STAT 545 HW 02

Elijah Willie

September 25, 2018

Introduction

In this document, I will be exploring the gapminder data further. I will be computing its class, how many variables, and how many obeservations. I will also be interested in the types of variable present and some summary statistics about them. I will also be looking at some variables in more detials by doing some graphical analyses. Hope you find this document eventful.

load in the required libraries

```
suppressMessages(library("tidyverse"))
library(gapminder)
```

Smell the test data

Summarize the data

```
knitr::kable(summary(gapminder))
```

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan: 12	Africa :624	Min. :1952	Min. :23.60	Min. :6.001e+04	Min. : 241.2
Albania : 12	Americas:300	1st Qu.:1966	1st Qu.:48.20	1st Qu.:2.794e+06	1st Qu.: 1202.1
Algeria : 12	Asia :396	Median :1980	Median :60.71	Median :7.024e+06	Median : 3531.8
Angola : 12	Europe :360	Mean :1980	Mean :59.47	Mean :2.960e+07	Mean : 7215.3
Argentina : 12	Oceania : 24	3rd Qu.:1993	3rd Qu.:70.85	3rd Qu.:1.959e+07	3rd Qu.: 9325.5
Australia : 12	NA	Max. :2007	Max. :82.60	Max. :1.319e+09	Max. :113523.1
(Other) :1632	NA	NA	NA	NA	NA

Get the type of the data

```
typeof(gapminder)
```

```
## [1] "list"
```

Get the data class

```
class(gapminder)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

Get the number of variables

```
ncol(gapminder)
```

```
## [1] 6
```

Get the number of observations

```
nrow(gapminder)
```

```
## [1] 1704
```

Get the types of variables

I will be using the sapply function that applies a function to each column of the dataframe and prints the output. For the function, I will be using the class function which returns the class of each of the columns of the gapminder data

```
knitr::kable(sapply(gapminder, class))
```

	x
country	factor
continent	factor
year	integer
lifeExp	numeric
pop	integer
gdpPercap	numeric

It is also interesting to do summaries of some variables stratified by a grouping variable.

I will group the gapminder dataset by continent

```
oldops <-options(tibble.width=Inf, tibble.print_max=Inf)
gm_byContinent <- group_by(gapminder,continent)
knitr::kable(head(gm_byContinent))
```

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134

Summarize the data using the grouped variable

```
knitr::kable(summary(gm_byContinent))
```

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan: 12	Africa :624	Min. :1952	Min. :23.60	Min. :6.001e+04	Min. : 241.2
Albania : 12	Americas:300	1st Qu.:1966	1st Qu.:48.20	1st Qu.:2.794e+06	1st Qu.: 1202.1
Algeria : 12	Asia :396	Median :1980	Median :60.71	Median :7.024e+06	Median : 3531.8
Angola : 12	Europe :360	Mean :1980	Mean :59.47	Mean :2.960e+07	Mean : 7215.3
Argentina : 12	Oceania : 24	3rd Qu.:1993	3rd Qu.:70.85	3rd Qu.:1.959e+07	3rd Qu.: 9325.5
Australia : 12	NA	Max. :2007	Max. :82.60	Max. :1.319e+09	Max. :113523.1
(Other) :1632	NA	NA	NA	NA	NA

We can aslo do a pairs plot for variables of interest in our dataset

```
library(GGally)
```

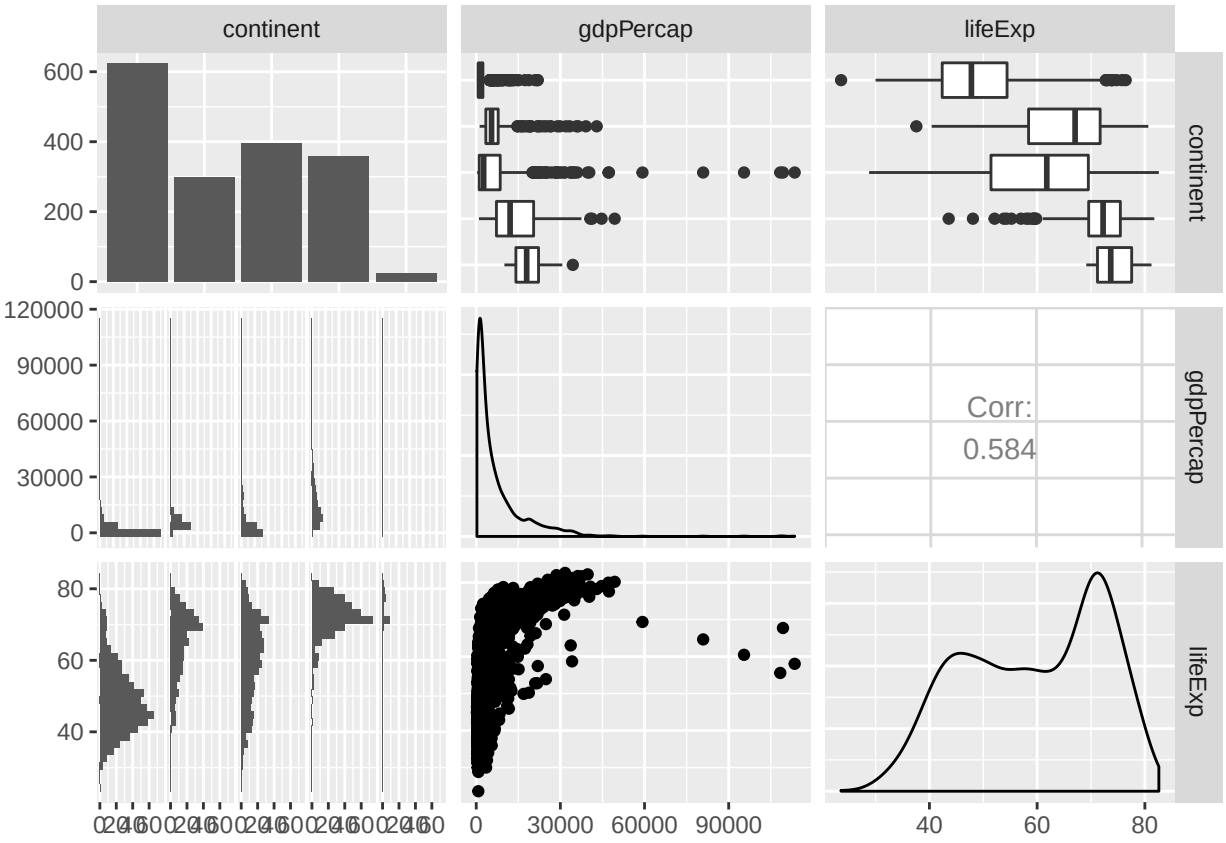
```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
gm_pairs <- select(gapminder, continent, gdpPercap, lifeExp)
ggpairs(gm_pairs)
```

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

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



Explore individual variables

For this analyses I will be picking continent, lifeExp, and gdpPercap

Get some data summary for gdpPercap for each continent

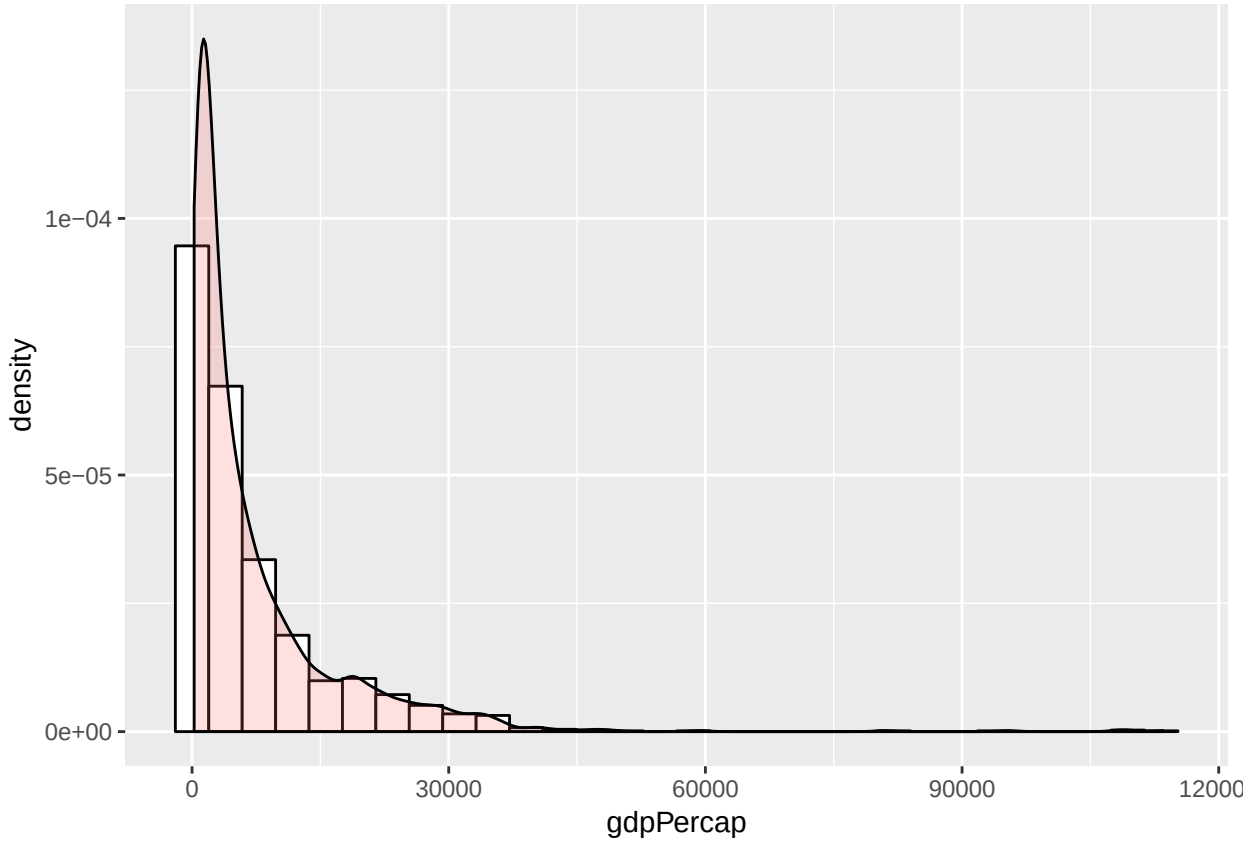
```
knitr::kable(summarize(gm_byContinent,min(gdpPercap),median(gdpPercap),
                       mean(gdpPercap),sd(gdpPercap),max(gdpPercap)))
```

continent	min(gdpPercap)	median(gdpPercap)	mean(gdpPercap)	sd(gdpPercap)	max(gdpPercap)
Africa	241.1659	1192.138	2193.755	2827.930	21951.21
Americas	1201.6372	5465.510	7136.110	6396.764	42951.65
Asia	331.0000	2646.787	7902.150	14045.373	113523.13
Europe	973.5332	12081.749	14469.476	9355.213	49357.19
Oceania	10039.5956	17983.304	18621.609	6358.983	34435.37

Overlay a histogram with a density plot for gdpPercap

