Exploratory Data Analysis

Sebnem Er

2021-03-04

Contents

Chapter 1

Introduction

As part of the MSc specializing in Data Science, this course aims to introduce the essential techniques for performing exploratory data analysis. These techniques are typically applied before formal modeling commences and allow the researcher to discover patterns, spot anomalies, test hypotheses and check assumptions with the help of summary statistics and graphical representations. Different types of data will be described and the appropriate exploratory data analysis techniques for each data type will be introduced. The course will distinguish between univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical techniques. Special attention will focus on the visualization of large data dets using appropriate software. Some of the topics to be covered include:

1) Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots).
2) Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
3) Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.
4) Plotting geocoded data and creating dashboards
5) Dimensionality reduction and clustering of similar observations

Resources

There are some really good free online textbooks by well known and respected teachers in this area – most of the material we need can be based on these three sources:

1. Exploratory Data Analysis with R (Roger Peng = RP): https://bookdown.org/rdpeng/exdata/

2. STA545: Data wrangling, exploration, and analysis with R (Jenny Bryan = JB): https://stat545.com/index.html
3. R for data science (Hadley Wickham = HW): https://r4ds.had.co.nz/

Chapter 2

EDA Lecture 1-2 R Examples

2.1 Example 2 - Gapminder

```
#install.packages("gapminder")
library(gapminder)
gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # ... with 1,694 more rows
```

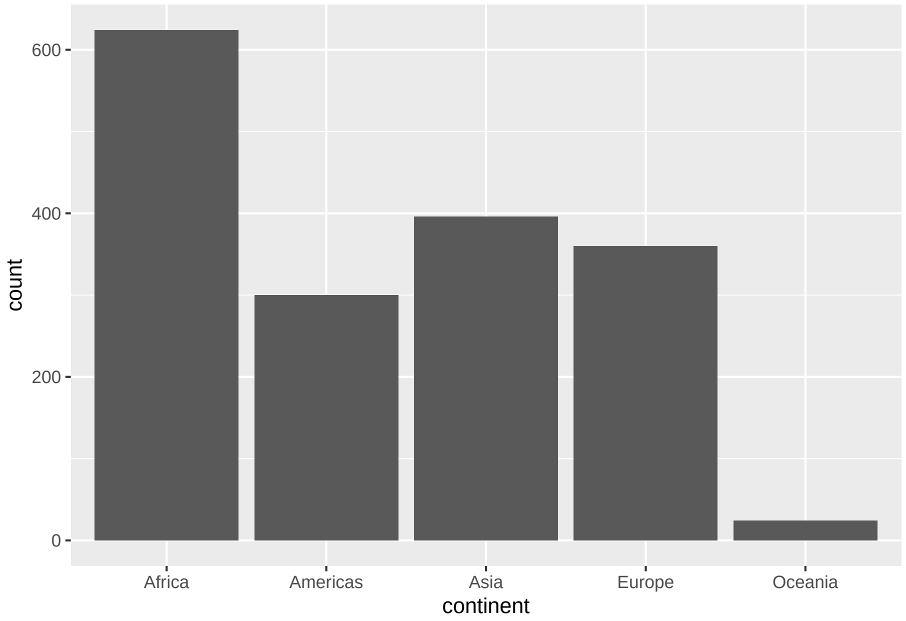
2.2 Frequency Distribution

```
library(ggplot2)
table(gapminder$continent)
```

```
##
##    Africa Americas     Asia   Europe  Oceania
##       624      300      396      360       24
```

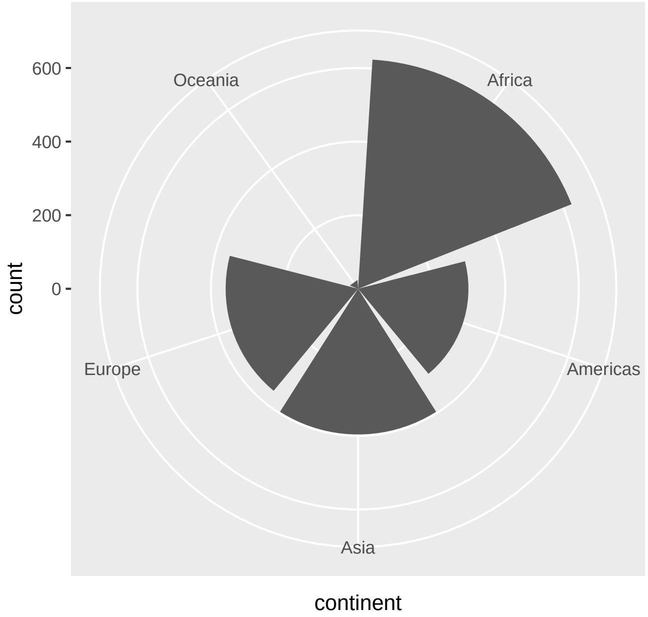
2.3 Bar Plot

```
library(ggplot2)
plot1 <- ggplot(gapminder, aes(x=continent)) + geom_bar()
plot1
```



