Untitled

Mike Marin

2019-09-17

The Gapminder Data

Variables Collected

Here, we can see the **variables** collected in the data set:

```
## [1] "country"   "continent" "year"      "lifeExp"   "pop"       "gdpPercap"
```

Some Basic Summaries

The data dimensions are as folows:

```
dim(gapminder)
```

```
## [1] 1704    6
```

The Full Data

To explore the full data, you can play around here:

```
# the datatable makes it easier to do something in HTML
datatable(as_tibble(gapminder))
```

Some Univariate Summaries

It is of interest to take a look at each of the variables, and provide a quick summary, both numerically as well as graphically

A Quick Glance

Following are basic univariate summaries for each of the variables:

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

```
##  Mean   :2.960e+07   Mean   :  7215.3
##  3rd Qu.:1.959e+07   3rd Qu.:  9325.5
##  Max.   :1.319e+09   Max.   :113523.1
##
```

Countires and Continents

Which countries are represented in the data?

- In total there are 142 countries represented. They are not all listed here for the sake of space.
- There are a total of 5 geopolitical continents recorded:
  - Africa
  - Asia
  - Europe
  - Americas (made up of North America and South America are combined)
  - Oceania (Australia, New Zealand, and surrounding island countries)
  - Antarctica is not included in the data

Years Recorded

The data was collected for a number of years. The years the data was collected for were:

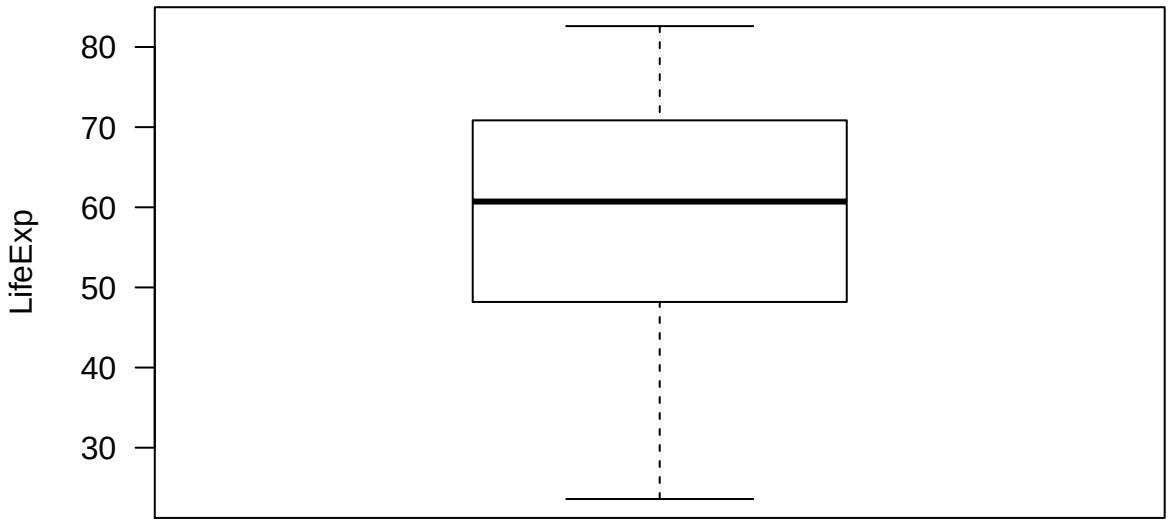
```
##  [1] "1952" "1957" "1962" "1967" "1972" "1977" "1982" "1987" "1992" "1997"
## [11] "2002" "2007"
```

Life Expectancy, GDPperCap, and Population

These are all numeric variables, and all likely related to one another. Following are basic summaries and plots for each:

Life Expectancy

```
boxplot(gapminder$lifeExp, ylab="LifeExp", las=1)
```



```r
summary(gapminder$lifeExp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.60   48.20   60.71   59.47   70.85   82.60
```