```
ggplot(gapminder, aes(x=gdpPercap)) + geom_histogram(aes(y=..density..), colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666")
```

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



Get some data summary for lifeExp

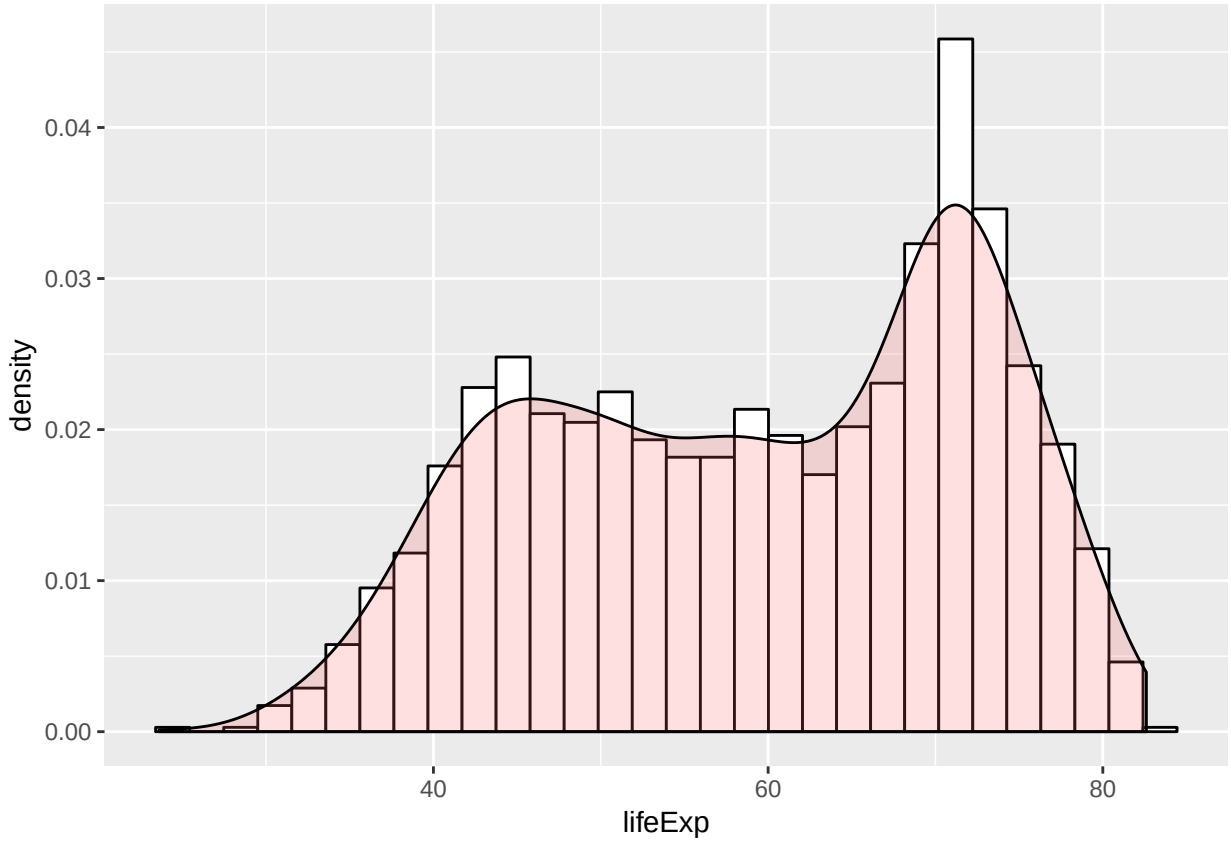
```
knitr::kable(summarize(gm_byContinent,min(lifeExp),
                       median(lifeExp),mean(lifeExp),sd(lifeExp),max(lifeExp)))
```

continent	min(lifeExp)	median(lifeExp)	mean(lifeExp)	sd(lifeExp)	max(lifeExp)
Africa	23.599	47.7920	48.86533	9.150210	76.442
Americas	37.579	67.0480	64.65874	9.345088	80.653
Asia	28.801	61.7915	60.06490	11.864532	82.603
Europe	43.585	72.2410	71.90369	5.433178	81.757
Oceania	69.120	73.6650	74.32621	3.795611	81.235

Overlay a histogram with a density plot for lifeExp