2.4 Pie Chart

```
plot1 + coord_polar()
```

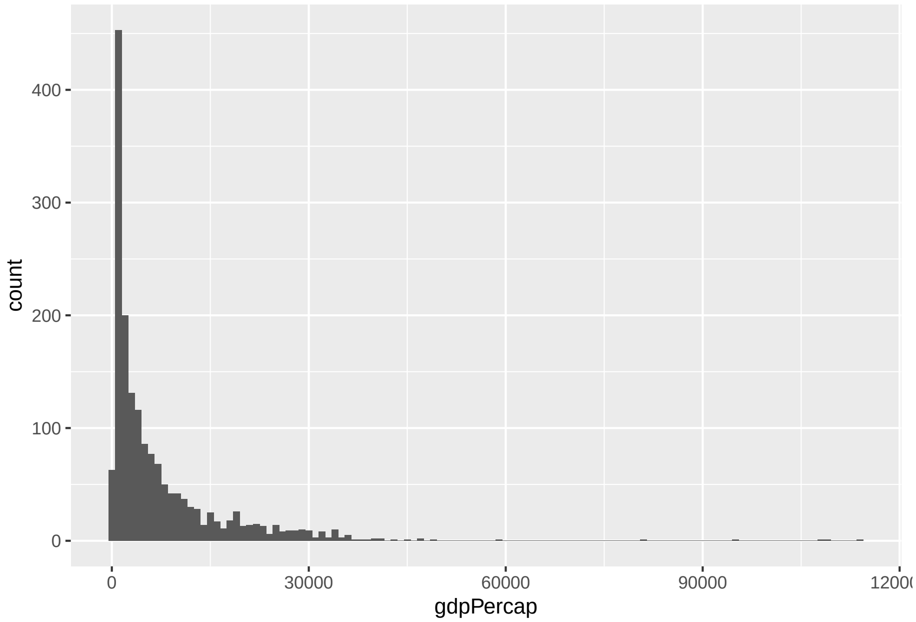


If you would like to have a regular pie chart, then you need to provide the frequency distribution.

2.5 Histogram

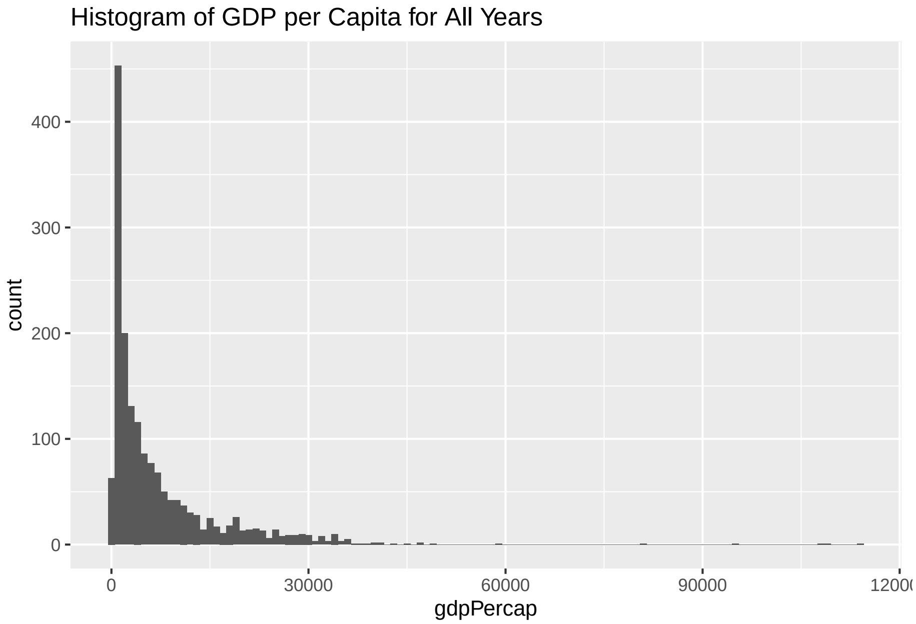
2.5.1 A Simple Histogram

```
library(ggplot2)
plot2 <- ggplot(gapminder,
                aes(x = gdpPercap))
plot2 + geom_histogram(binwidth = 1000)
```


