Part 2 Code

Carlos Sathler

2/5/2019

This document is created by Carlos Santhler for Project 1. Please also check for the Rmd file from Li Ai. Both Rmd files are combined to answer the questions in Project 1.

Question 1

```
library(gapminder)
library(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(broom)
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

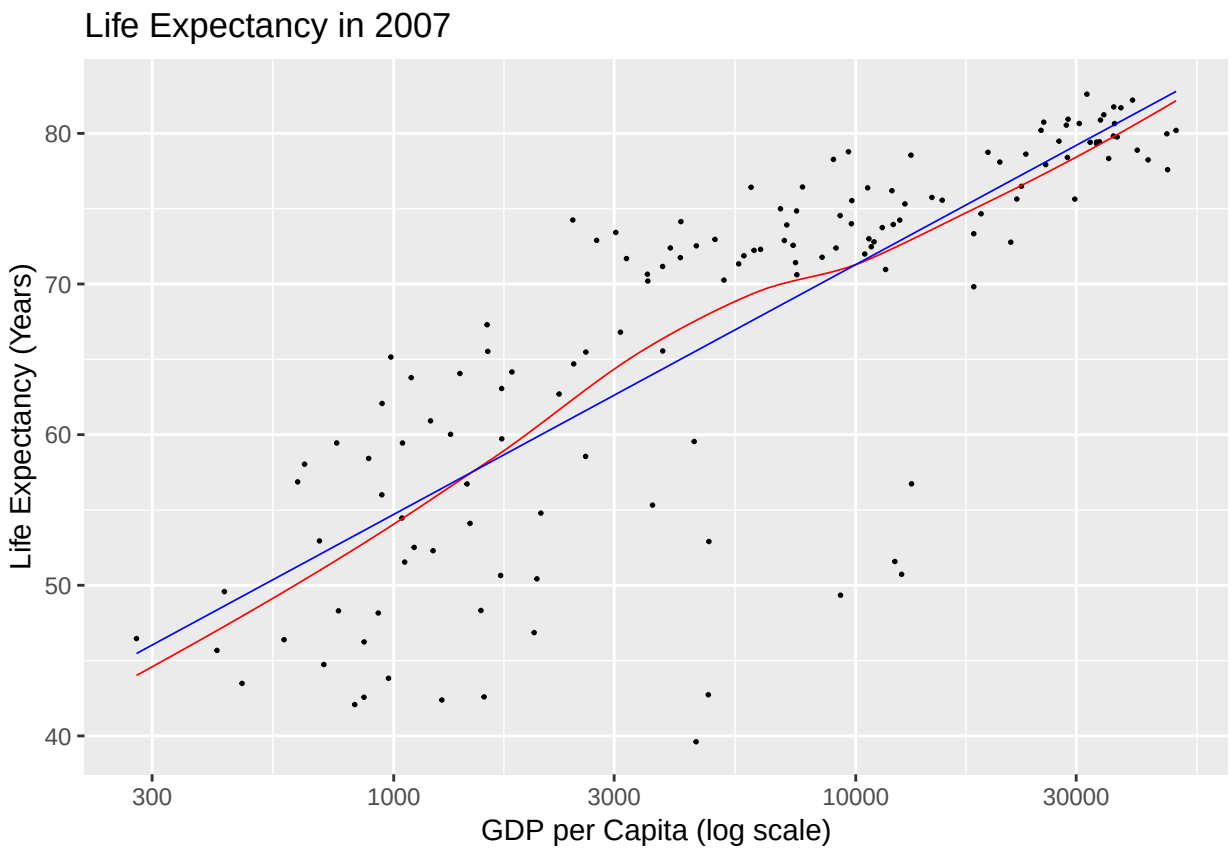
```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:ggpubr':
##
##     mutate
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
gapminder$pop = as.double(gapminder$pop)
gapminder$loggdpPercap = log(gapminder$gdpPercap)
gapminder.2007 = gapminder[gapminder$year == 2007,]
```

```
gg <- ggplot(gapminder.2007, aes(x=gdpPercap, y=lifeExp)) + geom_point(size=0.3) + scale_x_log10()
gg <- gg + geom_smooth(method="loess", se = F, size=0.3, color = 'red')
gg <- gg + geom_smooth(method="lm", se = F, size=0.3, color = 'blue')
gg <- gg + labs(title = "Life Expectancy in 2007", x = "GDP per Capita (log scale)", y = "Life Expectanc
gg
```




```
ggsave("Q1_1.png", device='pdf')
```

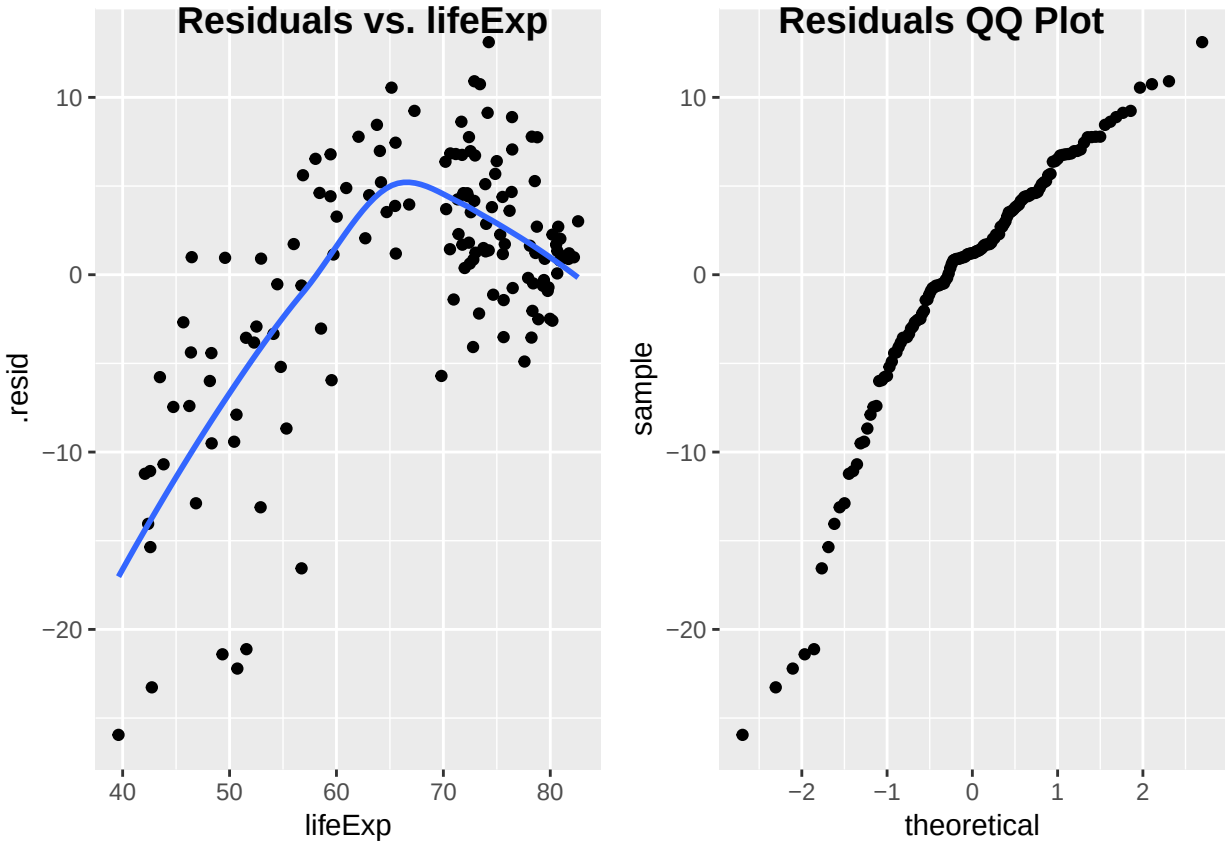
```
## Saving 6.5 x 4.5 in image
```

