si618hw1_report_wanjun

Jun Wang

2016/10/28

SI 618 Homework 1

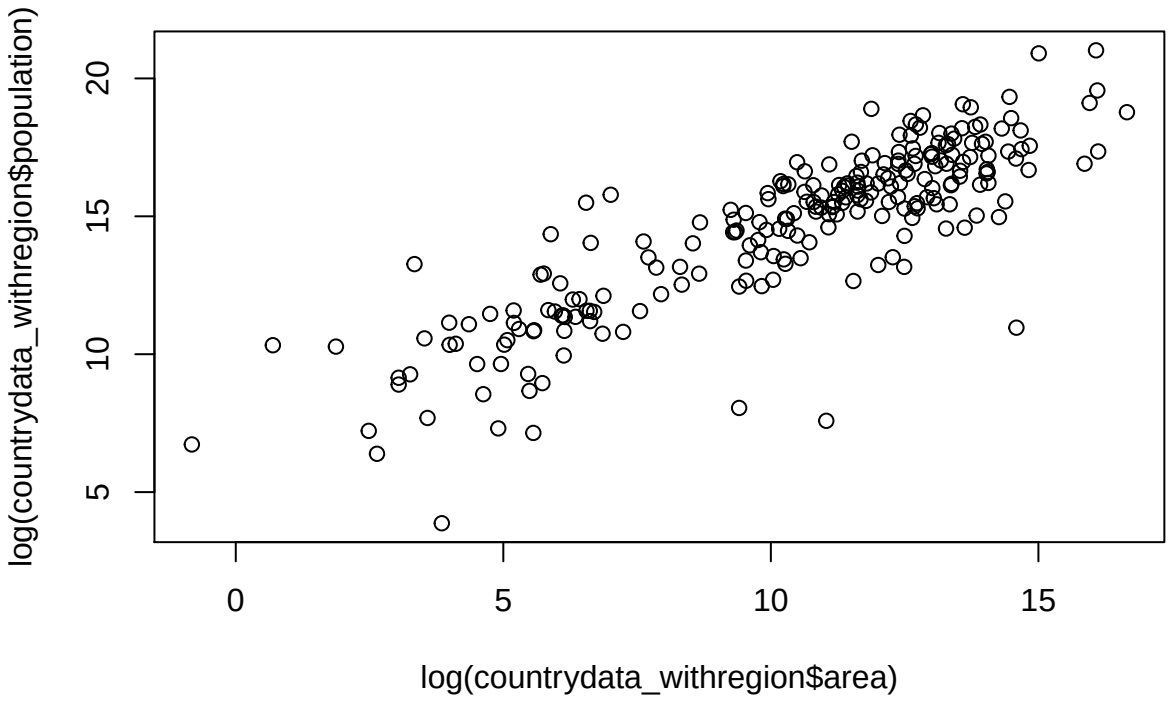
Step 1: Load data

First the provided TSV data file is loaded into R using the **read.table()** function. Here are the first 15 rows of the data frame:

```
##                country                          region      area
## 1          AFGHANISTAN                            Asia  652230.0
## 2              ALBANIA                          Europe   28748.0
## 3              ALGERIA                          Africa 2381741.0
## 4       AMERICAN SAMOA                         Oceania     199.0
## 5              ANDORRA                          Europe     468.0
## 6               ANGOLA                          Africa 1246700.0
## 7             ANGUILLA Central America & the Caribbean      91.0
## 8  ANTIGUA AND BARBUDA Central America & the Caribbean     442.6
## 9            ARGENTINA                   South America 2780400.0
## 10             ARMENIA                            Asia   29743.0
## 11               ARUBA Central America & the Caribbean     180.0
## 12           AUSTRALIA                         Oceania 7741220.0
## 13             AUSTRIA                          Europe   83871.0
## 14          AZERBAIJAN                            Asia   86600.0
## 15        BAHAMAS, THE Central America & the Caribbean   13880.0
##    population
## 1    30419928
## 2     3002859
## 3    37367226
## 4       54947
## 5       85082
## 6    18056072
## 7       15423
## 8       89018
## 9    42192494
## 10    2970495
## 11     107635
## 12   22015576
## 13    8219743
## 14    9493600
## 15     316182
```

