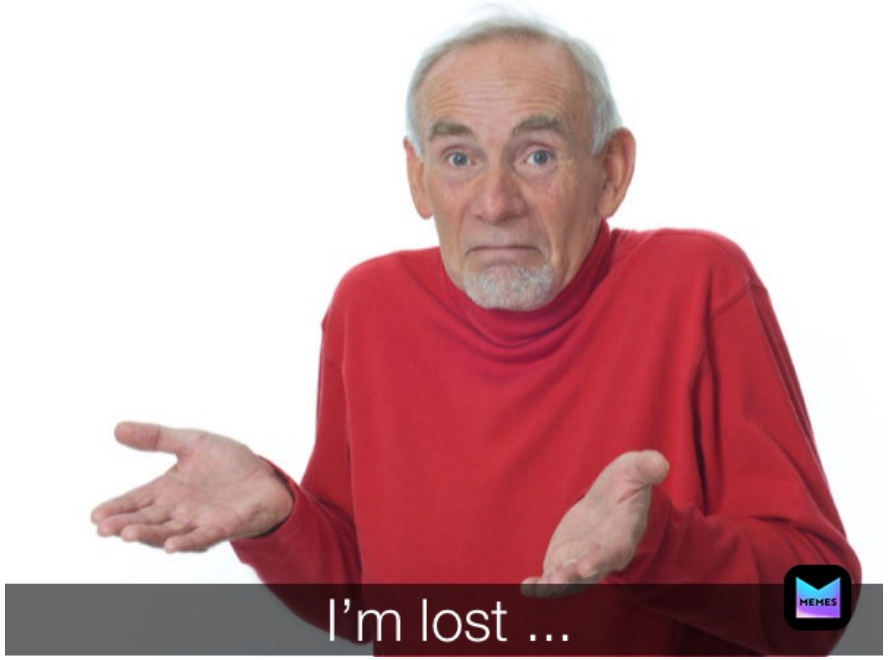
Data Science for Socioeconomists

Excercise 6 - Data visualization pt. 1

Lisa Wegner



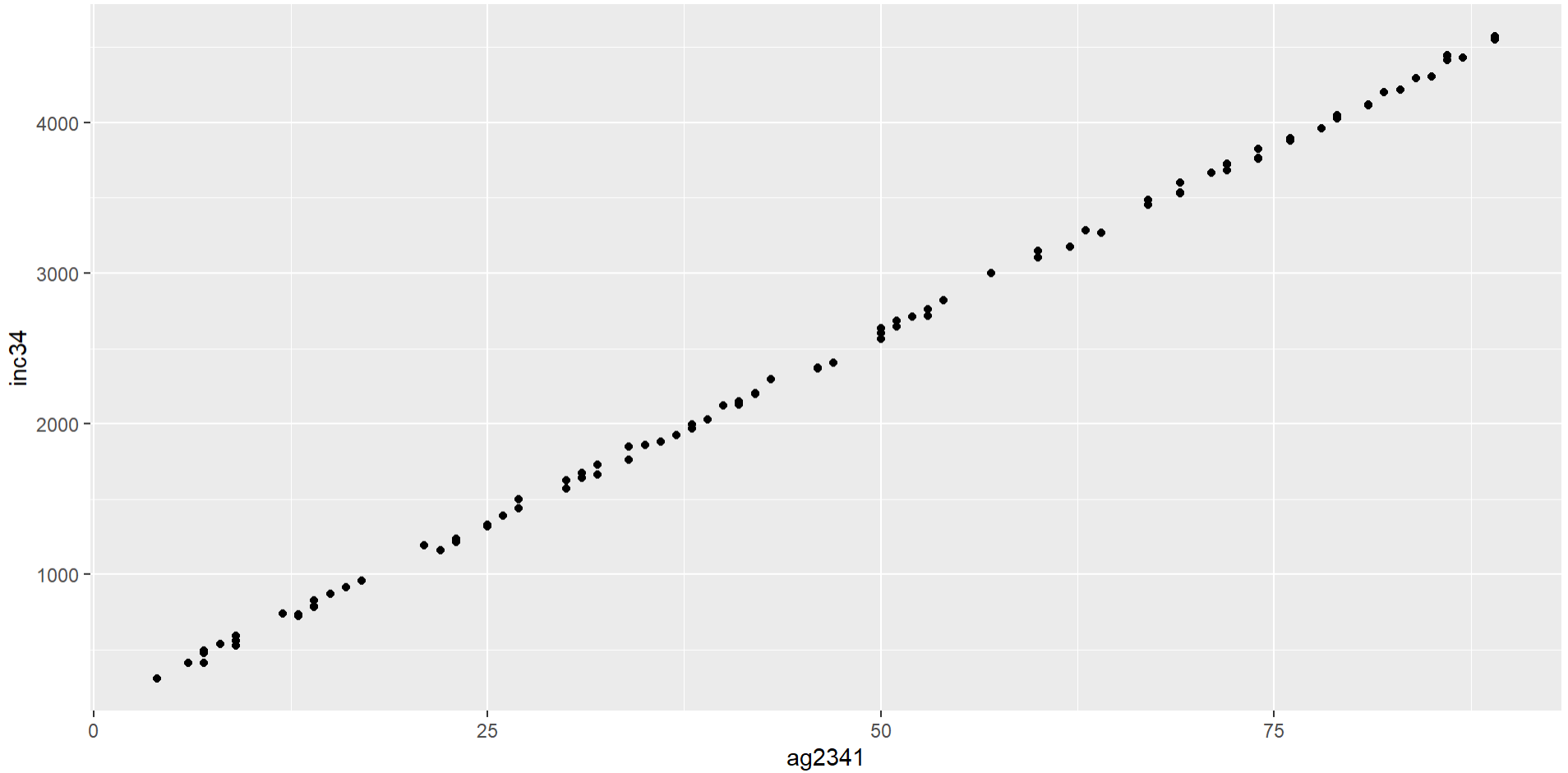
- We have created a link-list with all things helpful - this can be found under “learning materials” at OpenOlat
- The code of the lecture-, exercise- and solution slides can be copied! This way you can adapt it to your needs and don’t need to start from scratch.

