Final Project - EDA

Colm Kennedy, Shril Patel

2022-11-03

Libraries

```
library(ggplot2)
library(moderndive)
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
```

```
all <- read.csv('data/all.csv')

head(all)
```

```
##   X.1 X     country population lifeexp childmort income gdpcapita chdperwoman
## 1   1 1 Afghanistan   29200000    60.5     88.00   1960       543        5.82
## 2   2 2     Albania    2950000    78.1     13.30  10800      4090        1.65
## 3   3 3     Algeria   36000000    74.5     27.40  11000      4480        2.89
## 4   4 5      Angola   23400000    60.2    120.00   7690      3590        6.16
## 5   5 7   Argentina   40900000    75.9     14.40  23500     10400        2.37
## 6   6 9   Australia   22200000    82.1      4.77  45100     52000        1.93
##   healthspend   co2 water popdensity murder continent baby2
## 1        37.7  0.29  73.5      44.70 4940.0      Asia     0
## 2       241.0  1.56  92.9     108.00   68.4    Europe     1
## 3       178.0  3.28  95.0      15.10  447.0    Africa     0
## 4       123.0  1.24  67.5      18.70  978.0    Africa     0
## 5       742.0  4.57  99.3      14.90 2390.0  Americas     0
## 6      4780.0 18.40  99.9       2.88  308.0   Oceania     1
```

Exploratory Data Analysis

Population

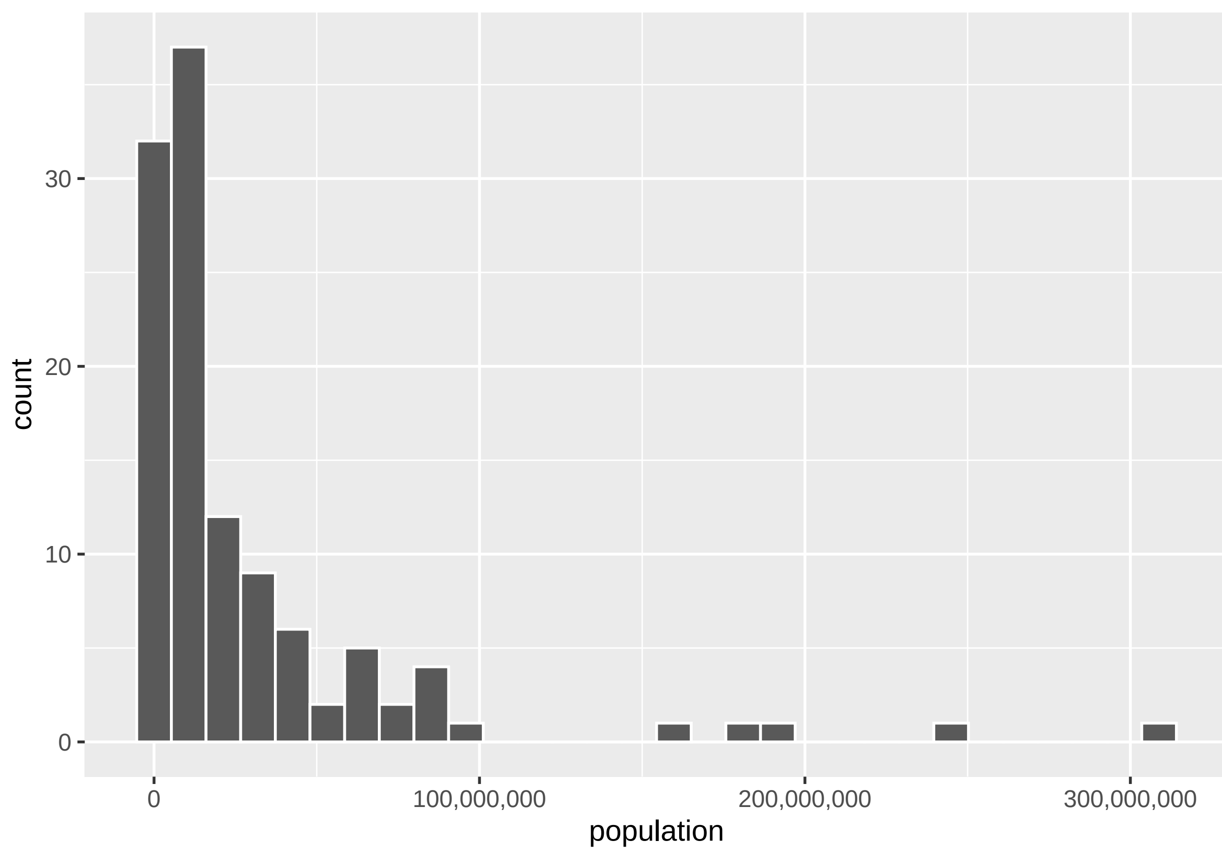
The populations are skewed right, meaning there are fewer high populations. Most populations lie between 4.6 million and 32 million. Population does not seem correlated with life expectancy, r = 0.05.

```
summary(all$population)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    180000   4565000  10900000  28945287  32350000 309000000
```

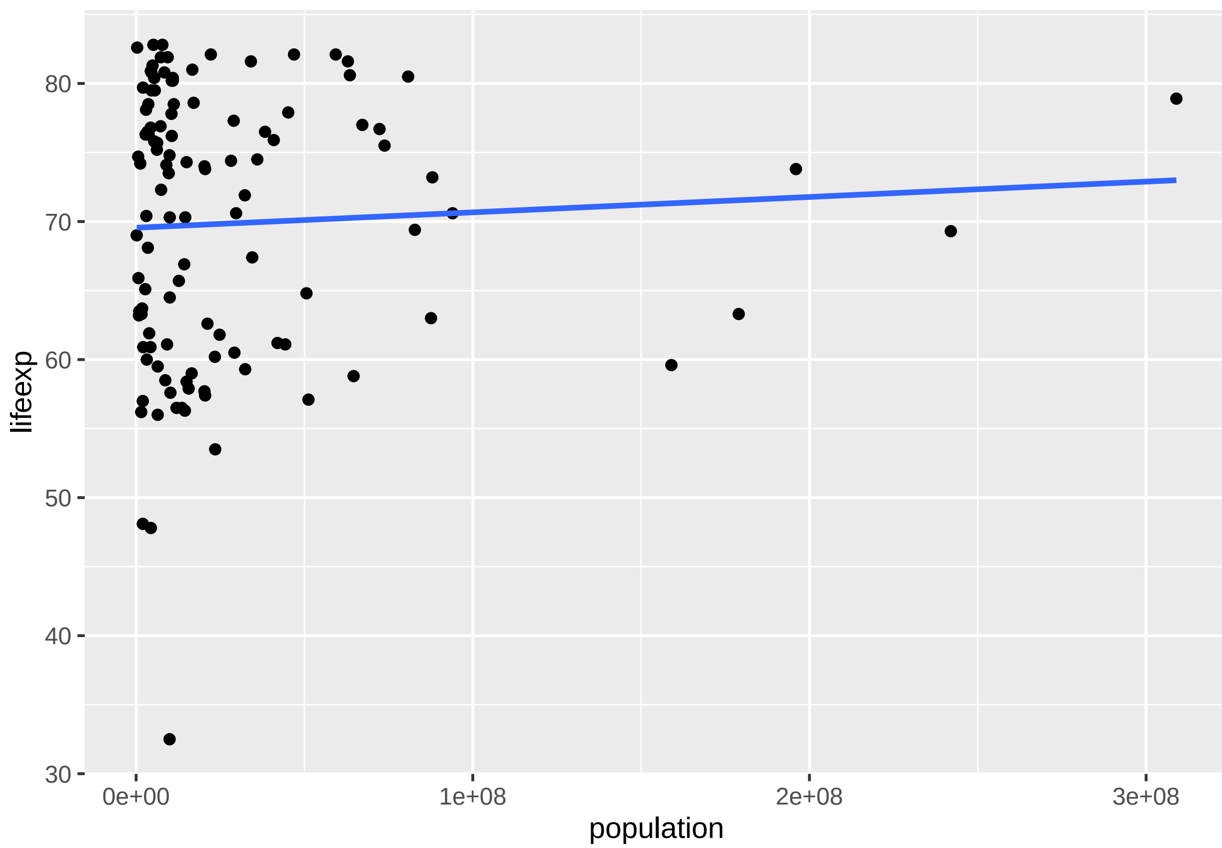
```
ggplot(data = all, mapping = aes(x = population)) +
  geom_histogram(color = 'white') +
  scale_x_continuous(labels = scales::comma)
```