```
ggplot(gapminder, aes(x=lifeExp)) + geom_histogram(aes(y=..density..), colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666")
```

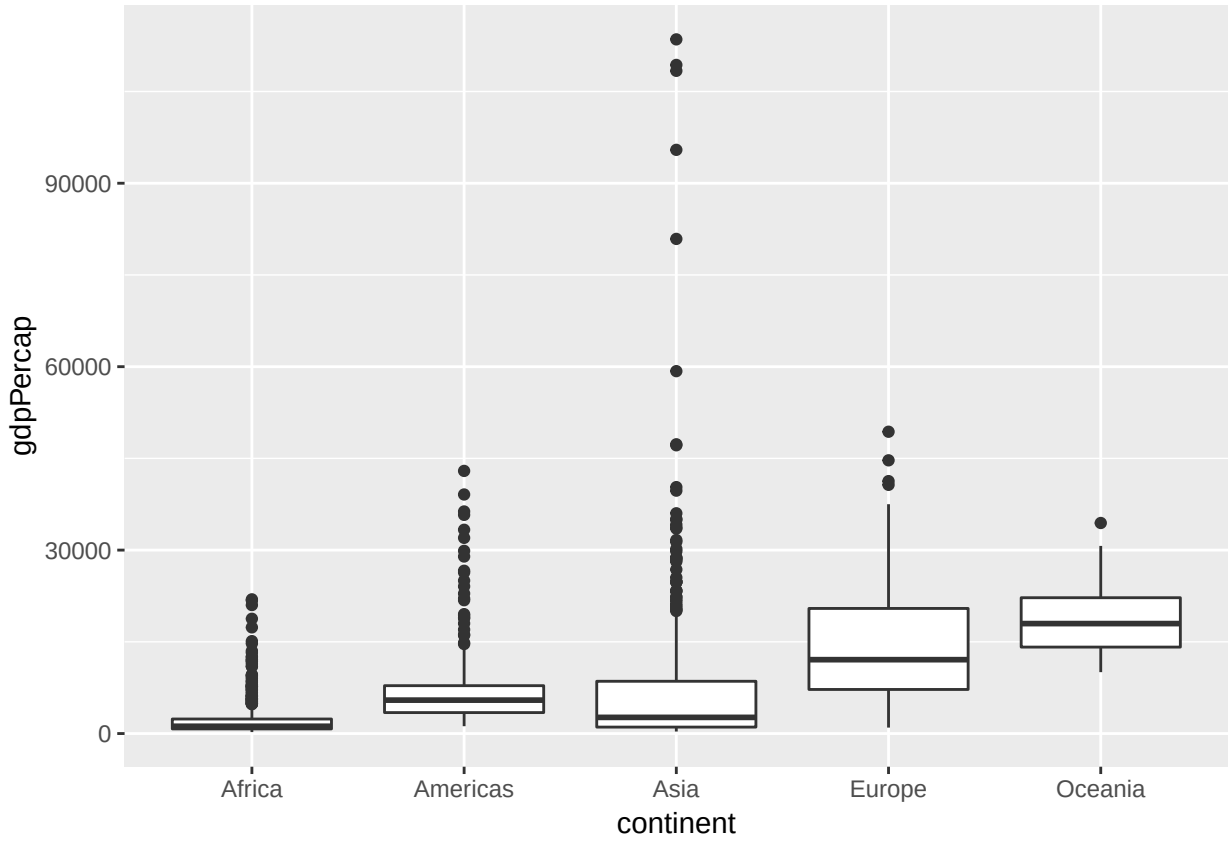
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Explore various plot types

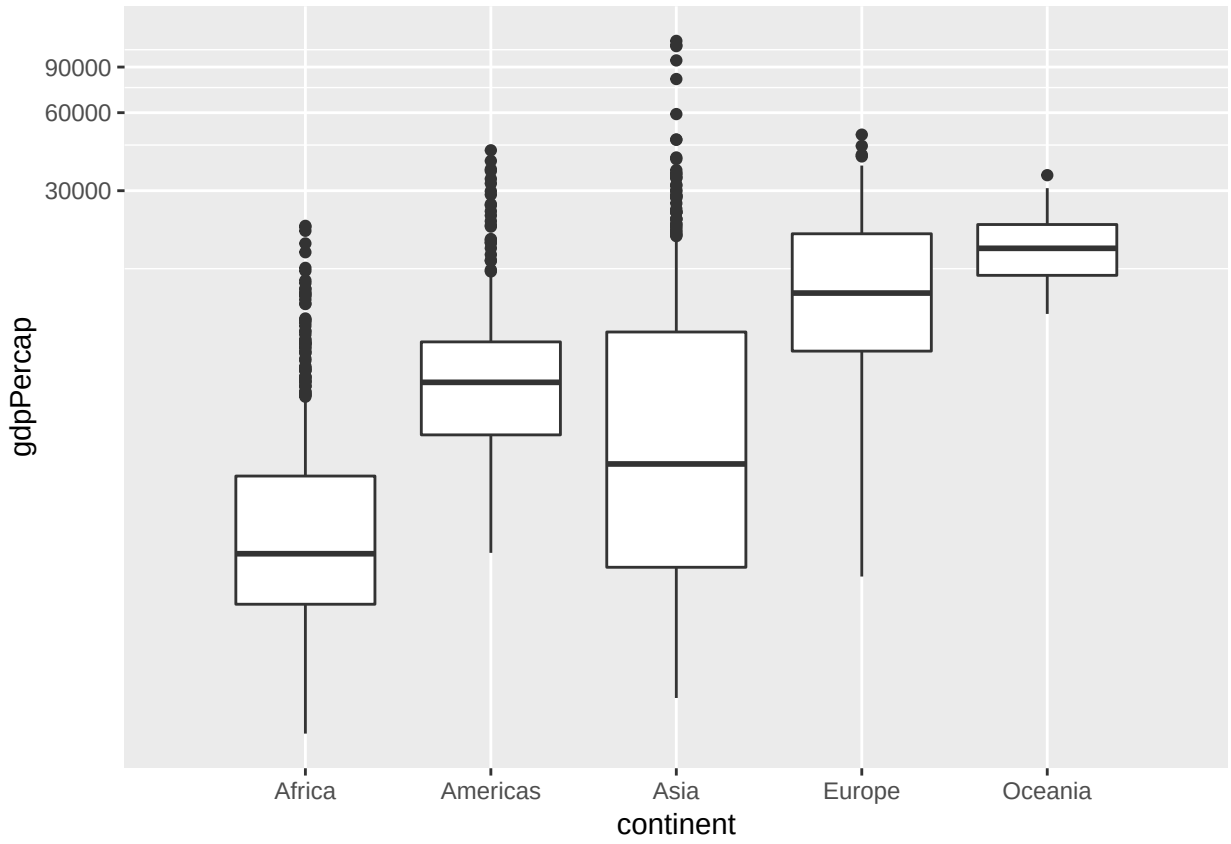
Boxplots

```
ggplot(gapminder, aes(continent, gdpPercap)) + geom_boxplot()
```



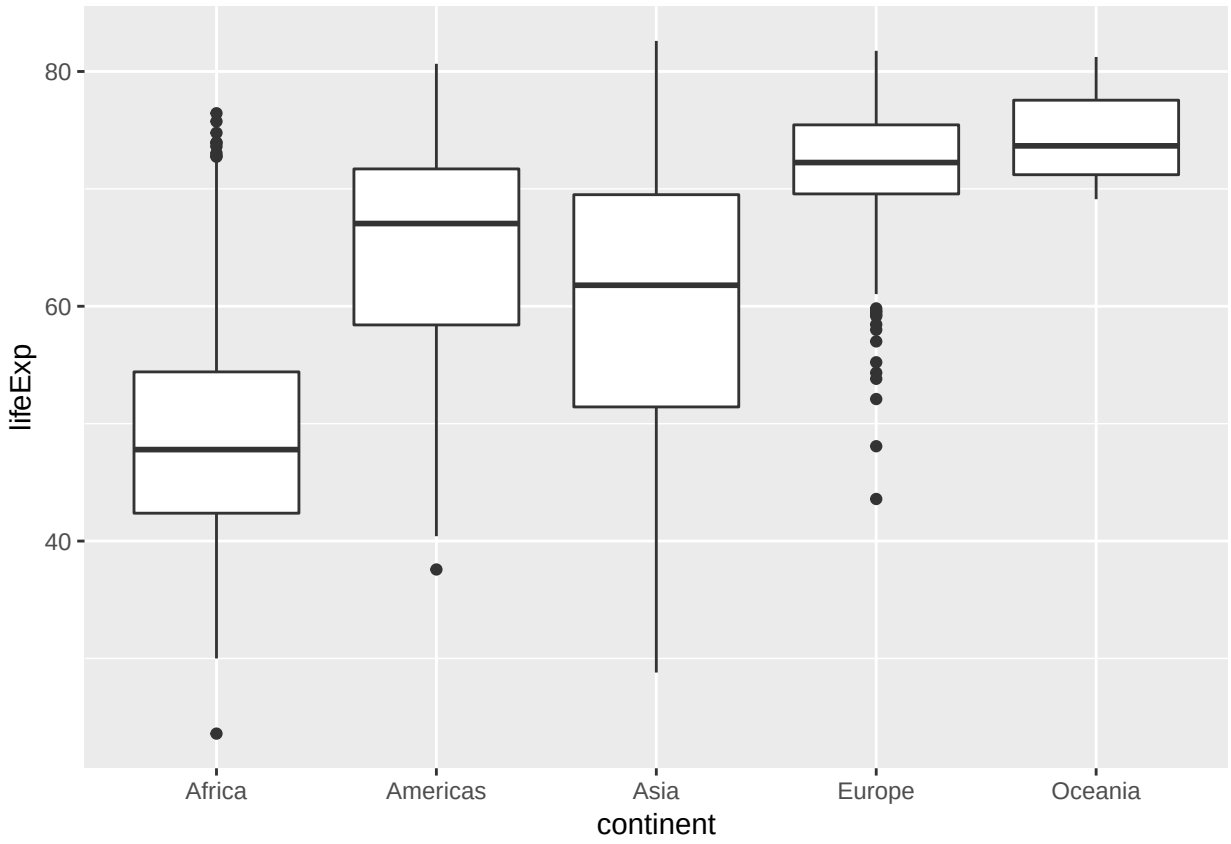
Now do a boxplot again, but with a log_10 transform of the gdpPercap variable

```
ggplot(gapminder, aes(continent, gdpPercap)) + coord_trans(y="log10") + geom_boxplot()
```



Boxplot for lifeExp