1. Axis labeling
2. scales
3. legends
4. readability
5. color usage
6. accessibility
7. Tasks for today

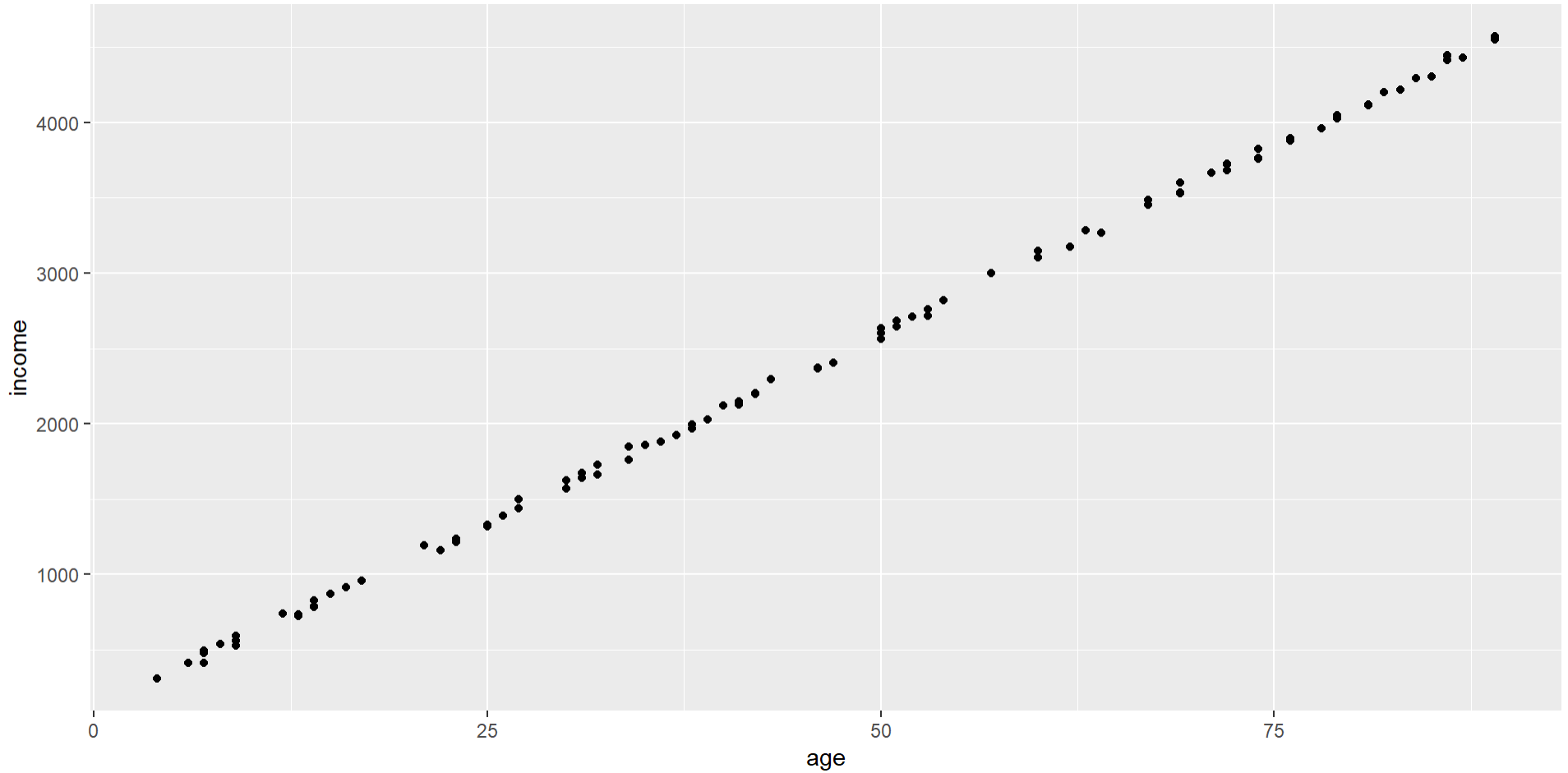
- We can look at:
  - labeling the axis
  - decision about units
  - aspect ratio (wide enhances changes on the x-axis, narrow and tall enhances changes on the y-axis)
  - grid spacings

We get the axis title of the variable we inserted - ugly names create ugly axis titles

```
library(sozoekds)
library(ggplot2)
test_dataset <- testdata
test_dataset$ag2341 <- test_dataset$age
test_dataset$inc34 <- test_dataset$income
 ggplot(data=test_dataset) +
   geom_point(aes(ag2341,inc34))
```