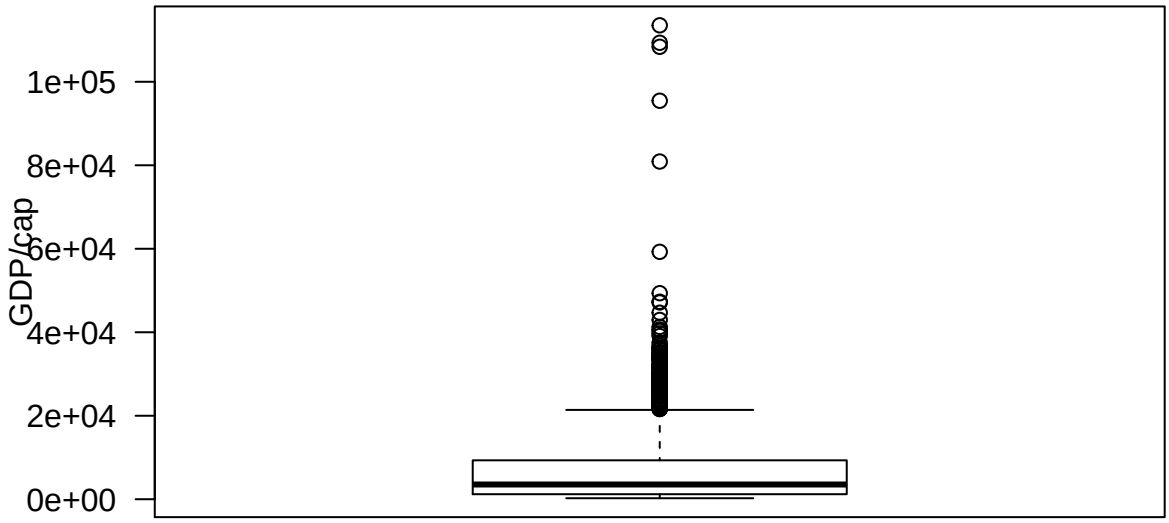
Wow, 23 years life expectancy?! Where did that come from?

```
## # A tibble: 1 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Rwanda  Africa     1992    23.6 7290203      737.
```

Oh, ok, we all know that there was a ***LOT*** going on there at that time!

GDP Per Capita

```r
boxplot(gapminder$gdpPercap, ylab="GDP/cap", las=1)
```

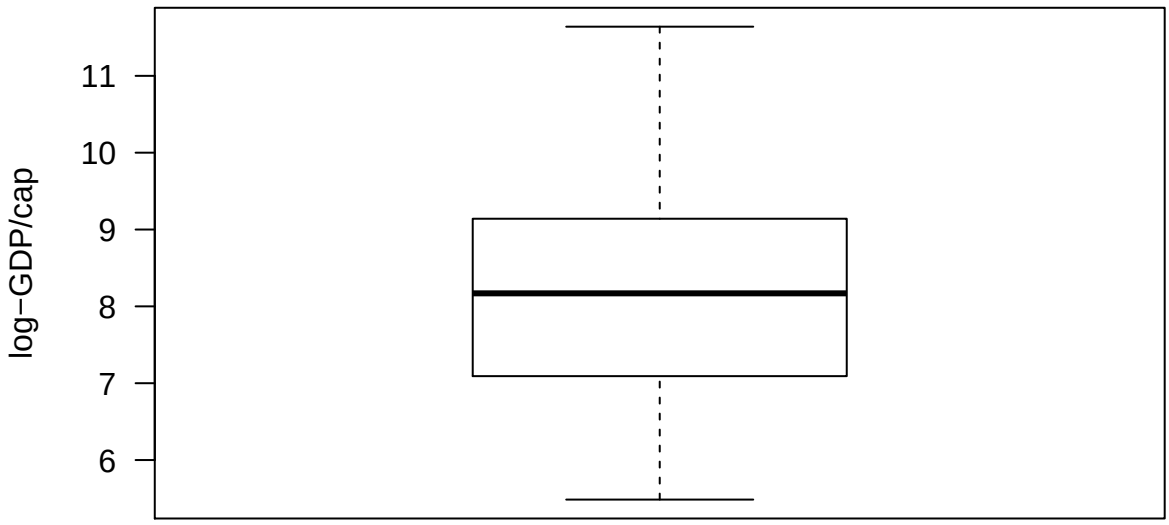



```r
summary(gapminder$gdpPercap)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    241.2   1202.1   3531.8   7215.3   9325.5 113523.1
```

