CaseStudy3

Teodor Chakarov, 12141198

2022-05-30

Contents

Case Study 3 - Data Vizualization **1**

Task 1 . . . 3

Task 2 . . . 4

Task 3 . . . 5

Task 4 . . . 7

Task 5 . . . 8

Task 6 . . . 9

Task 7 . . . 10

Task 8 . . . 11

Task 9 . . . 13

Case Study 3 - Data Vizualization

```
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(forcats)
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
data <- read.csv("final_results.csv", sep=";")

head(data)
```

```
##   official_name_en ISO3166.1.Alpha.2 ISO3166.1.Alpha.3
## 1      Afghanistan               AF               AFG
## 2          Albania               AL               ALB
## 3          Algeria               DZ               DZA
## 4           Angola               AO               AGO
## 5        Argentina               AR               ARG
## 6          Armenia               AM               ARM
##   Developed...Developing.Countries Region.Name                Sub.region.Name
## 1                       Developing        Asia                  Southern Asia
## 2                        Developed      Europe                Southern Europe
## 3                       Developing      Africa                Northern Africa
## 4                       Developing      Africa             Sub-Saharan Africa
## 5                       Developing    Americas Latin America and the Caribbean
## 6                       Developing        Asia                   Western Asia
##   Median_Age youth_unempl_rate above_avg_median_age above_avg_youth_unempl_rate
## 1       19,5              17,6                    No                         Yes
## 2       34,3              31,9                   Yes                         Yes
## 3       28,9              39,3                    No                         Yes
## 4       15,9              39,4                    No                         Yes
## 5       32,4              23,7                   Yes                         Yes
## 6       36,6              36,3                   Yes                         Yes
```

```
sum(is.na(data))
```

```
## [1] 0
```

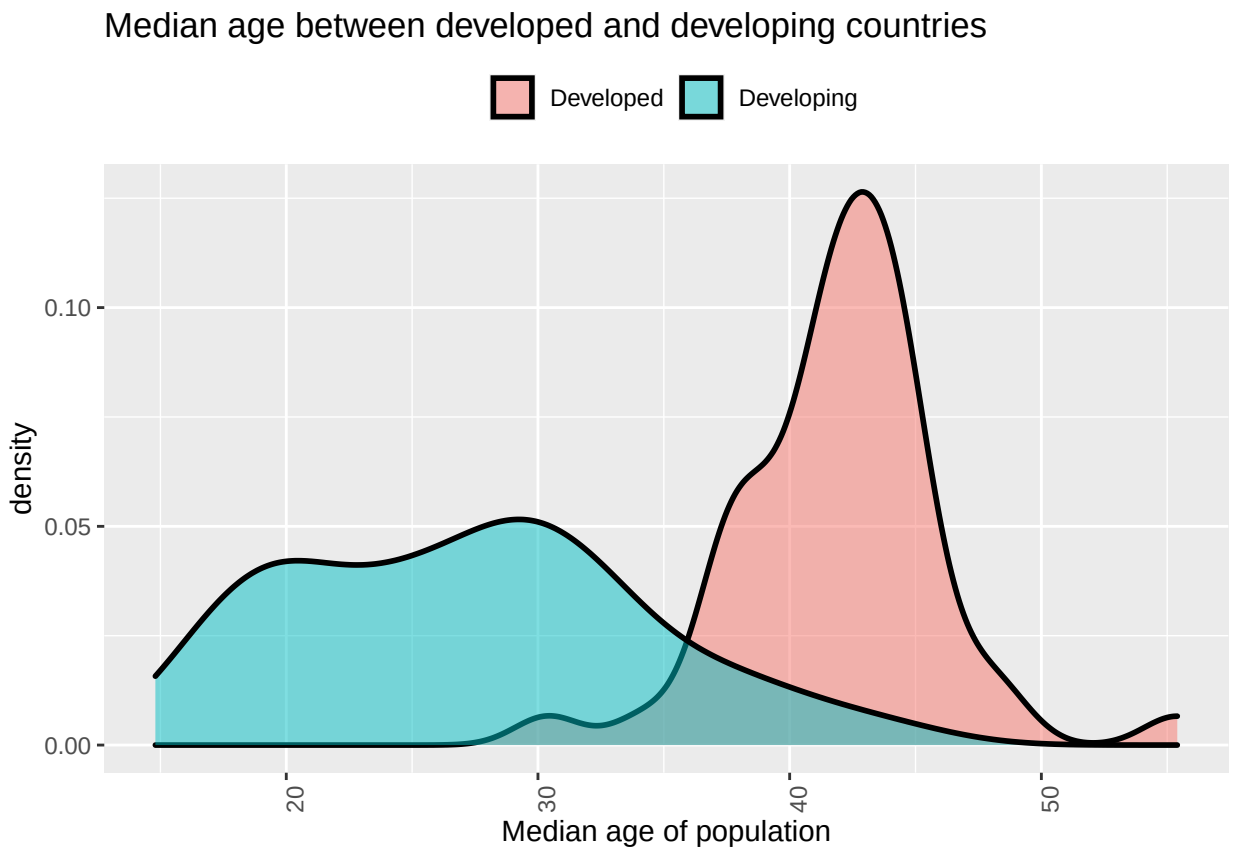
```
data$Median_Age<- gsub(",", ".", data$Median_Age)
data$Median_Age<-as.double(data$Median_Age)
```

```
developing<- data[data$Developed...Developing.Countries
 == 'Developing', ]
developed<- data[data$Developed...Developing.Countries
 == 'Developed', ]
head(developed)
```

```
##    official_name_en ISO3166.1.Alpha.2 ISO3166.1.Alpha.3
## 2           Albania                AL                ALB
## 7         Australia                AU                AUS
## 8           Austria                AT                AUT
## 14          Belarus                BY                BLR
## 15          Belgium                BE                BEL
## 18          Bermuda                BM                BMU
##    Developed...Developing.Countries Region.Name           Sub.region.Name
## 2                         Developed      Europe           Southern Europe
## 7                         Developed     Oceania Australia and New Zealand
## 8                         Developed      Europe            Western Europe
## 14                        Developed      Europe            Eastern Europe
## 15                        Developed      Europe            Western Europe
## 18                        Developed    Americas          Northern America
##    Median_Age youth_unempl_rate above_avg_median_age
## 2        34.3              31,9                  Yes
## 7        37.5              11,8                  Yes
## 8        44.5               9,4                  Yes
## 14       40.9              10,6                  Yes
## 15       41.6              15,8                  Yes
## 18       43.6              29,3                  Yes
##    above_avg_youth_unempl_rate
## 2                          Yes
## 7                           No
## 8                           No
## 14                          No
## 15                          No
## 18                         Yes
```

