Assignment2 STAT291

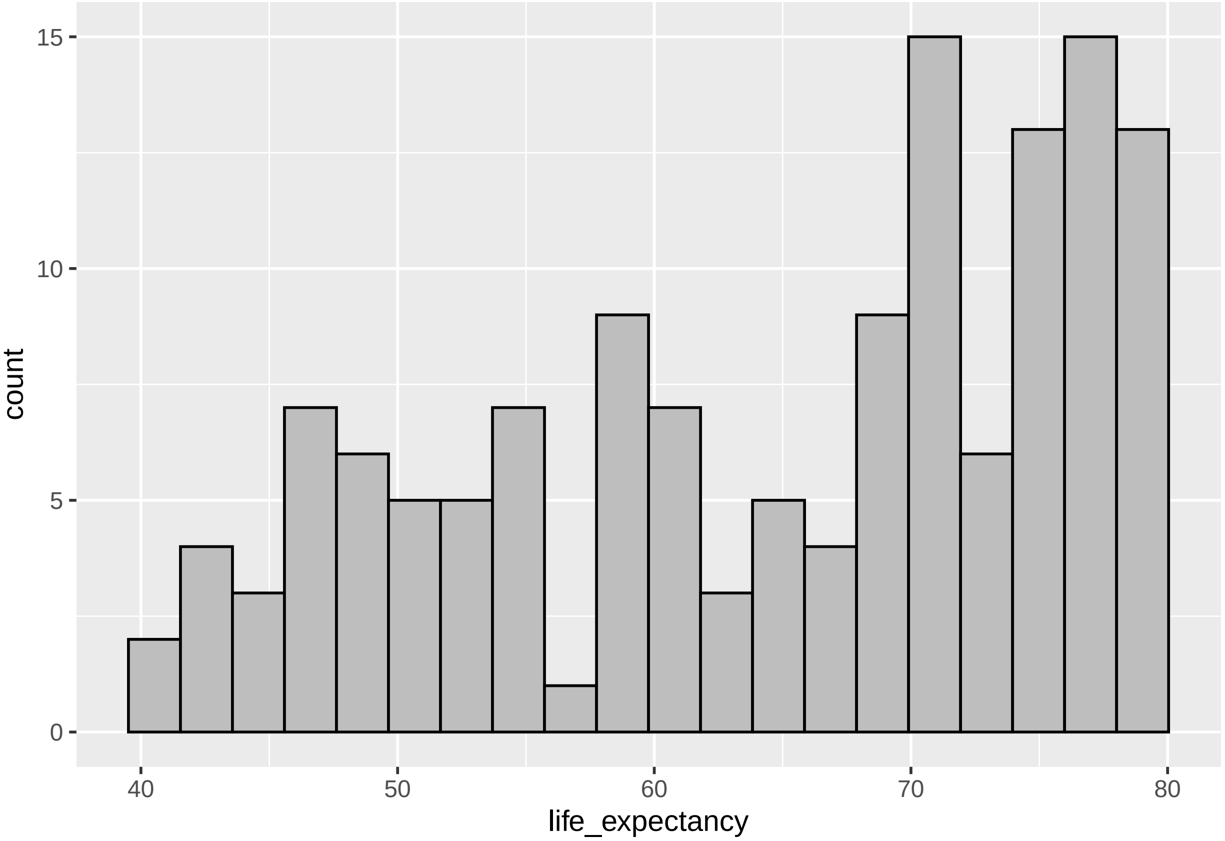
2023-02-09

By, Lucas Weston

Section 1: LIFE EXPECTANCY IN UNTIED STATES

Q1: Have a histogram of the life expectancy, describe the distribution of it

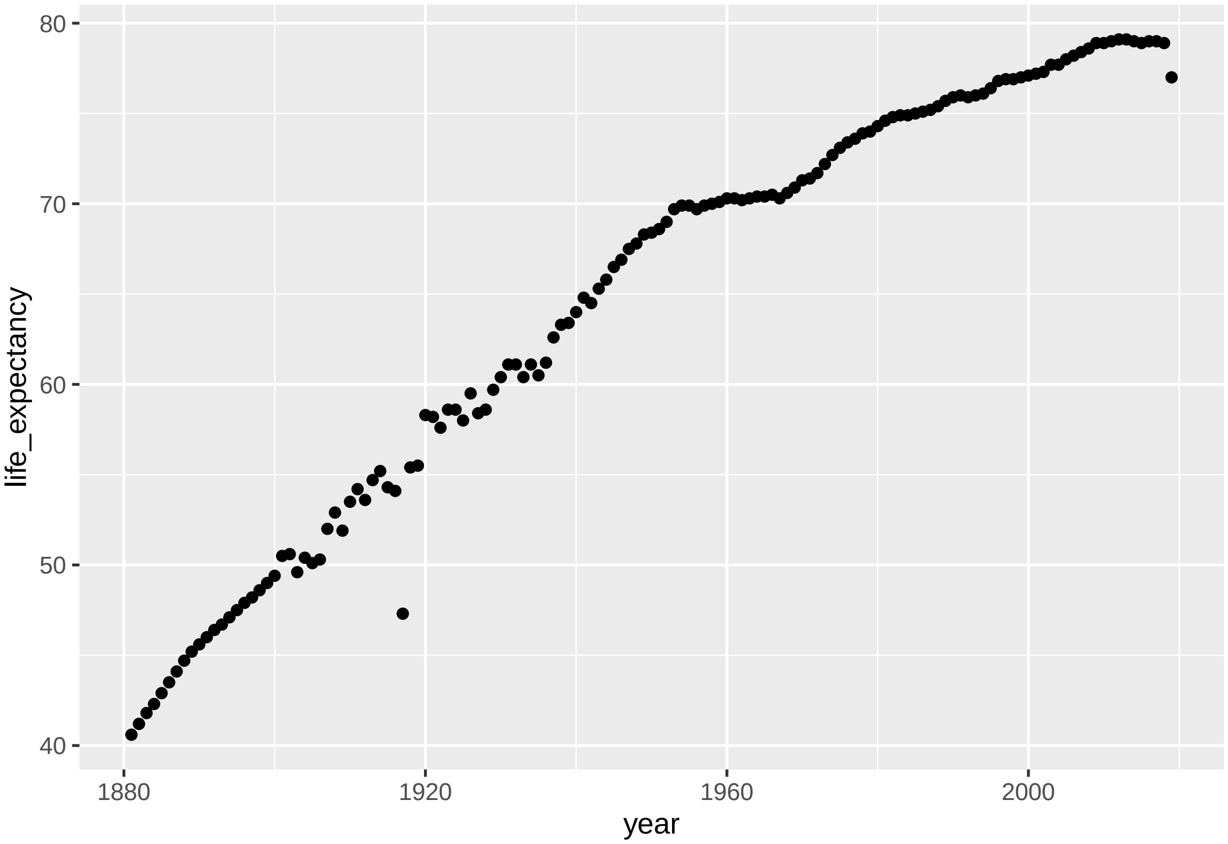
```
df <- read.csv("USlifehistory.csv")
ggplot(df, aes(x=life_expectancy)) + geom_histogram(bins = 20, color="black", fill="gray")
```



Based on the graph, we can see its a right leaning graph, due to the fact that the majority of the data exists >60.