And a quick note that since GDP/cap is a rate, it may be better to be exploring it on the log(ln)-scale

```r
boxplot(log(gapminder$gdpPercap), ylab="log-GDP/cap", las=1)
```

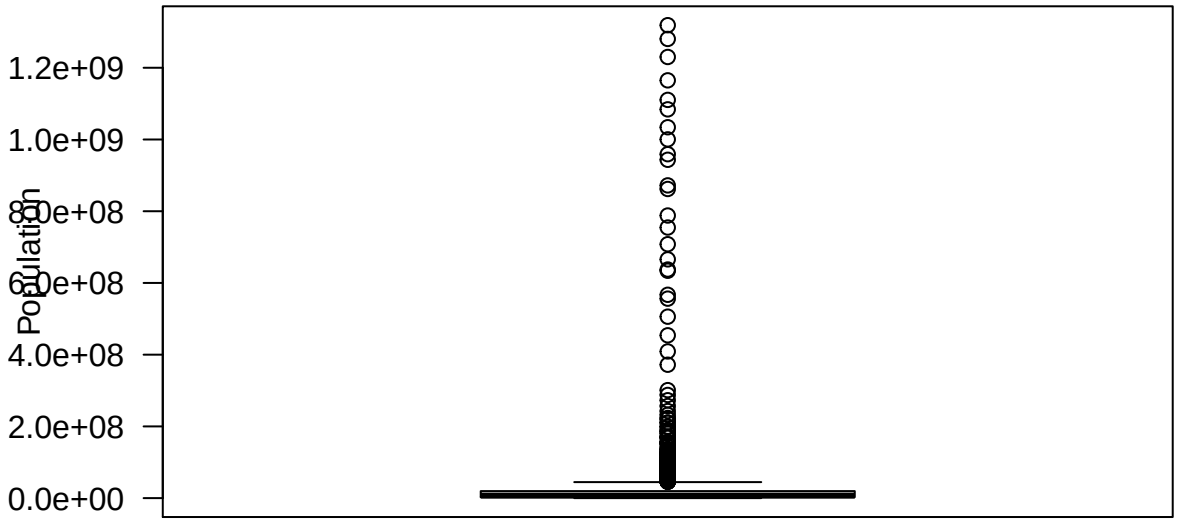



```
summary(log(gapminder$gdpPercap))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.485   7.092   8.170   8.159   9.141  11.640
```

Population

```
boxplot(gapminder$pop, ylab="Population", las=1)
```

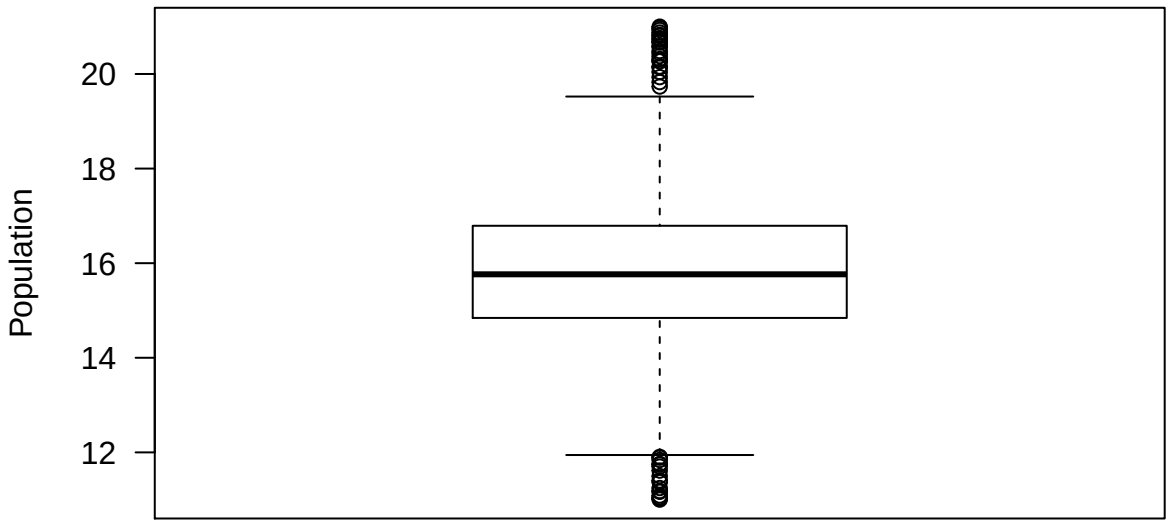



```
summary(gapminder$pop)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 6.001e+04 2.794e+06 7.024e+06 2.960e+07 1.959e+07 1.319e+09
```