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




```
ggplot(data = all, mapping = aes(x = population, y = lifeexp)) +
  geom_point()+
  geom_smooth(method = 'lm', se = FALSE)
```

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




```r
# Correlation
get_correlation(all, formula = lifeexp ~ population)
```

```
##          cor
## 1 0.05390685
```

```r
# very low corr life exp and pop
```

Child Mortality ***

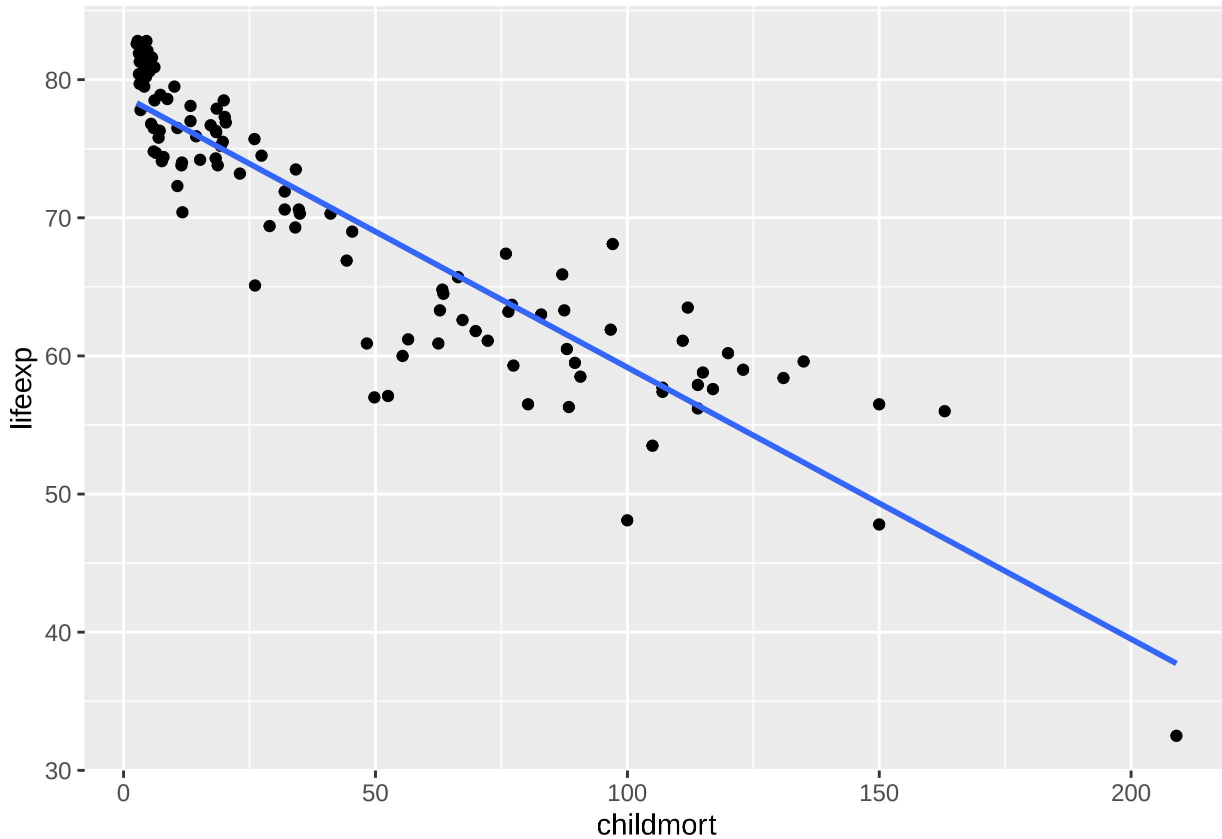
Child mortality and life expectancy have an extremely high negative correlation, r = -0.91. The plot illustrates a strong linear relationship. This is a very good indicator of life expectancy, and a great candidate for our model.

```r
summary(all$childmort)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.620   6.705  26.000  45.547  77.250 209.000
```

```r
ggplot(data = all, mapping = aes(x = childmort, y = lifeexp)) +
  geom_point()+
  geom_smooth(method = 'lm', se = FALSE)
```

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




```r
get_correlation(all, formula = lifeexp ~ childmort)
```

```
##          cor
## 1 -0.9145376
```

```r
# Extremely strong correlation
```

Income ***

Income and life expectancy are also highly correlated, r = 0.72. The relationship appears logarithmic, applying log to income appears to make the relationship linear. This is another good candidate for our model.

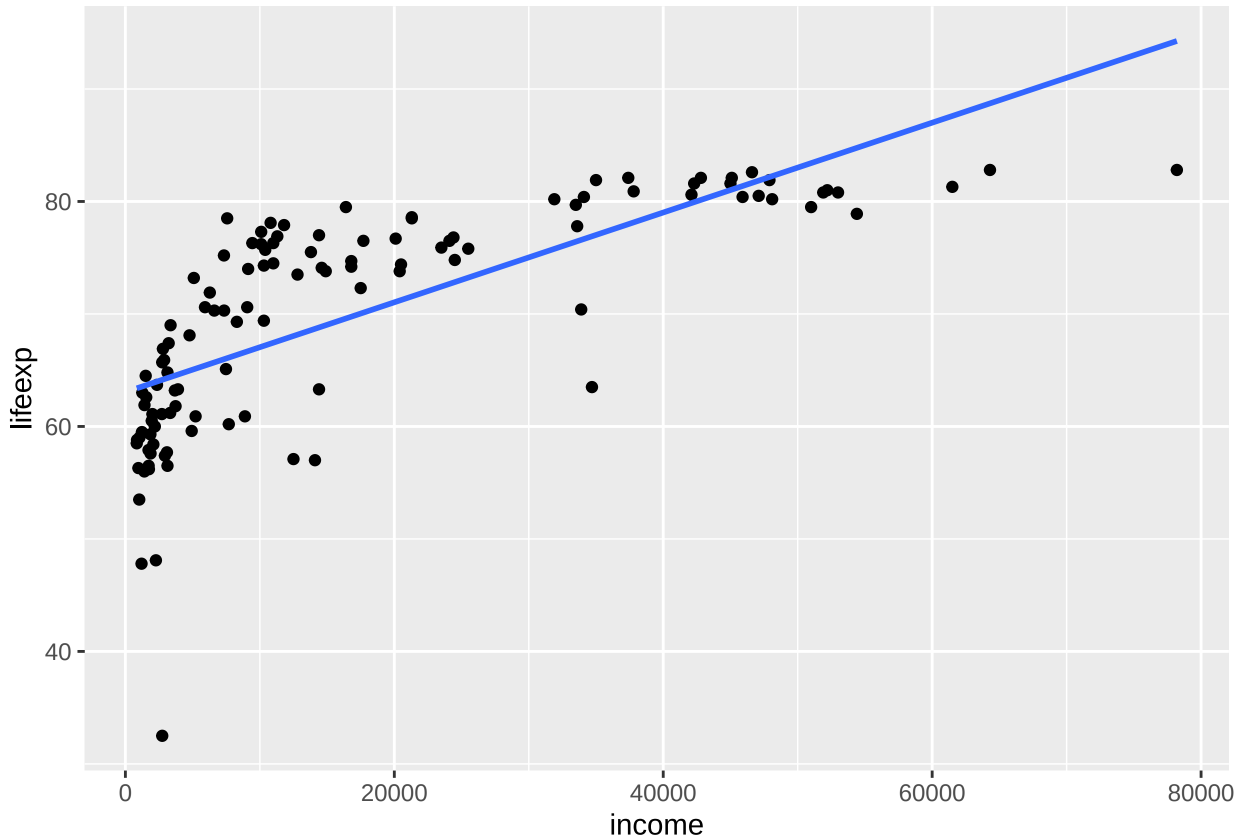
```r
# Income

summary(all$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     846    3015   10300   17089   24450   78200
```