```
lifeExp.lm = lm(lifeExp ~ loggdpPercap, data = gapminder.2007)
summary(lifeExp.lm)
```

```
##
## Call:
## lm(formula = lifeExp ~ loggdpPercap, data = gapminder.2007)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.947  -2.661   1.215   4.469  13.115
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.9496     3.8577   1.283    0.202
## loggdpPercap   7.2028     0.4423  16.283   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.122 on 140 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.652
## F-statistic: 265.2 on 1 and 140 DF,  p-value: < 2.2e-16
```

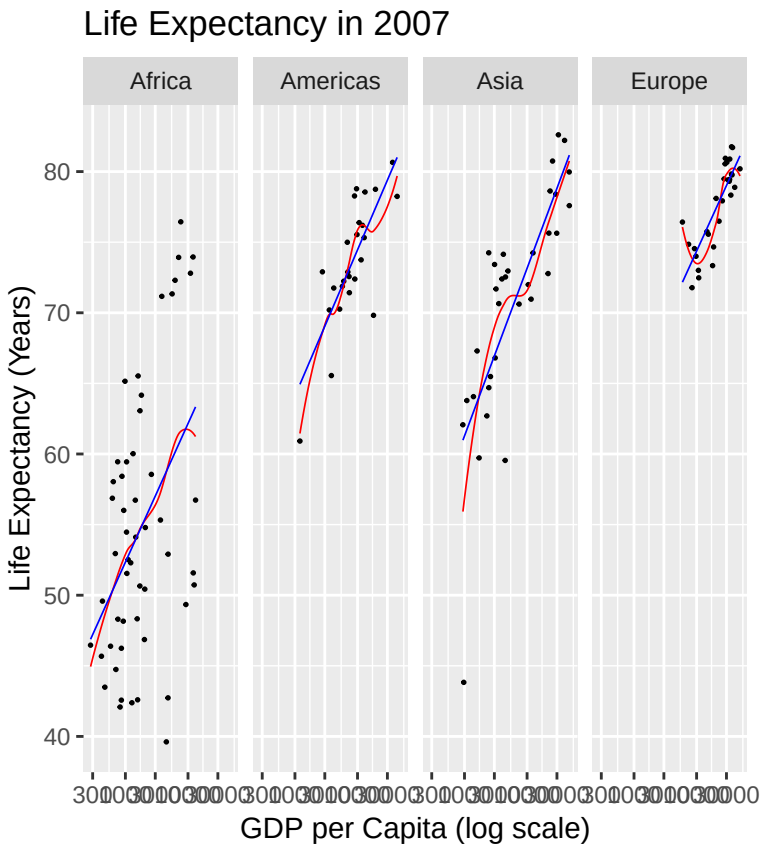
```
lifeExp.aug = augment(lifeExp.lm)
gg1 = ggplot(lifeExp.aug, aes(x = lifeExp, y = .resid)) + geom_point()
gg1 = gg1 + geom_smooth(method = "loess", se = FALSE)
gg2 = ggplot(lifeExp.lm, aes(sample = .resid)) + stat_qq()
figure <- ggarrange(gg1, gg2, labels = c("Residuals vs. lifeExp", "Residuals QQ Plot"),  nrow = 1, ncol
figure
```




```
ggsave("Q1_2.png", device='pdf')
```

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

```
gapminder.2007.further = gapminder.2007[gapminder.2007$continent!='Oceania',]
gg <- ggplot(gapminder.2007.further, aes(x=gdpPercap, y=lifeExp)) + geom_point(size=0.3) + scale_x_log1
gg <- gg + geom_smooth(method="loess", se = F, size=0.3, color = 'red')
gg <- gg + geom_smooth(method="lm", se = F, size=0.3, color = 'blue')
gg <- gg + labs(title = "Life Expectancy in 2007", x = "GDP per Capita (log scale)", y = "Life Expectan
gg <- gg + facet_wrap(~continent, 1)
gg
```




```
ggsave("Q1_3.png", device='pdf')
```

```
## Saving 4 x 4.5 in image
```

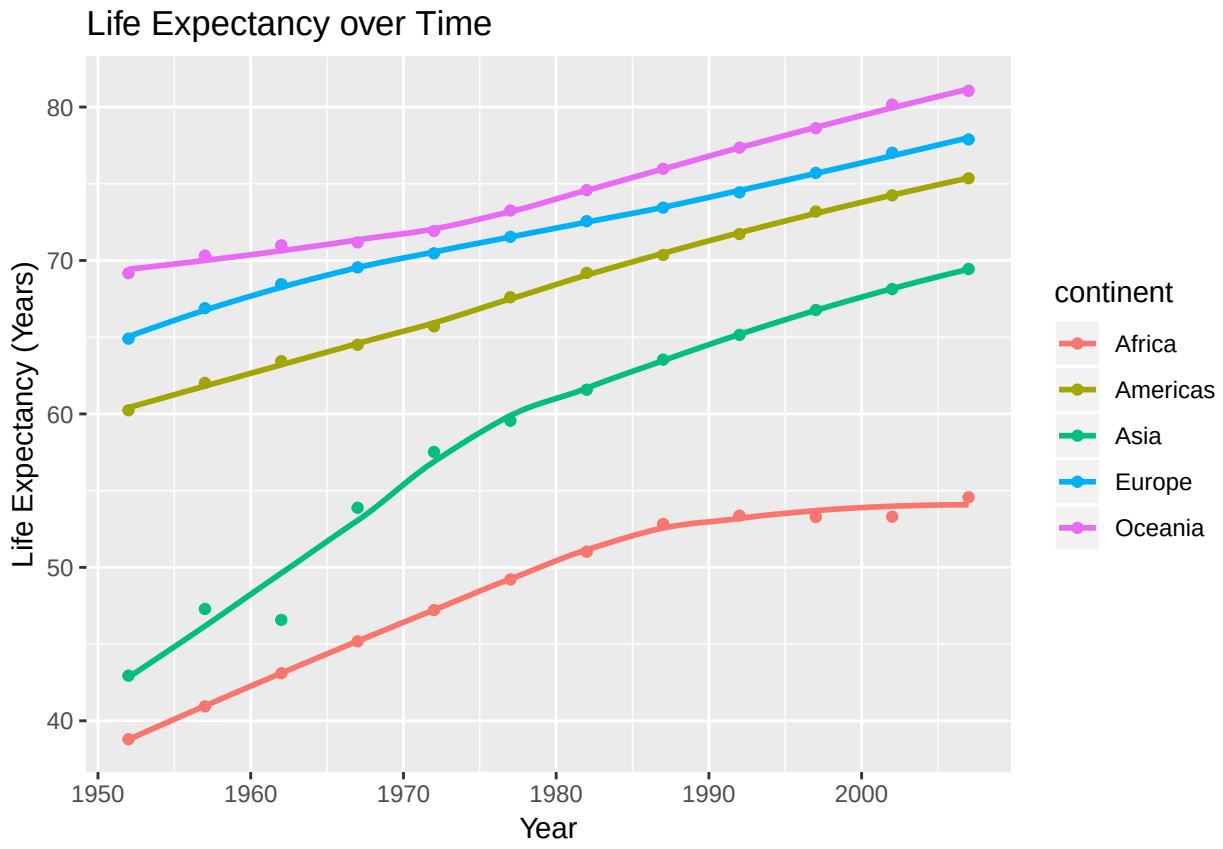
Question 2

```
# calculate lifeExp weighted average by continent
gapminder$lifeExp_w = gapminder$lifeExp * gapminder$pop
gapminder.continent.pop = aggregate(pop ~ continent+year, sum, data=gapminder)
names(gapminder.continent.pop) = c('continent', 'year', 'total_pop')
gapminder.continent.lifeExp = join(gapminder, gapminder.continent.pop)
```