Q2: Does it appear to be some linear relationship between life expectancy and the number of years since 1880 (using the scatterplot)? Is it a positive or negative trend?

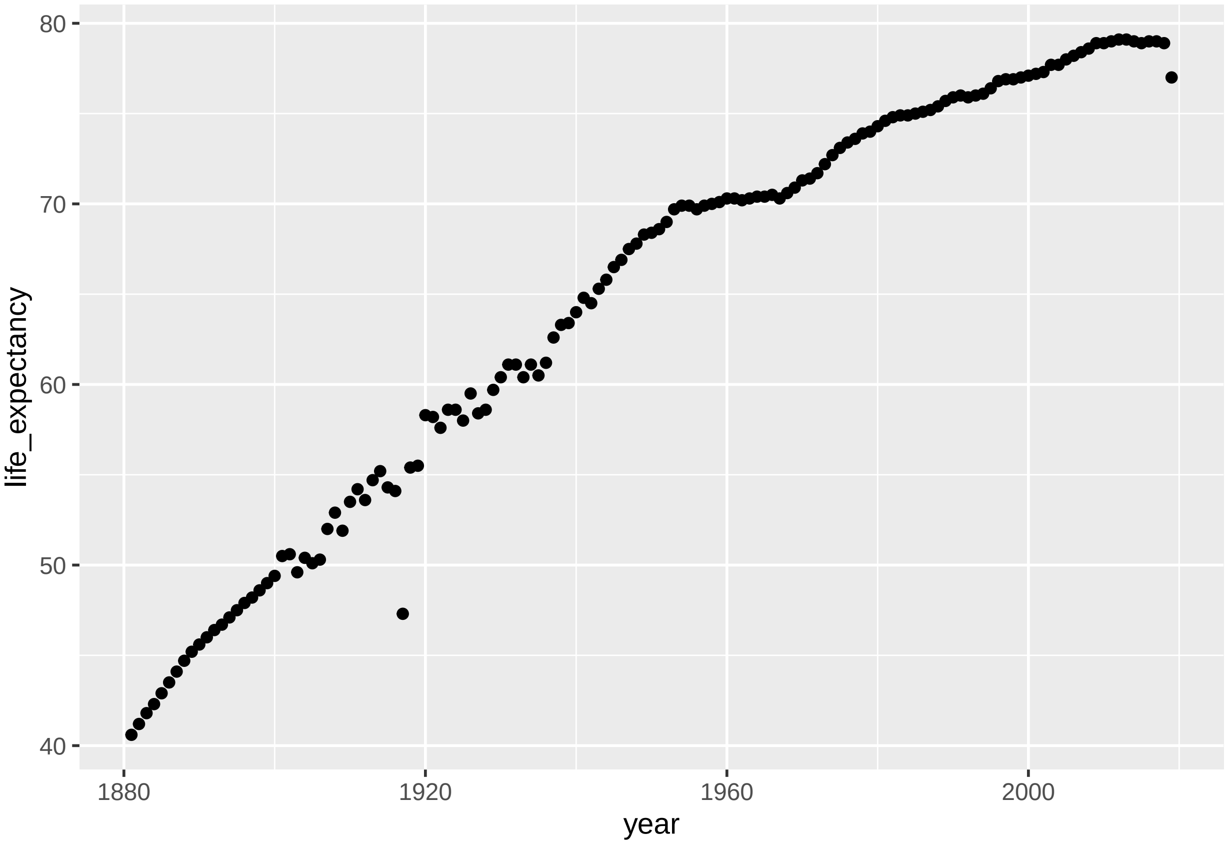
```
df <- read.csv("USlifehistory.csv")
ggplot(df, mapping = aes(x = year,y=life_expectancy)) + geom_point()
```



The data is linear and has a positive trend.

Q3: Are there any unusual points in that trend? What could be the possible reason for that?

```
df <- read.csv("USlifehistory.csv")
ggplot(df, mapping = aes(x = year,y=life_expectancy)) + geom_point()
```



There are 2 outliers along with a large breakup between 1900 and 1940 which could have been caused by the wars that happened during this time. |

## Q4: What is the correlation between life expectancy and number of years since 1880?
r df <- read.csv("USlifehistory.csv") cor(df$life_expectancy, df$year)
## [1] 0.9789403
The correlation value between life expectancy and the number of years since 1880 is 0.9789403

Q5: Run a simple regression. Is the model significant?

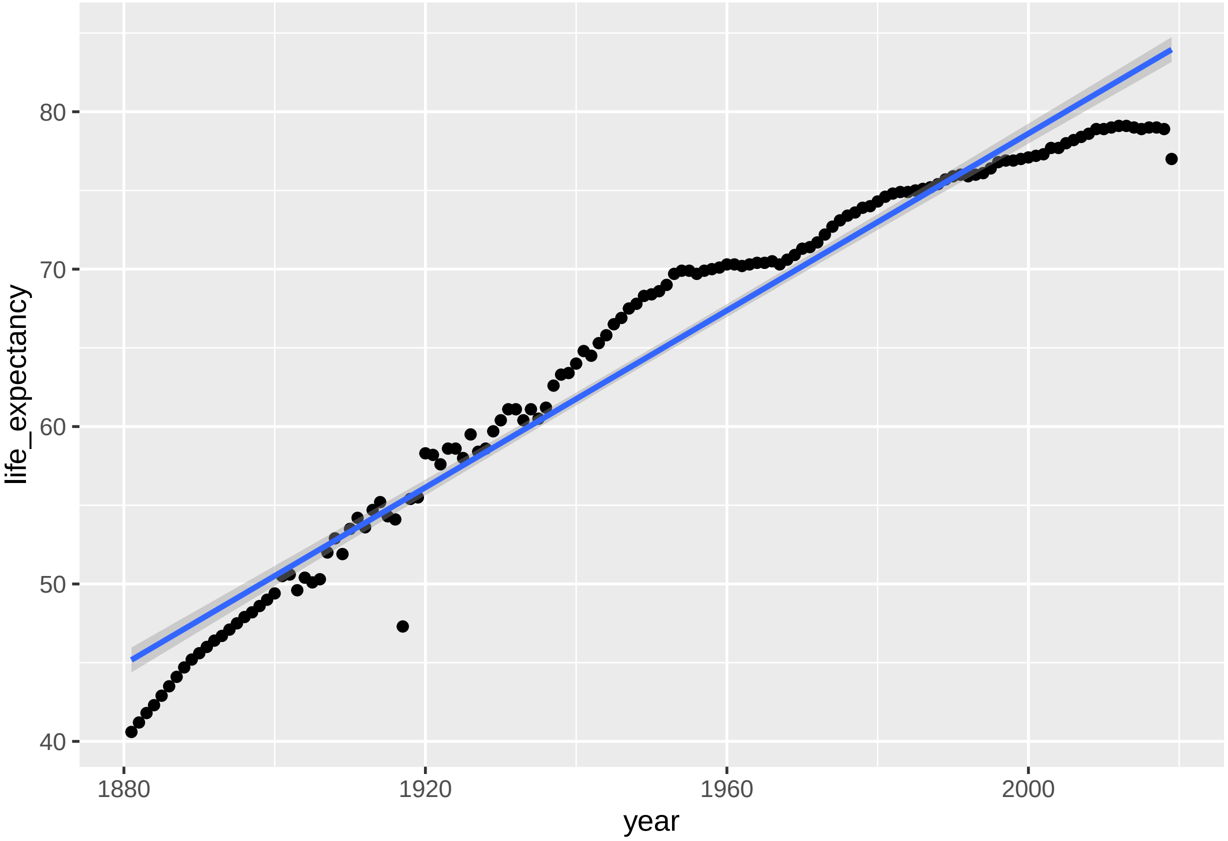
```
df <- read.csv("USlifehistory.csv")

model <- lm(life_expectancy ~ year, data = df)
model
```

```
##
## Call:
## lm(formula = life_expectancy ~ year, data = df)
##
## Coefficients:
## (Intercept)         year
##    -483.408        0.281
```

```
ggplot(df, aes(year, life_expectancy)) +
  geom_point() +
  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
confint(model)
```

```
##                   2.5 %       97.5 %
## (Intercept) -502.7174401 -464.0979404
## year           0.2711104    0.2909111
```

```
sigma(model)*100/mean(df$life_expectancy)
```

```
## [1] 3.668456
```