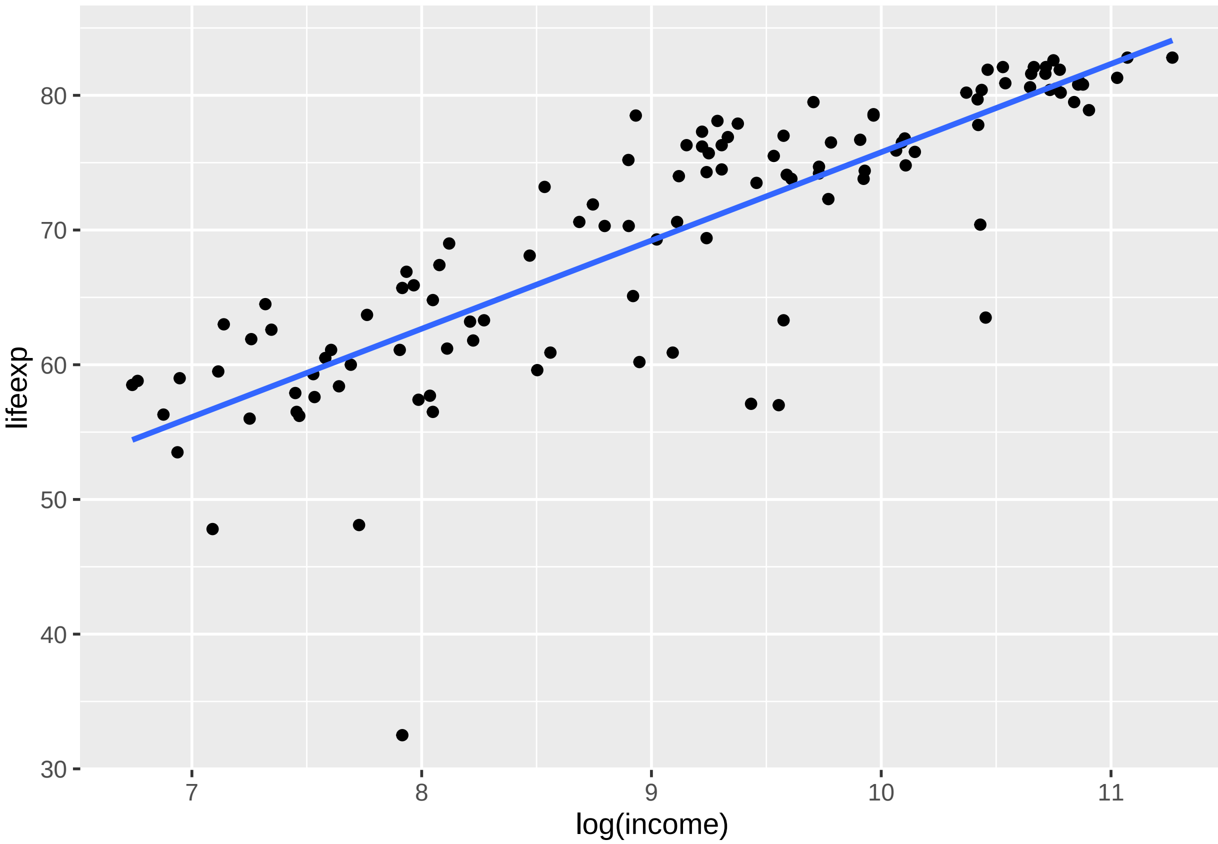
```r
ggplot(data = all, mapping = aes(x = income, y = lifeexp)) +
  geom_point()+
  geom_smooth(method = 'lm', se = FALSE)
```

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




```r
ggplot(data = all, mapping = aes(x = log(income), y = lifeexp)) +
  geom_point()+
  geom_smooth(method = 'lm', se = FALSE)
```

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




```
get_correlation(all, formula = lifeexp ~ income)
```

```
##        cor
## 1 0.723655
```

