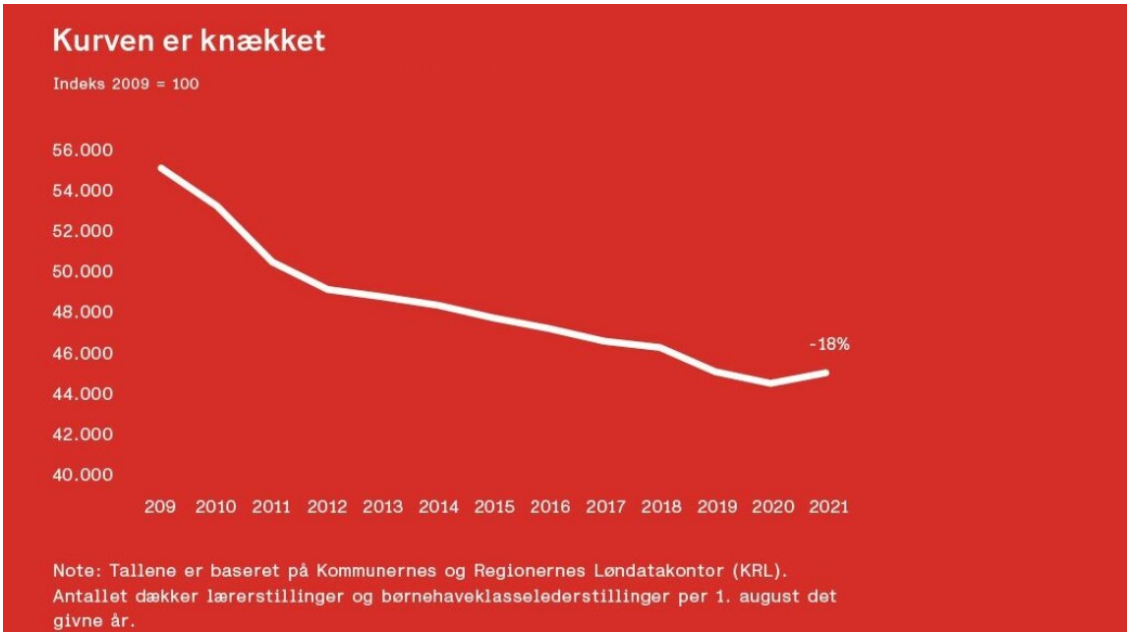
Data science exam portflio component II data visualization

Student-ID: 110560

1 Bad graph

Figure 1: *Bad graph*



Source: https://www.folkeskolen.dk/finanslov-folkeskolen-nr-15-2022-kommunal-okonomi/antallet-af-laerere-i-folkeskolen-er-steget/4675003

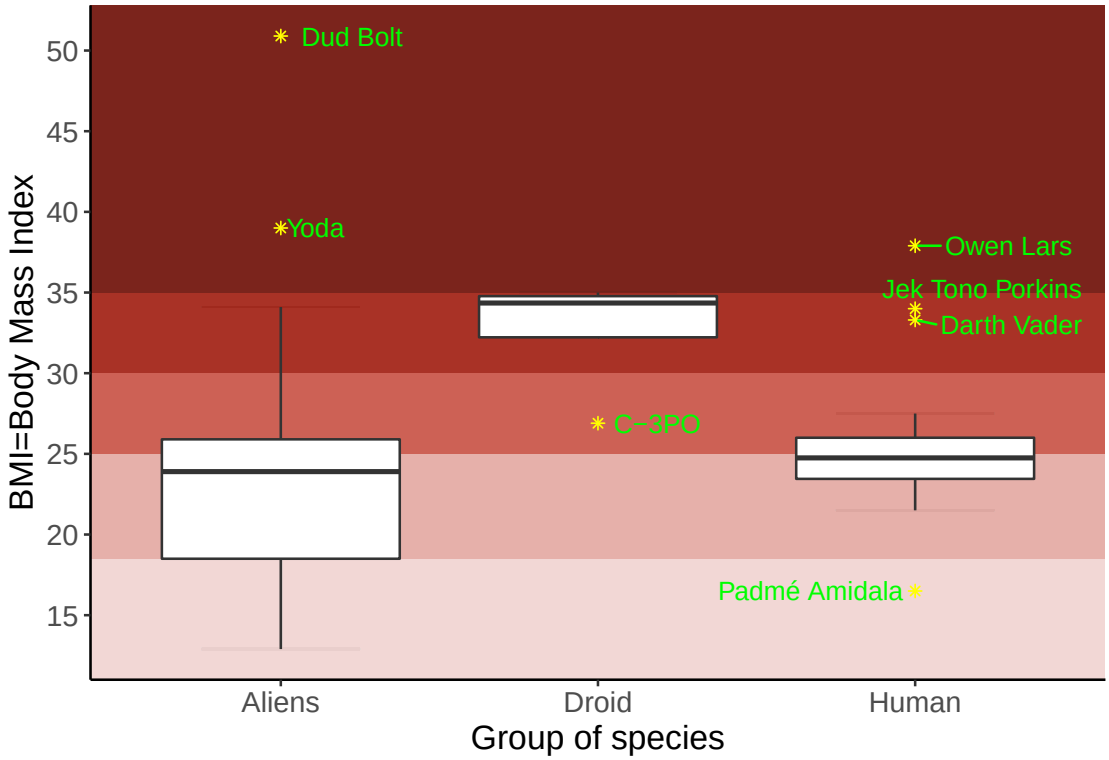
First some context for figure 1 for why this graph have been maded in the first place. The graph has been maded to underlined a pointe of fewer people employed as teachers in the public schools. Hence it should lower the quality of the school system. The title of figure 1 is that "The curve has been broken", and implying that more people are employed as teachers. So the idea and the graph itself is quite simple, but the graph in figure 1, is a bad graph in multiply ways. Obviously the graph itself with all the flaws, but also in what is should represent data and conclusion about. The only good thing that figure @ref(fig_badgraph) has, is it simple and quite easy to see what they inteded to do.