This is a simple regression model, I believe it is significant.

Q6: On average, what is the increase in life expectancy per year?

In order to find the average increase of life expectancy per year, we must take the first value on the data sheet which is 1881 with a life expectancy of 40.6 and the end value 2019 with a life expectancy of 77. We then divide 77 by 40.6 which is 1.9

We then subtract 1.9 by 1 and we get a value of .9 which is now our final value.

The average increase in life expectancy per year is .9 years.

##Q7: Predict the life expectancy in year 2023.

```
df <- read.csv("USlifehistory.csv")
lmModel <- lm(life_expectancy~year, data = df)
```

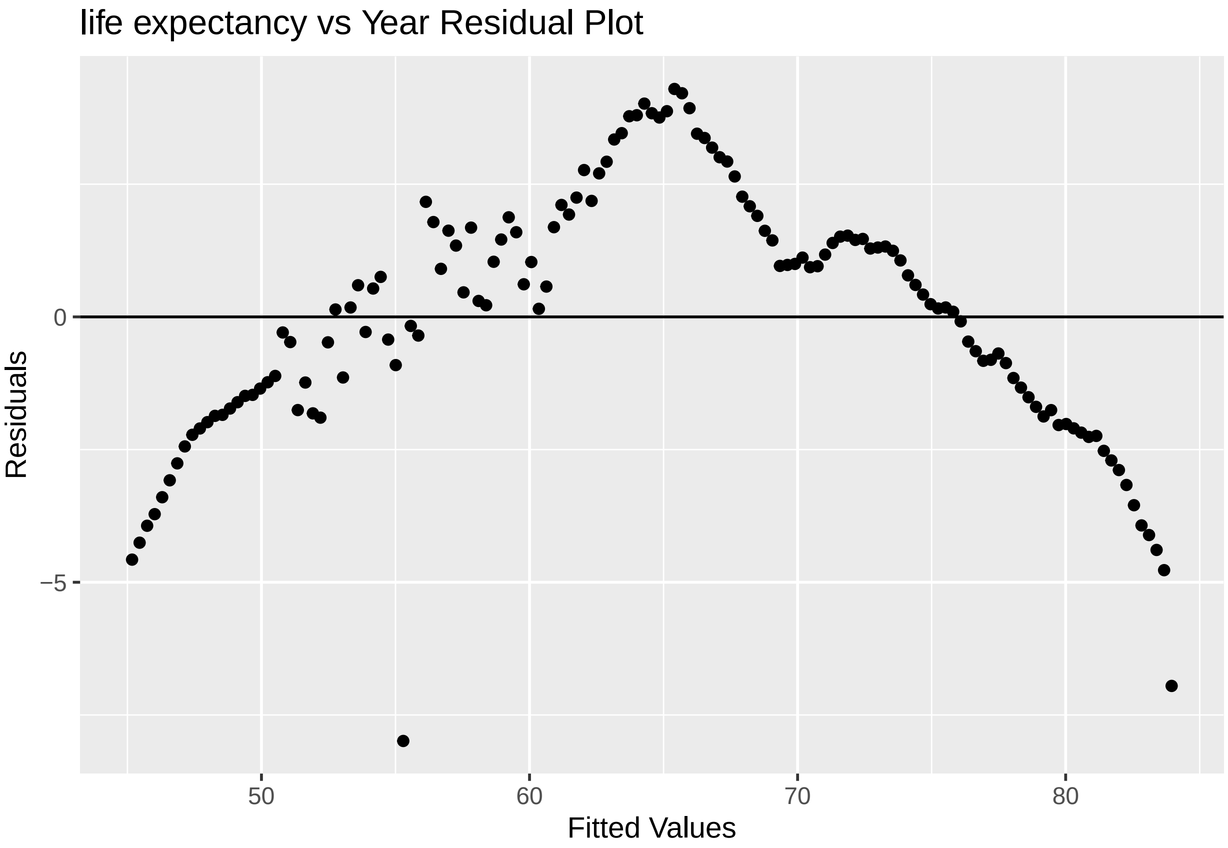
```
yuh_year <- data.frame(year = c(2020,2021,2022,2023))
linear_model <- lm(life_expectancy~year, data = df)
predict(linear_model, newdata = yuh_year)
```

```
##        1        2        3        4
## 84.23406 84.51507 84.79608 85.07710
```

