Case Study 3 AKSTA Statistical Computing

Hanna Kienast Iulia Mihaela Enache Kateryna Ponomarenko

2024-05-25

Load the data set you exported in the final Task of Case Study 2.

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.3.2
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
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(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.3
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.3.2

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```r
library(knitr)
require(stringr)
```

```
## Loading required package: stringr
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

In our csv the missing values are represented by dot character ".", so we will specify it when parsing.

```r
data <- read.csv("final_data.csv", sep = ";", stringsAsFactors = FALSE, na.strings = ".")

data <- data %>%
  mutate(across(where(is.character), trimws))

print(head(data))
```

```
##          ISO                   Country continent       sub.region Income_class
## 1 MC|MCO|492                    Monaco    Europe   Western Europe            H
## 2 JP|JPN|392                     Japan      Asia     Eastern Asia            H
## 3 PM|SPM|666 Saint Pierre and Miquelon  Americas Northern America         <NA>
## 4 DE|DEU|276                   Germany    Europe   Western Europe            H
## 5 IT|ITA|380                     Italy    Europe  Southern Europe            H
## 6 AD|AND|020                   Andorra    Europe  Southern Europe            H
##   Median_Age Youth_unemployment_rate Migration_rate
## 1       55.4                    26.6            8.3
## 2       48.6                     3.6            0.0
## 3       48.5                      NA           -7.7
## 4       47.8                     6.2            1.5
## 5       46.5                    32.2            3.2
## 6       46.2                      NA            0.0
```

```r
print(paste("Number of observations:", nrow(data)))
```

```
## [1] "Number of observations: 227"
```

Eliminate all observations with missing values in the income status variable.

```r
clean_data <- data %>%
  filter(!is.na(Income_class))


print(head(clean_data))
```

```
##          ISO Country continent      sub.region Income_class Median_Age
## 1 MC|MCO|492  Monaco    Europe  Western Europe            H       55.4
## 2 JP|JPN|392   Japan      Asia    Eastern Asia            H       48.6
## 3 DE|DEU|276 Germany    Europe  Western Europe            H       47.8
## 4 IT|ITA|380   Italy    Europe Southern Europe            H       46.5
```

```
## 5 AD|AND|020 Andorra    Europe Southern Europe            H       46.2
## 6 GR|GRC|300  Greece    Europe Southern Europe            H       45.3
##   Youth_unemployment_rate Migration_rate
## 1                    26.6            8.3
## 2                     3.6            0.0
## 3                     6.2            1.5
## 4                    32.2            3.2
## 5                      NA            0.0
## 6                    39.9            0.9
```

```
print(paste("Number of observations:", nrow(clean_data)))
```

```
## [1] "Number of observations: 184"
```

Also, we had a problem with missing ISO codes for Macedonia and Turkey in our exported in case study 2 data set, but these 2 countries are already eliminated from analyses, since they had missing values in the income status variable.