count
400
300
200
100
0
0
30000
60000
90000
12000
gdpPercap

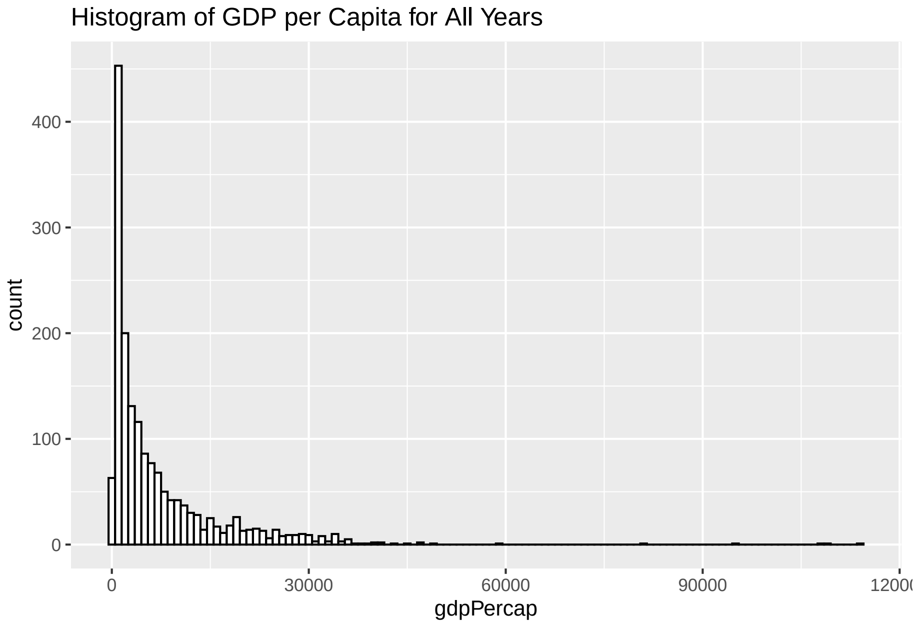
2.5.2 Histogram With a Title

```
plot2 +
  geom_histogram(binwidth = 1000) +
  labs(title = "Histogram of GDP per Capita for All Years")
```



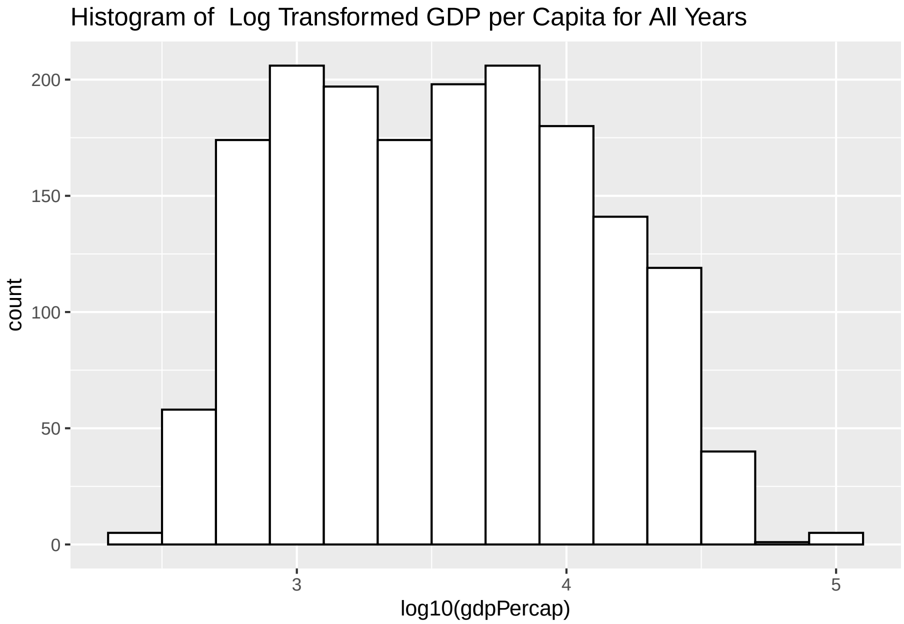
2.5.3 Histogram with Different Color Schemes:

```
plot2 +
  geom_histogram(binwidth = 1000, color="black", fill="white") +
  labs(title = "Histogram of GDP per Capita for All Years")
```



2.5.4 Histogram of Log Transformed Variable:

```
plot3 <- ggplot(gapminder,
                aes(x = log10(gdpPercap)))
plot3 +
  geom_histogram(binwidth = .2, color="black", fill="white") +
  labs(title = "Histogram of  Log Transformed GDP per Capita for All Years")
```