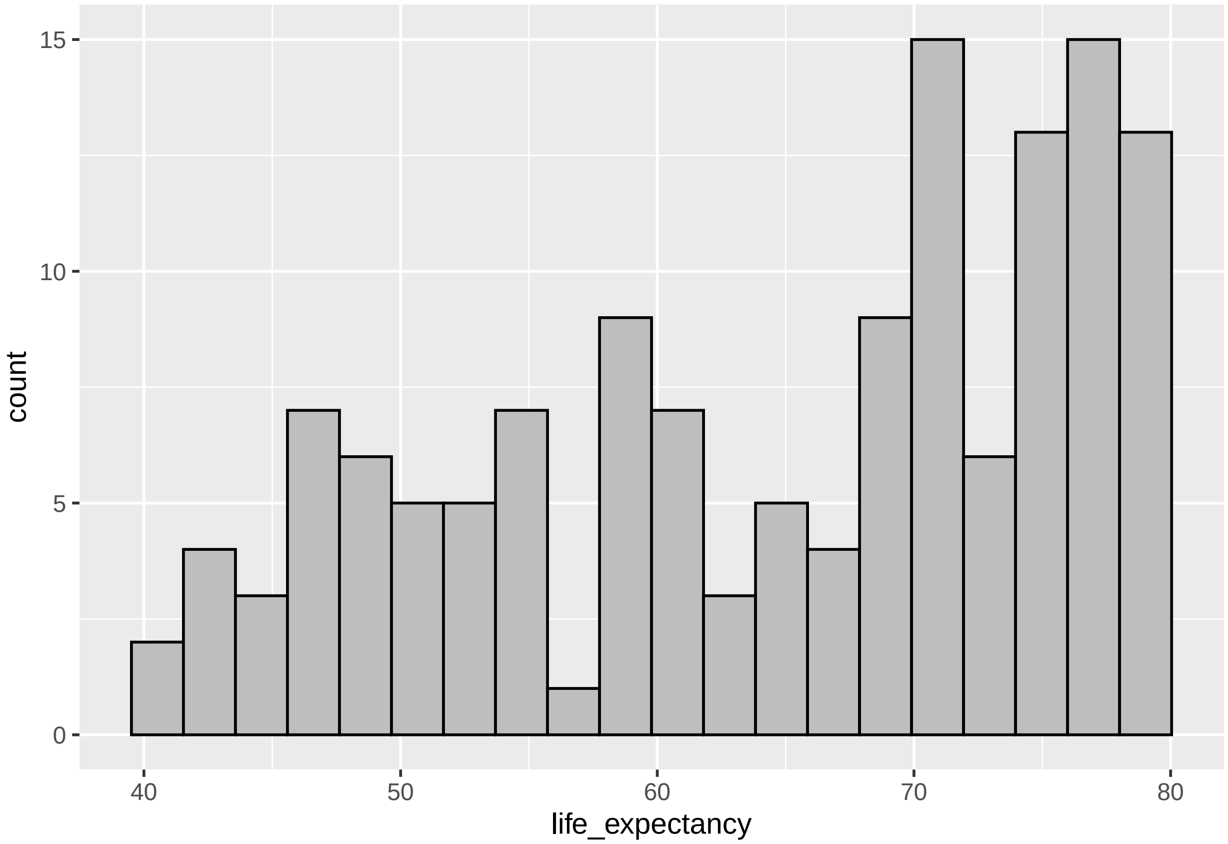
From this prediction model we can estimate that the life expectancy in the year 2023 is 85.07710.

##Q8: Have a residual plot of residual against number of years since 1880, and a histogram of the residual. Describe whether the residual seems to be random, explain why.

```
df <- read.csv("USlifehistory.csv")
rModel <- lm(life_expectancy~year, data = df)
ggplot(model, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title='life expectancy vs Year Residual Plot', x='Fitted Values', y='Residuals')
```



```
ggplot(model, aes(x=life_expectancy)) + geom_histogram(bins = 20, color="black", fill="gray")
```



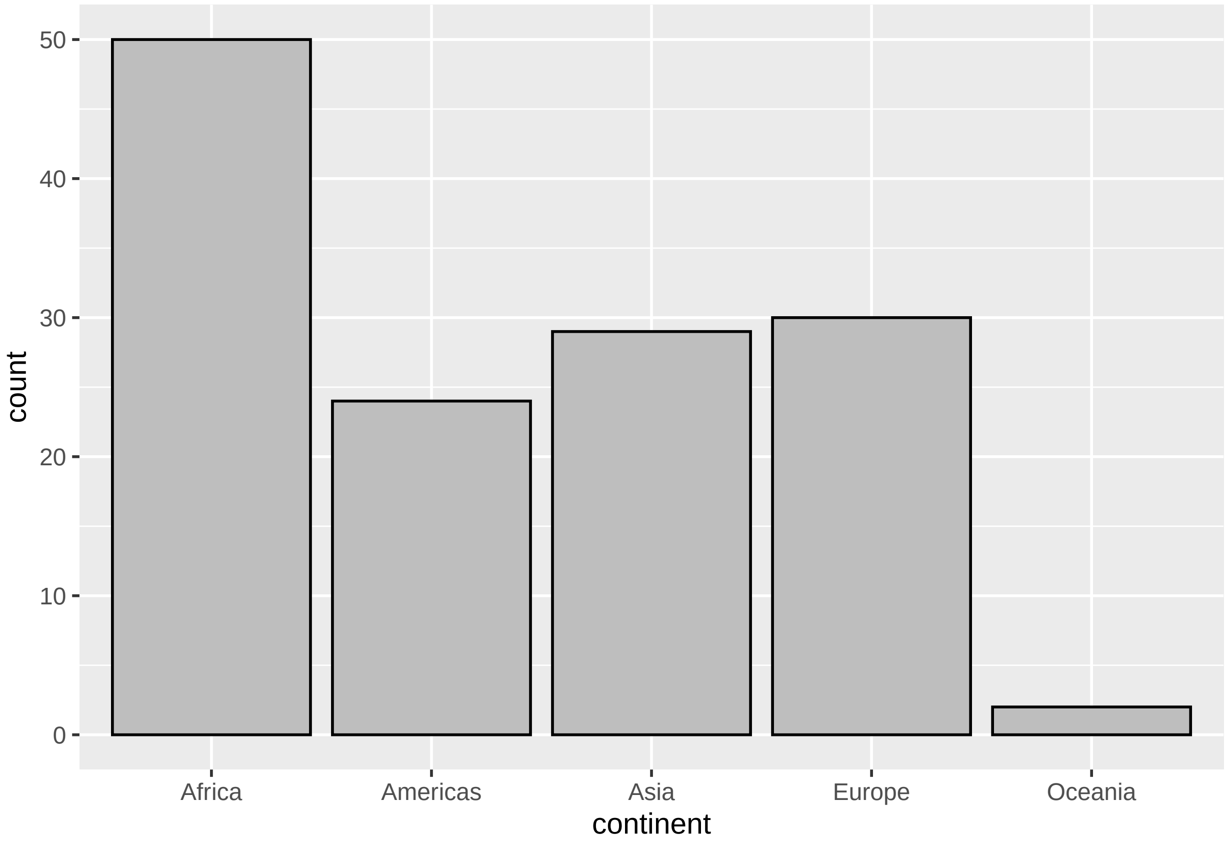
Here we have 2 plots, a residual plot based on the residual and a histogram based on the residual as well. The residual plot does not appear to be random either.