Task 1

```
data%>%
  ggplot(aes(x=Median_Age, color=Developed...Developing.Countries,
             fill=Developed...Developing.Countries)) +
  geom_density(alpha=0.5,size=1, color="black")+
  theme(legend.position = "top", legend.title = element_blank())+
  theme(axis.text.x = element_text(angle = 90))+
  xlab("Median age of population")+
  ggtitle("Median age between developed and developing countries")
```

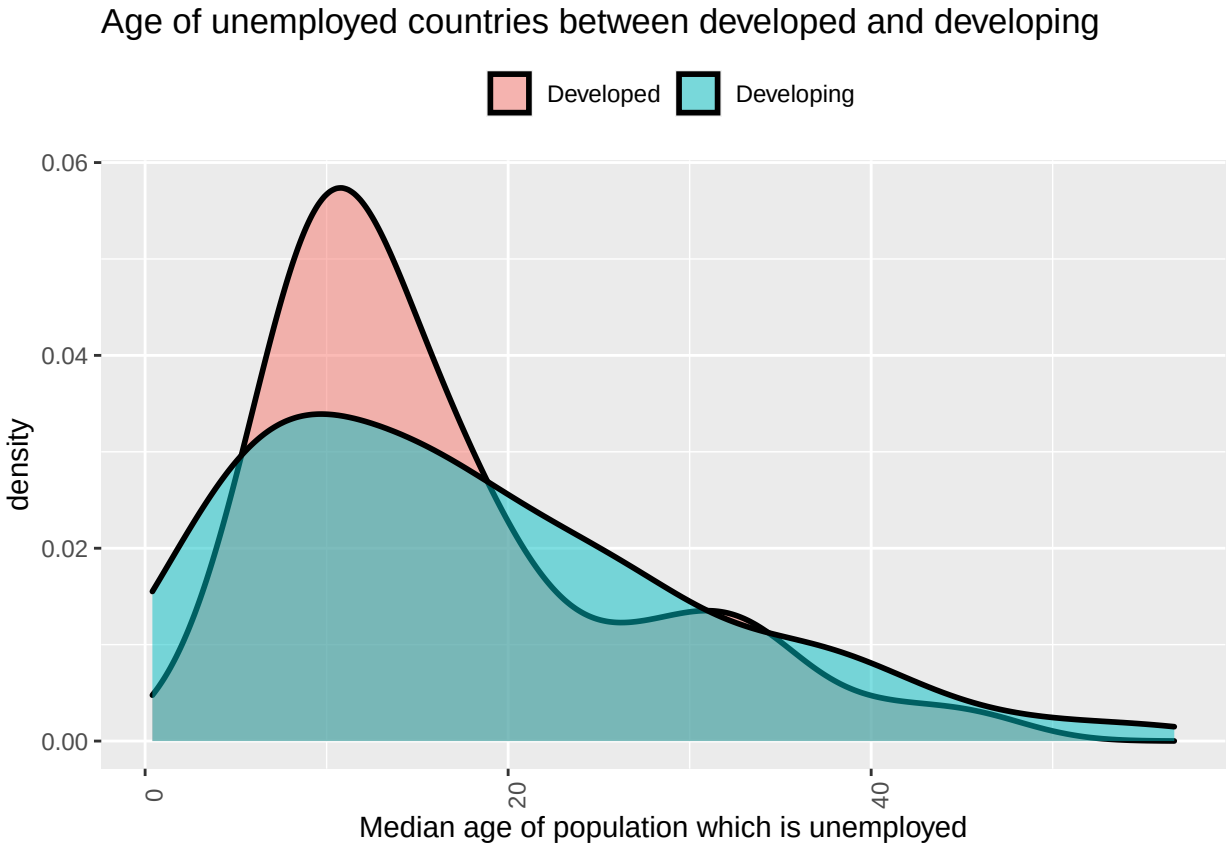


We can see that in the developed countries we have more people between 40 and 45 years old. While in developin countries we have younger people which are spread on longer range (20-30 years old).