2.5.5 Determine the Binwidth

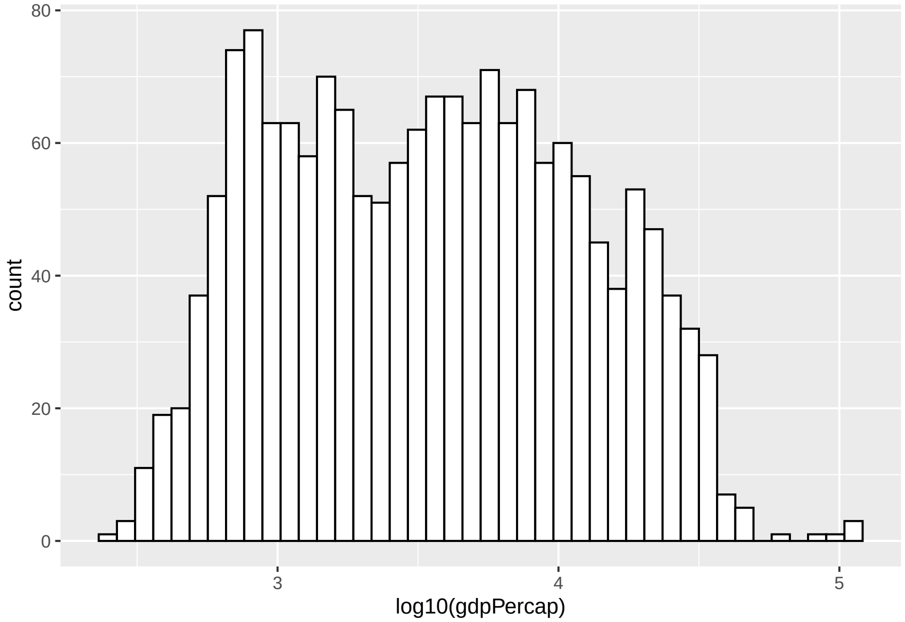
How do we determine the binwidth?

- Sturges' rule uses class intervals of length $L = \frac{x_{max}-x_{min}}{1+1.44*ln(n)}$
- Genstat rule uses uses class intervals of length: $L = \frac{x_{max}-x_{min}}{\sqrt{n}}$
- or a general rule

So we can create our own function for the binwidth:

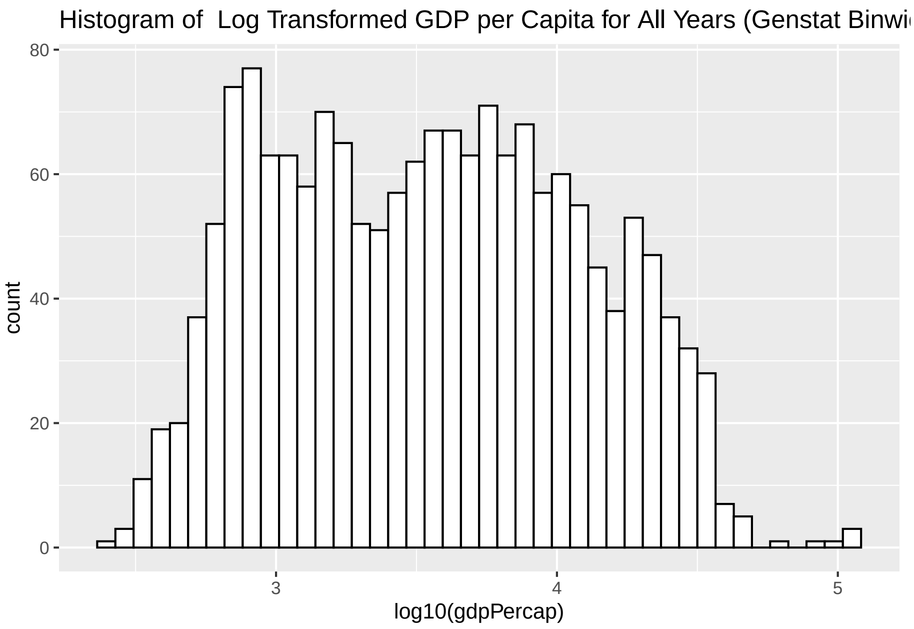
```
width_bin = function(x) (max(x)-min(x)) / sqrt(length(x))
manualbin = width_bin(log10(gapminder$gdpPercap))
```

```
plot3 +
  geom_histogram(binwidth = manualbin, color="black", fill="white")
```



or simply

```
plot3 +
  geom_histogram(binwidth = function(x) (max(x)-min(x)) / sqrt(length(x)), color="blac
    labs(title = "Histogram of  Log Transformed GDP per Capita for All Years (Genstat 
```