Again, it may be better to examine this on the log-scale, as populations tend to grow exponentially

```
boxplot(log(gapminder$pop), ylab="Population", las=1)
```




```
summary(log(gapminder$pop))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   14.84   15.76   15.77   16.79   21.00
```

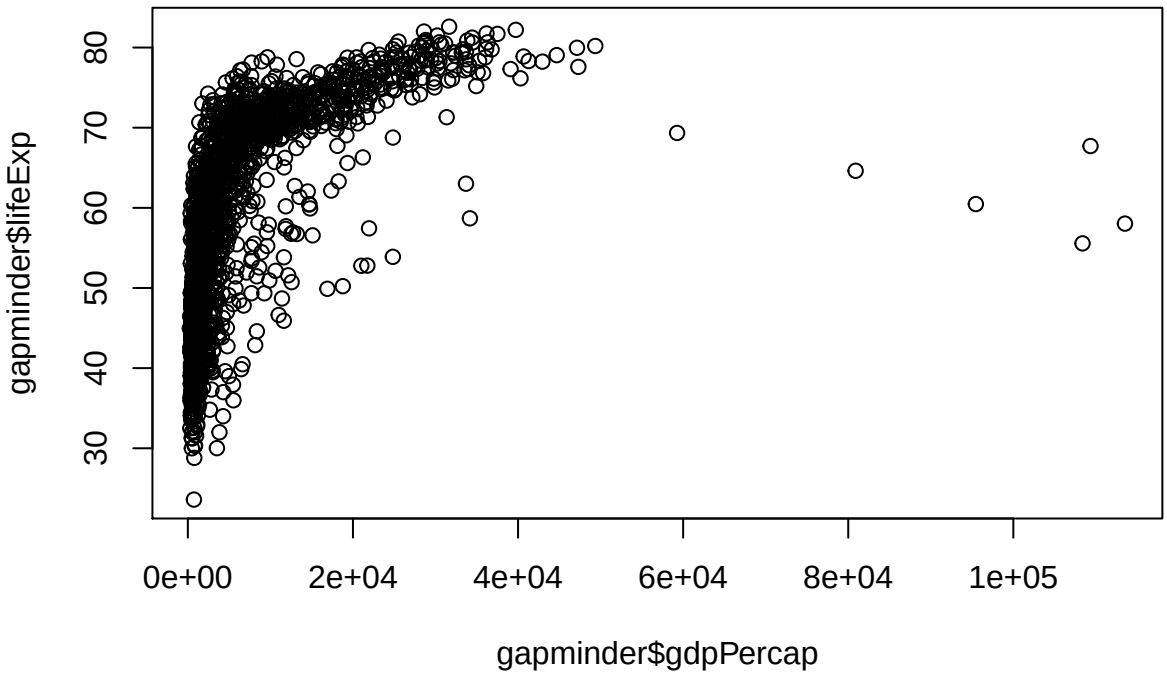
Some Bivariate Summaries

We may be interested in examining relationships between some of the variables. Following are a few of interest.

Note, we are ignoring the fact that the same countires have measurements taken for multiple years, for the time being...

Life Expectancy and GDP

It is reasonable to hypothesize that these would be related. Following is a visual examination of their relationship



Pearson's correlation:

```
cor(gapminder$gdpPercap, gapminder$lifeExp)
```

```
## [1] 0.5837062
```

It's not surprising to see they are associated. It also isn't surprising to see that it is a non-liear relationship.