```
ggplot(gapminder, aes(continent, lifeExp)) + geom_boxplot()
```



Now instead of transforming it everytime, I will transform it and store it in a new variable

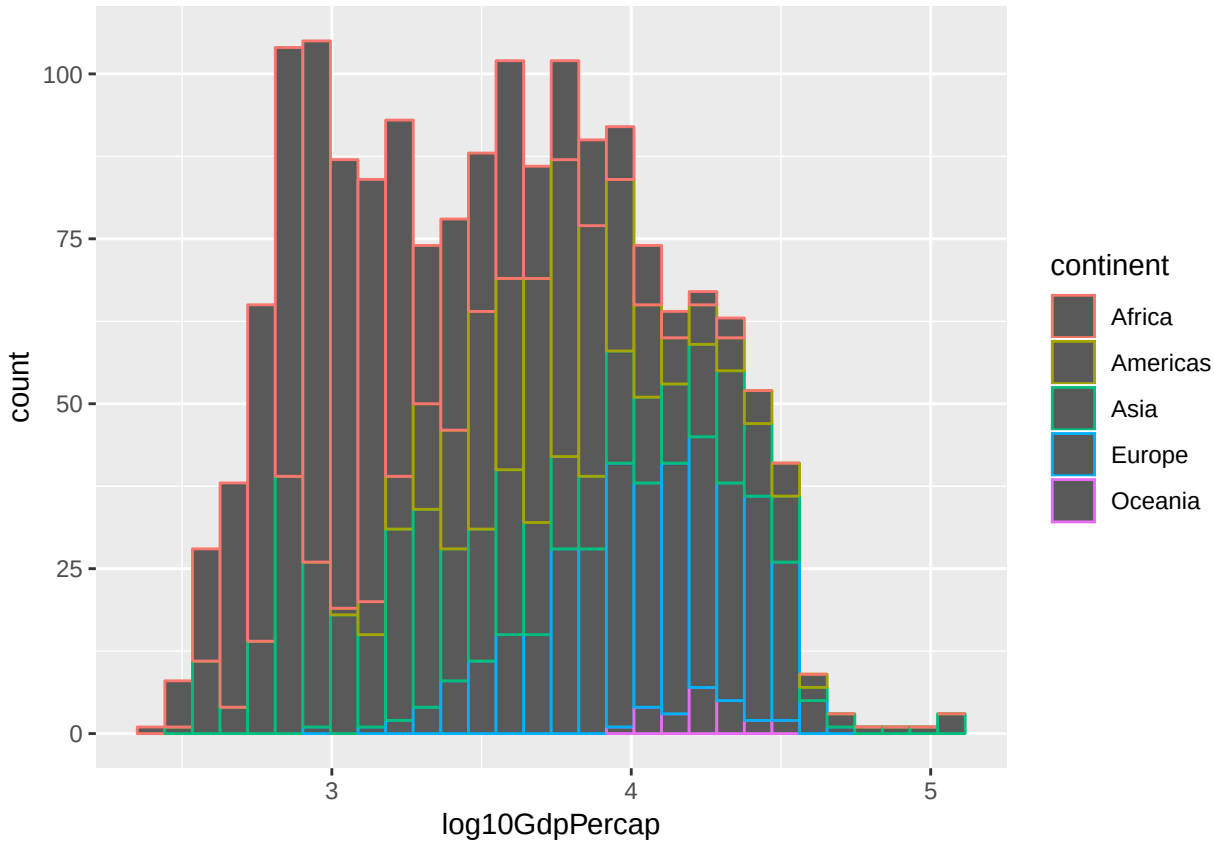
```
gapminder <- mutate(gapminder, log10GdpPercap =log10(gdpPercap))
```

Histograms

Plot a stacked historam of log(gdpPercap) as a function of continent

```
ggplot(gapminder,aes(x=log10GdpPercap, color=continent))+ geom_histogram()
```

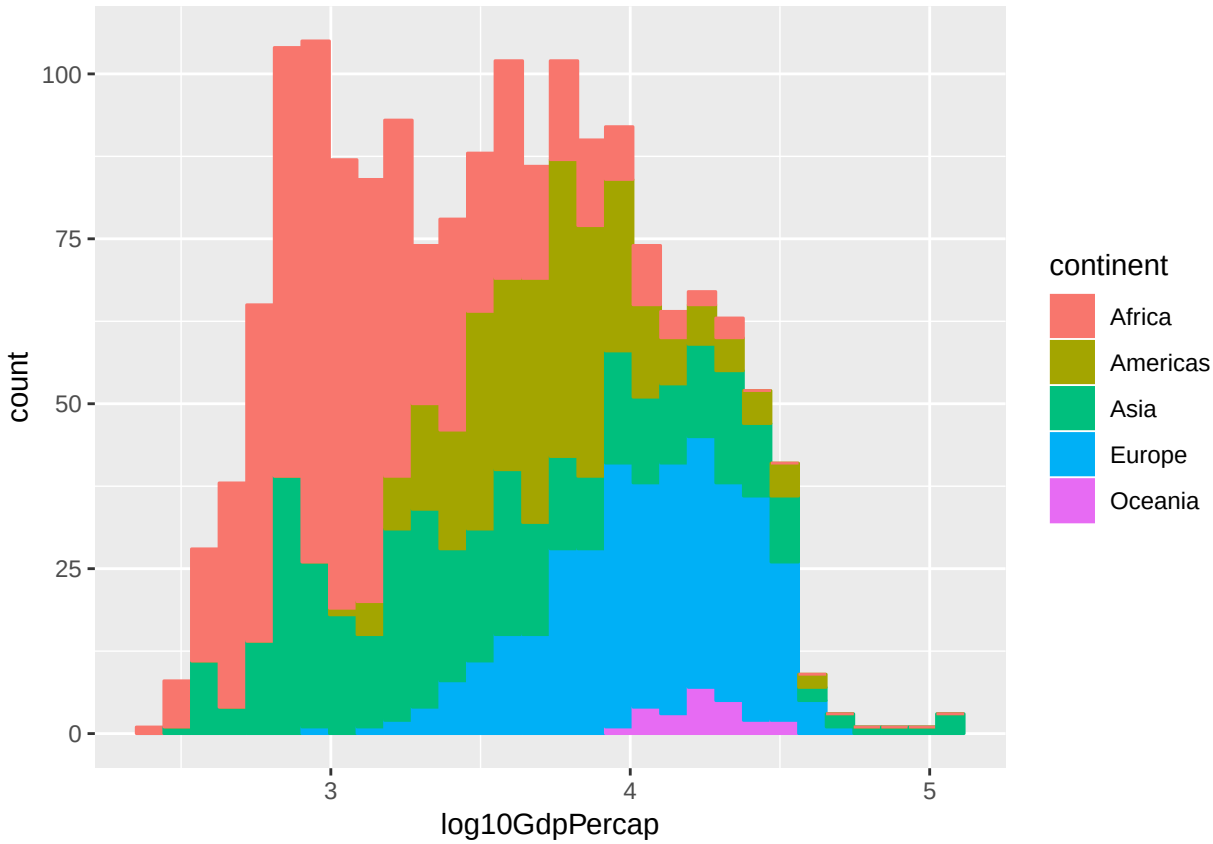
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



You can also fill the histogram bars with colors

```
ggplot(gapminder,aes(x=log10GdpPercap, color=continent))+ geom_histogram(aes(fill = continent))
```

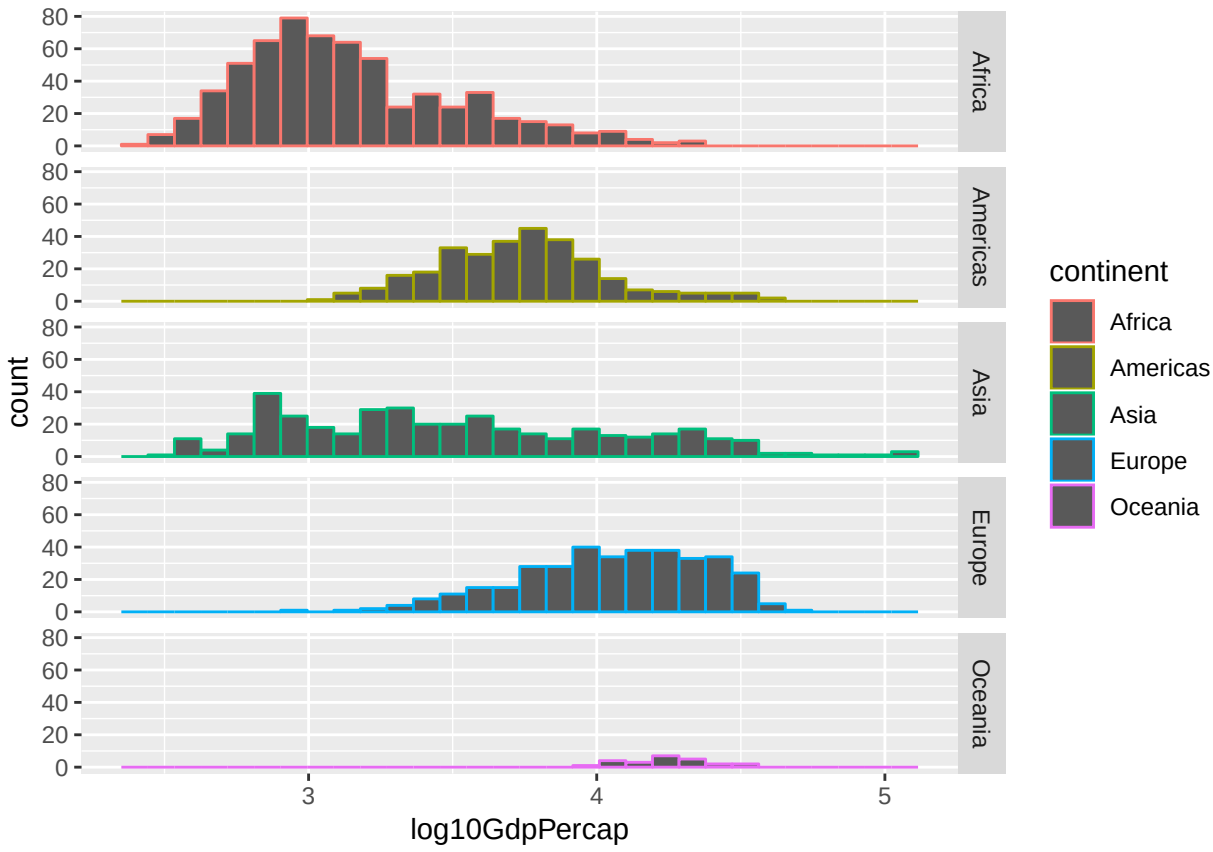
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



You can also stack the histograms if you do not wish to have them overlaying each other