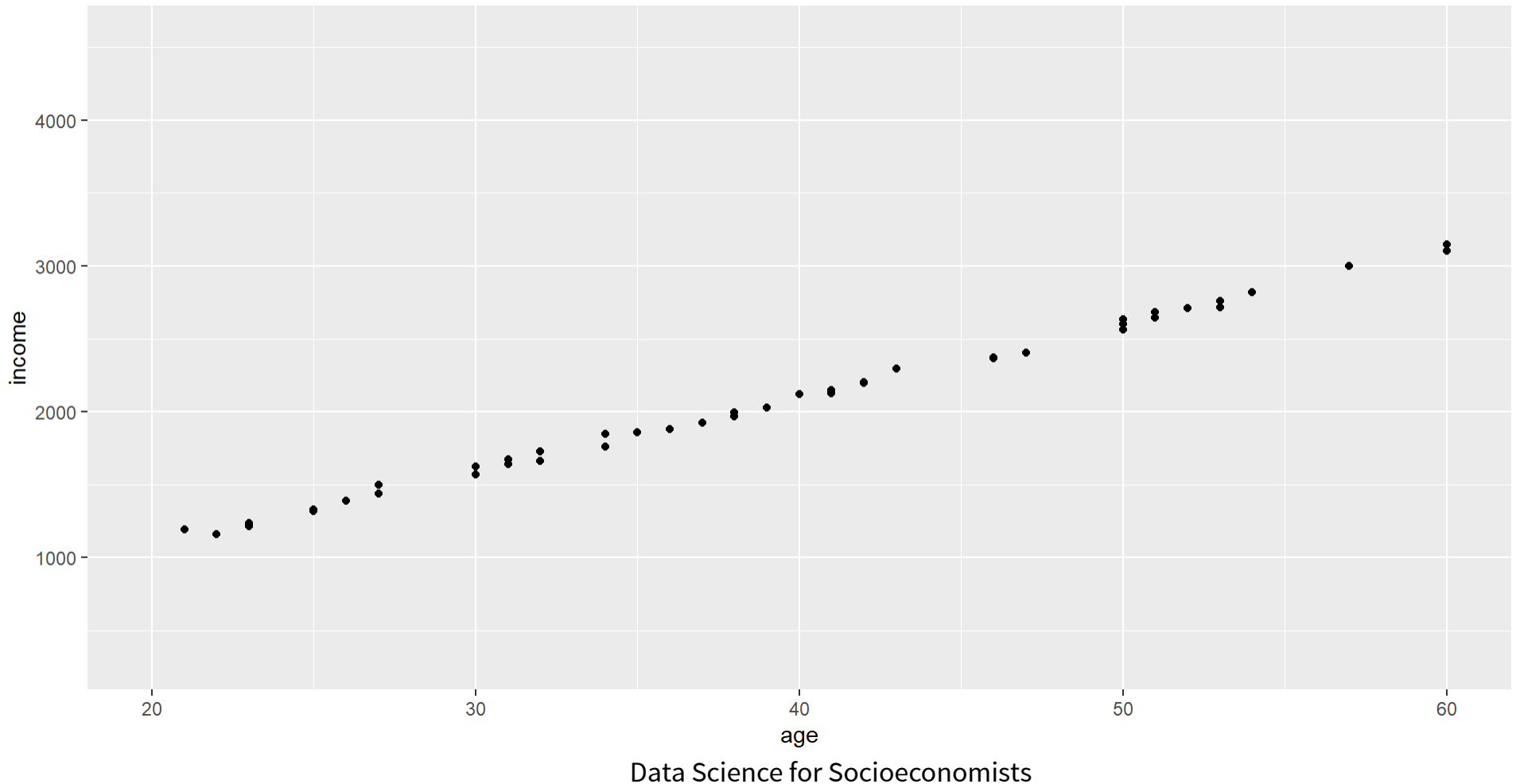



```
1  ggplot(data=test_dataset) +
2    geom_point(aes(ag2341,inc34)) +
3    xlab("age") +
4    ylab ("income")
```