```
## Joining by: continent, year
```

```
gapminder.continent.lifeExp$weighted_lifeExp = gapminder.continent.lifeExp$lifeExp_w /
                                               gapminder.continent.lifeExp$total_pop
gapminder.lifeExp.continent.year = aggregate(weighted_lifeExp ~ continent+year, sum,
                                             data=gapminder.continent.lifeExp)
```

```
gg = ggplot(gapminder.lifeExp.continent.year, aes(x = year, y = weighted_lifeExp, color=continent))
gg = gg + geom_point() + geom_smooth(method = "loess", se = F)
gg = gg + labs(title = "Life Expectancy over Time", x = "Year", y = "Life Expectancy (Years)")
gg
```

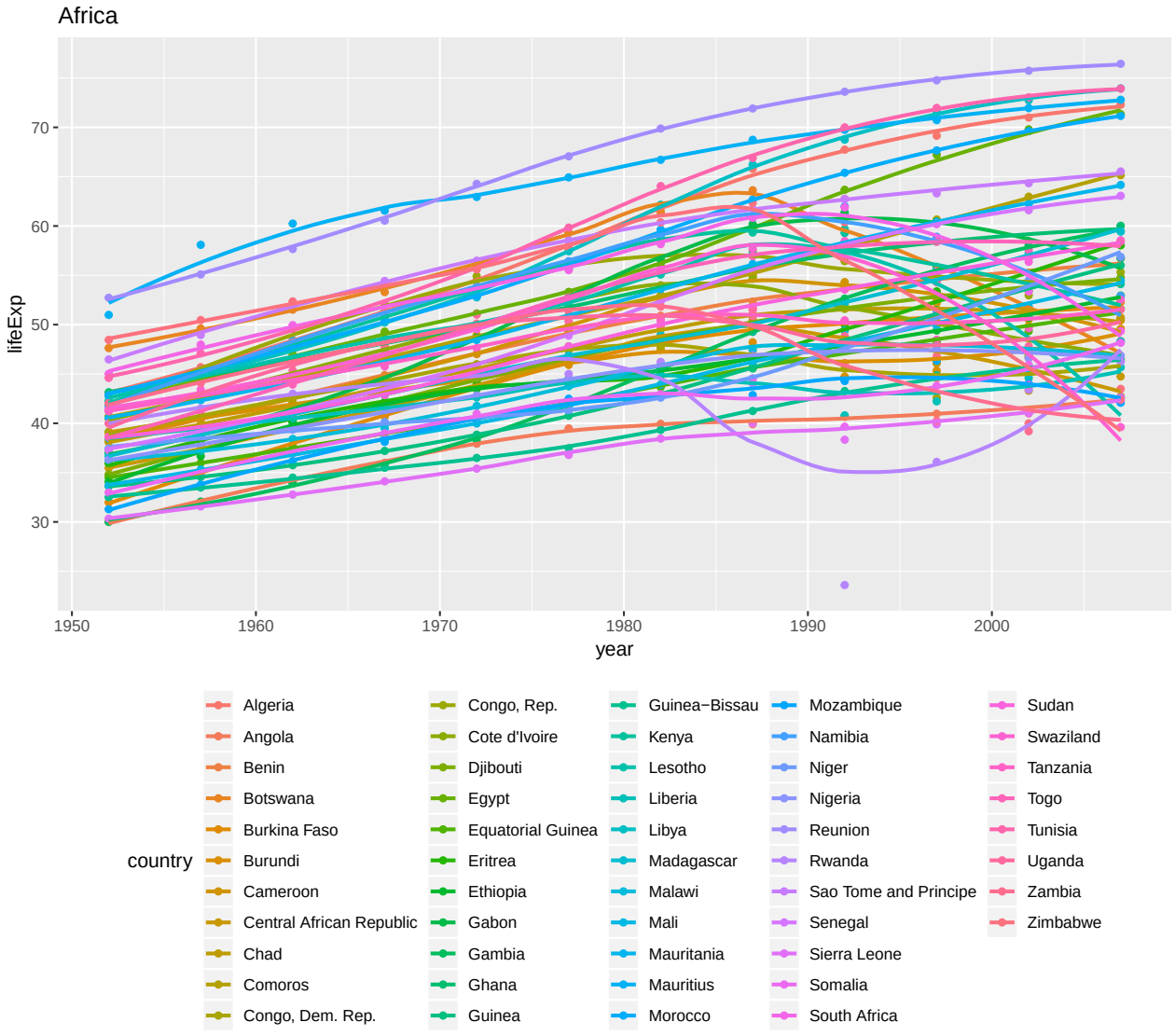



```
ggsave("Q1_4.png", device='pdf')
```

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