Task 2

```
data$youth_unempl_rate<- gsub(",", ".", data$youth_unempl_rate)
data$youth_unempl_rate<-as.double(data$youth_unempl_rate)
```

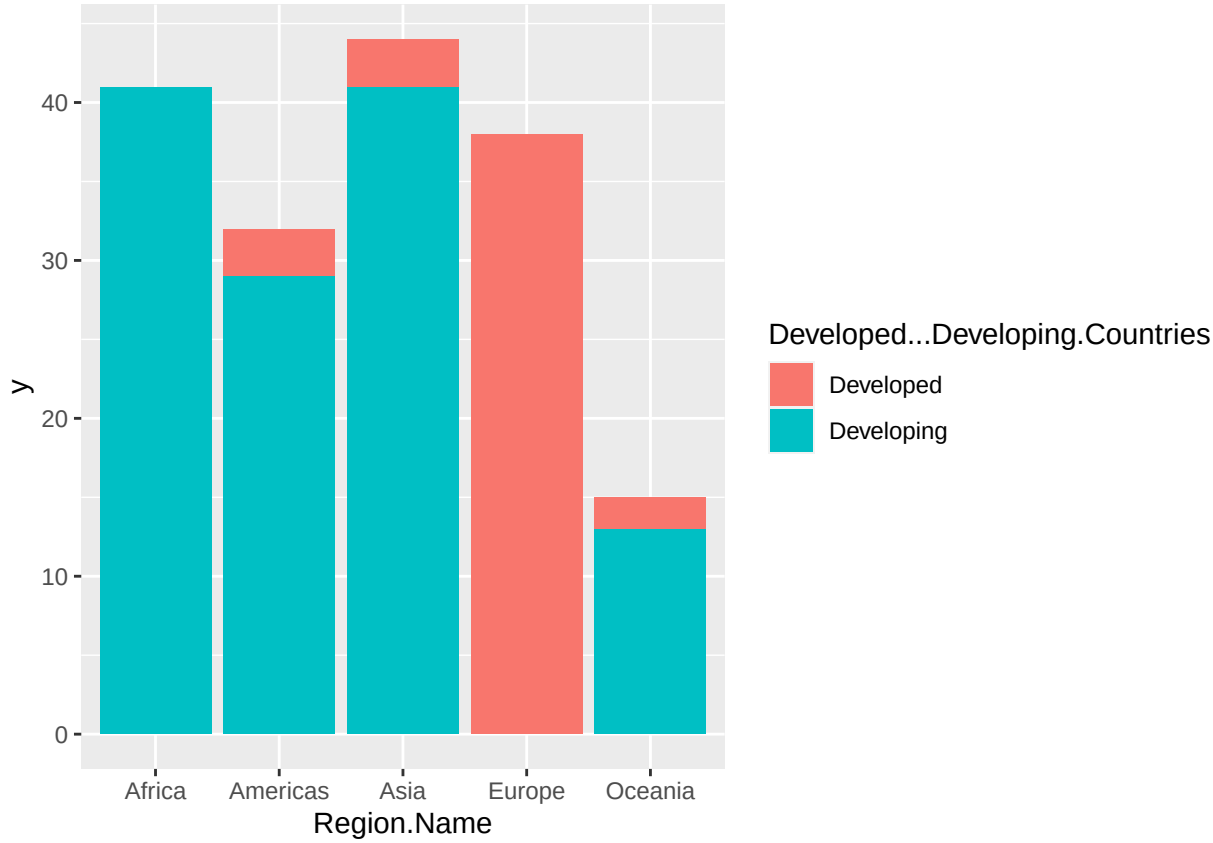
```
data%>%
  ggplot(aes(x=youth_unempl_rate, color=Developed...Developing.Countries,
             fill=Developed...Developing.Countries)) +
  geom_density(alpha=0.5,size=1, color="black")+
  theme(legend.position = "top", legend.title = element_blank())+
  theme(axis.text.x = element_text(angle = 90))+
  xlab("Median age of population which is unemployed")+
  ggtitle("Age of unemployed countries between developed and developing")
```