Step 2: Scatter plot of log transformed data

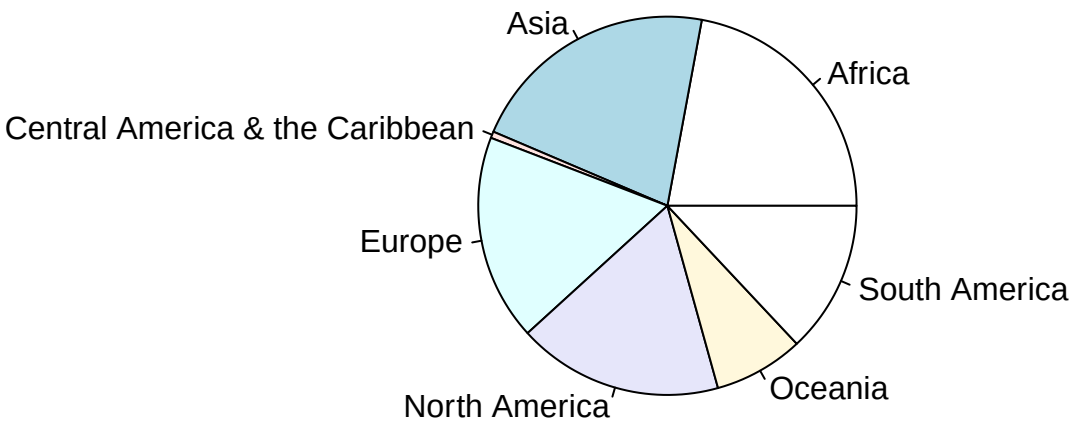
Natural logarithms of the area and the population of each country are computed and used to produce the following scatter plot using the **plot()** function.



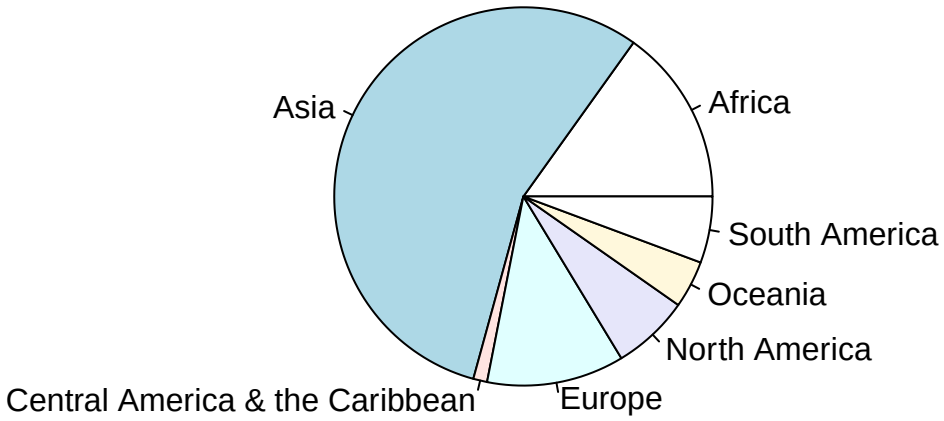
Step 3: Data aggregation by region

The areas and populations of all countries in a region are summed up using the **aggregate()** function, respectively. Then the following two pie charts are created using the **pie()** function.

Area of Regions



Population of Regions



Step 4: Visualization of Population per sq km of Regions

A new data frame is created to contain the population per sq km of each region using the **data.frame()** function. The data frame is then sorted by population per sq km in decreasing order with the help of the **order()** function. Finally, the following bar plot is created using the **barplot()** function.