```
ggplot(gapminder,aes(x=log10GdpPercap, color=continent))+ geom_histogram() + facet_grid(continent ~ .)
```

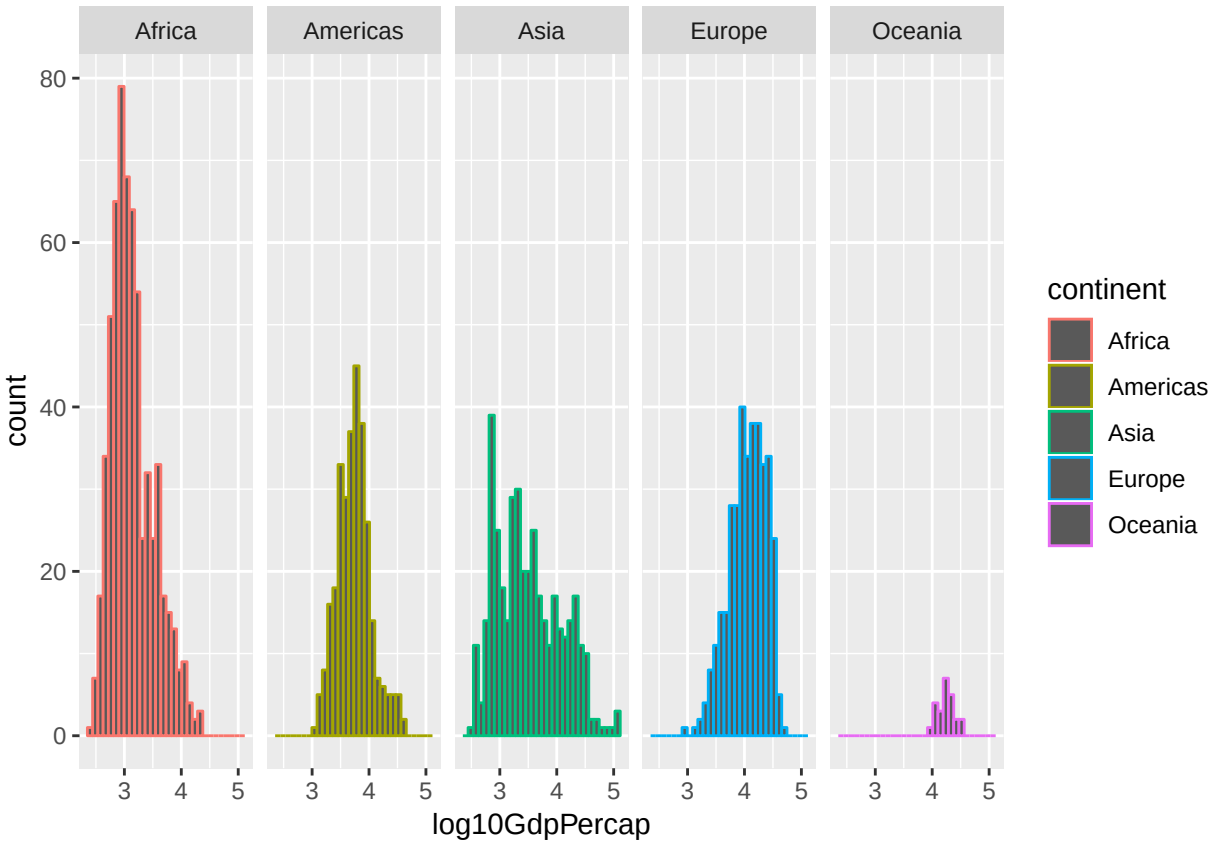
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



You can also stack the histograms sideways

```
ggplot(gapminder,aes(x=log10GdpPercap, color=continent))+ geom_histogram() + facet_grid(~continent)
```

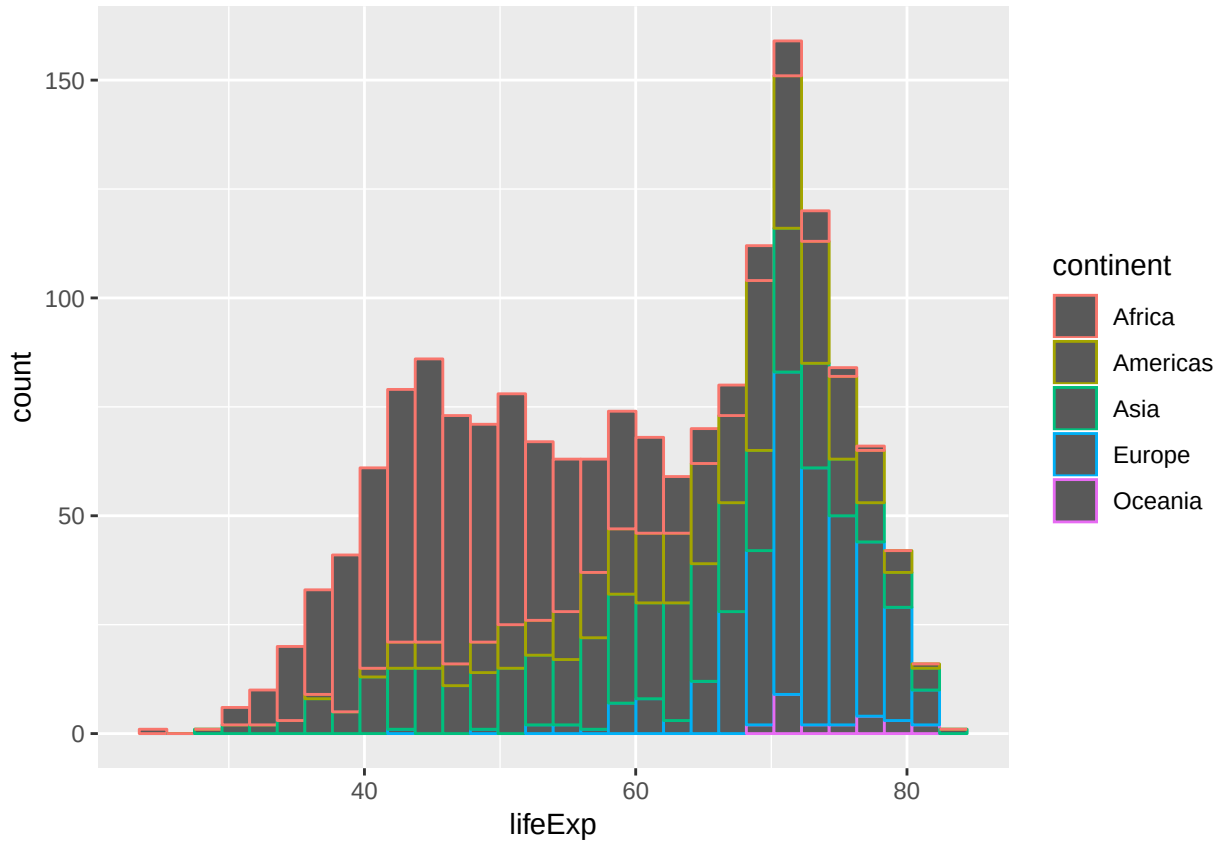
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Do the same for lifeExp

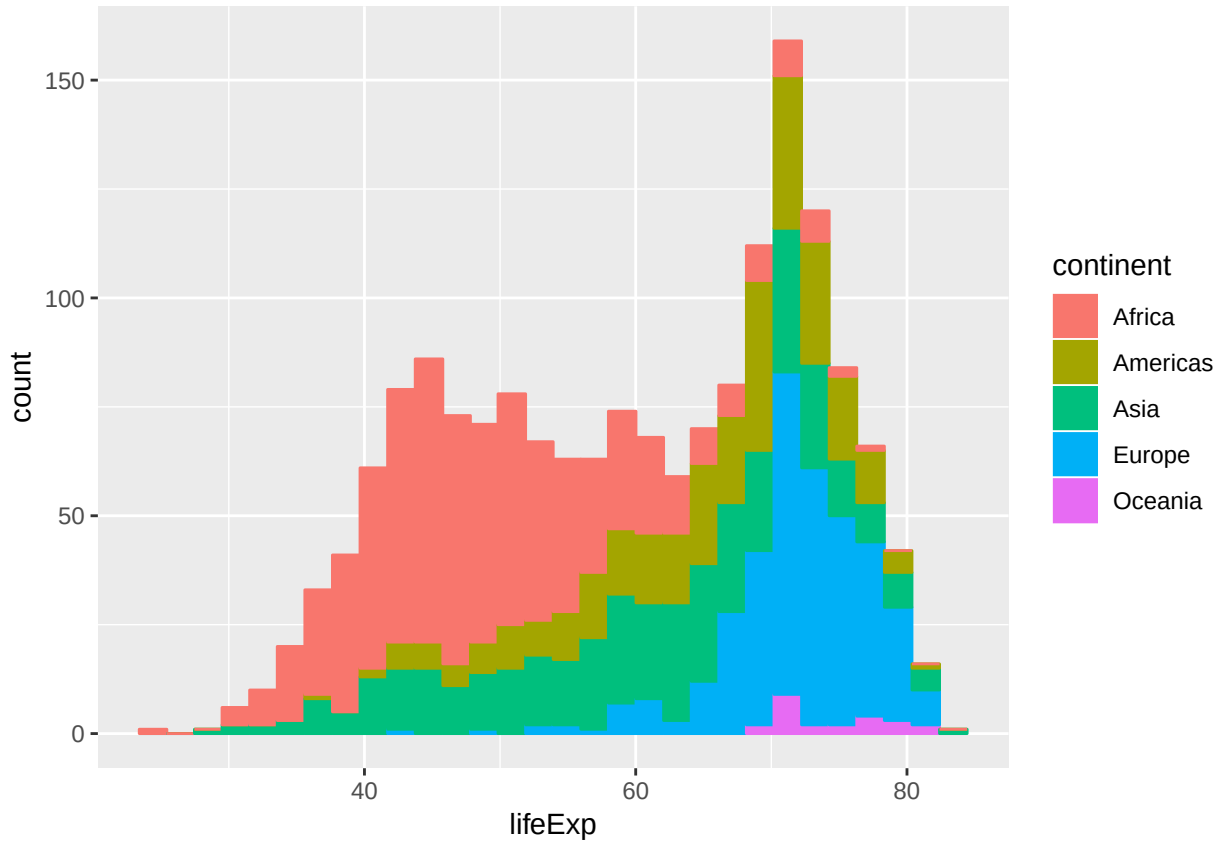
```
ggplot(gapminder,aes(x=lifeExp, color=continent))+ geom_histogram()
```

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



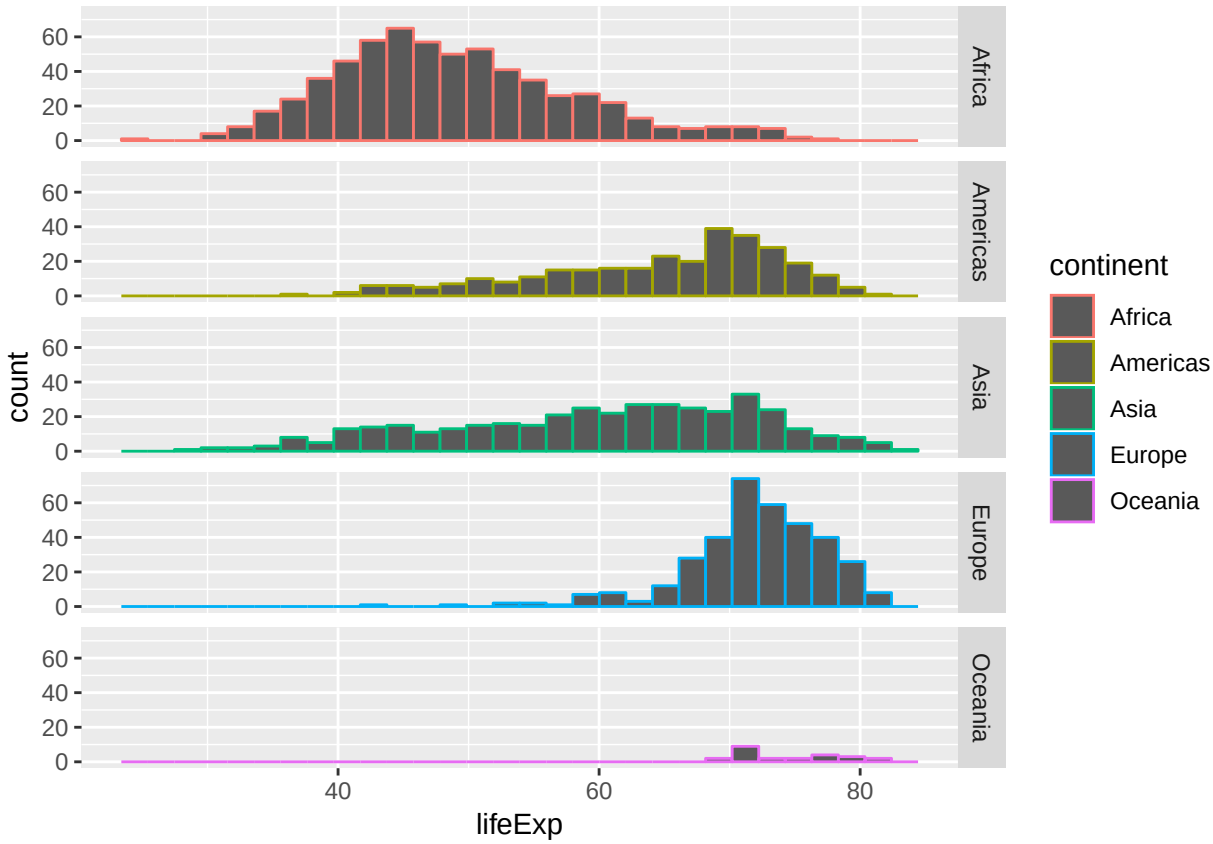