We can see that in developed countries, the unemployment is big in range 10 to 15. While in developing countries we have smaller peak at that range but more spread between other years.

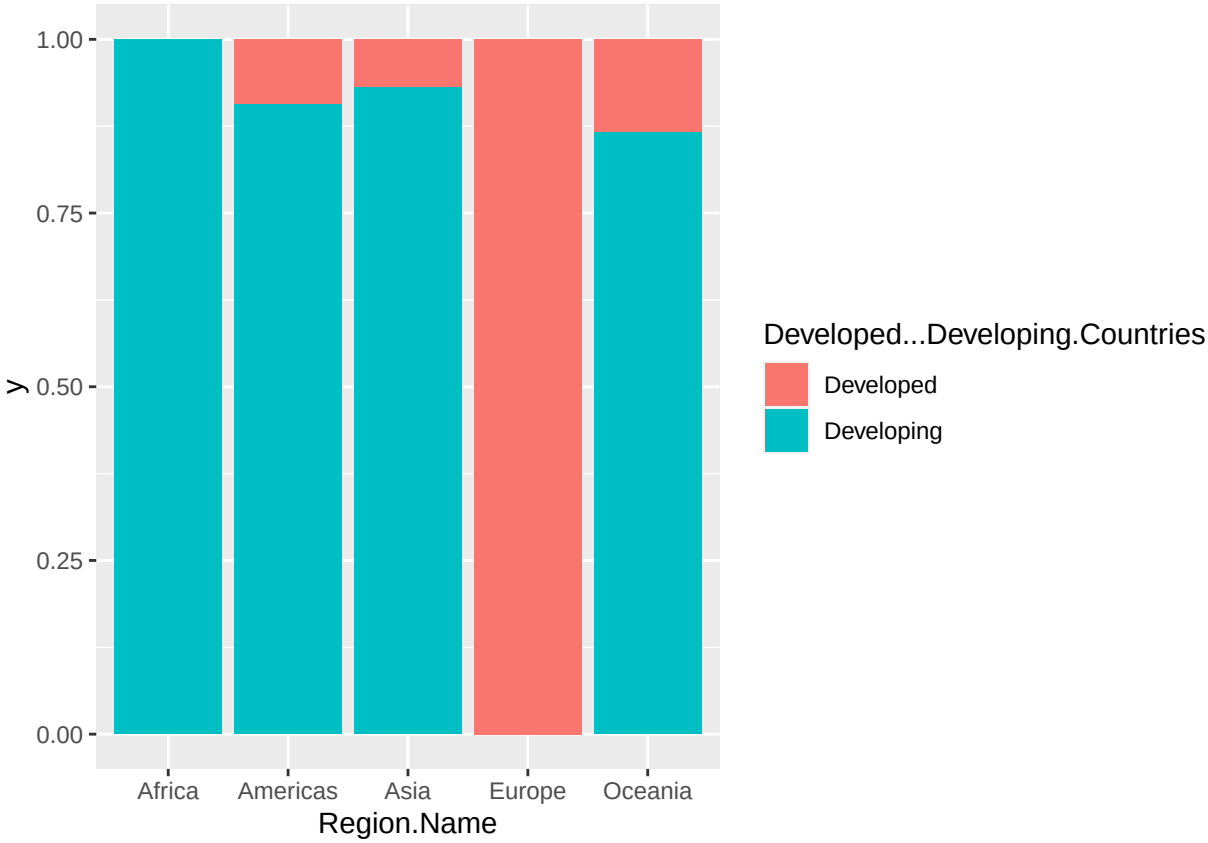
Task 3

```
selected<-data %>%
        select(Developed...Developing.Countries,Region.Name)
ggplot(data=selected,aes(x=Region.Name, y=1,
                                  fill=Developed...Developing.Countries))+
                                geom_bar(stat = 'identity')
```



