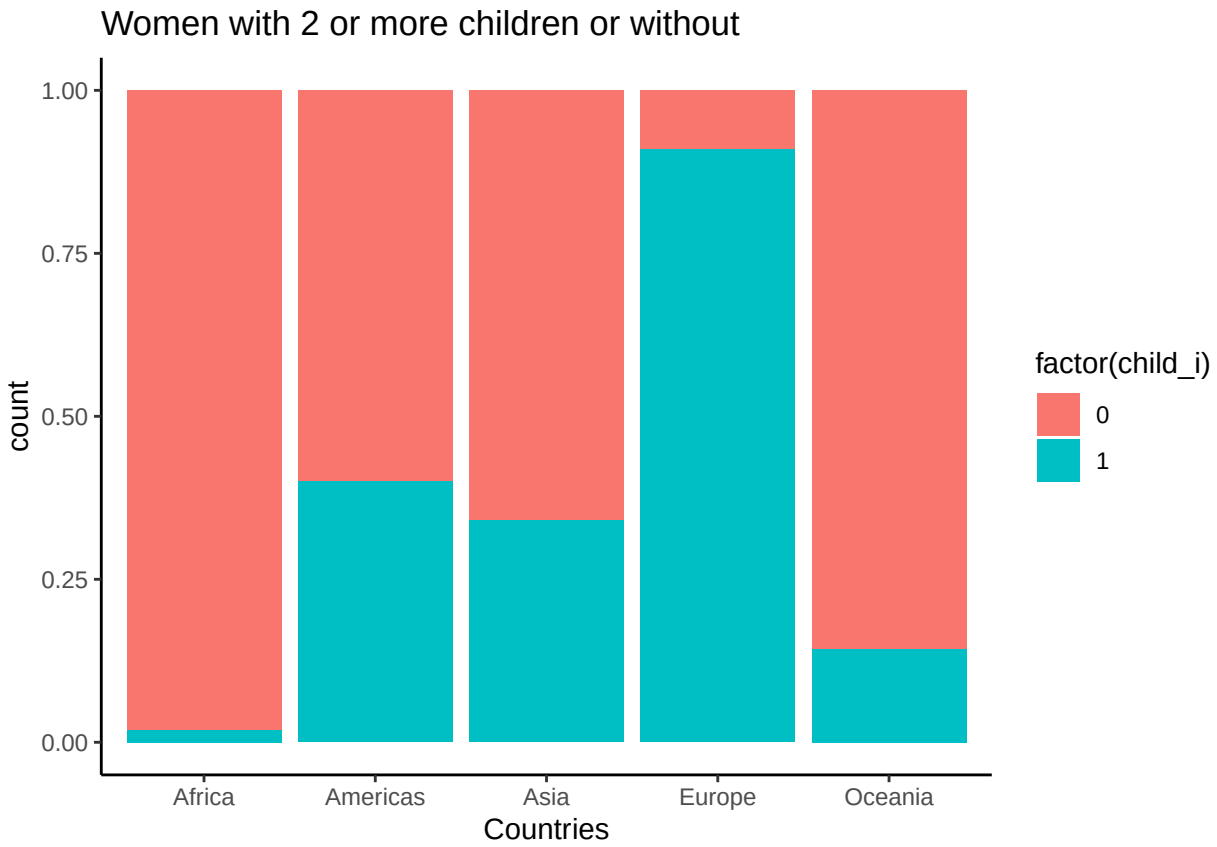
```
##   continent  n
## 1    Africa 54
## 2  Americas 35
## 3      Asia 50
## 4    Europe 44
## 5   Oceania 14
```

```
  # # 6 create plot to show numbver of countries
ggplot(data= c_counts, aes(x= continent, y=n)) + geom_bar(stat = "identity")+ labs(title = "Country Cou
```



This plot below shows that Africa and Asia have the highest country counts well continents like Oceania have singificantly lower counts which could affect the different stats, like murders, income, money per child, and more. Africa has the most countries and Asia has the second most countries.

```
#define an indicator of weather the avergae number of women per children is > 2
# 7 and 8
baby_c <- df %>%
  mutate(child_i = ifelse(babies_per_woman <= 2, 1 ,0))
# print(baby_c)
ggplot(data = baby_c, aes(x = continent, fill = factor(child_i))) + geom_bar(position="fill") + labs(ti
```



The percetnages of children <= 2 are not very similar, Europe and ociania have smaller smaller averages/ counts well other countries like Africa Asia and the Americas have more counts with less than 2 babies per women. A trend we can see is that most countries aren't very even(balanced). They either have more than 2 children majority of the time or less than 2 children majority of the time. there is very little overlap.