```
ggplot(gapminder,aes(x=lifeExp, color=continent))+ geom_histogram(aes(fill = continent))
```

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



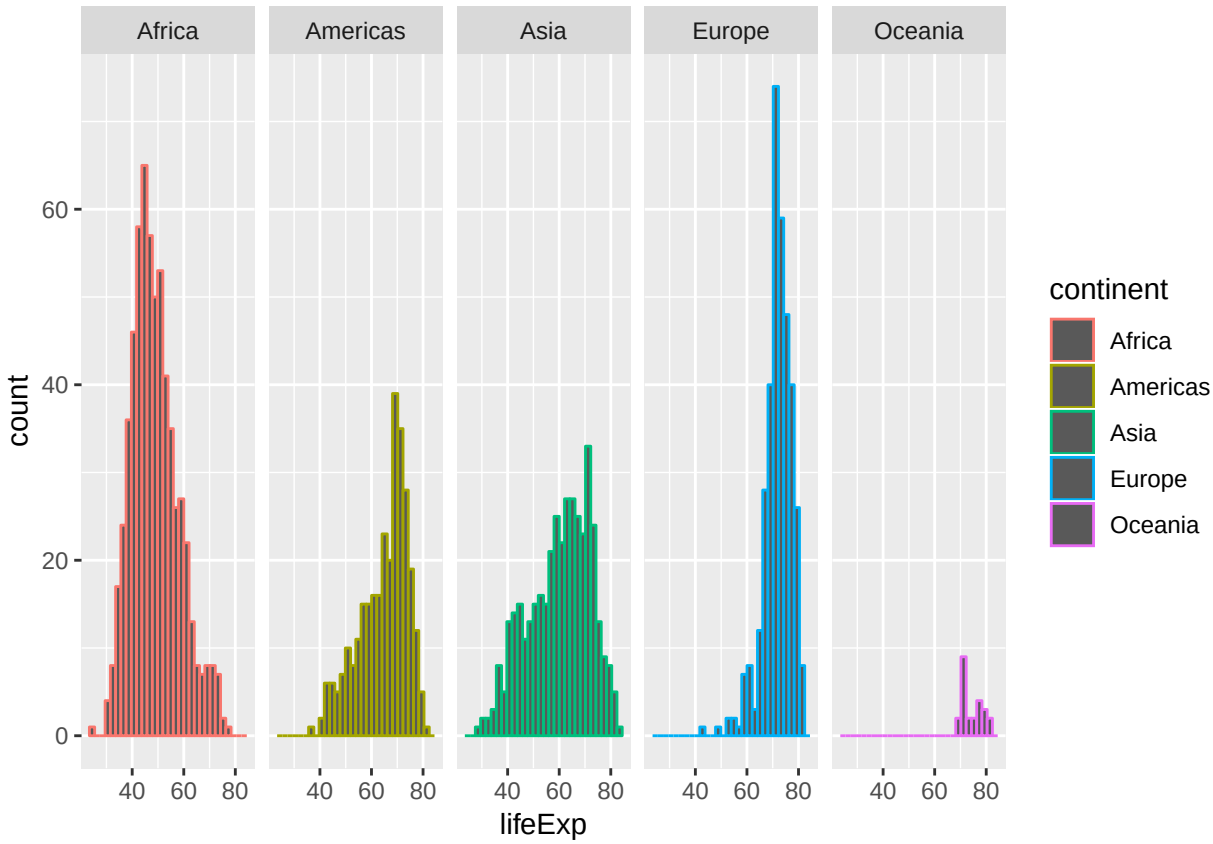