a. Median age in different income levels

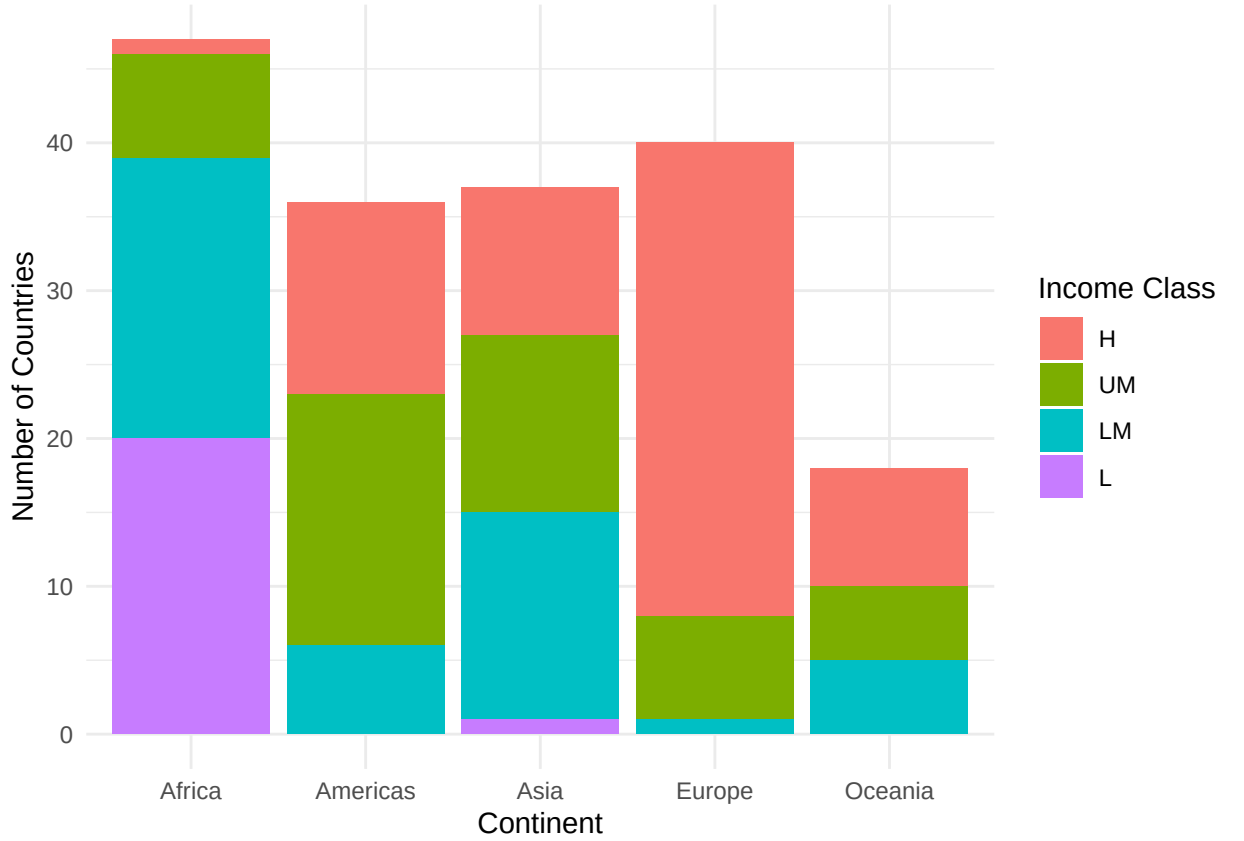
```
clean_data <- clean_data %>%
  mutate(Income_class = factor(Income_class, levels = c("H", "UM", "LM", "L")))

density_plot <- ggplot(clean_data, aes(x = Median_Age, fill = Income_class)) +
  geom_density(alpha = 0.5, color = "black") +
  scale_fill_manual(values = c("H" = "red", "UM" = "green", "LM" = "blue", "L" ="orange")) +
  labs(x = "Median age of population", fill = element_blank()) +
  theme(legend.position = "top", legend.title = element_blank())

print(density_plot)
```

The plot demonstrates that the low-income countries have the smallest spread of median age of the population, while countries of other income statutes have the wider spread. Also, high level income countries tend to have the highest median age among represented categories, indicating an older population, while the opposite situation is observed for countries of low income. Overall, we can spot a tendency that the higher income level country has, the higher median age of population it owns.

b. Income status in different continents

Using ggplot2, create a stacked barplot of absolute frequencies showing how the entities are split into continents and income status.