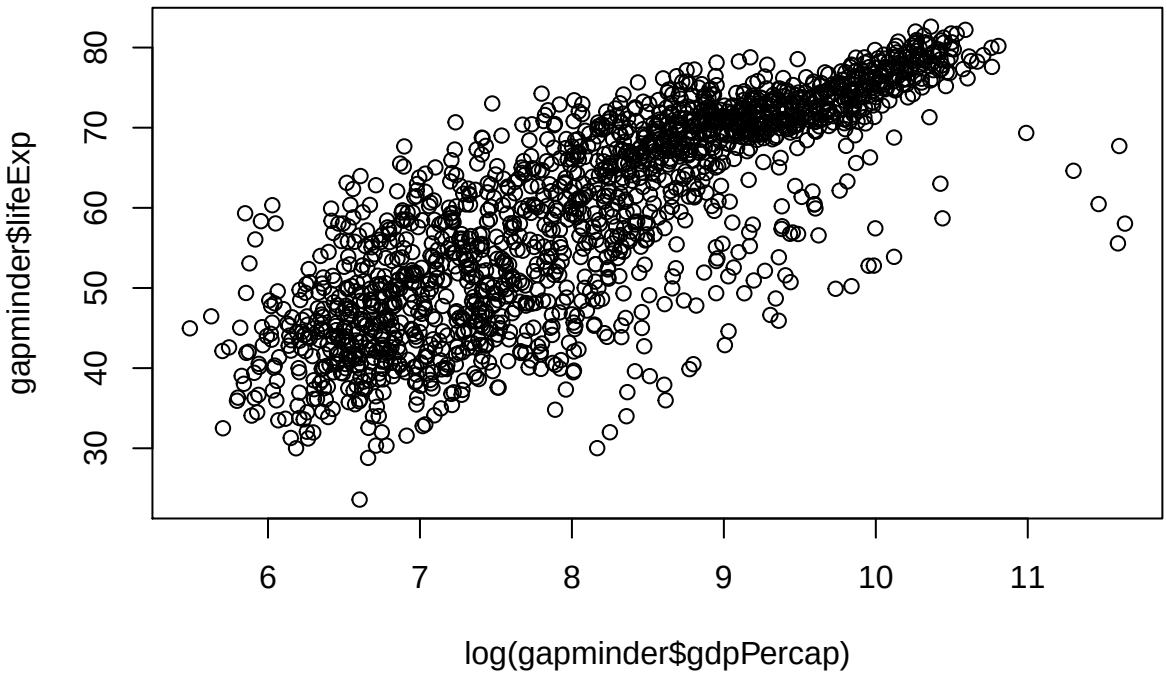
First, let's calculate Spearman's correlation to address the non-linearity, as it does appear a monotonic association:

```
cor(gapminder$gdpPercap, gapminder$lifeExp, method="spearman")
```

```
## [1] 0.8264712
```

yup, it looks much more reasonable

Let's look at the same plot, but this time using the log-GDP-per-capita, as it makes sense to examine it on this scale



That looks like it would be easier to model. Let's also take a quick look at Pearson's correlation when we use log-GDP...

```
cor(log(gapminder$gdpPercap), gapminder$lifeExp, method="pearson")
```