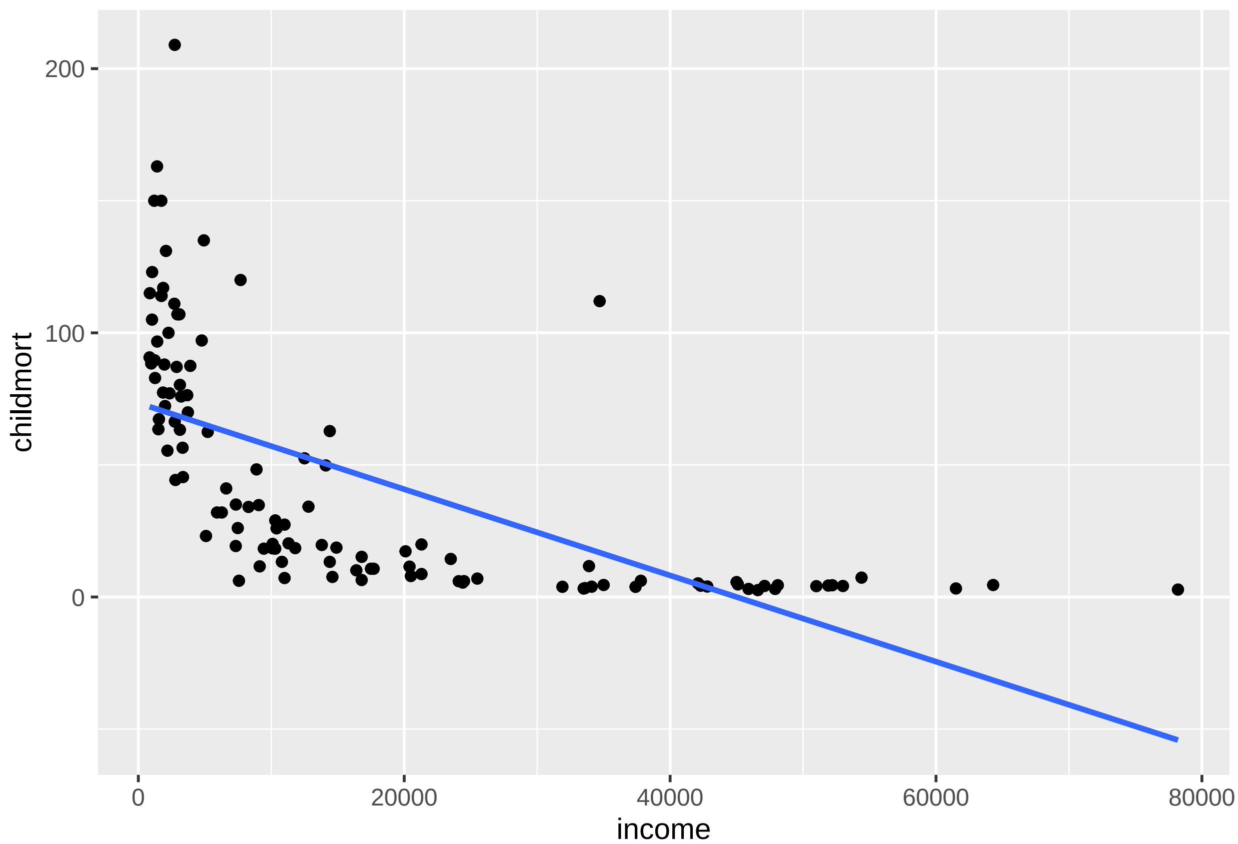
```
# Strong correlation
```

Income and Child Mortality

Income and child mortality appear to have a relationship with each other and may be interacting. They are negatively correlated, r = -0.64. After applying log to income, the relationship appears much more linear. It would be worth trying an interaction model with income and child mortality.

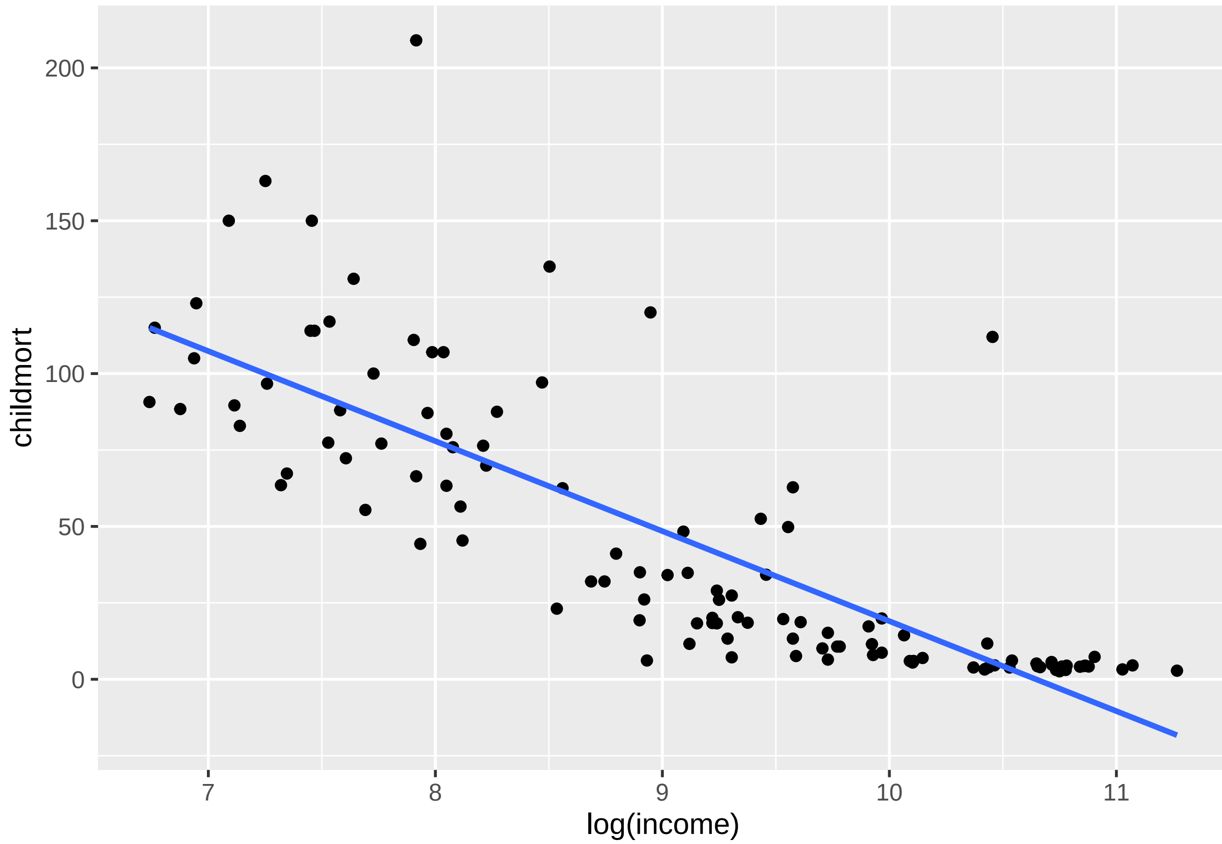
```
ggplot(data = all, mapping = aes(x = income, y = childmort)) +
  geom_point()+
  geom_smooth(method = 'lm', se = FALSE)
```

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




```
ggplot(data = all, mapping = aes(x = log(income), y = childmort)) +
  geom_point()+
  geom_smooth(method = 'lm', se = FALSE)
```

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



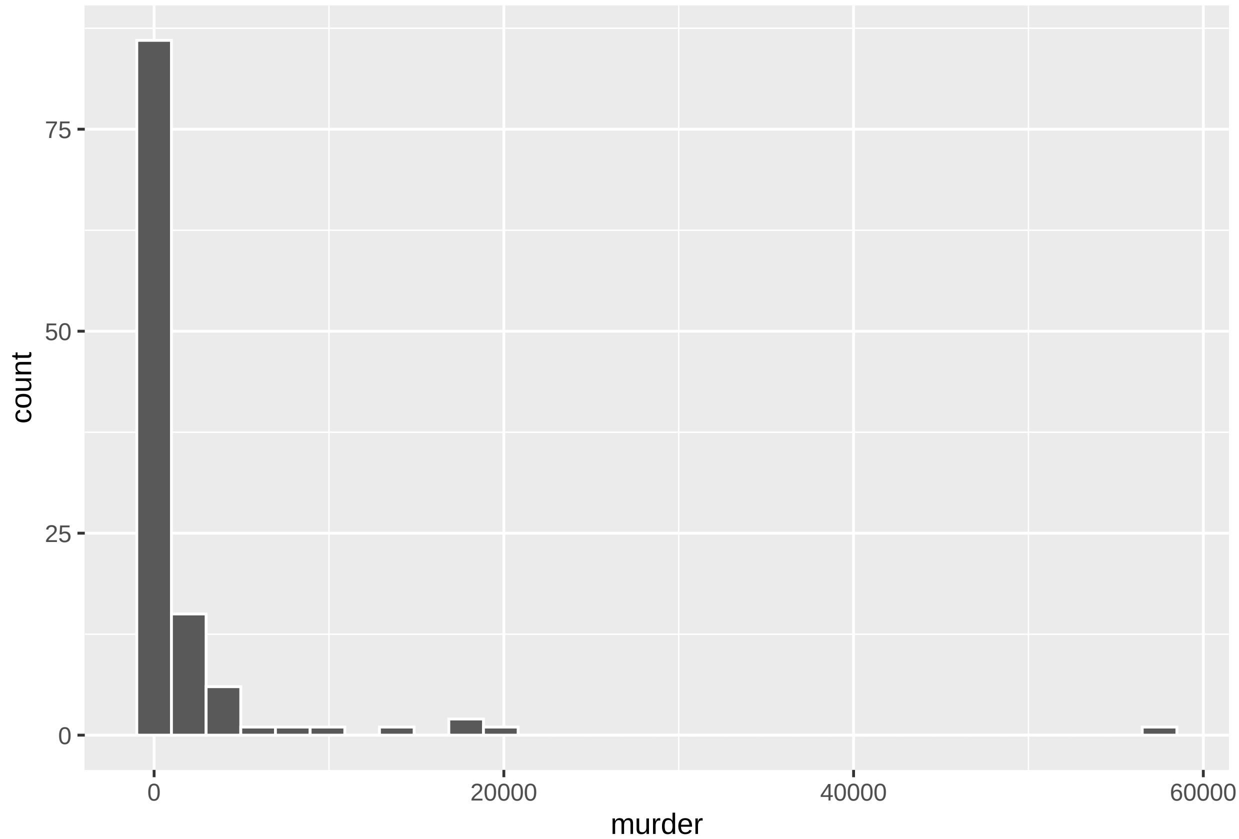

```
get_correlation(all, formula = childmort ~ income)
```

```
##          cor
## 1 -0.6359265
```