```
get_plot = function(continent) {
  cont.countries = as.character(unique(gapminder[gapminder$continent==continent,][,1])$country)
  gg = ggplot(subset(gapminder, country %in% cont.countries), aes(x = year, y = lifeExp,    color=countr
  gg = gg + geom_point() + geom_smooth(method = "loess", se = F)
  gg = gg + theme(legend.position="bottom",
```

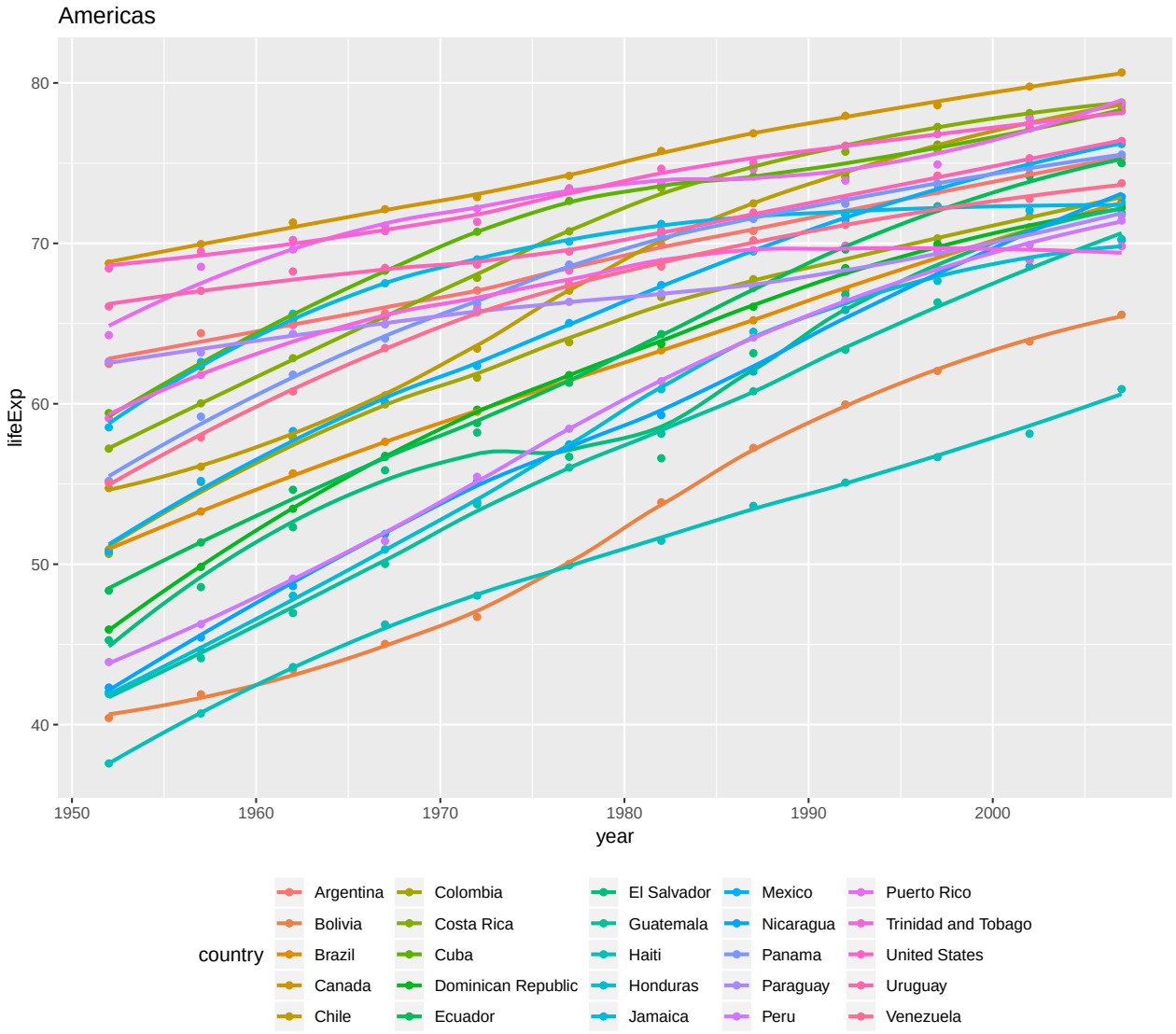
```
                plot.margin = margin(0,0,0,0, "cm"))
  return(gg)
}
get_plot("Africa")
```



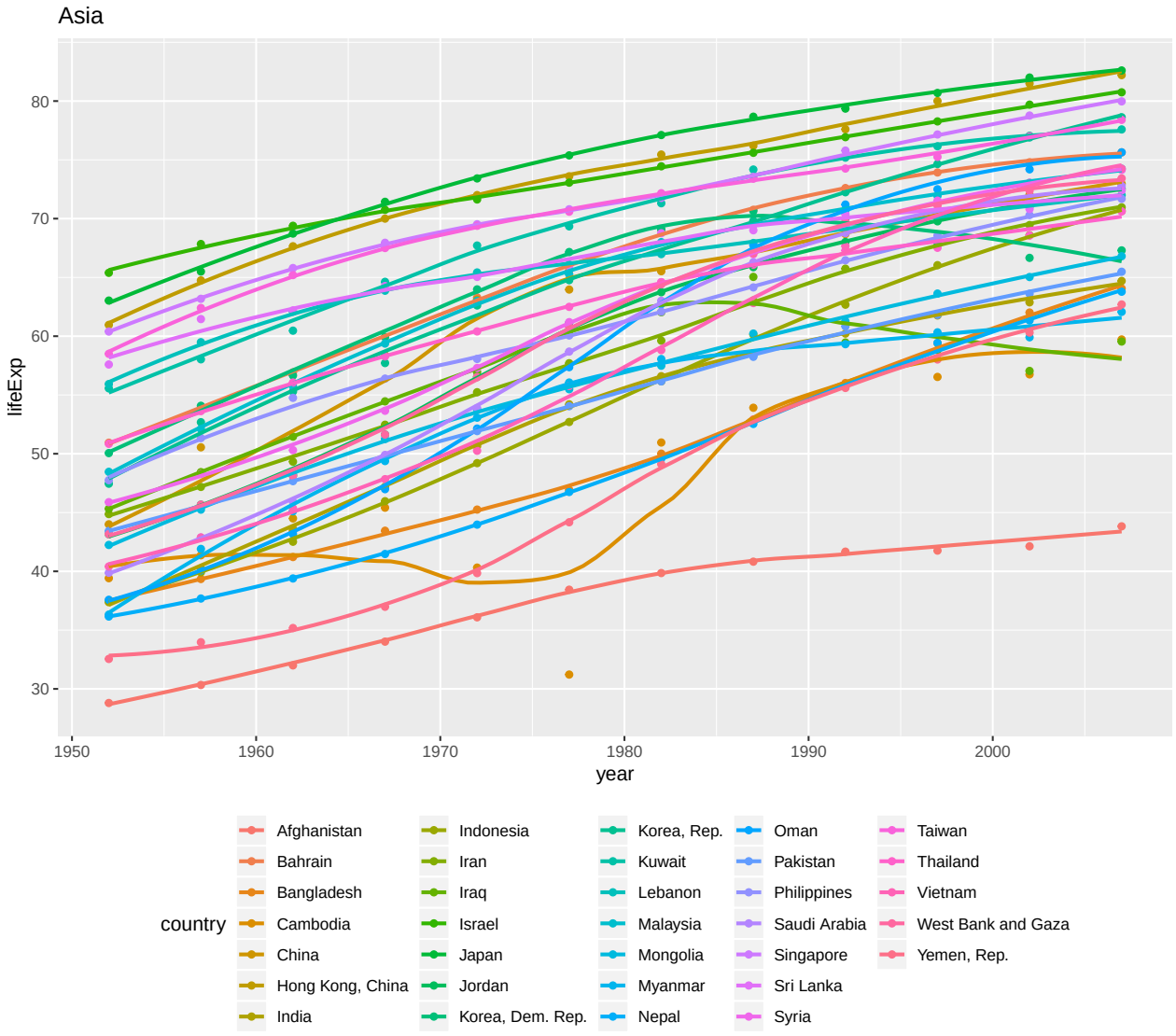