Section 2: LIFE EXPECTANCY IN THE WORLD

Q1: How many countries are there in each continent?

```
world <- read.csv("Worldlife100.csv")
ggplot(world, aes(x=continent)) + geom_histogram(stat = "count", bins = 20, color="black", fill="gray")
```

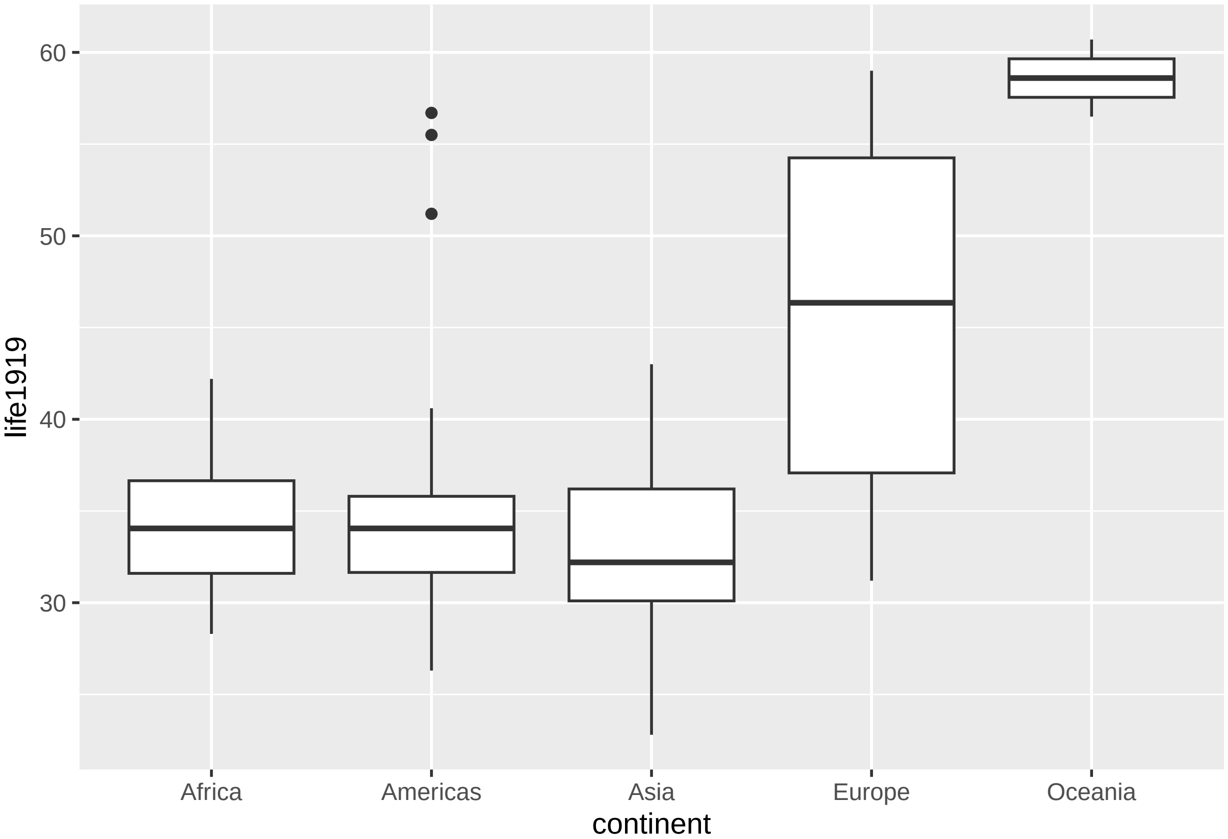
```
## Warning in geom_histogram(stat = "count", bins = 20, color = "black", fill =
## "gray"): Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```



We can see from the graph that Africa has the most countries out of all the continents.

Q2: Have a side-by-side boxplot of life expectancy in 1919 by continent and describe it.

```
world <- read.csv("Worldlife100.csv")
ggplot(world, aes(x=continent, y=life1919)) + geom_boxplot()
```



The graph described is a side by side comparison of Continents and their average age expectancy in 1919. From the graph, we can infer that the continents with the highest life expectancy are Oceania and Europe. Americas has a few outliers

---

## Q3: Have a histogram of life expectancy in 1919 by continent.

## Q4: Have a table summarizing the mean and median of life expectancy in 1919 in each continent.

```
r world <- read.csv("Worldlife100.csv") amMean<- mean(dfAmericas$life1919) asMean<-
mean(dfAsia$life1919) euMean<- mean(dfEurope$life1919) afMean<- mean(dfAfrica$life1919)
ocMean<- mean(dfOcean$life1919) means <- c(amMean, asMean, euMean, afMean, ocMean)
tab1<-table(means) tab1
```

```
## means ## 33.0137931034483           34.106 35.9166666666667 46.1866666666667 ##
1                1                1                1 ##            58.6 ##
1
```

```
r amMedian<- median(dfAmericas$life1919) asMedian<- median(dfAsia$life1919) euMedian<-
median(dfEurope$life1919) afMedian<- median(dfAfrica$life1919) ocMedian<-
median(dfOcean$life1919) medians<- c(amMedian, asMedian, euMedian, afMedian, ocMedian)
tab2<- table(medians) tab2
```

```
## medians ##  32.2 34.05 46.35  58.6 ##     1     2     1     1
```

```
r ##tab <- matrix(rep(2, times=10), ncol=5, byrow=TRUE) ##colnames(tab) <- c('Americas',
'Asia', 'Europe', 'Africa', ##'Oceania') ##rownames(tab) <- c('Mean', 'Median') ##tab <-
as.table(tab) ##tab
```

##Q5: Fit a regression model of life expectancy in 1919 on continent using default reference level. What is the estimated average life expectancy in each continent? Compare the results with the previous summary table. Are there any levels that are insignificant?

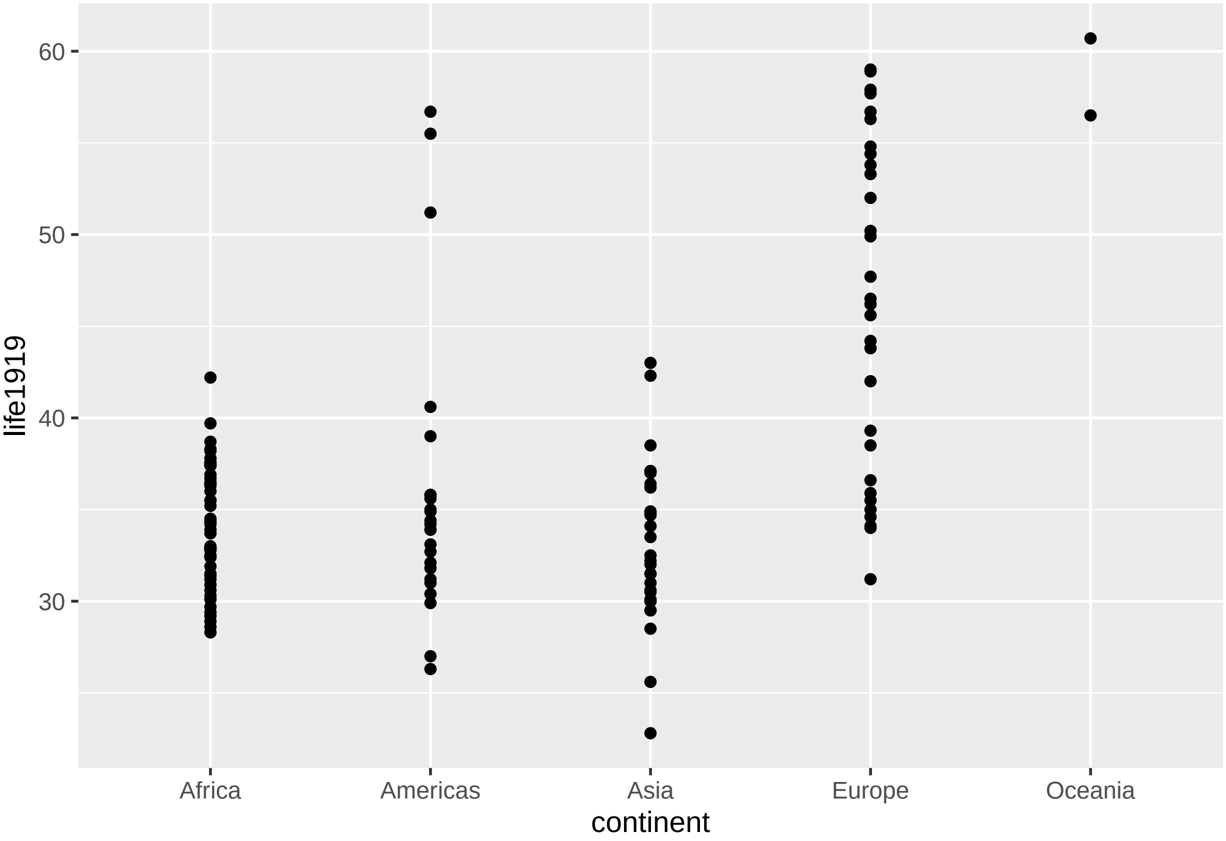
```
world <- read.csv("Worldlife100.csv")