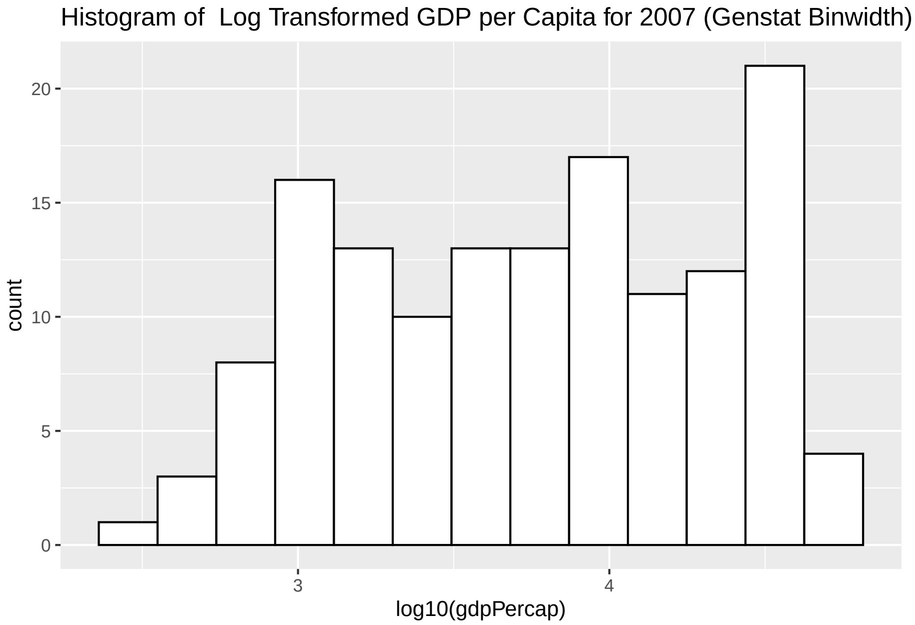


But you will notice that Gdp per capita variable includes all years, all continents, all countries!!!

2.5.6 Histogram for a Subset of Data

Log Transformed GDP per Capita for 2007:

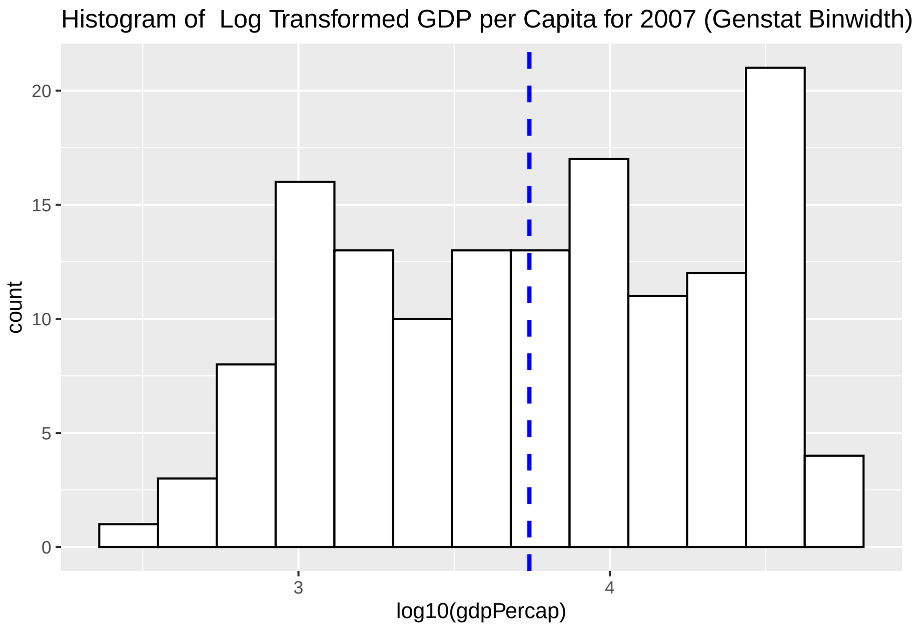
```
plot4 <- ggplot(subset(gapminder, year == 2007),
                aes(x = log10(gdpPercap)))
plot4 +
   geom_histogram(binwidth = function(x) (max(x)-min(x)) / sqrt(length(x)), color="black", fill='
  labs(title = "Histogram of  Log Transformed GDP per Capita for 2007 (Genstat Binwidth)")
```



2.5.7 Histogram with Overall Mean Line

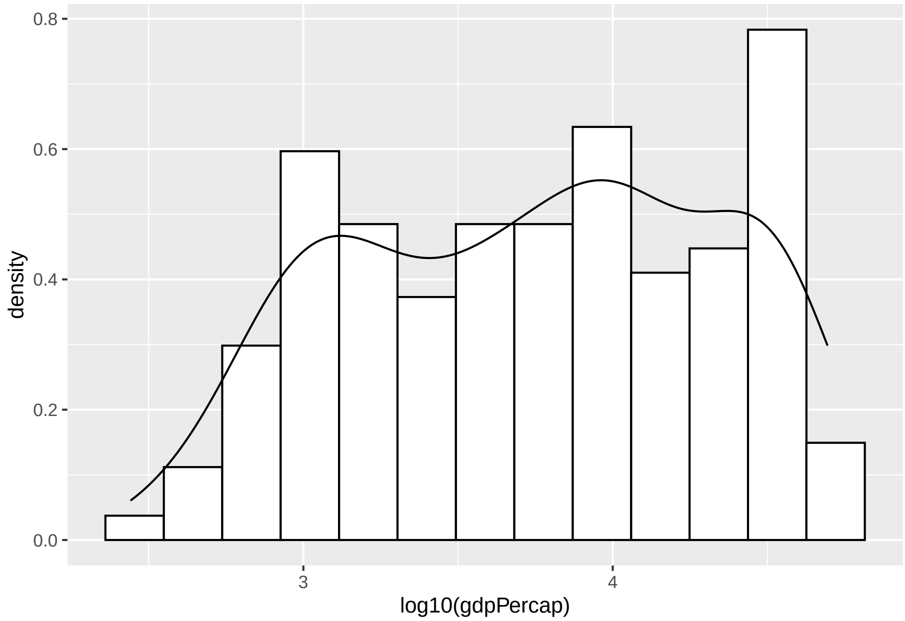
Log Transformed GDP per Capita for 2007 with the Overall Mean Line

```r
# Histogram with mean of log10(gdpPercap) on the plot
plot4 +
  geom_histogram(binwidth = function(x) (max(x)-min(x)) / sqrt(length(x)), color="blac
  geom_vline(aes(xintercept=mean(log10(gdpPercap))),
            color="blue", linetype="dashed", size=1) +
  labs(title = "Histogram of  Log Transformed GDP per Capita for 2007 (Genstat Binwidt
```