```
ggsave("Q1_5.png", device='pdf')
```

```
## Saving 9 x 8 in image
```

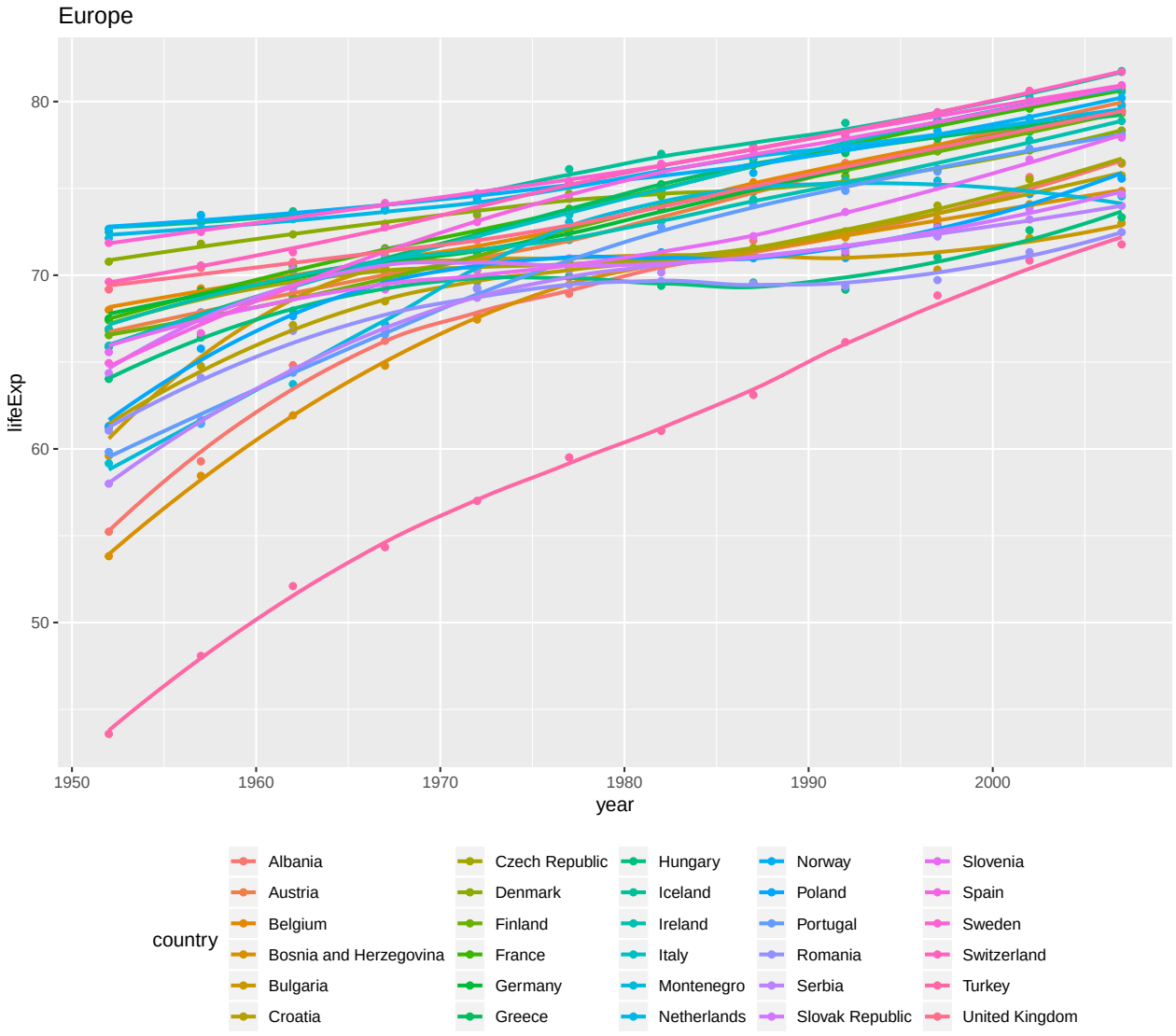
```
get_plot("Americas")
```



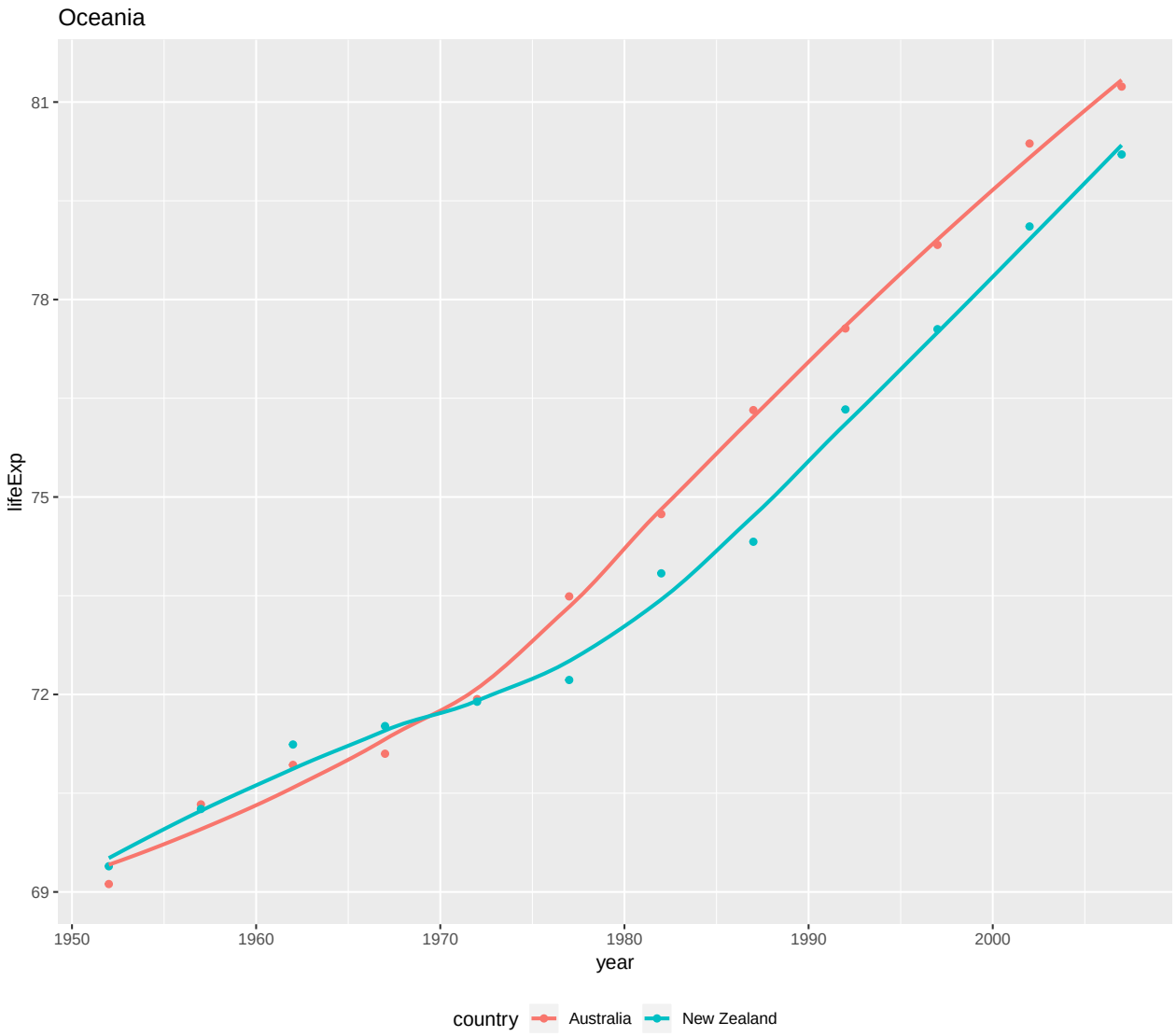