wModel <- lm(life1919 ~ continent, data = world)
wModel
```

```
##
## Call:
## lm(formula = life1919 ~ continent, data = world)
##
## Coefficients:
##       (Intercept)  continentAmericas      continentAsia    continentEurope
##            34.106              1.811             -1.092             12.081
##  continentOceania
##            24.494
```

```
ggplot(wModel, aes(continent, life1919)) +
  geom_point() +
  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
confint(wModel)
```

```
##                        2.5 %    97.5 %
## (Intercept)        32.394826 35.817174
## continentAmericas -1.194056  4.815389
## continentAsia     -3.916493  1.732079
## continentEurope    9.286332 14.875002
## continentOceania  15.768691 33.219309
```

```
sigma(wModel)*100/mean(world$life1919)
```

```
## [1] 16.42294
```

##Q6: Rerun the regression by using different reference levels.

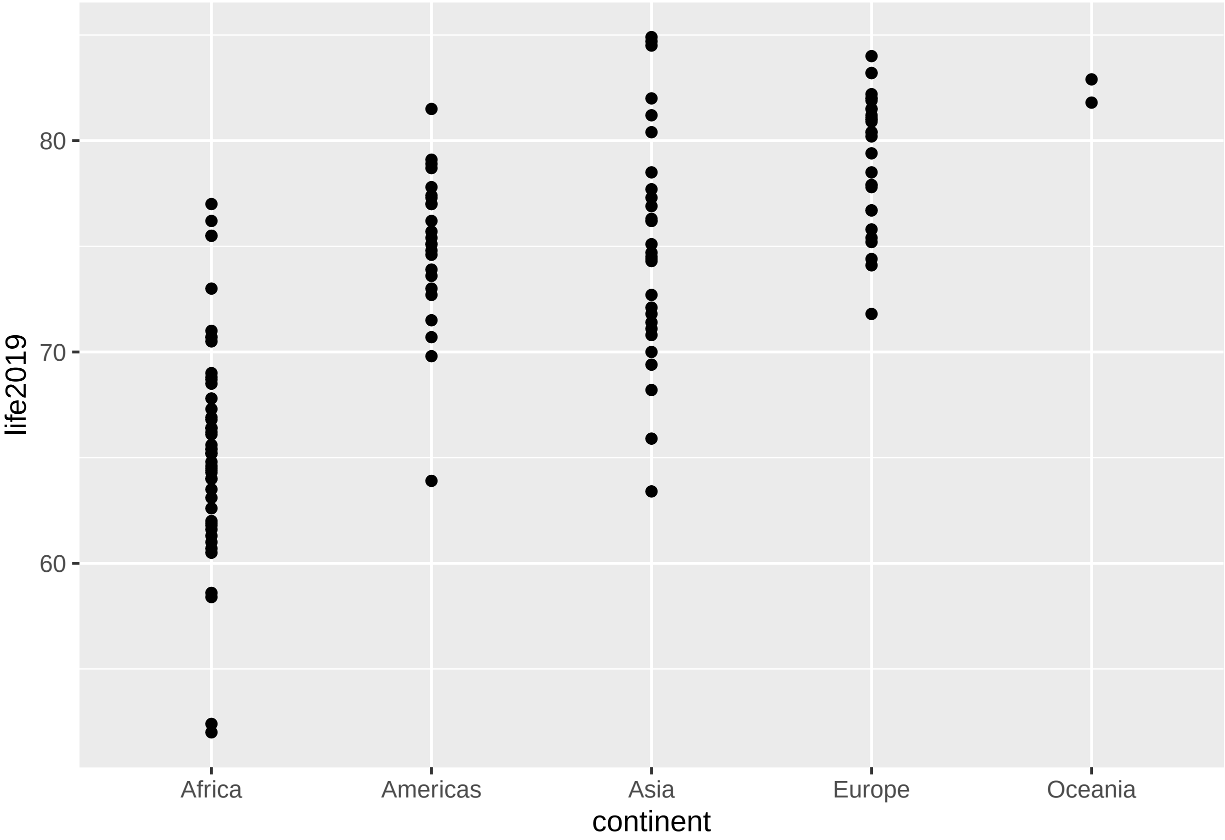
```
world <- read.csv("Worldlife100.csv")

wModel <- lm(life2019 ~ continent, data = world)
wModel
```

```
##
## Call:
## lm(formula = life2019 ~ continent, data = world)
##
## Coefficients:
##       (Intercept)  continentAmericas      continentAsia    continentEurope
##            65.588              9.441              9.467             13.722
##  continentOceania
##            16.762
```

```
ggplot(wModel, aes(continent, life2019)) +
  geom_point() +
  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
confint(wModel)
```

```
##                        2.5 %   97.5 %
## (Intercept)        64.311543 66.86446
## continentAmericas   7.199781 11.68255
## continentAsia       7.360384 11.57396
## continentEurope    11.637554 15.80645
## continentOceania   10.253320 23.27068
```