#Murder

```
ggplot(data=all, aes(x=murder)) + geom_histogram(color = "white")
```

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

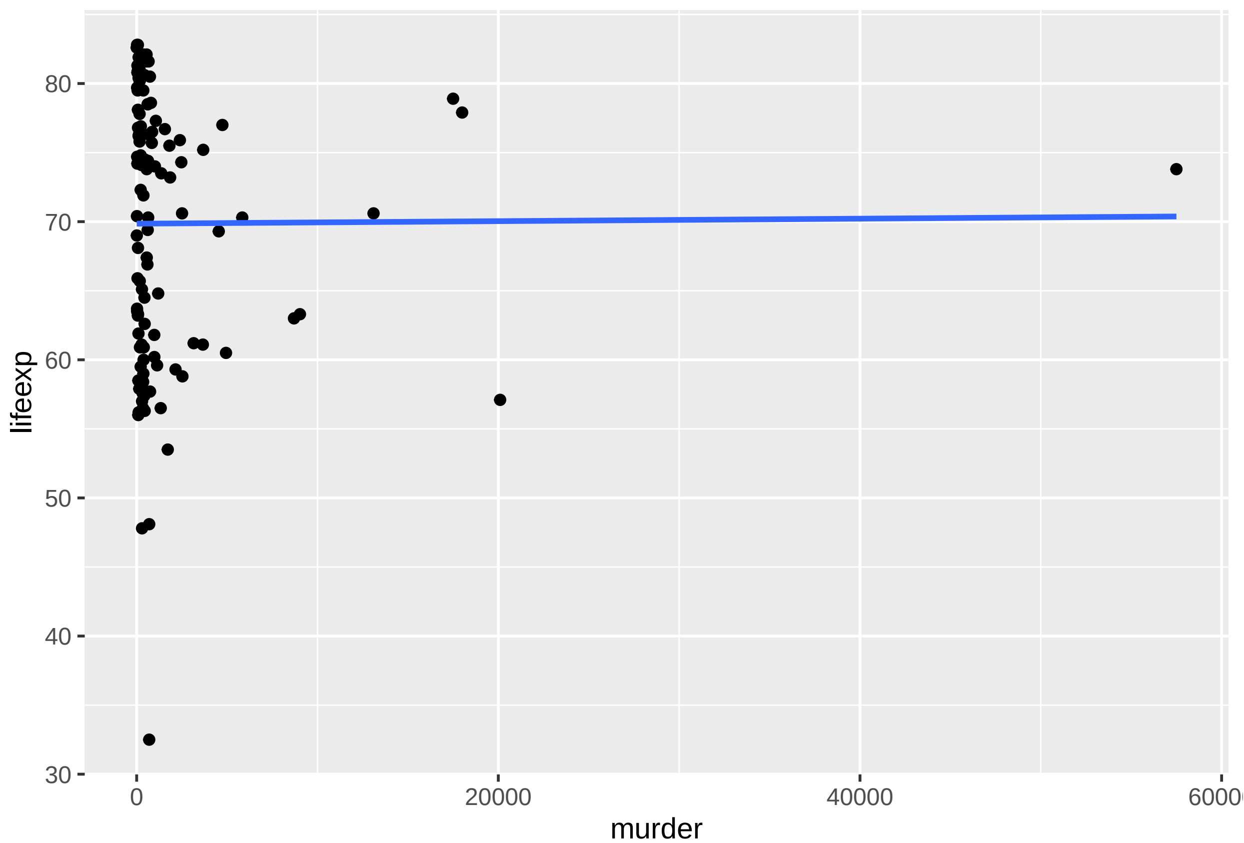



```
cor(all$lifeexp, all$murder)
```

```
## [1] 0.005740065
```

```
ggplot(data=all, aes(x=murder, y=lifeexp)) + geom_point() + geom_smooth(method = 'lm', se = FALSE)
```