```
get_plot("Asia")
```



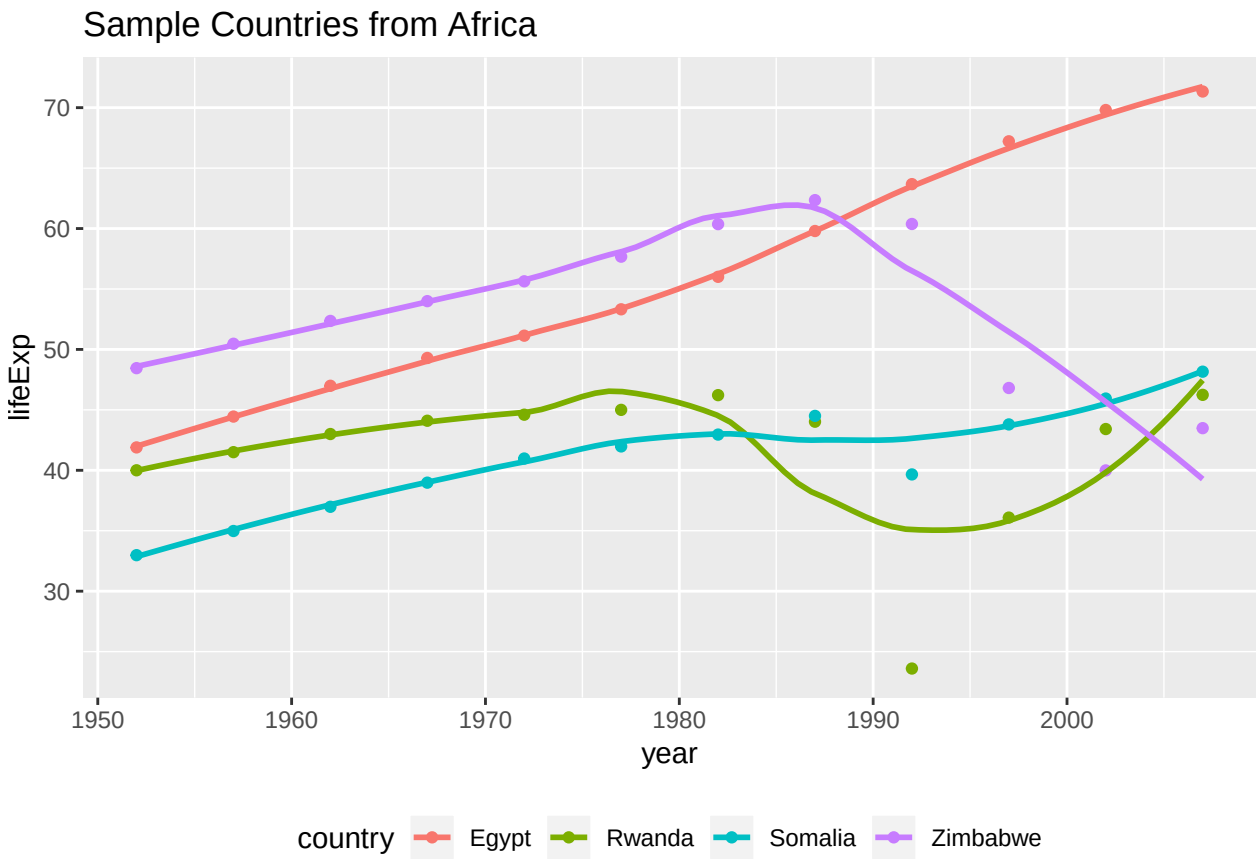

```
get_plot("Europe")
```




```
get_plot("Oceania")
```



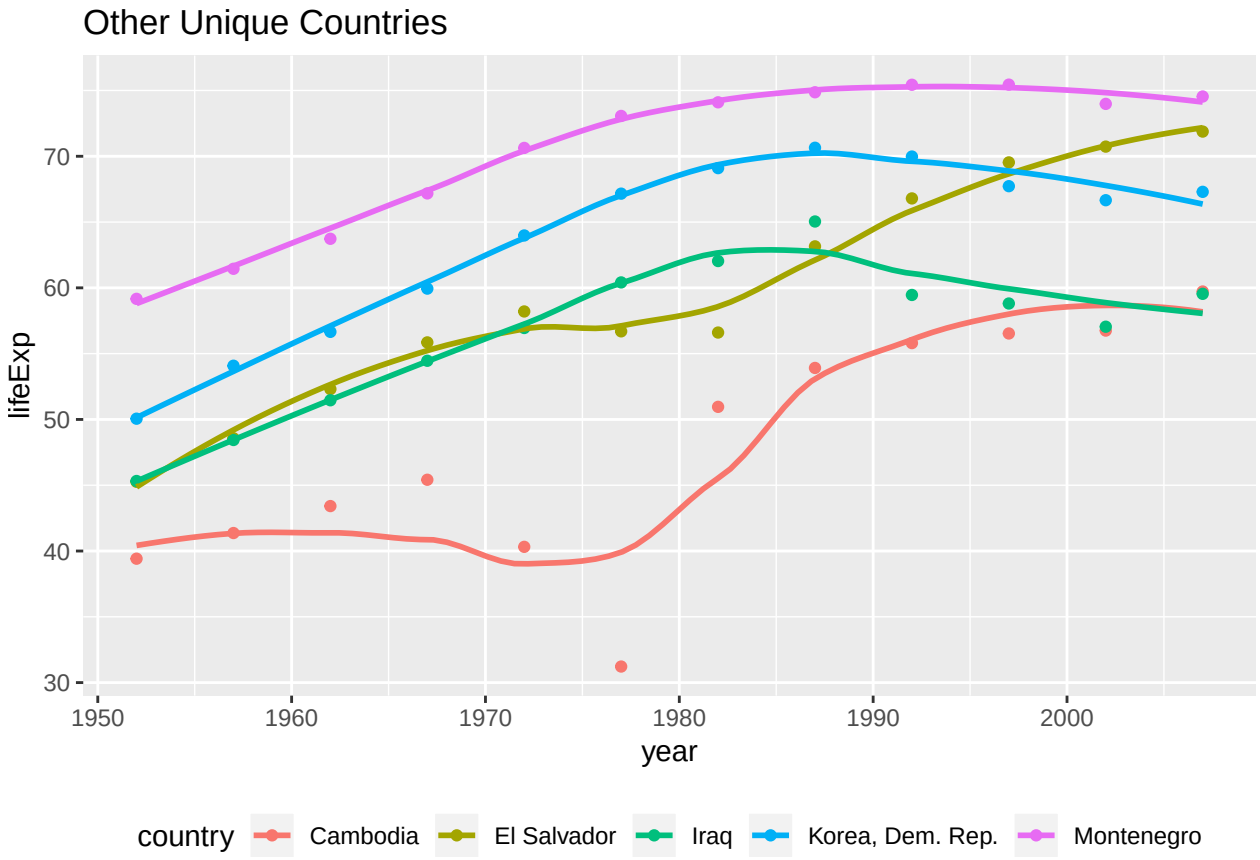

```
cont.countries = c('Zimbabwe','Rwanda','Somalia','Egypt')
gg = ggplot(subset(gapminder, country %in% cont.countries), aes(x = year, y = lifeExp,     color=country)
gg = gg + geom_point() + geom_smooth(method = "loess", se = F)
gg = gg + theme(legend.position="bottom", plot.margin = margin(0,0,0,0, "cm"))
gg
```




```
ggsave("Q1_6.png", device='pdf')
```

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

```
cont.countries = c('Cambodia','Iraq','Korea, Dem. Rep.', 'Montenegro','El Salvador')
gg = ggplot(subset(gapminder, country %in% cont.countries), aes(x = year, y = lifeExp,     color=country)
gg = gg + geom_point() + geom_smooth(method = "loess", se = F)
gg = gg + theme(legend.position="bottom", plot.margin = margin(0,0,0,0, "cm"))
gg
```


Other Unique Countries
lifeExp
70
60
50
40
30
1950
1960
1970
1980
1990
2000
year
country
Cambodia
El Salvador
Iraq
Korea, Dem. Rep.
Montenegro

Question 3

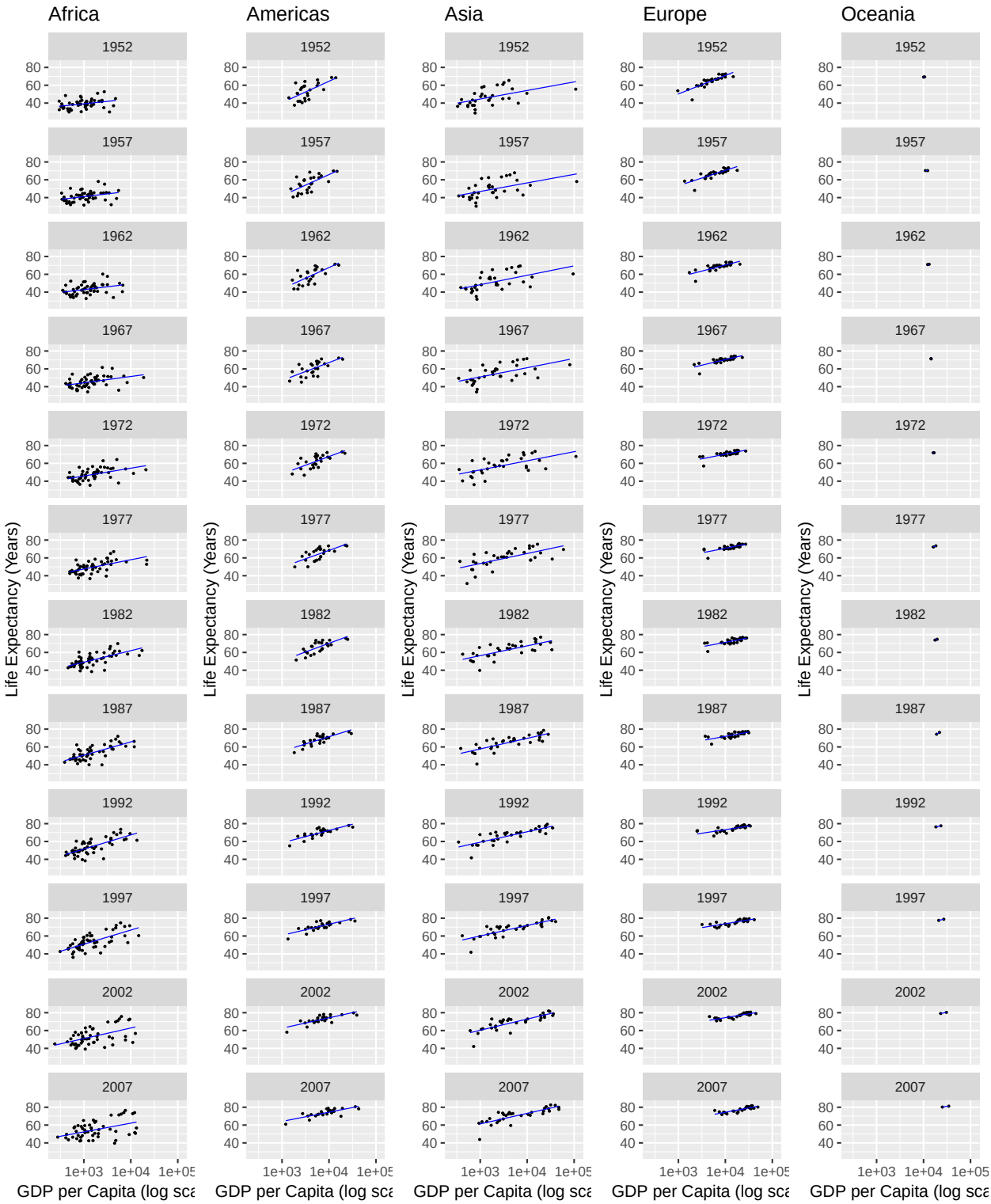
```
get_plot = function(continent, nrows=3) {
  #gg = ggplot(subset(gapminder.lifeExp.country.year, country %in% cont.countries), aes(x = year, y = l
  #gg = gg + geom_point() + geom_smooth(method = "loess", se = F)
  #gg = gg + theme(legend.position="bottom", plot.margin = margin(0,0,0,0, "cm"))