```
sigma(wModel)*100/mean(world$life2019)
```

```
## [1] 6.284318
```

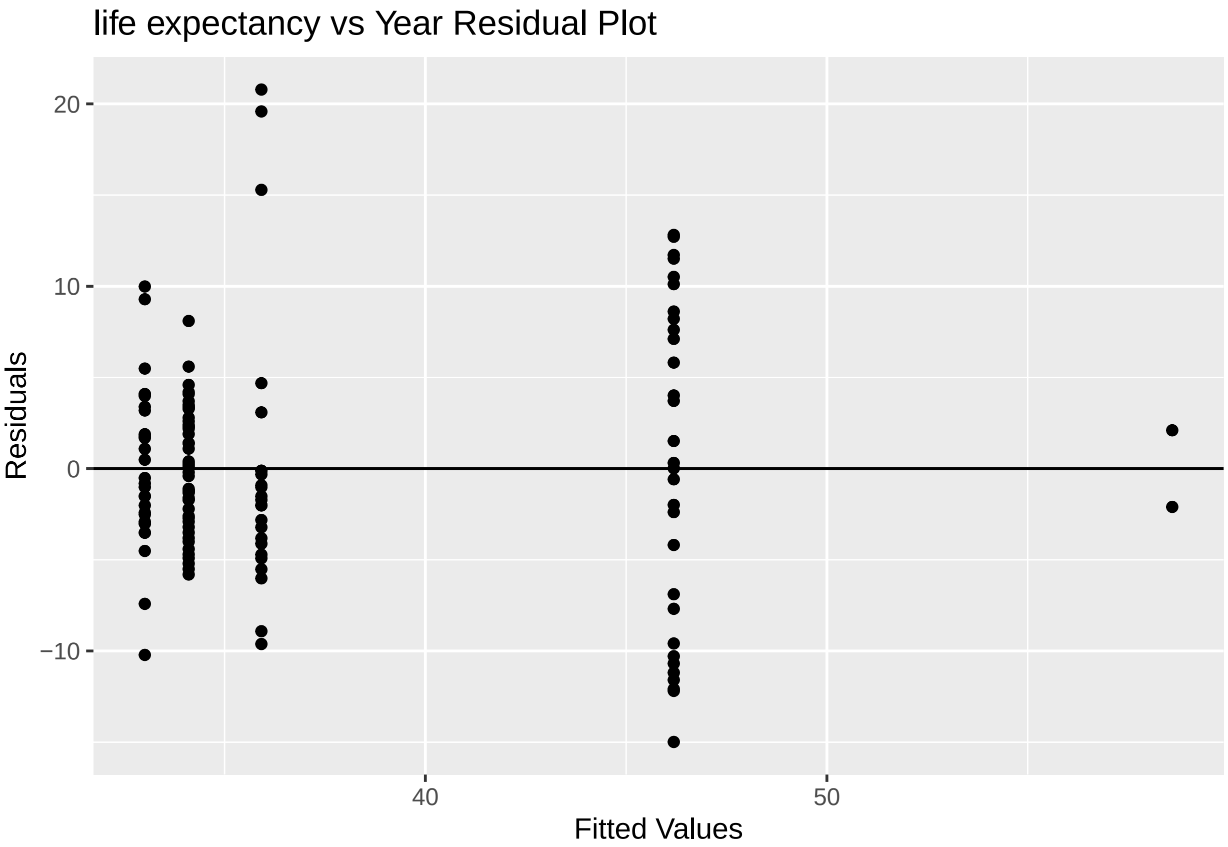
Same regression model, but with different values

##Q7: If we want to regroup the 5 levels in continent to have a new continent indicator, how will you regroup based on the output previously?

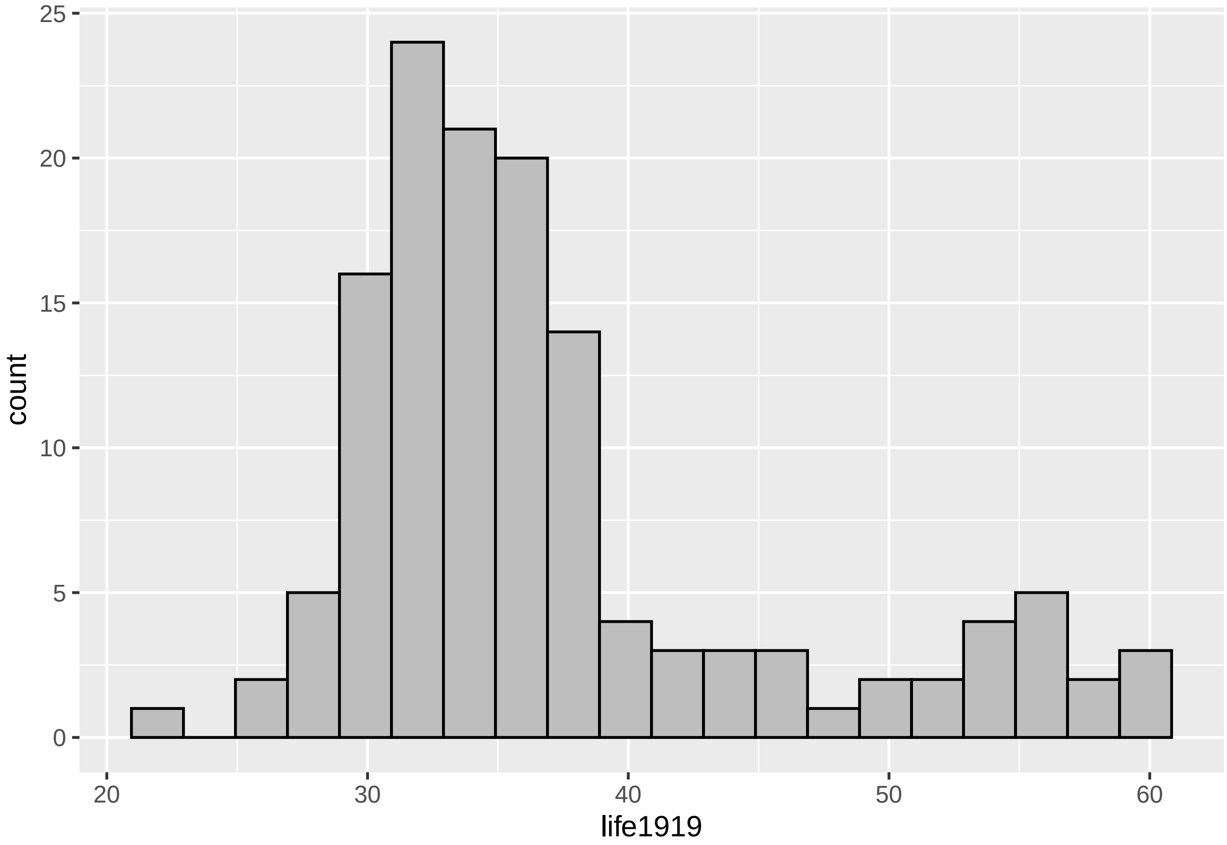
In order to regroup the 5 levels in the continent table to have a new indicator, we would need to regroup them into a new data frame which we can then use to alter the regression table method.

##Q8: Run the model using your new continent indicator and get the histogram of residual. Describe the residual. What is the percentage of total variability in life expectancy that can be explained through the linear model using this new continent indicator?