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

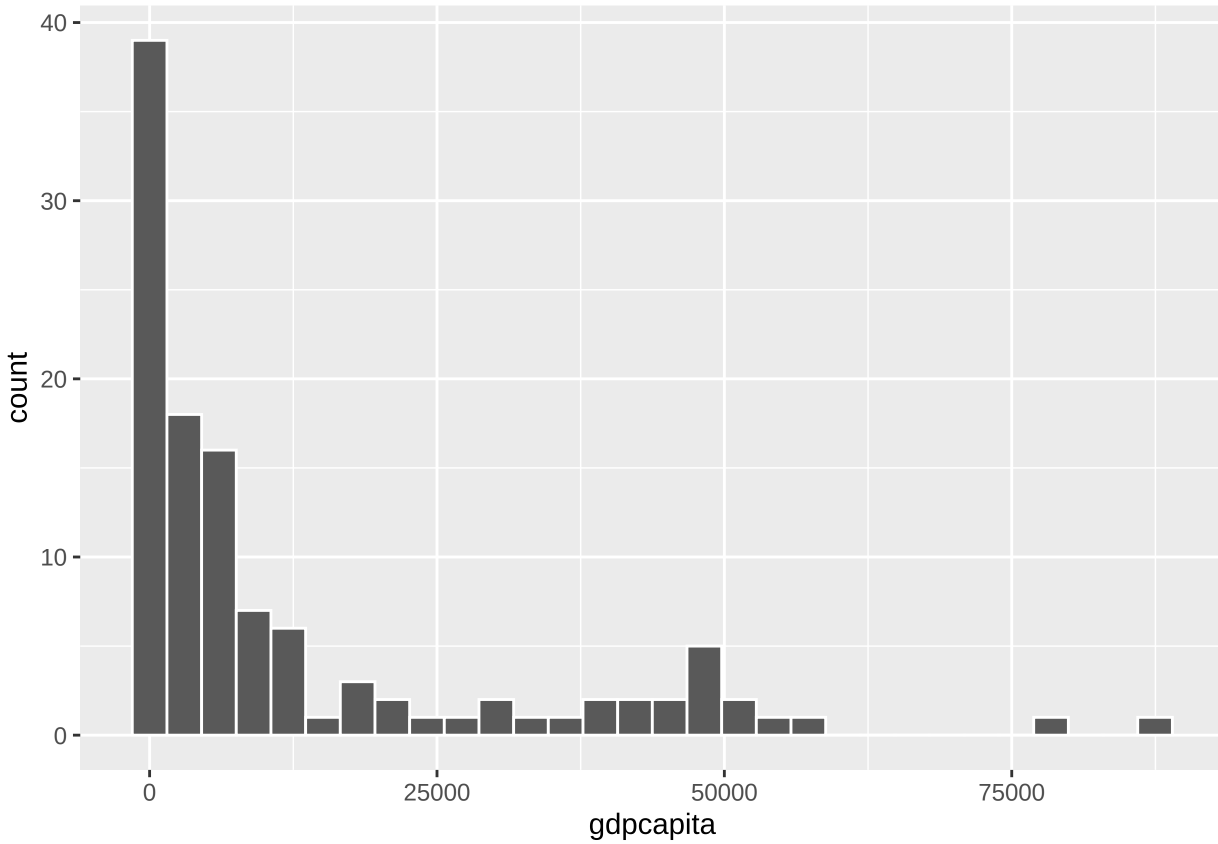


#Analysis: After observing the correlation coefficient between lifeexp and murder it was clear that the relationship between the two was very weak. As a result, when plotted on a scatterplot the projected line is almost a horizontal line. Although the murder variable does not have as big an impact on lifeexp, murder may be closely related to another variable to create a influential factor for lifeexp. Further analysis with its colinearity with other variables would be needed.

#GDPCapita

```
ggplot(data=all, aes(x=gdpcapita)) + geom_histogram(color = "white")
```

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

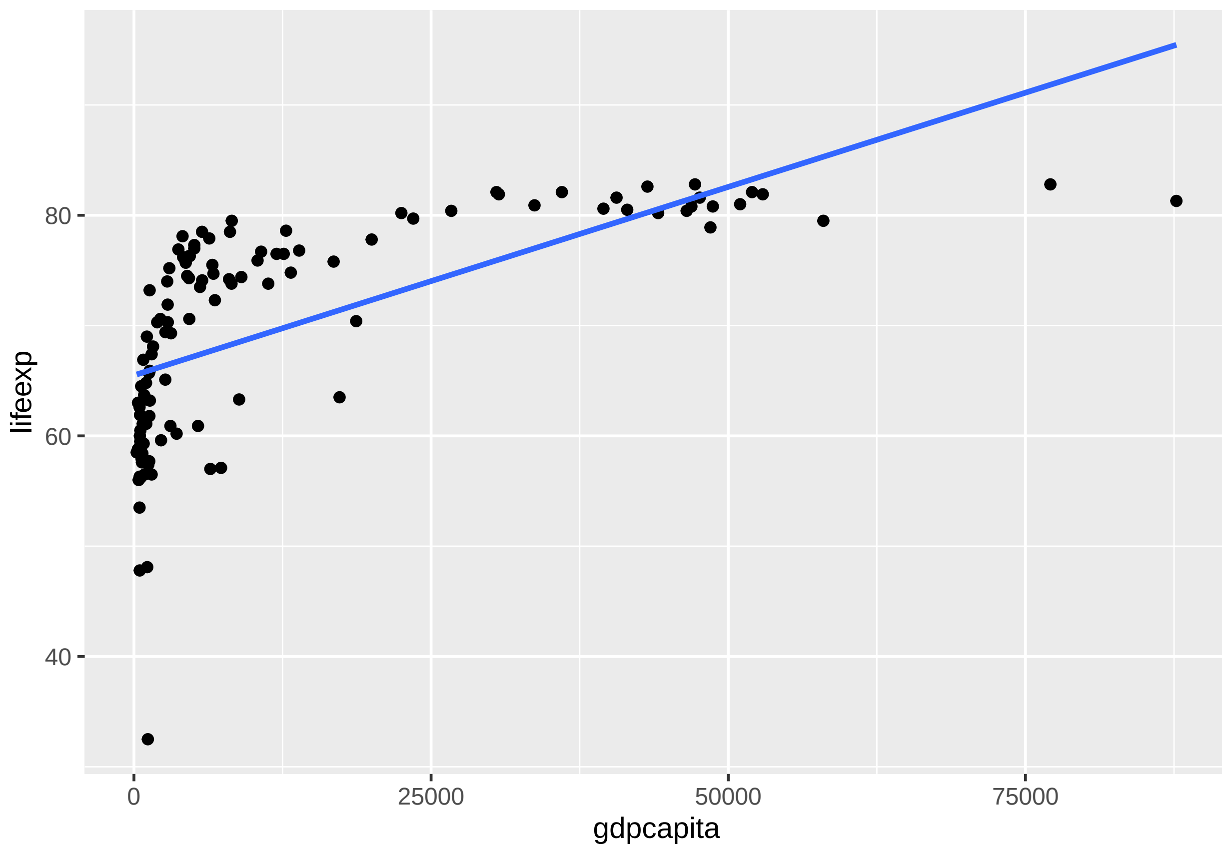



```
cor(all$lifeexp, all$gdpcapita)
```

```
## [1] 0.6381357
```

```
ggplot(data=all, aes(x=gdpcapita, y=lifeexp)) + geom_point() + geom_smooth(method = 'lm', se = FALSE)
```

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




```
model1 <- lm(data=all, lifeexp~gdpcapita)
get_regression_table(model1)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     65.5     0.865      75.7       0     63.8     67.2
## 2 gdpcapita      0       0           8.81      0      0        0
```

```
get_regression_summaries(model1)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.407         0.402  56.5  7.52  7.58      77.6       0     1   115
```

```
all2 <- all %>% mutate(lifeexp=log(lifeexp), gdpcapita=log(gdpcapita))
model2 <- lm(data=all2, lifeexp~gdpcapita)
get_regression_table(model2)
```

