

# Exercises - Visualising amounts and (nested) proportions

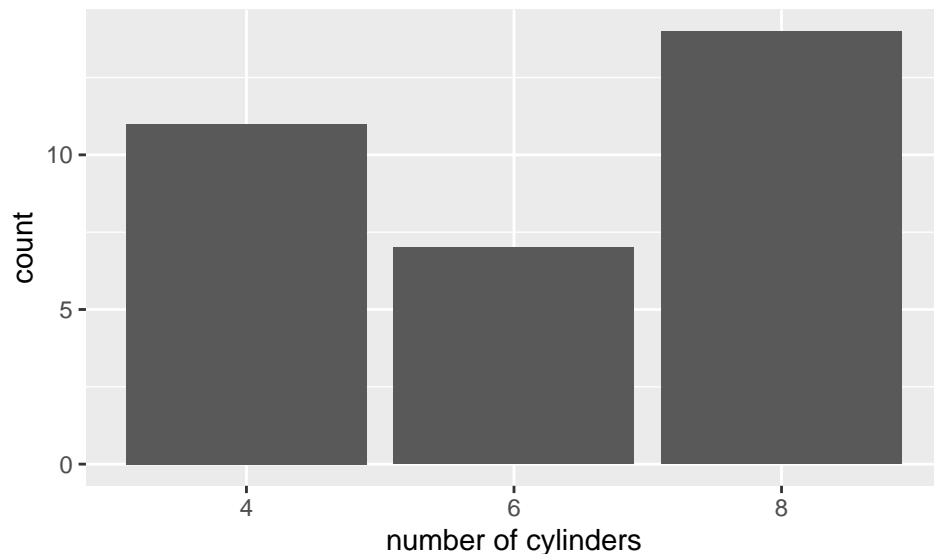
## Data Visualisation

### Visualising amounts for one categorical variable

#### Exercise 1: Bar charts with counts

Use the `mtcars` data set to perform the following exercises. This data set contains the specifications of several cars.

- Make a bar chart that shows the distribution of car frequency based on number of cylinders. Give sensible labels to the axes, as in the example below.

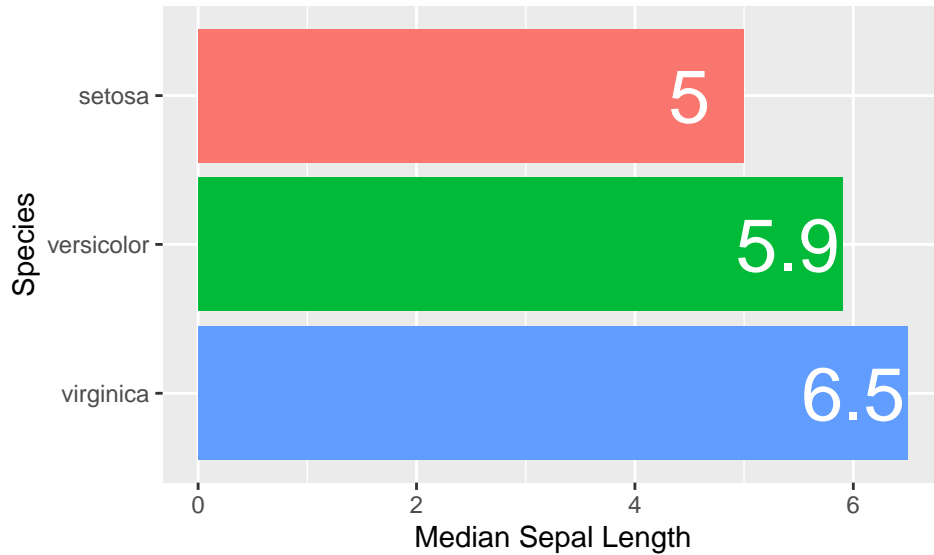


- Next, flip the bar chart such that the number of cylinders are on the y-axis and the counts on the x-axis.
- Next, reorder the bars in terms of the number of cars in each category.

#### Exercise 2: Bar charts with summary statistics

For this exercise we will use the `iris` data set that includes the sepal and petal lengths and widths of flowers of three species.

- Create a bar chart that displays the mean sepal length for each of the species and give sensible labels to the axes. Flip the bar chart so we can easily read the species names.
- Now reorder the bars such that the species with the shortest sepals is shown at the top and the longest at the bottom.
- Give each bar a different color, but remove the legend as color is a bit redundant.
- Add each bar's value (i.e. the sepal length means) to the plot and make sure that they are in the area of the bar. As an example, see the plot below.