Here we can see how many countries are developed and developing in each region.

```
ggplot(data=selected,aes(x=Region.Name, y=1,
                                  fill=Developed...Developing.Countries))+
  geom_bar(position="fill", stat="identity")
```



This is the same but its with relative frequencies. The values are shrinked to 1.

Task 4

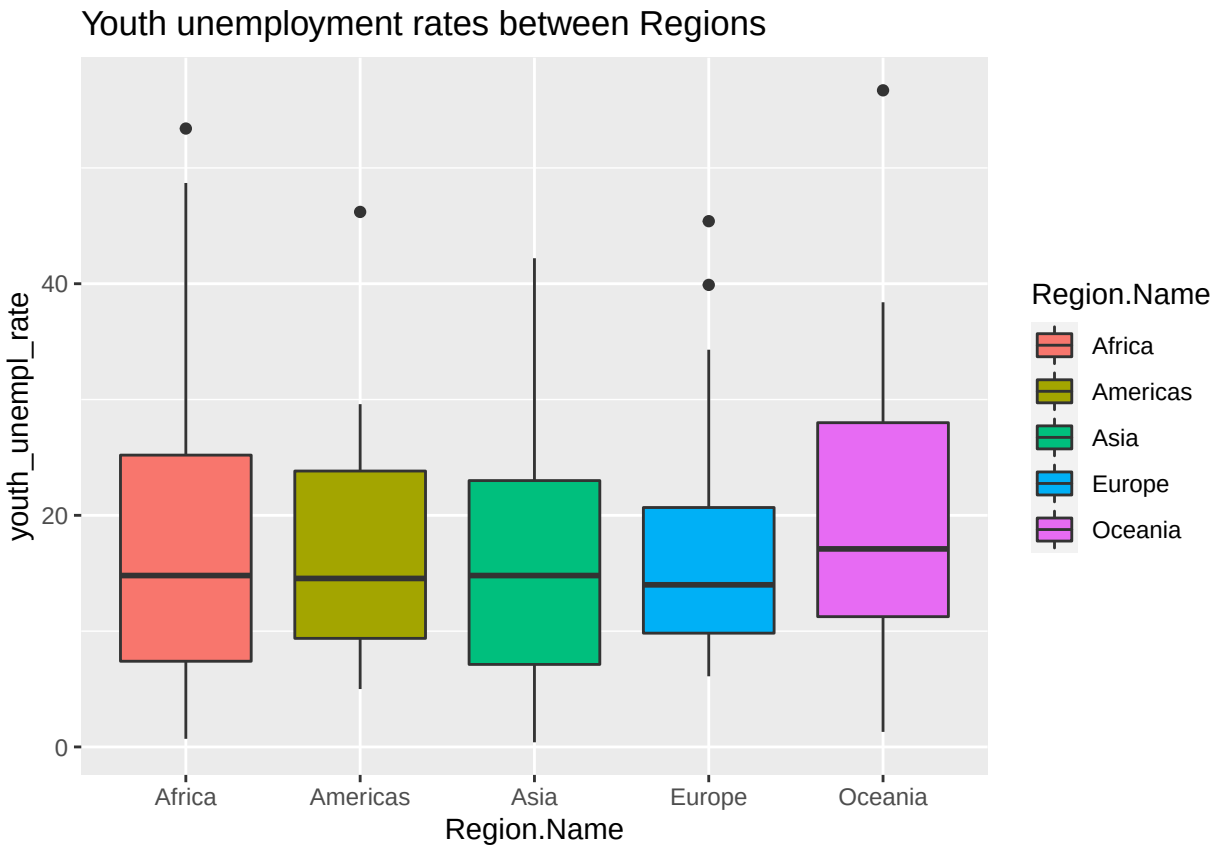
```
ggplot(data=data, aes(x=youth_unempl_rate, y=Median_Age, color=Developed...Developing.Countries))+
geom_point(size=2) +
stat_smooth(method = "lm", geom = "smooth")+
  ggtitle("Uneployment rate and median age")
```

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

We can see that in developed countries, we have higher median age values within unemployment rate between 0 and 20. Also values for developing countries are more spread than developed once.