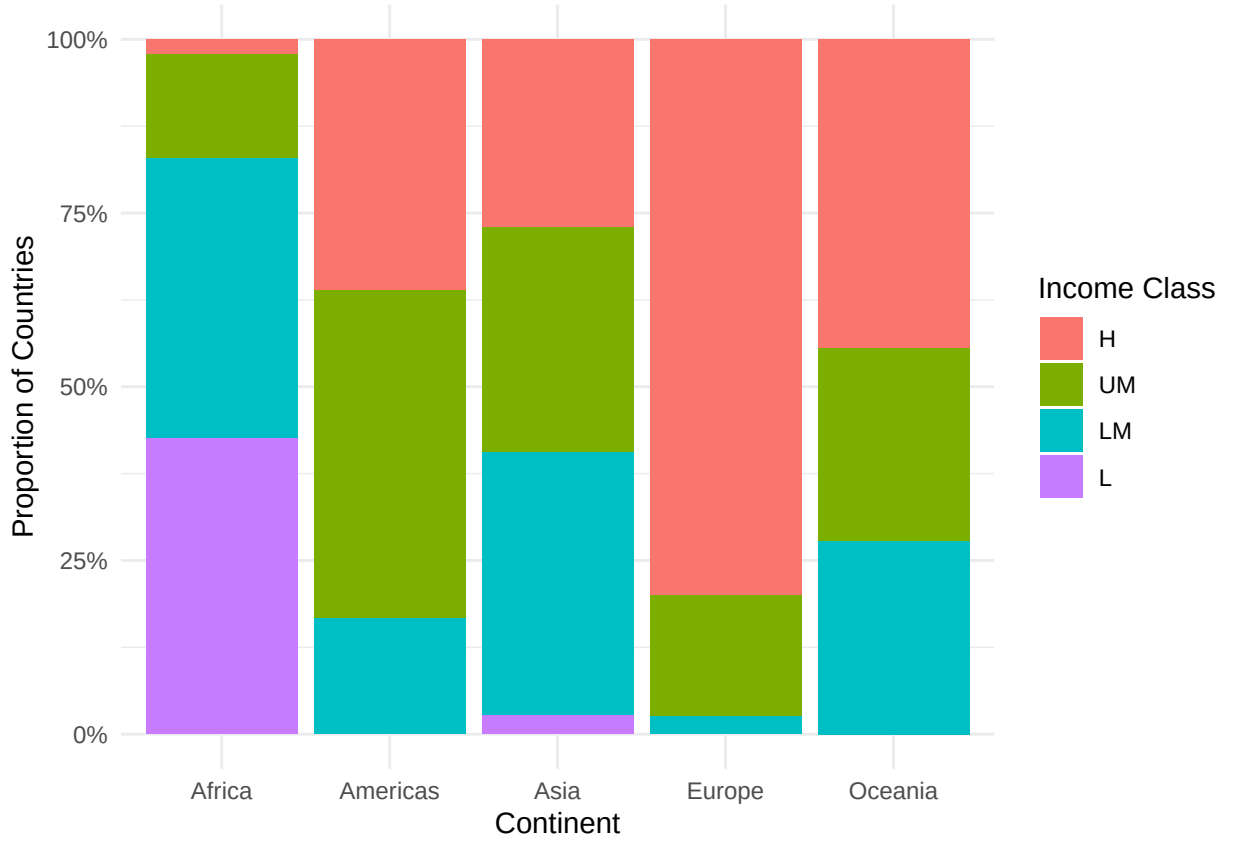
```
clean_data_present_continent <- clean_data[!is.na(clean_data$continent), ]

absolute_freq_barplot <- ggplot(clean_data_present_continent, aes(x = continent,
                      fill = Income_class)) + geom_bar(position = "stack") +
  labs(x = "Continent", y = "Number of Countries", fill = "Income Class") +
                                                        theme_minimal()

print(absolute_freq_barplot)
```



We can see that Africa has the greatest number of low-income countries and the lowest proportion of high-income countries among all other continents. Additionally to Africa, only Asia has one low-income country represented. Europe has the highest number of high-income countries and only one lower-middle income country. Oceania has equal number of lower- and upper- middle countries (10 each) but the highest number (17) is of high-income countries. Prevailing countries in Amerias is upper-middle income countries.

Create another stacked barplot of relative frequencies (height of the bars should be one).