## Visualising amounts for two categorical variable

### Exercise 3: Stacked and grouped bar charts

Use the `happy` data set from the `GGally` package to perform the following exercises. The data is a small sample of variables related to happiness from the General Social Survey (GSS). The GSS is a yearly cross-sectional survey of Americans, run from 1972.

The variables in this dataset relate to happiness, and we will be using the two variables `happy` and `marital`:

- `happy`: happiness: very happy, pretty happy, not too happy.
- `marital`: marital status: married, never married, divorced, widowed, separated.

There is quite some missing data, so only use the cases that have no missing values on the variables that you are looking at.

- Create stacked bar charts that show the distribution of marital status within each happiness level.
- Create grouped bar charts that show the distribution of marital status within each happiness level.
- Create a multipanel version of the previous plot using `facet_grid`.

### Exercise 4: Heatmaps

Make a heat map that indicates the number of internet users in several countries throughout the years, as was shown in the lecture video. The data from many countries are available in `internet.rda`. To reduce the data, make a selection of  $\pm 15$ -20 countries, based on which you find most interesting. Play around with the colors; what colors do you think are most insightful?

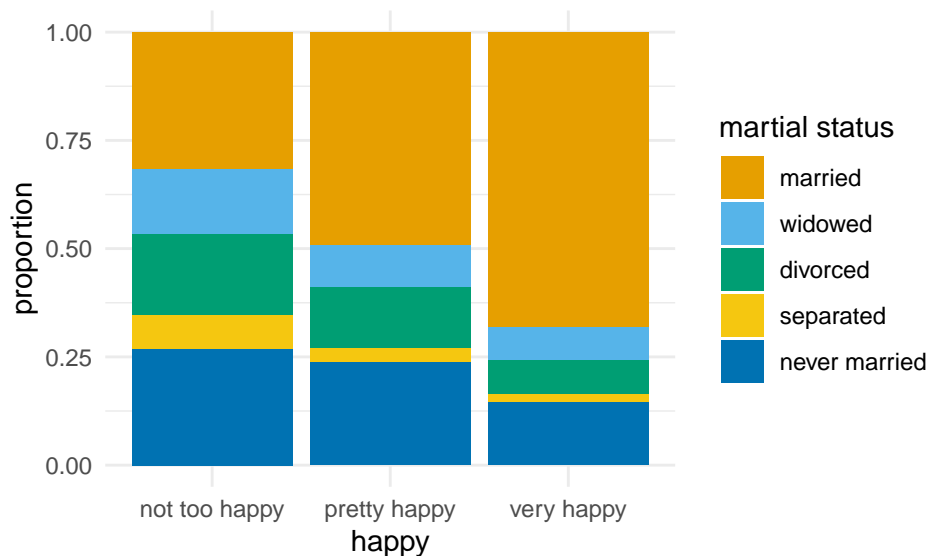
## Visualising proportions

Again use the `happy` data from the `Ggally` package, as we used in Exercise 3.

### Exercise 5: Stacked bar charts

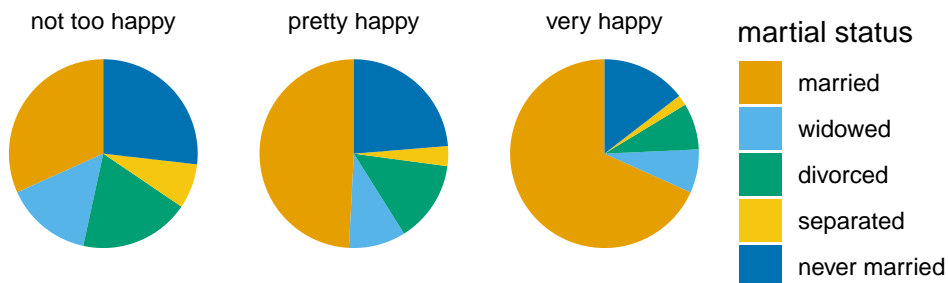
- Create stacked bar chart that shows the distribution of marital status in the data set.
- Now create the same stacked bar chart, but make sure that the bar length indicates the proportions (i.e. number between 0 and 1) instead of the counts. To accomplish this, use the argument `position = position_fill()` in the `geom_bar` function.

