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 income status variable.

As a reminder, the data set contains information on

- median age
- youth unemployment rate
- net migration rate (difference between the number of persons entering and leaving a country during the year per 1,000 persons – based on midyear population)

for most world entities in 2020. The data was downloaded from https://www.cia.gov/the-world-factbook/about/archives/. Additional information on continent, subcontinent and income status was appended to the dataset in Case Study 2.

Tasks:

a. Median age in different income levels

Using **ggplot2**, create a density plot of the median age grouped by income status groups. The densities for the different groups are superimposed in the same plot rather than in different plots. Ensure that you order the levels of the income status such that in the plots the legend is ordered from High (H) to Low (L).

- The color of the density lines is black.
- The area under the density curve should be colored differently among the income status levels.
- 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.

b. Income status in different continents

Investigate how the income status is distributed in the different continents.

- Using **ggplot2**, create a stacked barplot of absolute frequencies showing how the entities are split into continents and income status. Comment the plot.
- Create another stacked barplot of relative frequencies (height of the bars should be one). Comment the plot.

- Create a mosaic plot of continents and income status using base R functions.
- Briefly comment on the differences between the three plots generated to investigate the income distribution among the different continents.

c. Income status in different subcontinents

For Asia, investigate further how the income status distribution is in the different subcontinents. Use one of the plots in b. for this purpose. Comment on the results.

d. Net migration in different continents

- Using **ggplot2**, create parallel boxplots showing the distribution of the net migration rate in the different continents.
- Prettify the plot (change y-, x-axis labels, etc).
- Identify which country in Asia constitutes the largest negative outlier and which country in Asia constitutes the largest positive outlier.
- Comment on the plot.

e. Net migration in different subcontinents

The graph in d. clearly does not convey the whole picture. It would be interesting also to look at the subcontinents, as it is likely that a lot of migration flows happen within the continent.

- Investigate the net migration in different subcontinents using again parallel boxplots. Group the boxplots by continent (hint: use `facet_grid` with `scales = "free_x"`).
- Remember to prettify the plot (rotate axis labels if needed).
- Describe what you see.

f. Median net migration rate per subcontinent.

The plot in task e. shows the distribution of the net migration rate for each subcontinent. Here you will work on visualizing only one summary statistic, namely the median.

For each subcontinent, calculate the median net migration rate. Then create a plot which contains the sub-regions on the y-axis and the median net migration rate on the x-axis.

- As geoms use points.
- Color the points by continent – use a colorblind friendly palette (see e.g., here).
- Rename the axes.
- Using `fct_reorder` from the **forcats** package, arrange the levels of subcontinent such that in the plot the lowest (bottom) subcontinent contains the lowest median net migration rate and the upper most region contains the highest median net migration rate.
- Comment on the plot. E.g., what are the regions with the most influx? What are the regions with the most outflux?

g. Median youth unemployment rate per subcontinent

For each subcontinent, calculate the median youth unemployment rate. Then create a plot which contains the sub-regions on the y-axis and the median unemployment rate on the x-axis.

- Use a black and white theme (`?theme_bw()`)

- As geoms use bars. (hint: pay attention to the statistical transformation taking place in `geom_bar()` – look into argument `stat="identity"`)
- Color the bars by continent – use a colorblind friendly palette.
- Make the bars transparent (use `alpha = 0.7`).
- Rename the axes.
- Using `fct_reorder` from the **forcats** package, arrange the levels of subcontinent such that in the plot the lowest (bottom) subcontinent contains the lowest median youth unemployment rate and the upper most region contains the highest median youth unemployment rate.
- Comment on the plot. E.g., what are the regions with the highest vs lowest youth unemployment rate?

h. Median youth unemployment rate per subcontinent – with error bars

The value displayed in the barplot in g. is the result of an aggregation, so it might be useful to also plot error bars, to have a general idea on how precise the median unemployment is. This can be achieved by plotting the error bars which reflect the standard deviation or the interquartile range of the variable in each of the subcontinents.

Repeat the plot in h. but include also error bars which reflect the 25% and 75% quantiles. You can use `geom_errorbar` in **ggplot2**.

i. Relationship between median age and net migration rate

Using **ggplot2**, create a plot showing the relationship between median age and net migration rate.

- Color the geoms based on the income status.
- Add a regression line for each development status (using `geom_smooth()`).

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

j. Relationship between youth unemployment and net migration rate

Create a plot as in Task f. but for youth unemployment and net migration rate. Comment briefly.

k. Merging population data

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: World Bank))
- Inspect the data and check whether the join worked well.

l. Scatterplot of median age and net migration rate in Europe

Make a scatterplot of median age and net migration rate for the countries of Europe. * Scale the size of the points according to each country's population.

- For better visibility, use a transparency of `alpha=0.7`.
- Remove the legend.
- Comment on the plot.

m. Interactive plot

On the merged data set from Task k., using function `ggplotly` from package **plotly**
re-create the scatterplot in Task l., but this time for all countries. Color the points according to their continent.

When hovering over the points the name of the country, the values for median age, net migration rate, and population should be shown. (Hint: use the aesthetic `text = Country`. In `ggplotly` use the argument `tooltip = c("text", "x", "y", "size")`).

n. Parallel coordinate plot

In **parallel coordinate plots** each observation or data point is depicted as a line traversing a series of parallel axes, corresponding to a specific variable or dimension. It is often used for identifying clusters in the data.

One can create such a plot using the **GGally** R package. You should create such a plot where you look at the three main variables in the data set: median age, youth unemployment rate and net migration rate. Color the lines based on the income status. Briefly comment.

o. World map visualisation

Using the package **rworldmap**, create a world map of the median age per country. Use the vignette https://cran.r-project.org/web/packages/rworldmap/vignettes/rworldmap.pdf to find how to do this in R.