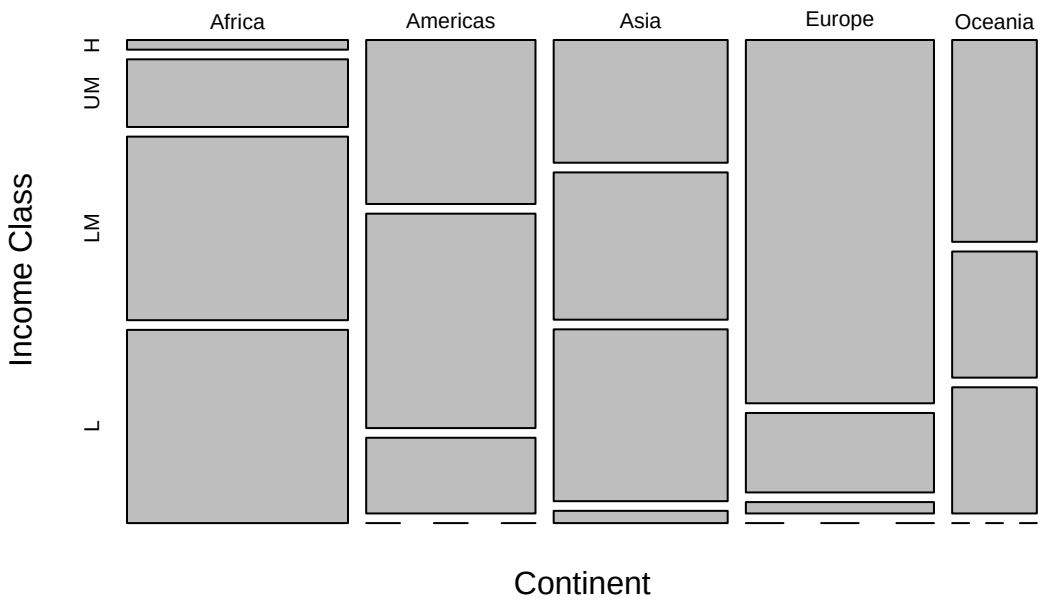
```
relative_freq_plot <- ggplot(clean_data_present_continent, aes(x = continent, fill = Income_class)) +
  geom_bar(position = "fill") +
  labs(x = "Continent", y = "Proportion of Countries", fill = "Income Class") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()

print(relative_freq_plot)
```



This plot gives us a better understanding of proportion of each income class countries for each continent. Now it is easier to determine that for example over 75% countries in Europe are high-income countries, while in Asia - around 26%

Create a mosaic plot of continents and income status using base R functions.

```
mosaicplot(table(clean_data_present_continent$continent,
                 clean_data_present_continent$Income_class),
           main = "Mosaic Plot of Continents and Income Status",
           xlab = "Continent",
           ylab = "Income Class")
```