2.5.8 Histogram with Density plot

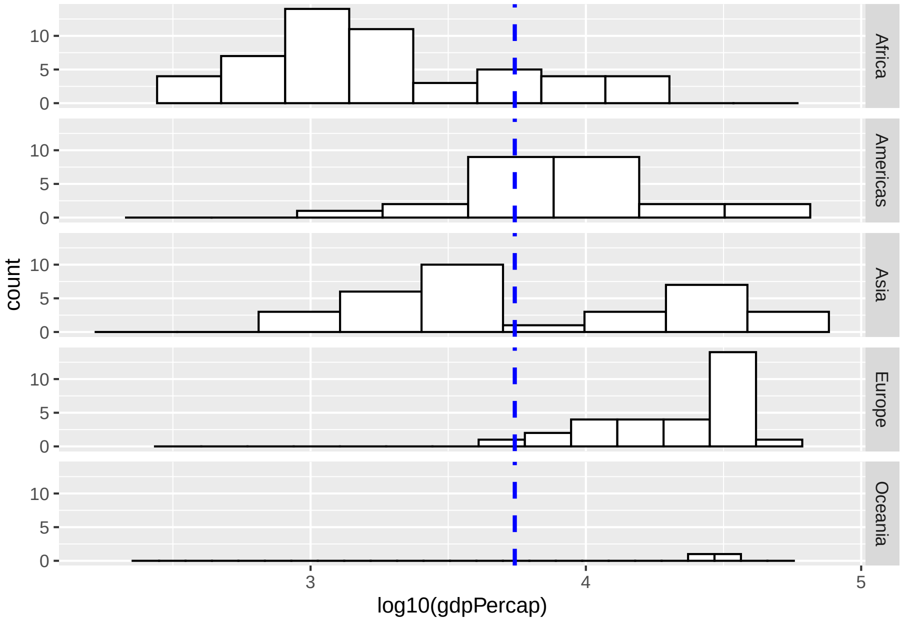
```
# Histogram with density plot
ggplot(subset(gapminder, year == 2007),
                aes(x = log10(gdpPercap))) +
  geom_histogram(aes(y=..density..), binwidth = function(x) (max(x)-min(x)) / sqrt(length(x)),col
  geom_density(alpha=0, fill="#FF6666") #alpha for transparency, if alpha = 0, no fill
```



2.5.9 Histogram with Facets

How about looking at the differences among different continents?

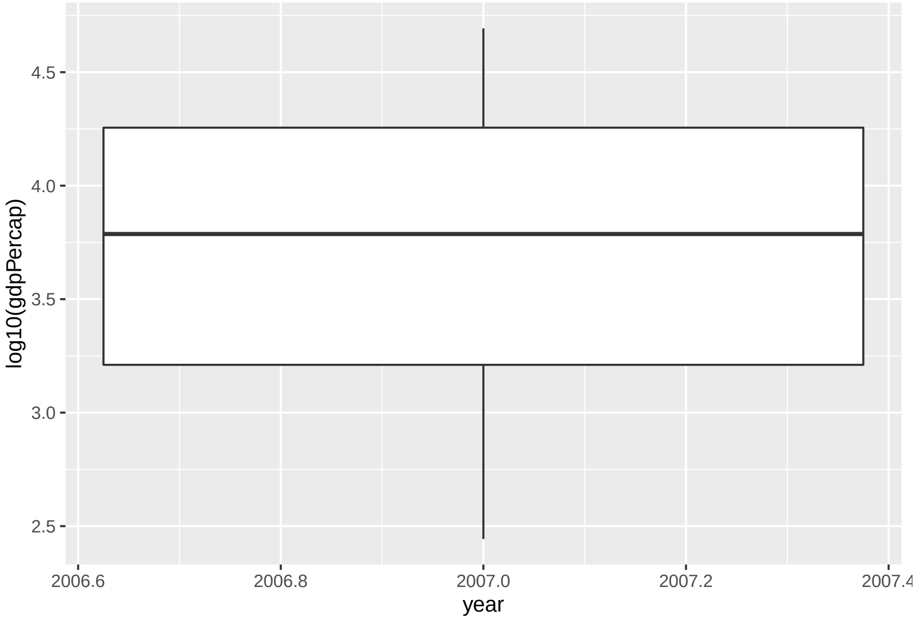
```
# Histogram with mean of log10(gdpPercap) on the plot
plot4 +
  geom_histogram(binwidth = function(x) (max(x)-min(x)) / sqrt(length(x)), color="black
  geom_vline(aes(xintercept=mean(log10(gdpPercap))),
             color="blue", linetype="dashed", size=1) +
  facet_grid(continent ~ .)
```