  cont.countries = as.character(unique(gapminder[gapminder$continent==continent,][,1])$country)
  gg <- ggplot(subset(gapminder, country %in% cont.countries), aes(x=gdpPercap, y=lifeExp))
  gg <- gg + geom_point(size=0.3) + scale_x_log10(limits=c(240, 114000)) + ylim(25, 85)
  #
  gg <- gg + geom_smooth(method="lm", se = F, size=0.3, color = 'blue')
  #}
  gg <- gg + labs(title = continent, x = "GDP per Capita (log scale)", y = "Life Expectancy (Years)")
  gg <- gg + facet_wrap(~year, nrows)
  return(gg)
}
gg1 = get_plot("Africa",12)
gg2 = get_plot("Americas",12)
gg3 = get_plot("Asia",12)
gg4 = get_plot("Europe",12)
gg5 = get_plot("Oceania",12)
```

```
figure <- ggarrange(gg1, gg2, gg3, gg4, gg5, nrow = 1, ncol = 5)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
figure
```




```
ggsave("Q1_7.png", device='pdf')
```

```
## Saving 9 x 11 in image
```

```
get_cors = function(continent) {
  gapminder.continent = gapminder[gapminder$continent==continent,c(3,4,6)]
```

```
  print(paste(continent, 'Correlation Matrix:', sep=' '))
  x = cor(gapminder.continent)
  print(x)
}
cor_Africa = get_cors('Africa')
```

```
## [1] "Africa Correlation Matrix:"
##                 year   lifeExp gdpPercap
## year      1.0000000 0.5465842 0.1600793
## lifeExp   0.5465842 1.0000000 0.4256076
## gdpPercap 0.1600793 0.4256076 1.0000000
```

```
cor_Americas = get_cors('Americas')
```

```
## [1] "Americas Correlation Matrix:"
##                 year   lifeExp gdpPercap
## year      1.0000000 0.6801813 0.3063167
## lifeExp   0.6801813 1.0000000 0.5583655
## gdpPercap 0.3063167 0.5583655 1.0000000
```

```
cor_Asia = get_cors('Asia')
```

```
## [1] "Asia Correlation Matrix:"
##                 year   lifeExp gdpPercap
## year      1.0000000 0.6600265 0.1372517
## lifeExp   0.6600265 1.0000000 0.3820476
## gdpPercap 0.1372517 0.3820476 1.0000000
```

```
cor_Europe = get_cors('Europe')
```

```
## [1] "Europe Correlation Matrix:"
##                 year   lifeExp gdpPercap
## year      1.0000000 0.7060212 0.6087531
## lifeExp   0.7060212 1.0000000 0.7807831
## gdpPercap 0.6087531 0.7807831 1.0000000
```

```
cor_Oceania = get_cors('Oceania')
```

```
## [1] "Oceania Correlation Matrix:"
##                 year   lifeExp gdpPercap
## year      1.0000000 0.9767640 0.9255503
## lifeExp   0.9767640 1.0000000 0.9564738
## gdpPercap 0.9255503 0.9564738 1.0000000
```

```
# get correlation for all continents
print('All Continents')
```

```
## [1] "All Continents"
```

```
cor_All = cor(gapminder[,c(3,4,6)])
print(cor_All)
```

```
##                 year   lifeExp gdpPercap
## year      1.0000000 0.4356112 0.2273181
## lifeExp   0.4356112 1.0000000 0.5837062
## gdpPercap 0.2273181 0.5837062 1.0000000
```

```
# gather correlation between lifeExp and year, and lifeExp and gdpPercap, for each continent
correlations = rbind(cor_Africa[2,c(1,3)], cor_Americas[2,c(1,3)], cor_Asia[2,c(1,3)],
                     cor_Europe[2,c(1,3)], cor_Oceania[2,c(1,3)], cor_All[2,c(1,3)])
# gather corrleation between year and gdpPercap, for each continent
correlations2 = rbind(cor_Africa[1,3], cor_Americas[1,3], cor_Asia[1,3], cor_Europe[1,3],
                      cor_Oceania[1,3], cor_All[1,3])
correlations = cbind(correlations, correlations2)
correlations = data.frame(correlations)
correlations = cbind(correlations, c('Africa','Americas','Asia','Europe','Oceania','All Continents'))
names(correlations) = c('lifeExp_year','lifeExp_gdpPercap','year_gdpPercap','continent')
correlations = correlations[,c(4,1,2,3)]
correlations
```

```
##        continent lifeExp_year lifeExp_gdpPercap year_gdpPercap
## 1         Africa    0.5465842         0.4256076      0.1600793
## 2       Americas    0.6801813         0.5583655      0.3063167
## 3           Asia    0.6600265         0.3820476      0.1372517
## 4         Europe    0.7060212         0.7807831      0.6087531
## 5        Oceania    0.9767640         0.9564738      0.9255503
## 6 All Continents    0.4356112         0.5837062      0.2273181
```

```
cont = 'Africa'
print(cont)
```

```
## [1] "Africa"
```

```
summary(lm(lifeExp ~ loggdpPercap+year+I(year^2), data=subset(gapminder, continent==cont)))
```

```
##
## Call:
## lm(formula = lifeExp ~ loggdpPercap + year + I(year^2), data = subset(gapminder,
##     continent == cont))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8905  -3.7923  -0.0136   3.6544  14.9709
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.639e+04  3.816e+03  -4.295 2.03e-05 ***
## loggdpPercap  4.753e+00  3.033e-01  15.670  < 2e-16 ***
## year          1.633e+01  3.856e+00   4.234 2.64e-05 ***
## I(year^2)    -4.061e-03  9.738e-04  -4.170 3.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.397 on 620 degrees of freedom
## Multiple R-squared:  0.5136, Adjusted R-squared:  0.5113
## F-statistic: 218.3 on 3 and 620 DF,  p-value: < 2.2e-16
```

```
cont = 'Americas'
print(cont)
```

```
## [1] "Americas"
```

```
summary(lm(lifeExp ~ loggdpPercap+I(loggdpPercap^2)+year, data=subset(gapminder, continent==cont)))
```