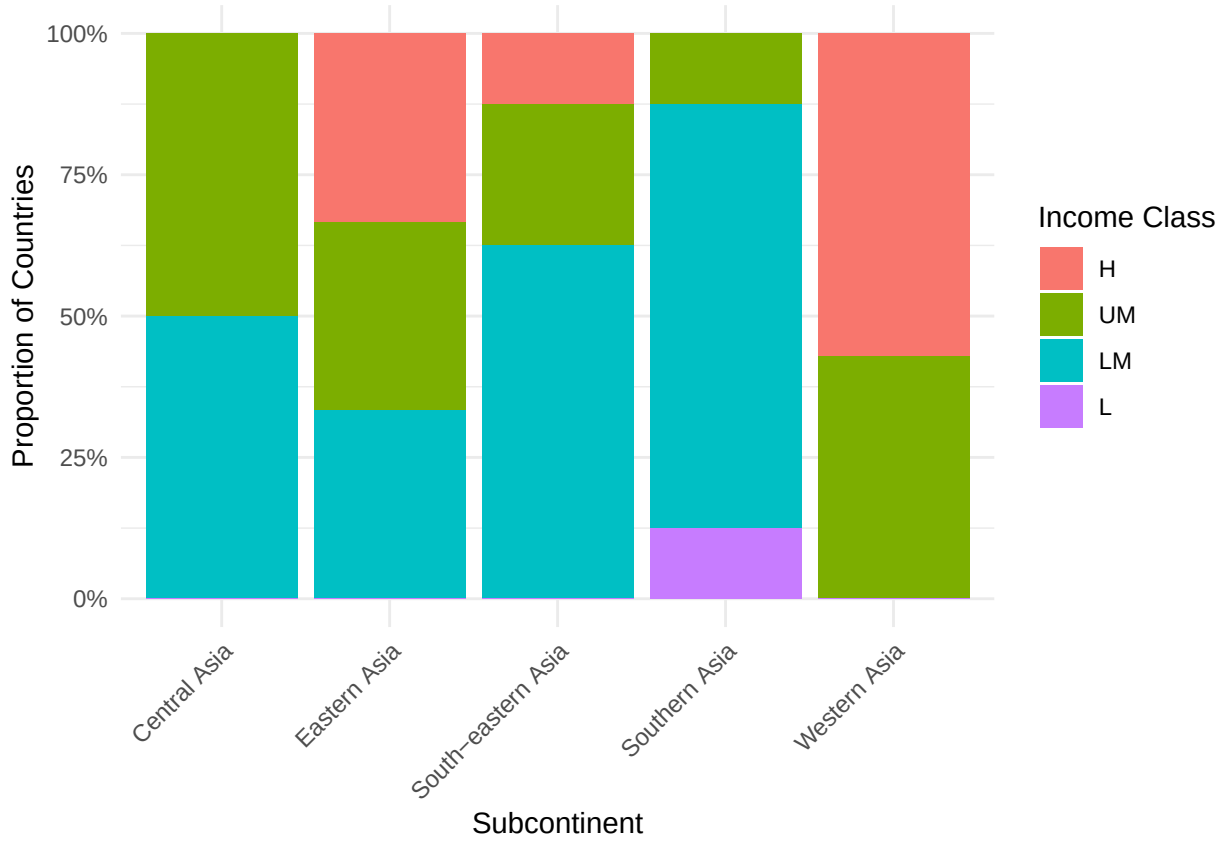


Overall, each plot offers a different perspective on the data, making it easier to understand the distribution of income statuses across continents. The barplot of absolute frequencies provides absolute viewpoints, allowing to detect the exact number of countries on some continent in specific category. The barplot of relative frequencies provides relative viewpoints, allowing to detect the proportion of countries of some status on the continent. The mosaic plot adds an additional layer of understanding by visualizing the relationship between the two categorical variables ().

c. Income status in different subcontinents

```
asia_data <- clean_data_present_continent[clean_data_present_continent$continent
                                          == "Asia", ]

relative_freq_asia <- prop.table(table(asia_data$sub.region,
                                       asia_data$Income_class), margin = 1)

relative_freq_df_asia <- as.data.frame(relative_freq_asia)
names(relative_freq_df_asia) <- c("Subcontinent", "Income_class", "Proportion")


relative_freq_plot_asia <- ggplot(relative_freq_df_asia, aes(x = Subcontinent,
                                      y = Proportion, fill = Income_class)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(x = "Subcontinent", y = "Proportion of Countries", fill = "Income Class")+
  scale_y_continuous(labels = scales::percent) +
  theme_minimal() +
```