- c. Now make a series of stacked bars indicating the proportion of marital status within each happiness level, like in the figure below.



### Exercise 6: Pie charts

- Create a pie chart that shows the distribution of marital status in the data set. Remember to use the `void` theme to remove grid lines.
- Now create this pie chart multiple times to represent the distribution of marital status within each happiness level, as below. Tip: use `facet_grid`.



### Exercise 7: Side-by-side bar charts

- Create a side-by-side chart that shows the distribution of marital status in the data set.
- Use `facet_grid` to create a series of side-by-side charts that shows the distribution of marital status in each happiness level.

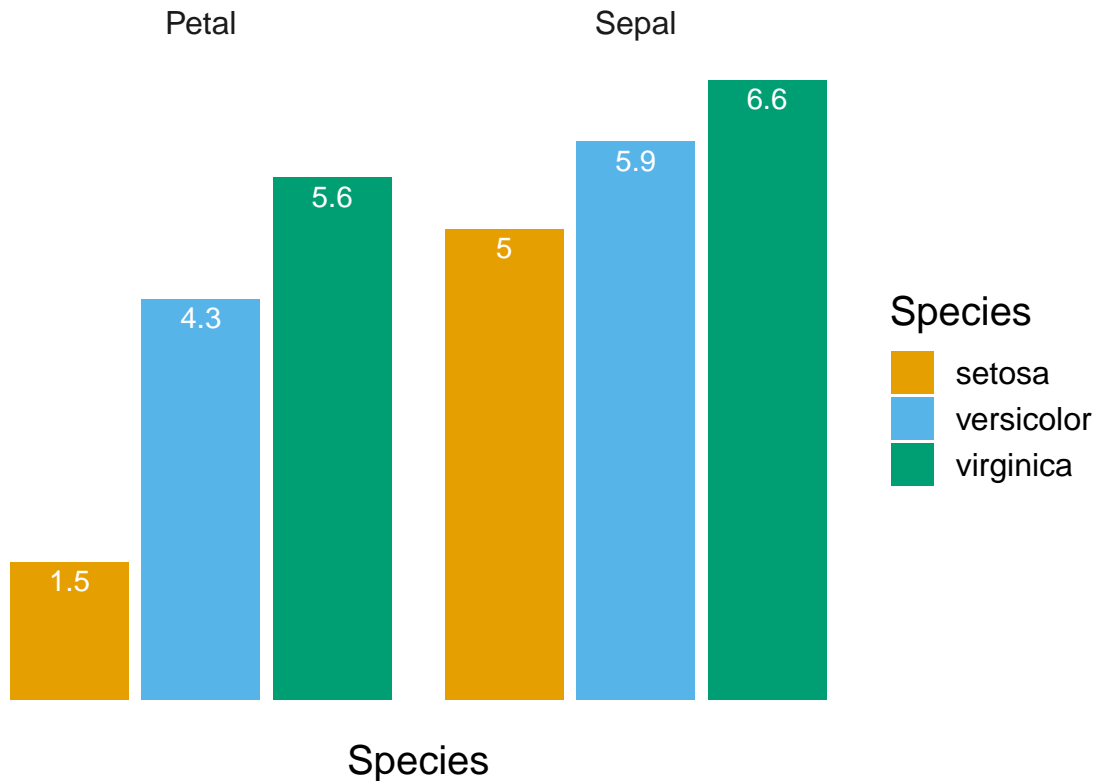
## Challenges!

### Remake the plot

The plot below was made with the `iris` data set, but then in long format (`long_iris`) with one row per measurement. Remake the plot as closely as possible.

# A comparison of three iris species

on their mean petal and sepal lengths



Data source: R

## Make a data story

The data set `1976-2020-president_simplified.csv` contains election data of US presidents from 1976-2020 (Data & Lab, 2017). However, it contains only the votes for the candidates of the Democratic and Republican party, as all other votes were removed from the data to simplify it.

Pretend that you are an analyst hired by a campaign manager of a presidential candidate of one of the two parties (you can decide which one). The campaign manager is trying to plan a campaign trip around the US. For this, they want to know which states may need some convincing to vote for the party's candidate via a campaign.

Create graph(s) that gives an overview of the votes in each state over a certain period. The idea is that the graph(s) could guide the campaign; what states should they focus on? You're free to choose the periods, but include several years. Also include the data of multiple states in your graphs.

*Note: some data preparations may be required!*