Task 5

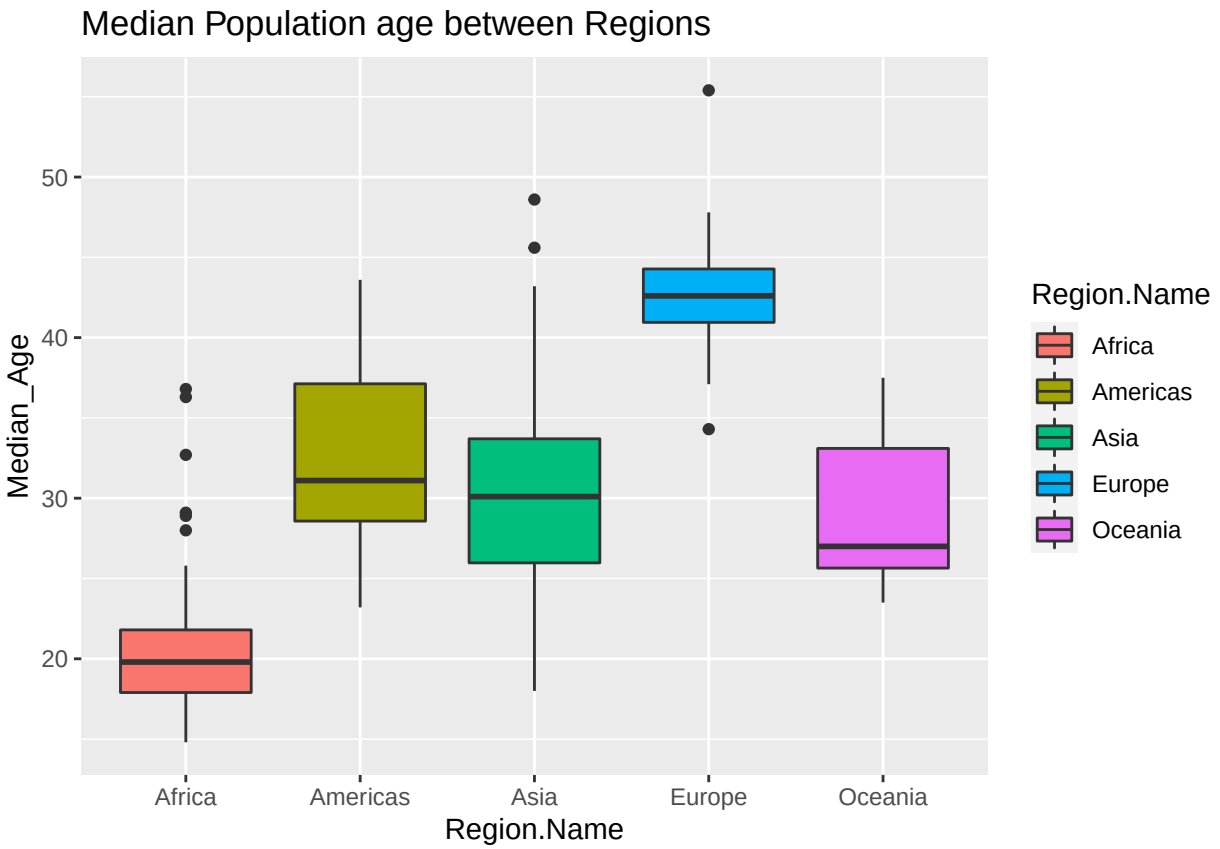
```
data$Region.Name <- factor(data$Region.Name)
ggplot(data=data, aes(y= youth_unempl_rate, x=Region.Name, fill=Region.Name ) ) + geom_boxplot() + ggti
```



Seeing the boxplots I don't see big differences in the average value of unemployment rates between each region. We can see outliers and short distribution in Europe and large in Oceania and Africa.