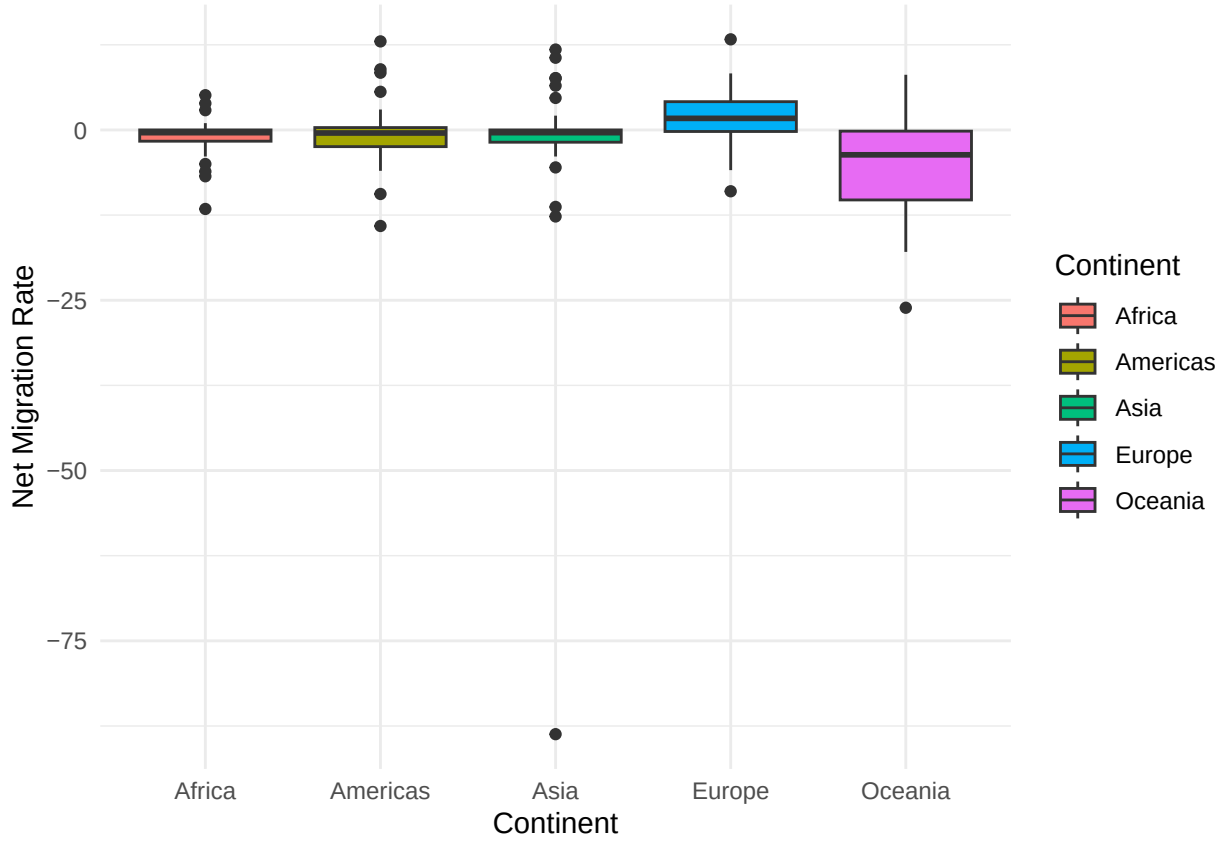
```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(relative_freq_plot_asia)
```



Plot demonstrates that the highest proportion of high-income countries is in Western Asia, while there are no high-income countries in Southern and Central Asia at all. The prevailing countries in Southern Asia are of lower-middle income level. In Central Asia there is an even distribution between lower- and upper-middle income countries.

d. Net migration in different continents

```
boxplot_migration <- ggplot(clean_data_present_continent, aes(x = continent,
                      y=Migration_rate, fill = continent)) +  geom_boxplot() +
  labs(x = "Continent", y = "Net Migration Rate", fill = "Continent") +
  theme_minimal()

asia_largest_negative_outlier <- asia_data[which.max(asia_data$Migration_rate),
                                           "Country"]
asia_largest_positive_outlier <- asia_data[which.min(asia_data$Migration_rate),
                                           "Country"]

print(boxplot_migration)
```




```
cat("Largest Negative Outlier in Asia:", asia_largest_negative_outlier, "\n")
```

```
## Largest Negative Outlier in Asia: Singapore
```

```
cat("Largest Positive Outlier in Asia:", asia_largest_positive_outlier, "\n")
```