```
ggplot(gapminder,aes(x=lifeExp, color=continent))+ geom_histogram() + facet_grid(continent ~ .)
```

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



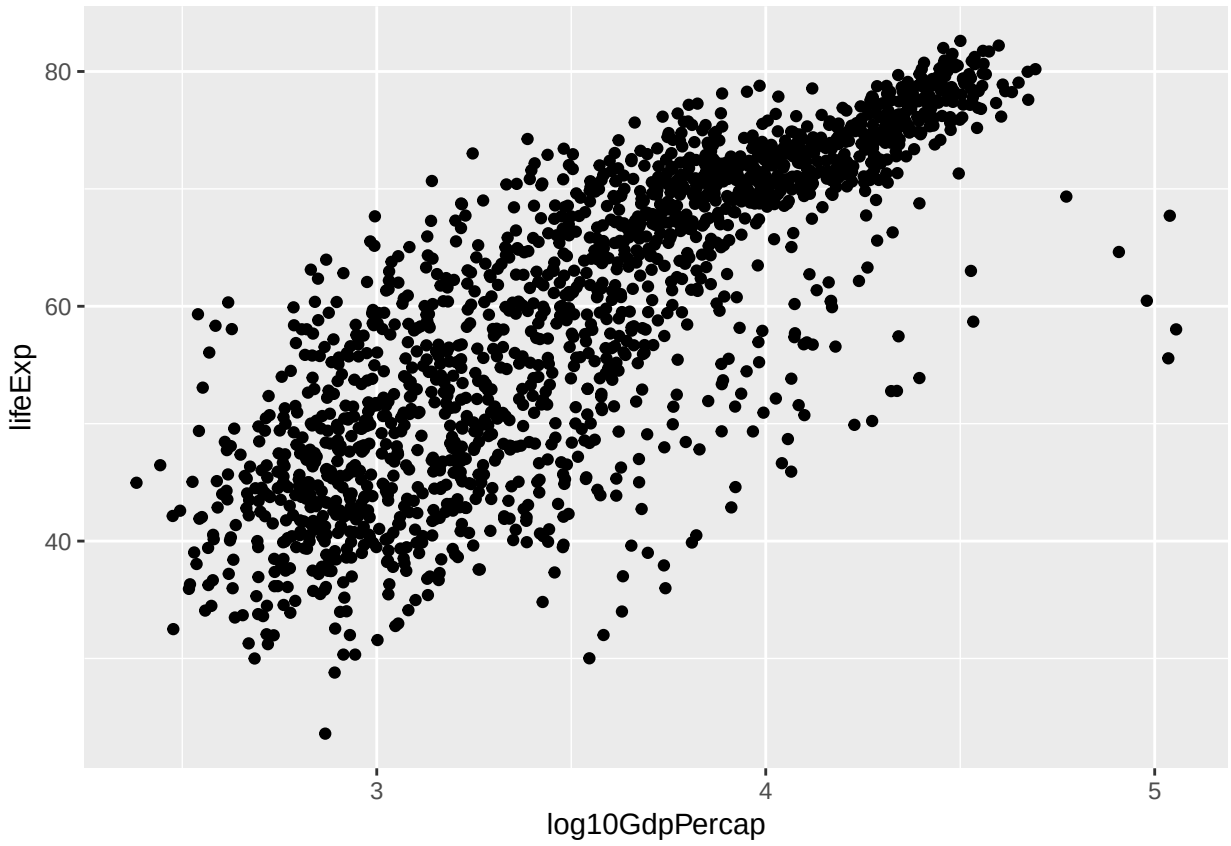

```
ggplot(gapminder,aes(x=lifeExp, color=continent))+ geom_histogram() + facet_grid(~continent)
```

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



Do a scatterplot of logGdpPercap vs lifeExp

```
my_plot <- ggplot(gapminder, aes(log10GdpPercap, lifeExp))
my_plot + geom_point()
```

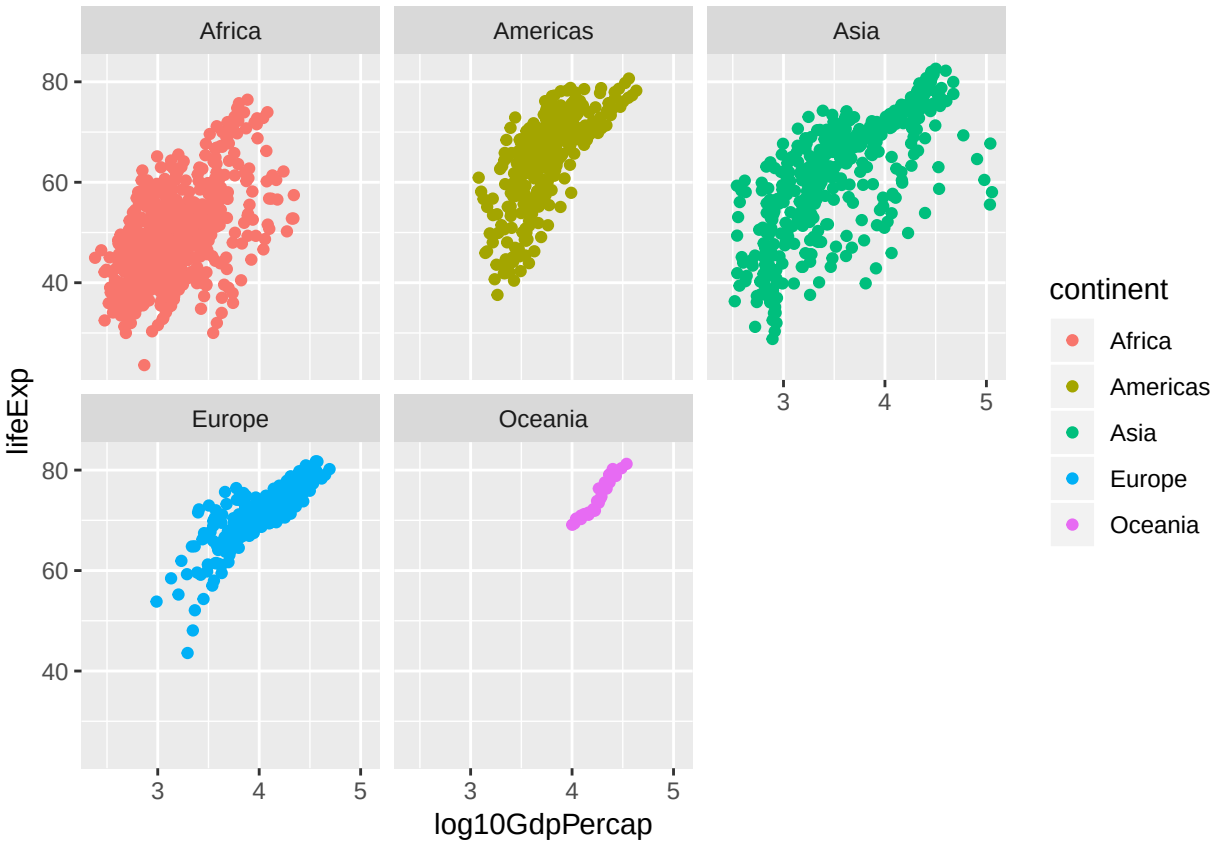


Now color the plot by continent

```
my_plot + geom_point(aes(color = continent))
```

you can even do a facet wrap with countries

```
my_plot + geom_point(aes(color = continent)) + facet_wrap(~continent)
```



Use select and piping functions

I will explore the life expectancy variable for the year 2007. First filter the data to just 2007.

```
gapminder07 <- filter(gapminder, year == 2007)
 knitr::kable(head(gapminder07))
```

country	continent	year	lifeExp	pop	gdpPercap	log10GdpPercap
Afghanistan	Asia	2007	43.828	31889923	974.5803	2.988818
Albania	Europe	2007	76.423	3600523	5937.0295	3.773569
Algeria	Africa	2007	72.301	33333216	6223.3675	3.794025
Angola	Africa	2007	42.731	12420476	4797.2313	3.680991
Argentina	Americas	2007	75.320	40301927	12779.3796	4.106510
Australia	Oceania	2007	81.235	20434176	34435.3674	4.537005

Calculate median life expectancy, first overall, and then by continent.

```
knitr::kable(summarize(gapminder07, median(lifeExp)))
```

median(lifeExp)
71.9355

```
by_cont <- group_by(gapminder07, continent)
knitr::kable(summarise(by_cont, median(lifeExp)))
```

continent	median(lifeExp)
Africa	52.9265
Americas	72.8990
Asia	72.3960
Europe	78.6085
Oceania	80.7195

We can compute the median life expectancies.

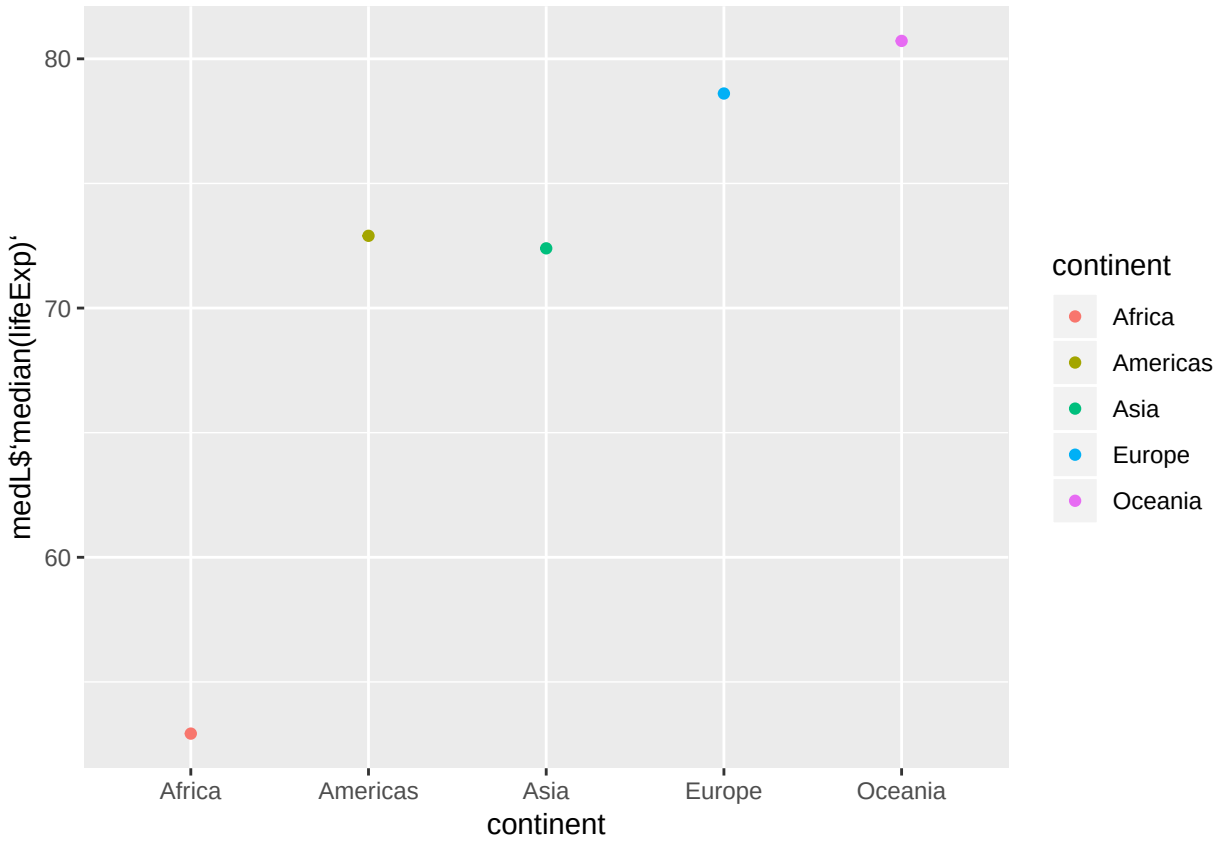
```
medL <- summarize(by_cont, median(lifeExp))
```

We can also commpute the median life expectancy using chaining and piping

