Case Study 3: Visualization

AKSTA Statistical Computing

The .Rmd and preferably .html instead of .pdf should be uploaded in TUWEL by the deadline. Refrain from using explanatory comments in the R code chunks but write them as text instead. Points will be deducted if the .PDF is not in a decent form.

Data

Load the data set you exported in the final Task of Case Study 2. Eliminate all observations with missing values in the development status variable.

As a reminder, this data contains 2020 information on

- median age
- youth unemployment rate

for most world entities. Additional information related to the region, sub-region and development status is also provided for the entities.

Tasks:

1. Using **ggplot2**, create a density plot of the median age in the developing countries and another superimposed density plot of the median age in the developed countries.

- The color of the density lines is black.
- The area under the density curve should be colored differently among developed vs. developing countries.
- For the colors, choose a transparency level of 0.5 for better visibility.
- Position the legend at the top center of the plot and give it no title (hint: use `element_blank()`).
- Rename the x axis as "Median age of population"

Comment briefly on the plot.

2. Using **ggplot2**, create a plot as in task 1 for the youth unemployment variable. Comment briefly on the plot.
3. Using **ggplot2**, create a stacked barplot of absolute frequencies showing how the entities are split into regions and development status. Create another stacked barplot of relative frequencies (height of the bars should be one). Comment briefly on the plots.

4. Using **ggplot2**, create a plot showing the relationship between median age and youth unemployment rate.

- Color the geoms based on the development status.
- Add a regression line for each development status.

Comment on the plot. Do you see any relationship between the two variables? Do you see any difference among the development status?

5. Using base R or **ggplot2** create parallel boxplots of the youth unemployment variables for each region. Do you see any striking differences?
6. Create a plot as in Task 5 but for the median population age. Comment briefly.
7. For each sub-region, calculate the median unemployment rate. Then create a plot which contains the sub-regions on the y-axis and the median unemployment rate on the x-axis.

- As geoms use points.
- Color the points by continent (i.e., Region) - use a colorblind friendly palette (see http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/).
- Rename the axes.
- Using `fct_reorder` from the **forcats** package, arrange the levels of subregion such that in the plot the lowest (bottom) region contains the lowest youth unemployment rate and the upper most region contains the highest youth unemployment rate.

8. Go online and find a data set which contains the 2020 population for the countries of the world together with ISO codes. Download this data and merge it to the dataset you are working on in this case study using a left join (A possible source: https://data.worldbank.org/indicator/SP.POP.TOTL?end=2020&start=2020)
9. On the merged data set from Task 8, using function `ggplotly` from package **plotly**
re-create the scatterplot in Task 4 (without the regression lines), but this time make the size of the points proportional to the population. When hovering over the points the name of the country, the values for the age and the youth unemployment and population should be shown. (Hint: use the aesthetic `text = Country`. In `ggplotly` use the argument `tooltip = c("text", "x", "y", "size")`).