Task 6

```
ggplot(data=data, aes(y= Median_Age, x=Region.Name, fill=Region.Name ) ) + geom_boxplot() + ggtitle("Me
```



Here th difference is bigger than the previous. We see low meadan populaiton age in Africa then Oceania. In Africa the distribution of the values is small which means that the population is with young people. We see almost equal mean age between Americans and Asia, but America has Older population. Europe has the oldes population in comparison to others.

Task 7

```
cbbPalette <- c("#000000","#E69F00", "#56B4E9", "#009E73", "#F0E442")

data  %>%
      group_by(Sub.region.Name, Region.Name) %>%
      summarise_at(vars(Median_Age), list(Avg = mean)) %>%
      mutate(Sub.region.Name=fct_reorder(Sub.region.Name, Avg))%>%
      ggplot(aes(x=Avg, y=Sub.region.Name) ) +
      geom_point(size=3, aes(color=Region.Name)) +
      scale_fill_manual(values=cbbPalette)+
      xlab("Average Age") +
      ylab("Sub Region") +
      ggtitle("Average Age in each Sub region")
```

We can see the youngest people are in sub region of Sub-Saharan Africa and as expected the oldest are in Western Europe.

Task 8

```
new_data <- read.csv("new_data.csv")
head(new_data)
```

```
##                ï..Country.Name Country.Code    Indicator.Name Indicator.Code
## 1                        Aruba          ABW Population, total    SP.POP.TOTL
## 2 Africa Eastern and Southern          AFE Population, total    SP.POP.TOTL
## 3                  Afghanistan          AFG Population, total    SP.POP.TOTL
## 4  Africa Western and Central          AFW Population, total    SP.POP.TOTL
## 5                       Angola          AGO Population, total    SP.POP.TOTL
## 6                      Albania          ALB Population, total    SP.POP.TOTL
##       X1960     X1961     X1962     X1963     X1964     X1965     X1966
## 1     54208     55434     56234     56699     57029     57357     57702
## 2 130836765 134159786 137614644 141202036 144920186 148769974 152752671
## 3   8996967   9169406   9351442   9543200   9744772   9956318  10174840
## 4  96396419  98407221 100506960 102691339 104953470 107289875 109701811
## 5   5454938   5531451   5608499   5679409   5734995   5770573   5781305
## 6   1608800   1659800   1711319   1762621   1814135   1864791   1914573
##       X1967     X1968     X1969     X1970     X1971     X1972     X1973
## 1     58044     58377     58734     59070     59442     59849     60236
```

```
## 2 156876454 161156430 165611760 170257189 175100167 180141148 185376550
## 3  10399936  10637064  10893772  11173654  11475450  11791222  12108963
## 4 112195950 114781116 117468741 120269044 123184308 126218502 129384954
## 5   5774440   5771973   5803677   5890360   6041239   6248965   6497283
## 6   1965598   2022272   2081695   2135479   2187853   2243126   2296752
##       X1974     X1975     X1976     X1977     X1978     X1979     X1980
## 1     60527     60653     60586     60366     60102     59972     60097
## 2 190800796 196409937 202205766 208193045 214368393 220740384 227305945
## 3  12412960  12689164  12943093  13171294  13341199  13411060  13356500
## 4 132699537 136173544 139813171 143615715 147571063 151663853 155882270
## 5   6761623   7023994   7279630   7533814   7790774   8058112   8341290
## 6   2350124   2404831   2458526   2513546   2566266   2617832   2671997
##       X1981     X1982     X1983     X1984     X1985     X1986     X1987
## 1     60561     61341     62213     62826     63024     62645     61838
## 2 234058404 240999134 248146290 255530063 263161451 271050065 279184536
## 3  13171679  12882518  12537732  12204306  11938204  11736177  11604538
## 4 160223588 164689764 169279422 173991851 178826553 183785612 188868567
## 5   8640478   8952971   9278104   9614756   9961993  10320116  10689247
## 6   2726056   2784278   2843960   2904429   2964762   3022635   3083605
##       X1988     X1989     X1990     X1991     X1992     X1993     X1994
## 1     61072     61033     62152     64623     68240     72495     76705
## 2 287524258 296024639 304648010 313394693 322270073 331265579 340379934
## 3  11618008  11868873  12412311  13299016  14485543  15816601  17075728
## 4 194070079 199382783 204803865 210332267 215976366 221754806 227692136
## 5  11068051  11454784  11848385  12248901  12657361  13075044  13503753
## 6   3142336   3227943   3286542   3266790   3247039   3227287   3207536
##       X1995     X1996     X1997     X1998     X1999     X2000     X2001
## 1     80324     83211     85450     87280     89009     90866     92892
## 2 349605660 358953595 368440591 378098393 387977990 398113044 408522129
## 3  18110662  18853444  