```
##
## Call:
## lm(formula = lifeExp ~ loggdpPercap + I(loggdpPercap^2) + year,
##     data = subset(gapminder, continent == cont))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2954  -3.1572  -0.0011   3.5209  14.4587
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -653.66771   45.88192 -14.247  < 2e-16 ***
## loggdpPercap        38.90398    7.34017   5.300 2.27e-07 ***
## I(loggdpPercap^2)   -1.80122    0.41763  -4.313 2.20e-05 ***
## year                 0.26152    0.01761  14.854  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.905 on 296 degrees of freedom
## Multiple R-squared:  0.7273, Adjusted R-squared:  0.7245
## F-statistic: 263.1 on 3 and 296 DF,  p-value: < 2.2e-16
```

```
cont = 'Asia'
print(cont)
```

```
## [1] "Asia"
```

```
summary(lm(lifeExp ~ loggdpPercap+year, data=subset(gapminder, continent==cont)))
```

```
##
## Call:
## lm(formula = lifeExp ~ loggdpPercap + year, data = subset(gapminder,
##     continent == cont))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7152  -3.9880   0.0174   4.3554  14.2066
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -642.21972   39.76532  -16.15   <2e-16 ***
## loggdpPercap    4.82259    0.26975   17.88   <2e-16 ***
## year            0.33511    0.02042   16.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.636 on 393 degrees of freedom
## Multiple R-squared:  0.6888, Adjusted R-squared:  0.6872
## F-statistic: 434.8 on 2 and 393 DF,  p-value: < 2.2e-16
```

```
cont = 'Europe'
print(cont)
```

```
## [1] "Europe"
```

```
summary(lm(lifeExp ~ loggdpPercap+year, data=subset(gapminder, continent==cont)))
```

```
##
## Call:
## lm(formula = lifeExp ~ loggdpPercap + year, data = subset(gapminder,
##     continent == cont))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.8827  -1.2324  -0.0545   1.3309   6.4445
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.720e+02  1.730e+01  -9.941   <2e-16 ***
## loggdpPercap  4.940e+00  2.189e-01  22.566   <2e-16 ***
## year          9.988e-02  9.299e-03  10.741   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.477 on 357 degrees of freedom
## Multiple R-squared:  0.7933, Adjusted R-squared:  0.7921
## F-statistic: 685.1 on 2 and 357 DF,  p-value: < 2.2e-16
```

```
cont = 'Oceania'
print(cont)
```

```
## [1] "Oceania"
```

```
summary(lm(lifeExp ~ year, data=subset(gapminder, continent==cont)))
```

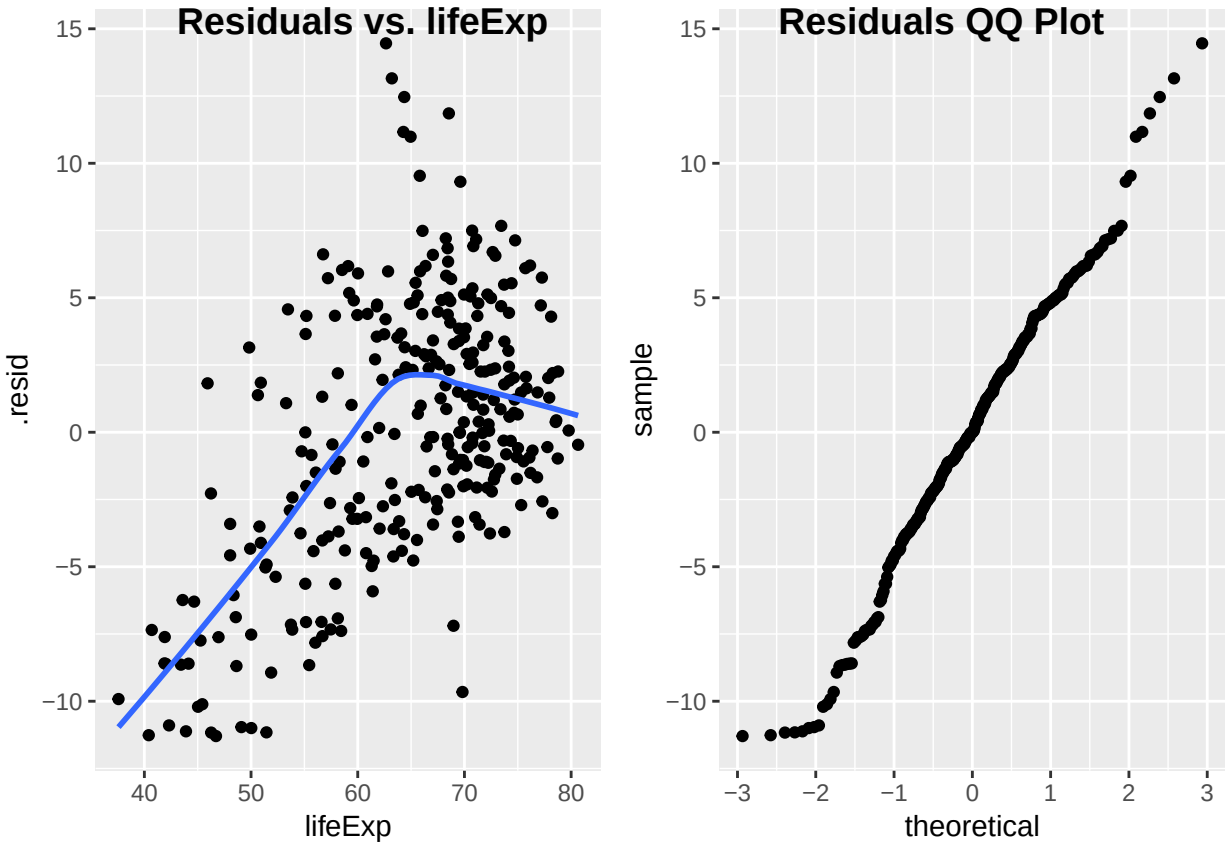
```
##
## Call:
## lm(formula = lifeExp ~ year, data = subset(gapminder, continent ==
##     cont))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58325 -0.60451  0.07398  0.62027  1.31266
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.419e+02  1.947e+01  -17.56 1.99e-14 ***
## year         2.103e-01  9.836e-03   21.38 3.30e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8317 on 22 degrees of freedom
## Multiple R-squared:  0.9541, Adjusted R-squared:  0.952
## F-statistic:   457 on 1 and 22 DF,  p-value: 3.299e-16
```

```
cont = 'All Continents'
summary(lm(lifeExp ~ loggdpPercap+I(loggdpPercap^2)+year+I(year^2), data=gapminder))
```

```
##
## Call:
## lm(formula = lifeExp ~ loggdpPercap + I(loggdpPercap^2) + year +
##     I(year^2), data = gapminder)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.688  -3.604   1.007   4.306  18.379
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.359e+03  2.453e+03  -2.593   0.0096 **
## loggdpPercap       1.829e+01  1.746e+00  10.476  < 2e-16 ***
## I(loggdpPercap^2) -6.413e-01  1.059e-01  -6.054 1.73e-09 ***
## year               6.177e+00  2.478e+00   2.492   0.0128 *
## I(year^2)         -1.509e-03  6.260e-04  -2.410   0.0160 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.793 on 1699 degrees of freedom
## Multiple R-squared:  0.7241, Adjusted R-squared:  0.7234
## F-statistic:  1115 on 4 and 1699 DF,  p-value: < 2.2e-16
```

```
lifeExp.lm = lm(lifeExp ~ loggdpPercap+I(loggdpPercap^2)+year, data=subset(gapminder, continent=='Ameri
lifeExp.aug = augment(lifeExp.lm)
gg1 = ggplot(lifeExp.aug, aes(x = lifeExp, y = .resid)) + geom_point()
gg1 = gg1 + geom_smooth(method = "loess", se = FALSE)
gg2 = ggplot(lifeExp.lm, aes(sample = .resid)) + stat_qq()
figure <- ggarrange(gg1, gg2, labels = c("Residuals vs. lifeExp", "Residuals QQ Plot"),  nrow = 1, ncol
figure
```



```r
ggsave("Q1_8.png", device='pdf')
```

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