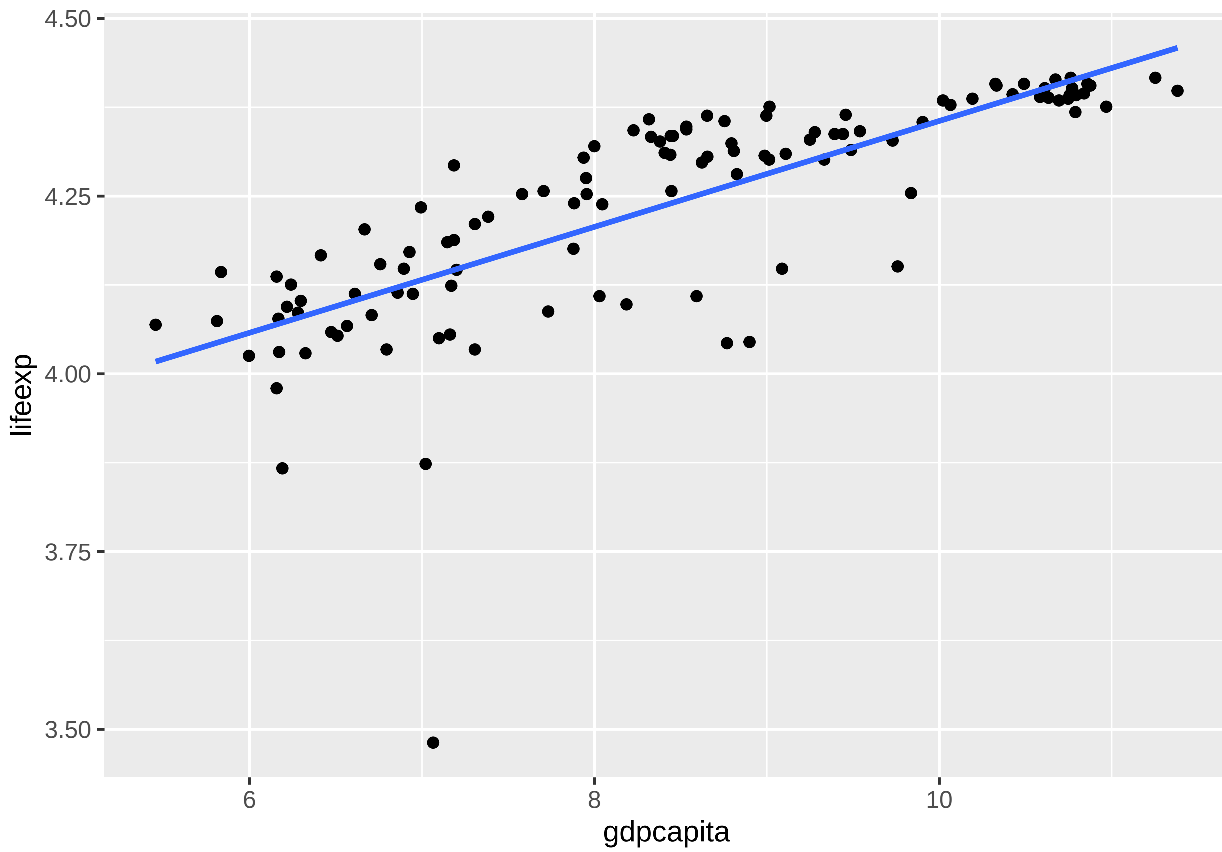
```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    3.61      0.05       71.5       0    3.51     3.71
## 2 gdpcapita    0.074     0.006      12.6       0    0.063    0.086
```

```
get_regression_summaries(model2)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared     mse   rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>   <dbl>  <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.584          0.58 0.00965 0.0983 0.099      159.       0     1   115
```

```
ggplot(data=all2, aes(x=gdpcapita, y=lifeexp)) + geom_point() + geom_smooth(method = 'lm', se = FALSE)
```

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

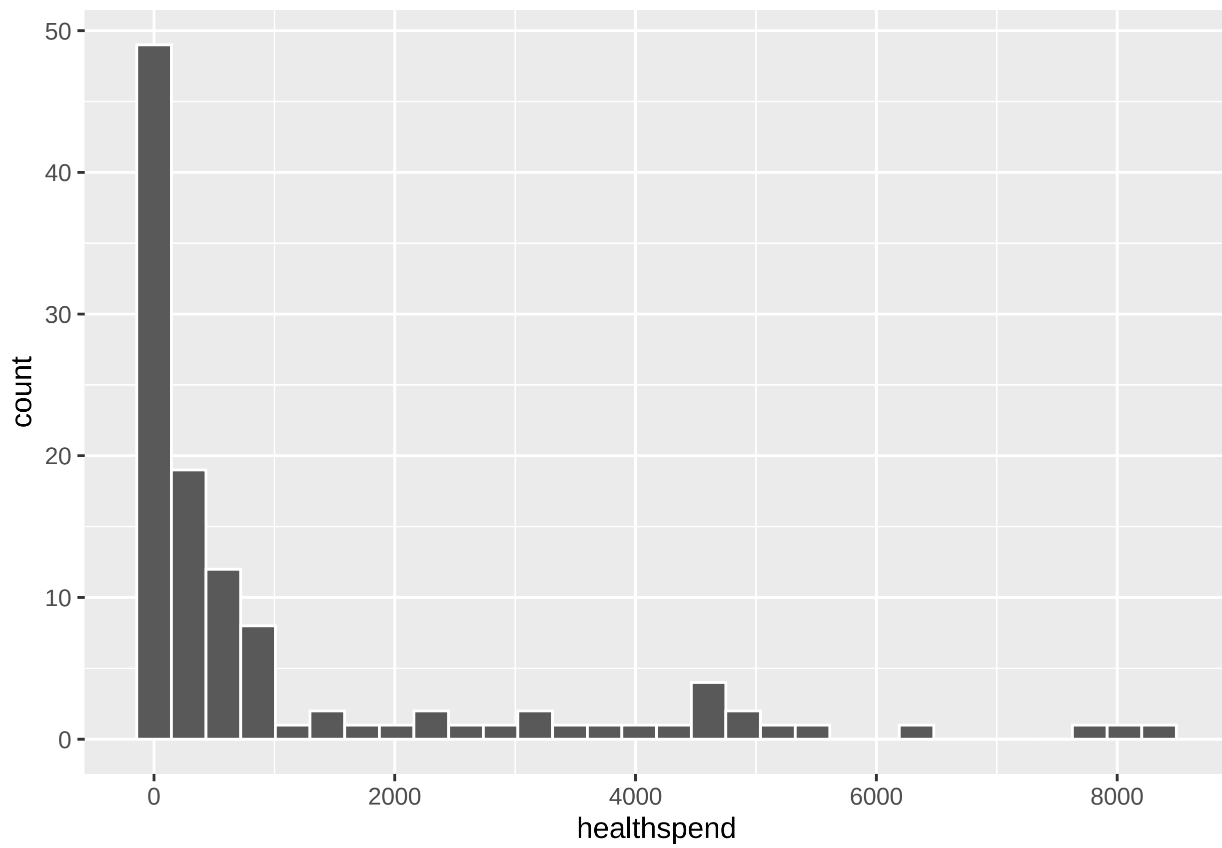


#Analysis: After taking a look at the relatively hight correlation coefficient between gdpcapita and lifeexp, I saw that plotting a scatterplot with a regression line of lifeexp on gdpcapita showed that the pattern of points followed a exponential curve rather than a linear line. So after taking a look at the log of lifeexp on log of gdpcapita, the scatterplot shows that the points more closely follow the regression line. The relationship between gdpcapita and lifeexp is a positive one that shows that as gdpcapita increases, so does lifeexp.

#HealthSpend

```
ggplot(data=all, aes(x=healthspend)) + geom_histogram(color = "white")
```