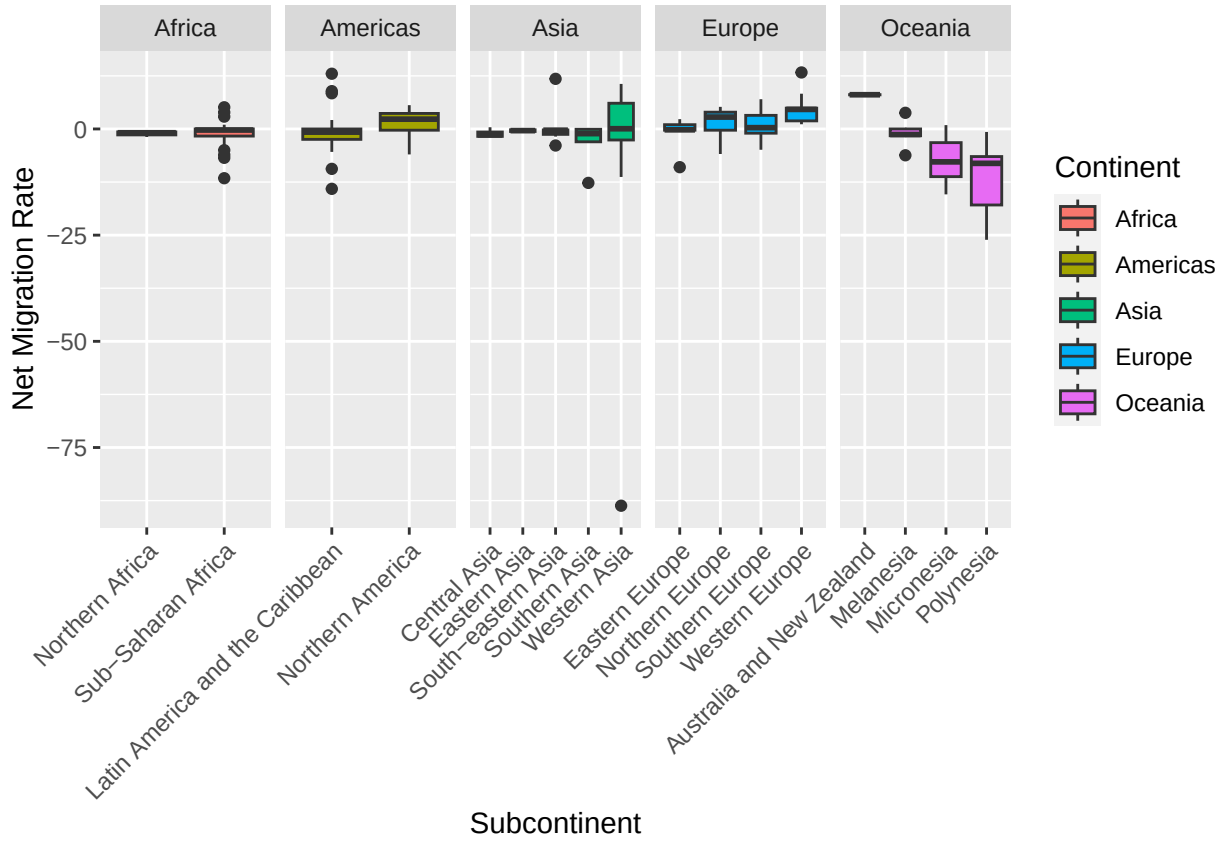
```
## Largest Positive Outlier in Asia: Lebanon
```

We can see that Asia has some severe outliers, which were detected (Singapore and Lebanon), while other countries also have outlying values but not to such high extent as Asia. All continents have more countries with negative migration rate except Europe where number of countries with positive and negative migration rates is balanced. Prevailing number of countries in Oceania have negative migration rate.

e. Net migration in different subcontinents

```
subcontinent_migration <- clean_data_present_continent %>%
  filter(!is.na(sub.region)) %>%
  filter(!is.na(Migration_rate))


migration_boxplot <- ggplot(subcontinent_migration, aes(x = sub.region,
  y = Migration_rate, fill = continent)) + geom_boxplot() +
```

```
  facet_grid(~continent, scales = "free_x") +
  labs(x = "Subcontinent", y = "Net Migration Rate", fill = "Continent") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(migration_boxplot)
```



Now we see more information about which regions on the continent experience prevailing negative/positive migration rate. For example, now we can clearer detect that Micronesia and Polynesia in majority have countries with negative migration rate, and that migration rate in Northern America is in general higher than the in Latin America and the Caribbean.