2.6 Boxplots

```
# Histogram with mean of log10(gdpPercap) on the plot
plot5 <- ggplot(subset(gapminder, year == 2007),
                aes(x = year, y = log10(gdpPercap)))
# if x axis variable is numeric, then one single boxplot
# if x axis variable is categorical, then works like facets

plot5 +
  geom_boxplot() #+ coord_flip()
```

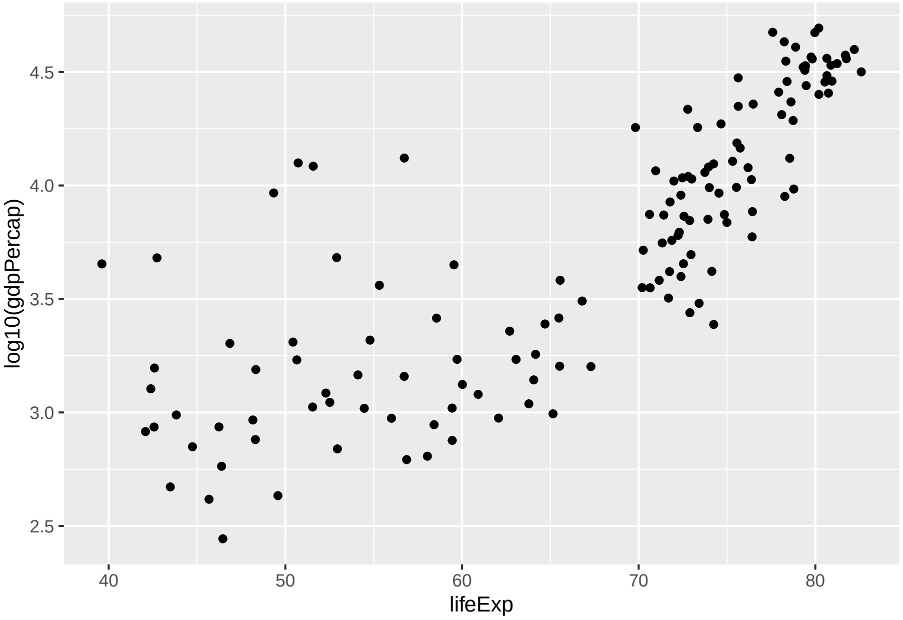


Try with "continent" variable.

2.7 Scatter Plots

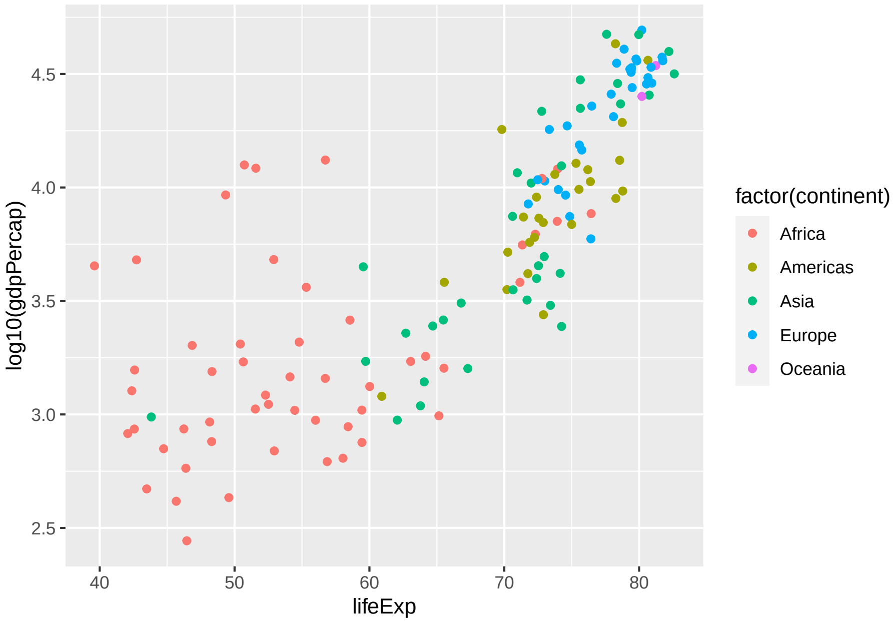
2.7.1 A Simple Scatter Plot

```
plot6 <- ggplot(subset(gapminder, year == 2007),
                aes(x = lifeExp, y = log10(gdpPercap)))
plot6 +
  geom_point()
```