```
world <- read.csv("Worldlife100.csv")
rwModel <- lm(life1919~continent, data = world)
ggplot(rwModel, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title='life expectancy vs Year Residual Plot', x='Fitted Values', y='Residuals')
```

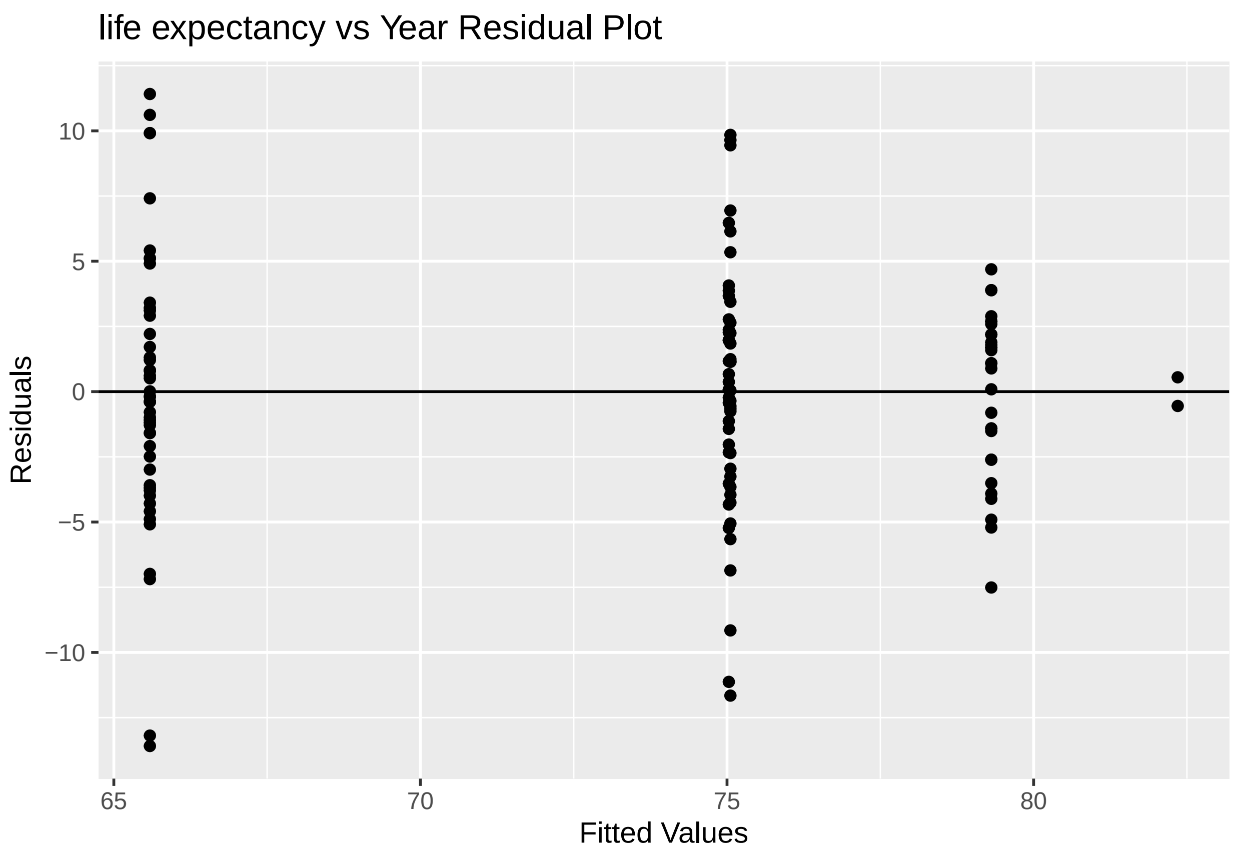


```
ggplot(rwModel, aes(x=life1919)) + geom_histogram(bins = 20, color="black", fill="gray")
```

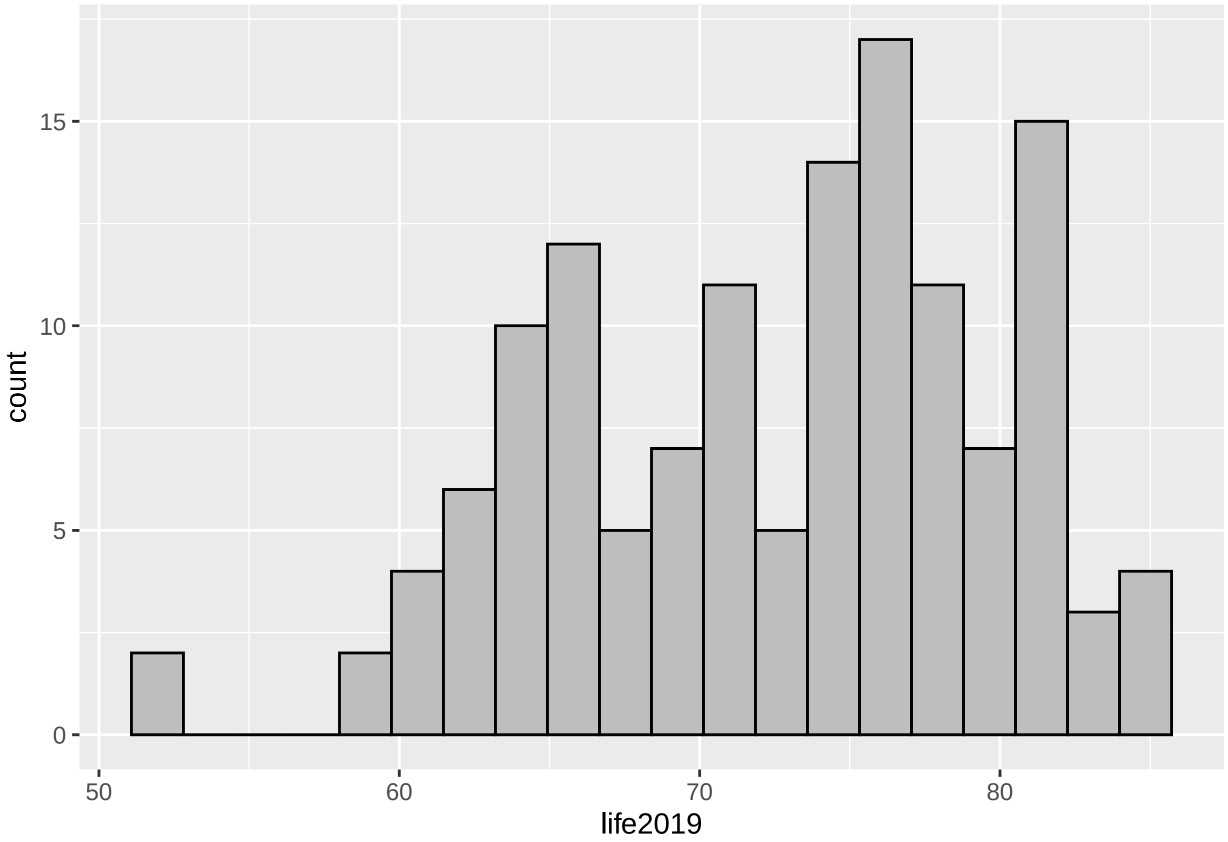


##Q9: Repeat the previous steps using life expectancy in 2019 as the dependent variable.

```
world <- read.csv("Worldlife100.csv")
rwModel <- lm(life2019~continent, data = world)
ggplot(rwModel, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(title='life expectancy vs Year Residual Plot', x='Fitted Values', y='Residuals')
```



```
ggplot(rwModel, aes(x=life2019)) + geom_histogram(bins = 20, color="black", fill="gray")
```



##Q10: Describe whether you see any difference happened in these 100 years.

Yes I see differences. The biggest difference that happened within the last 100 years is that the overall life expectancy across all continents increased by nearly 20 years.