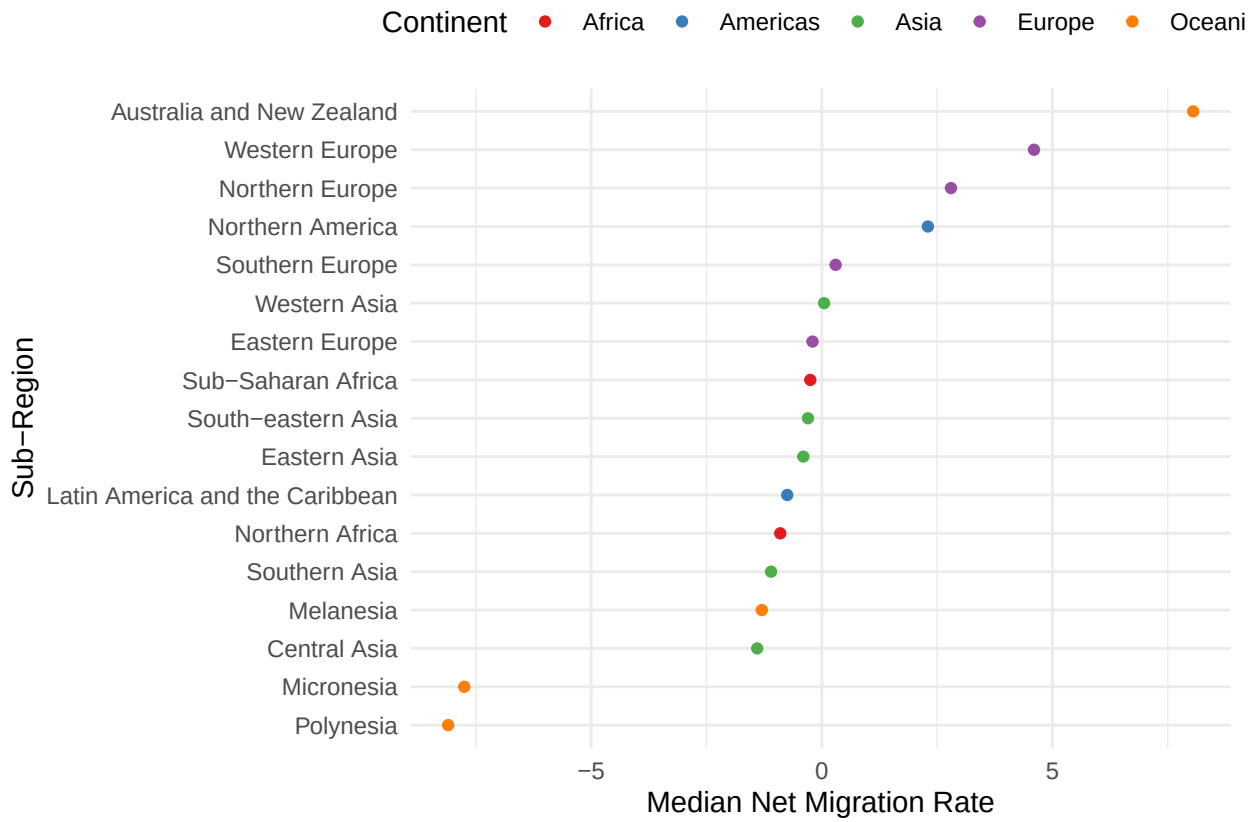
f. Median net migration rate per subcontinent

```
library(forcats)
```

```
## Warning: package 'forcats' was built under R version 4.3.2
```

```
median_migration <- clean_data_present_continent %>%
  group_by(sub.region) %>%
  summarise(median_migration_rate = median(Migration_rate, na.rm = TRUE),
            continent = first(continent)) %>%
  ungroup() %>%
  mutate(sub.region = fct_reorder(sub.region, median_migration_rate))
```

```r
median_migration_plot <- ggplot(median_migration, aes(x = median_migration_rate,
                                        y = sub.region, color = continent)) +
  geom_point() +
  labs(x = "Median Net Migration Rate", y = "Sub-Region", color = "Continent") +
  scale_color_brewer(palette = "Set1") +
  theme_minimal() +
  theme(legend.position = "top")

print(median_migration_plot)
```



The plot shows us that majority of regions have median net migration rate between -2 and 0. The most negative median migration rate is possessed by Polynesia and Micronesia subregions, while the most positive by Australia and New Zealand. Western and Northern Europe also has relatively high positive median migration rate.