```
medL.2 <- gapminder %>%
  filter(year == 2007) %>%
  group_by(continent) %>%
  summarise(medLifeExp = median(lifeExp))
```

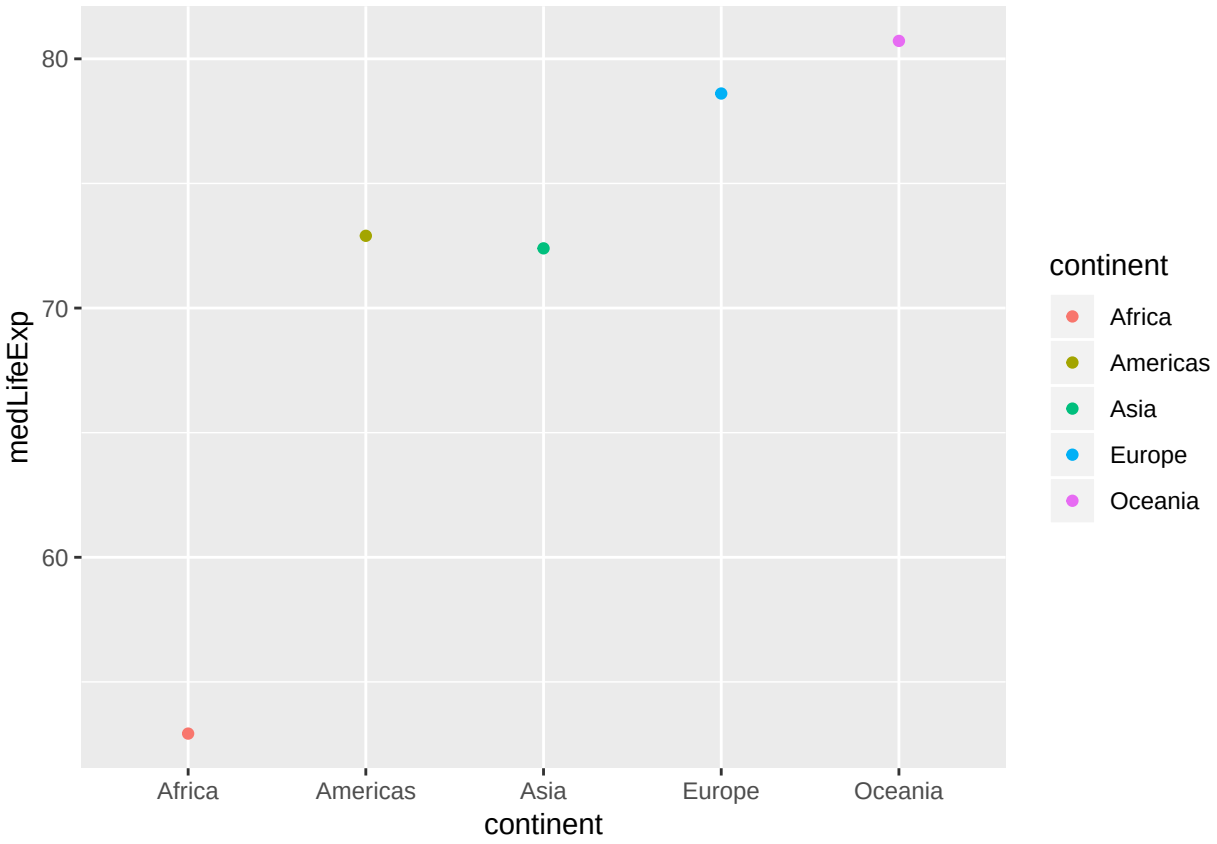
We can visualize the median life expectancies

```
ggplot(medL, aes(continent, y = medL$`median(lifeExp)`)) + geom_point(aes(color = continent))
```



We can actually combine all of this into a set of chaining and piping command that does the plot at the end

```
gapminder %>%
  filter(year == 2007) %>%
  group_by(continent) %>%
  summarise(medLifeExp = median(lifeExp)) %>%
  ggplot(aes(continent, y = medLifeExp)) + geom_point(aes(color = continent))
```



But I want to do more

Evaluate the given code

```
result <- filter(gapminder, country == c("Rwanda", "Afghanistan"))
 knitr::kable(result)
```

country	continent	year	lifeExp	pop	gdpPercap	log10GdpPercap
Afghanistan	Asia	1957	30.332	9240934	820.8530	2.914265
Afghanistan	Asia	1967	34.020	11537966	836.1971	2.922309
Afghanistan	Asia	1977	38.438	14880372	786.1134	2.895485
Afghanistan	Asia	1987	40.822	13867957	852.3959	2.930641
Afghanistan	Asia	1997	41.763	22227415	635.3414	2.803007
Afghanistan	Asia	2007	43.828	31889923	974.5803	2.988818
Rwanda	Africa	1952	40.000	2534927	493.3239	2.693132
Rwanda	Africa	1962	43.000	3051242	597.4731	2.776318
Rwanda	Africa	1972	44.600	3992121	590.5807	2.771279
Rwanda	Africa	1982	46.218	5507565	881.5706	2.945257
Rwanda	Africa	1992	23.599	7290203	737.0686	2.867508
Rwanda	Africa	2002	43.413	7852401	785.6538	2.895231

The analyst did not. The proper way to do it is shown below. As we can see, the difference is that we get double the number of rows so initially the analyst only obtains haft the total result, but with the second attempt below, the full result is obtained.

```
result2 <- filter(gapminder, country =="Rwanda"| country ==  "Afghanistan")
 knitr::kable(result2)
```

country	continent	year	lifeExp	pop	gdpPercap	log10GdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453	2.891786
Afghanistan	Asia	1957	30.332	9240934	820.8530	2.914265
Afghanistan	Asia	1962	31.997	10267083	853.1007	2.931000
Afghanistan	Asia	1967	34.020	11537966	836.1971	2.922309
Afghanistan	Asia	1972	36.088	13079460	739.9811	2.869221
Afghanistan	Asia	1977	38.438	14880372	786.1134	2.895485
Afghanistan	Asia	1982	39.854	12881816	978.0114	2.990344
Afghanistan	Asia	1987	40.822	13867957	852.3959	2.930641
Afghanistan	Asia	1992	41.674	16317921	649.3414	2.812473
Afghanistan	Asia	1997	41.763	22227415	635.3414	2.803007
Afghanistan	Asia	2002	42.129	25268405	726.7341	2.861375
Afghanistan	Asia	2007	43.828	31889923	974.5803	2.988818
Rwanda	Africa	1952	40.000	2534927	493.3239	2.693132
Rwanda	Africa	1957	41.500	2822082	540.2894	2.732626
Rwanda	Africa	1962	43.000	3051242	597.4731	2.776318
Rwanda	Africa	1967	44.100	3451079	510.9637	2.708390
Rwanda	Africa	1972	44.600	3992121	590.5807	2.771279
Rwanda	Africa	1977	45.000	4657072	670.0806	2.826127
Rwanda	Africa	1982	46.218	5507565	881.5706	2.945257
Rwanda	Africa	1987	44.020	6349365	847.9912	2.928391
Rwanda	Africa	1992	23.599	7290203	737.0686	2.867508
Rwanda	Africa	1997	36.087	7212583	589.9445	2.770811
Rwanda	Africa	2002	43.413	7852401	785.6538	2.895231
Rwanda	Africa	2007	46.242	8860588	863.0885	2.936055