The flaws is many in figure 1, and starting with the x-axis, which should be the year of 2009 to 2021. But the first observation is in 209, which should have been 2009. The note in figure 1 says that the data comes from register of salaries, so the years should be correct. The y-axis is started at 40.000 and ends with 56.000. There is no unit of the y-axis, and the title of the graph do not indicate what the units are. The choosen range of the y-axis makes the decline looks bigger, than if the range have been at absolute minimum at 0 to 56.000. By the choice of range, make the figure 1 a lie-factor of $\frac{(44.000-55.500)}{(0-55.500)} = 20.91\%$. Furthermore the label of the y-axis says "Indeks 2009 = 100", which is a way to scale numbers to each other, if there was multiply variables. But since the y-axis is way past 100, and 2009 do not exist at the x-axis, the index become meaningless. This label is just adding to confusing and understanding. At the end of the curve there is a label with the number "-18%". This number do not reference to anything or given an explanation. You have to guess that it should be the percentage change since 2009. This values could have been seen if it the graph have been the index as label of the y-axis suggest. At the right side of the graph there is a huge gap, where the graph is not utilized, and then it is just a waste of space.

When looking at the information that the graph should provided, the development does not make sense if it is not related other things. It would omit any confounding variabel, since it only shows the number of teachers in figure 1. In this case the confounding variable could be number of pupils or classes, and then the teacher could be compare to the development in pupils and/or the number of classes. If the quality of teaching should have gone down the ratio between $\frac{teachers}{pupils\ or\ classes}$ should go down. So the curve of the index for teachers should have been compared to an index of pupils or an index of classes. So with these curves the level of information would have grown. The figure 1 is a figure that cannot stand alone to drawn the suggested conclusion.

2 Good graph

Figure 2: *Obesity among Star Wars characters*



To answer the second part of the visualization component the graph in figure 2 have been constructed. The graph investigate the obesity of characters in the Star Wars universe. It investigate whom of the characters that have a high *Body Mass Index*(**BMI**), how the distribution of observations within each species is, and all this is compared to the thresholds of obesity defined by the human species. All data is from the *Star Wars* dataset build into the *tidyverse* package in R. Figure 2 has been build by constructing the **BMI** by the equation:

$$**BMI** = \frac{weight\ in\ kilo}{(height\ in\ meters)^2}$$

and the group of species are build to compare aliens, droids and humans with each other. In the star wars dataset the groups of droids and humans are defined as unique species, while the alien group is rest of the species. Within that category there are different species, such as Wookies, Ewok, Gungan, etc. All types of aliens are put together since the number each species are too small to form its own group.

The **BMI** is chosen as an indicator of obesity, since it takes the characters' height into account when comparing the weight. To give the best comparisons of **BMI** across groups of species a boxplot have been chosen, since it gives the distribution of the group, and only one observation is removed from the distribution, since it was an extreme value. The observation was *Jabba the Hutt* with a **BMI** at 443.4, which would have distorted the boxplot too much, so it was better to remove it. The distributions are based at 31, 4, 22 observations in the group of aliens, group of droids, and group of humans. The box plot also indicates what is the normal range of **BMI** within each group of species.

The severity of obesity are related to the human scale of obesity, and the different categories of obesity are painted a shade of red, with a darker color as indicator for severe obesity. A **BMI** at or higher than 25 indicates overweight and **BMI** at 30 or higher indicates obesity.

In figure 2 droids are generally obese, but mostly because they are made of metal. Humans are mostly in the range of normalweight and overweight, with some obese characters, and Padmé as underweight. Within the group of aliens most of them are normalweight, but interestingly are Yoda obese.