19357126  19737770  20170847  20779957  21606992
## 4 233807627 240114179 246613750 253302310 260170348 267214544 274433894
## 5  13945205  14400722  14871572  15359600  15866871  16395477  16945753
## 6   3187784   3168033   3148281   3128530   3108778   3089027   3060173
##       X2002     X2003     X2004     X2005     X2006     X2007     X2008
## 1     94992     97016     98744    100028    100830    101226    101362
## 2 419223717 430246635 441630149 453404076 465581372 478166911 491173160
## 3  22600774  23680871  24726689  25654274  26433058  27100542  27722281
## 4 281842480 289469530 297353098 305520588 313985474 322741656 331772330
## 5  17519418  18121477  18758138  19433604  20149905  20905360  21695636
## 6   3051010   3039616   3026939   3011487   2992547   2970017   2947314
##       X2009     X2010     X2011     X2012     X2013     X2014     X2015
## 1    101452    101665    102050    102565    103165    103776    104339
## 2 504604672 518468229 532760424 547482863 562601578 578075373 593871847
## 3  28394806  29185511  30117411  31161378  32269592  33370804  34413603
## 4 341050537 350556886 360285439 370243017 380437896 390882979 401586651
## 5  22514275  23356247  24220660  25107925  26015786  26941773  27884380
## 6   2927519   2913021   2905195   2900401   2895092   2889104   2880703
##       X2016     X2017     X2018     X2019     X2020 X2021  X
## 1    104865    105361    105846    106310    106766    NA NA
## 2 609978946 626392880 643090131 660046272 677243299    NA NA
## 3  35383028  36296111  37171922  38041757  38928341    NA NA
## 4 412551299 423769930 435229381 446911598 458803476    NA NA
## 5  28842482  29816769  30809787  31825299  32866268    NA NA
## 6   2876101   2873457   2866376   2854191   2837849    NA NA
```

```
selected_data <- new_data %>% select("ï..Country.Name", "Country.Code", "X2020")
head(selected_data)
```

```
##               ï..Country.Name Country.Code     X2020
## 1                       Aruba          ABW    106766
## 2 Africa Eastern and Southern          AFE 677243299
## 3                 Afghanistan          AFG  38928341
## 4  Africa Western and Central          AFW 458803476
## 5                      Angola          AGO  32866268
## 6                     Albania          ALB   2837849
```

```
final_data <-left_join(data,selected_data,by = c("ISO3166.1.Alpha.3"="Country.Code"))
head(final_data)
```

```
##   official_name_en ISO3166.1.Alpha.2 ISO3166.1.Alpha.3
## 1      Afghanistan               AF               AFG
## 2          Albania               AL               ALB
## 3          Algeria               DZ               DZA
## 4           Angola               AO               AGO
## 5        Argentina               AR               ARG
## 6          Armenia               AM               ARM
##   Developed...Developing.Countries Region.Name                Sub.region.Name
## 1                       Developing        Asia                  Southern Asia
## 2                        Developed      Europe                Southern Europe
## 3                       Developing      Africa                Northern Africa
## 4                       Developing      Africa             Sub-Saharan Africa
## 5                       Developing    Americas Latin America and the Caribbean
## 6                       Developing        Asia                     Western Asia
##   Median_Age youth_unempl_rate above_avg_median_age above_avg_youth_unempl_rate
## 1       19.5              17.6                    No                          Yes
## 2       34.3              31.9                   Yes                          Yes
## 3       28.9              39.3                    No                          Yes
## 4       15.9              39.4                    No                          Yes
## 5       32.4              23.7                   Yes                          Yes
## 6       36.6              36.3                   Yes                          Yes
##   ï..Country.Name    X2020
## 1     Afghanistan 38928341
## 2         Albania  2837849
## 3         Algeria 43851043
## 4          Angola 32866268
## 5       Argentina 45376763
## 6         Armenia  2963234
```

Task 9

```
result <- final_data %>%
  ggplot(aes(x=youth_unempl_rate, y=Median_Age, color=Developed...Developing.Countries,
             size = X2020, text=official_name_en))+
  geom_point()

ggplotly(result, tooltip = c("text", "Median_age", "youth_unempl_rate", "X2020"))
```

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

Again the same scatter plot as Task 4 but this time we have 3rd dimension based on the populaiton of the country in 2020.