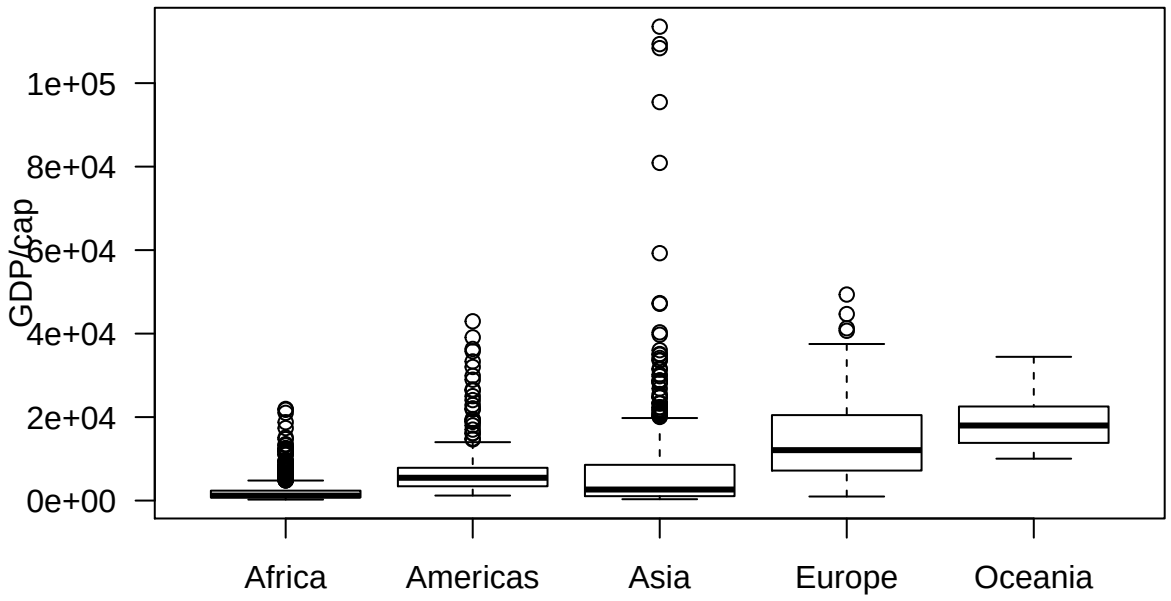
```
## [1] 0.8076179
```

Let's not get carried away... we will stop here

Continent and GDP

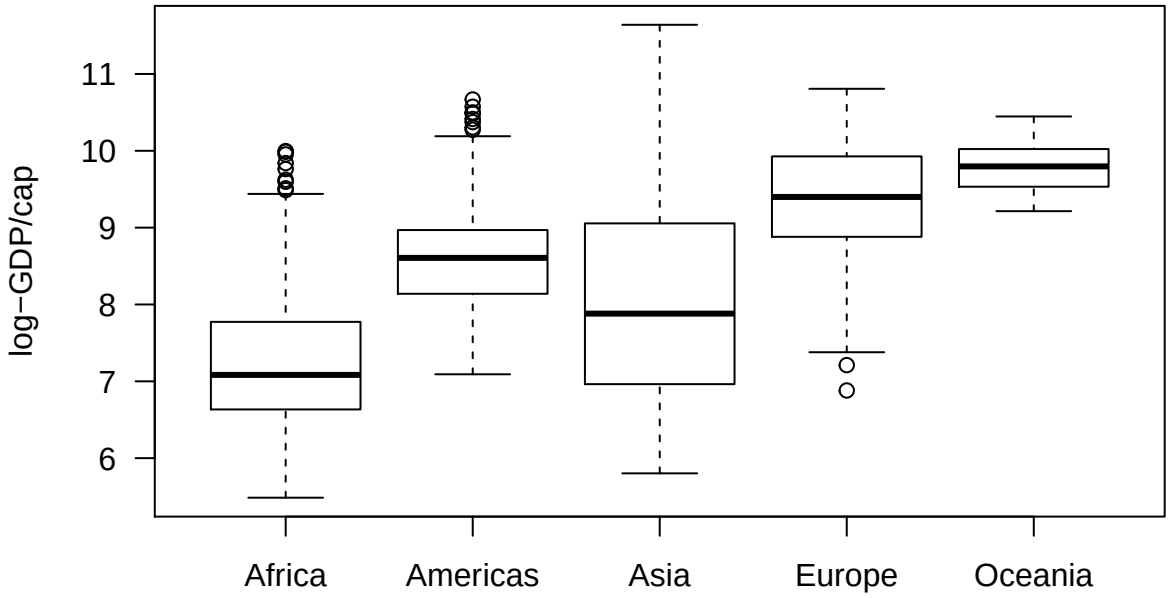
It is also reasonable to explore if GDP per captia varies by continent (again, ignoring the fact that we have measurements for multiple years)

Let's look at a plot of this



We can see an associaition. And as before, it may be more reasonable to examine GDP per capita on the *log-scale* as this variable is a **rate**

Let's look at that here...



No More...

I wont get carried away, as the point of this exercise is working with R Markdown and Git(Hub), and not the actual data analysis...although, this is a ***VERY*** interesting dataset, and I will be exploring it further on my own, for myself :)