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

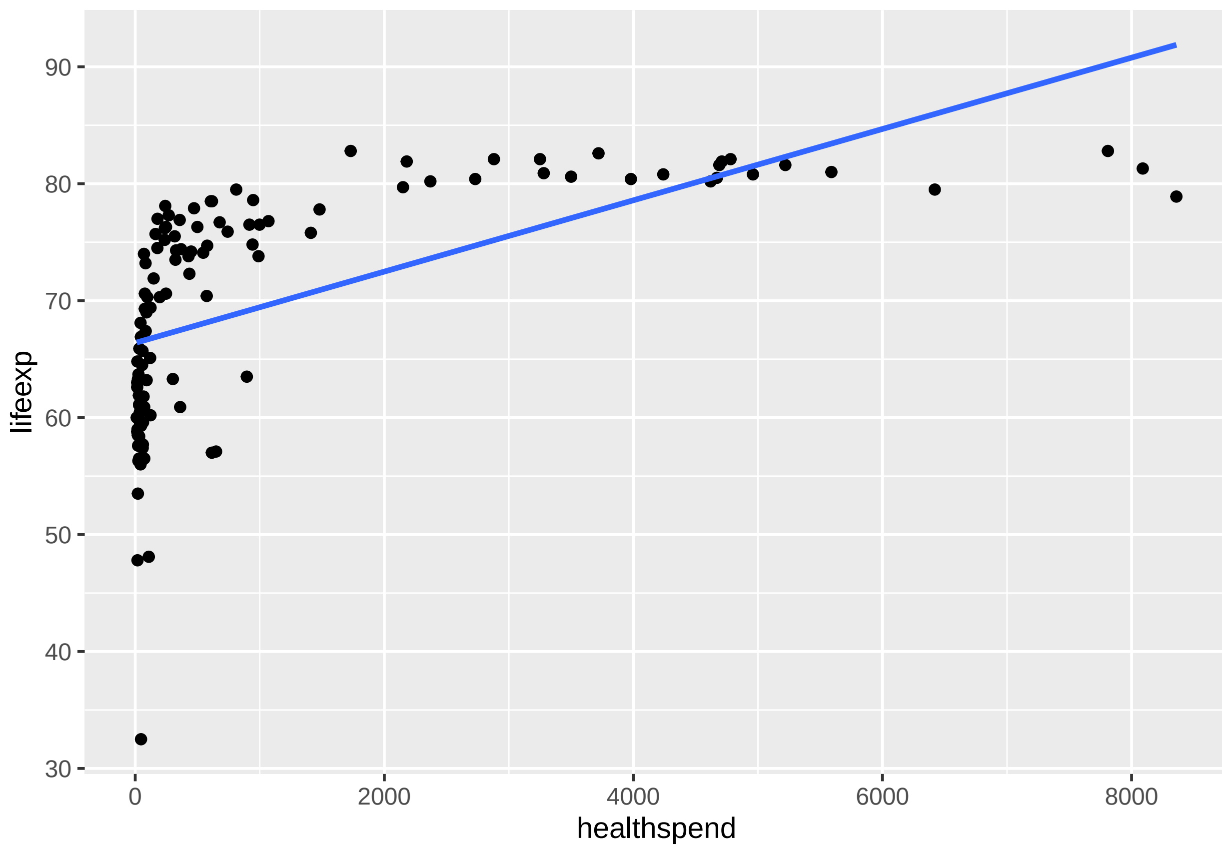



```
cor(all$lifeexp, all$healthspend)
```

```
## [1] 0.5916694
```

```
ggplot(data=all, aes(x=healthspend, y=lifeexp)) + geom_point() + geom_smooth(method = 'lm', se = FALSE)
```

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




```r
model3 <- lm(data=all, lifeexp~healthspend)
get_regression_table(model3)
```

```
## # A tibble: 2 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     66.4       0.865     76.8        0    64.7     68.1
## 2 healthspend    0.003     0          7.80       0     0.002    0.004
```

```r
get_regression_summaries(model3)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1      0.35         0.344  62.0  7.87  7.94      60.9       0     1   115
```

```r
all3 <- all %>% mutate(lifeexp=log(lifeexp), healthspend=log(healthspend))
model4 <- lm(data=all3, lifeexp~healthspend)
get_regression_table(model4)
```

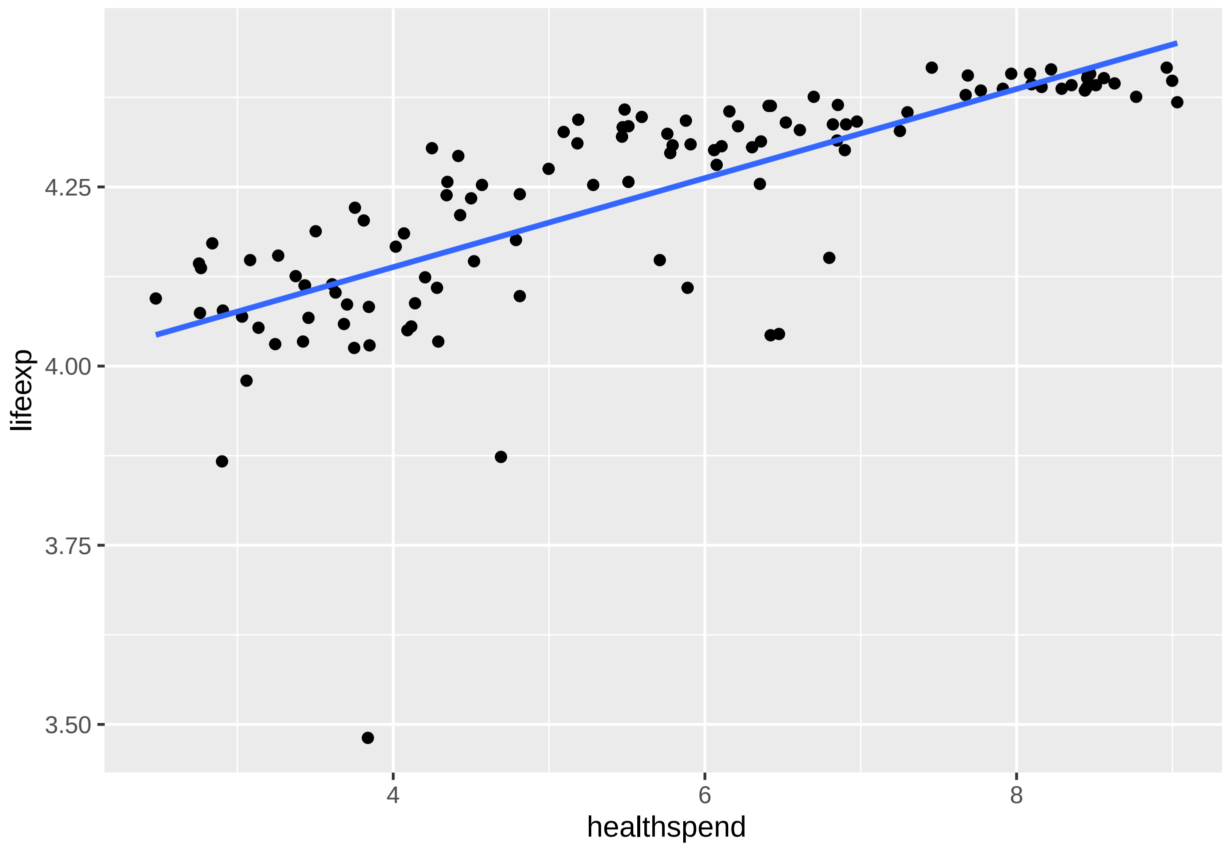
```
## # A tibble: 2 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      3.89      0.03      130.        0     3.83     3.95
## 2 healthspend    0.062     0.005      12.2       0     0.052    0.072
```

```r
get_regression_summaries(model4)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared    mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>  <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.569         0.565 0.0100 0.100 0.101      149.       0     1   115
```

```r
ggplot(data=all3, aes(x=healthspend, y=lifeexp)) + geom_point() + geom_smooth(method = 'lm', se = FALSE)
```

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



#Analysis: The correlation coefficient between lifeexp and healthspend is 0.592, which shows that there is a positive relationship between healthspend and lifeexp. Further examining this relationship, the scatterplot of the relationship shows a exponential curve of the data points. After applying the log() function to lifeexp and healthspend, we can see more clearly how the data points on the plot appear to be closer to the projected positive regression line.