2.7.2 Scatter Plot with Labellings

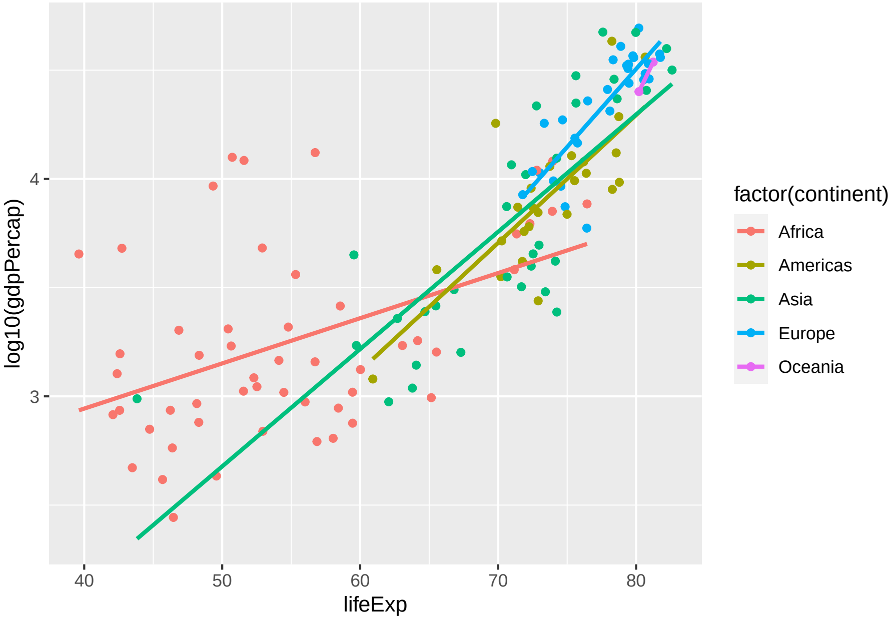
```
plot6 +
  geom_point(aes(colour = factor(continent)))
```



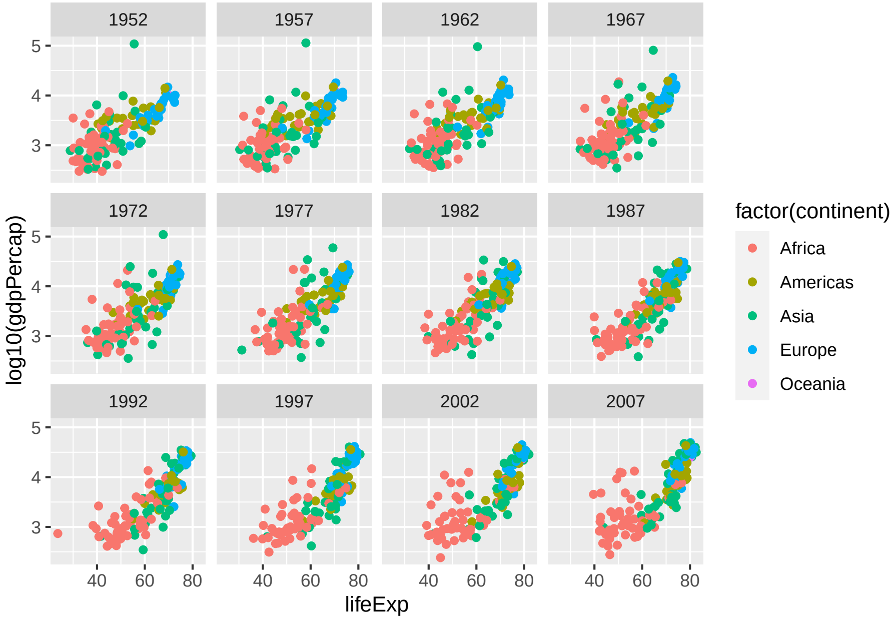
2.7.3 Scatter Plot with Linear Lines for Different Groups

```
plot6 +
  geom_point(aes(colour = factor(continent))) +
  geom_smooth(aes(group = continent, colour = factor(continent)), lwd = 1, se = FALSE, method = '
```

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



```
plot7 <- ggplot(gapminder,
                aes(x = lifeExp, y = log10(gdpPercap)))
plot7 +
  geom_point(aes(colour = factor(continent))) +
  facet_wrap(~ year) # scales = "free_x"
```



For more check "ggplot2: Elegant Graphics for Data Analysis (Use R!)" by (Hadley Wickham)