4000
3000
2000
1000
income
0
25
50
75
age

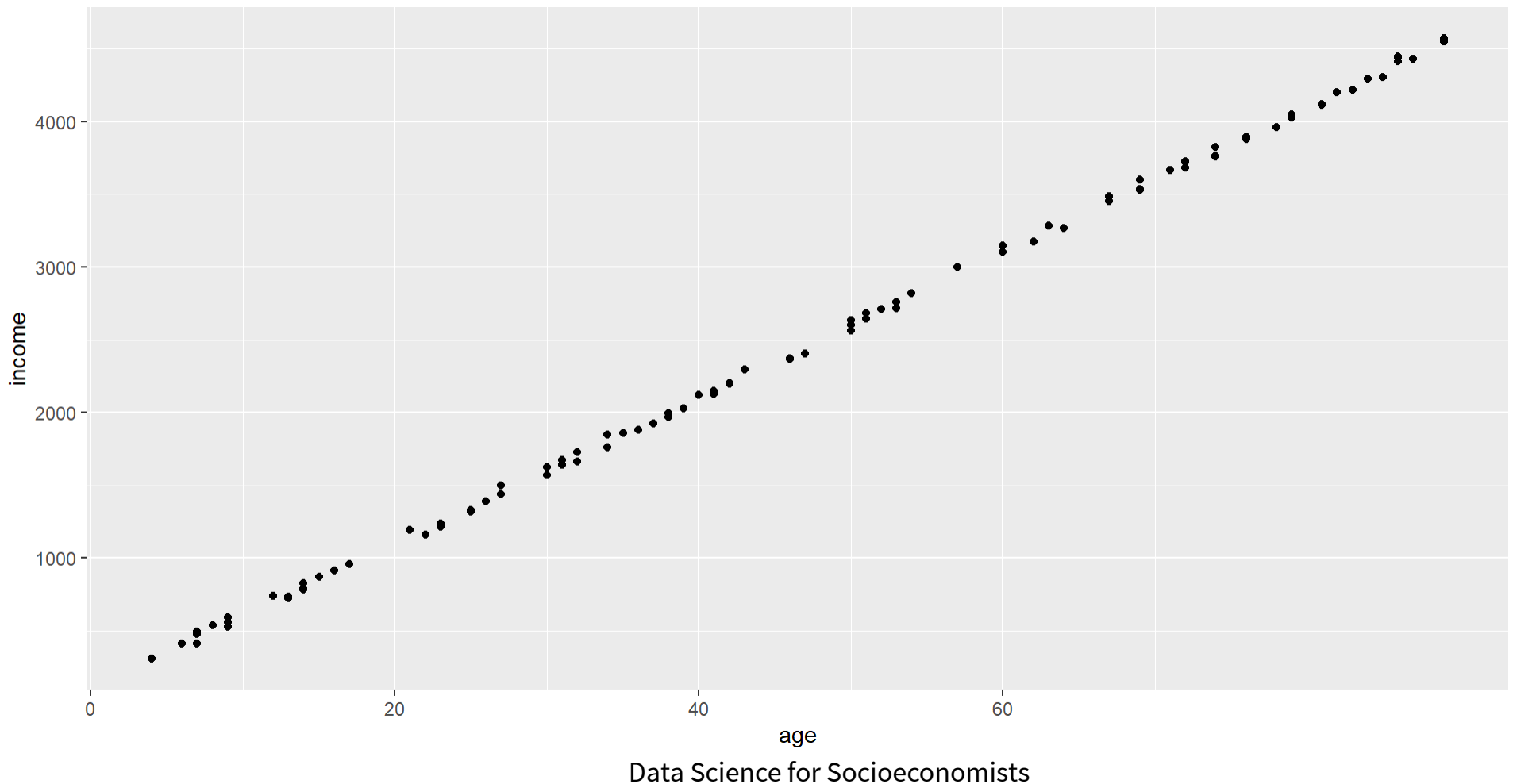
changing the range of values that are plotted

```
1   ggplot(data=test_dataset) +
2     geom_point(aes(ag2341,inc34)) +
3     xlab("age") +
4     ylab ("income") +
5      xlim(20, 60)
```

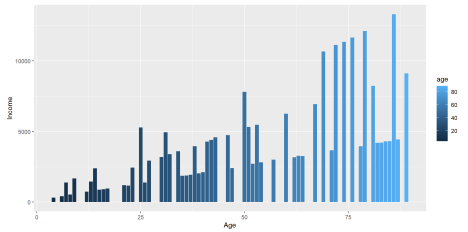


setting "tick marks"

```
1   ggplot(data=test_dataset) +
2     geom_point(aes(ag2341,inc34)) +
3     xlab("age") +
4     ylab ("income") +
5  scale_x_continuous(breaks=c(0, 20, 40, 60))
```



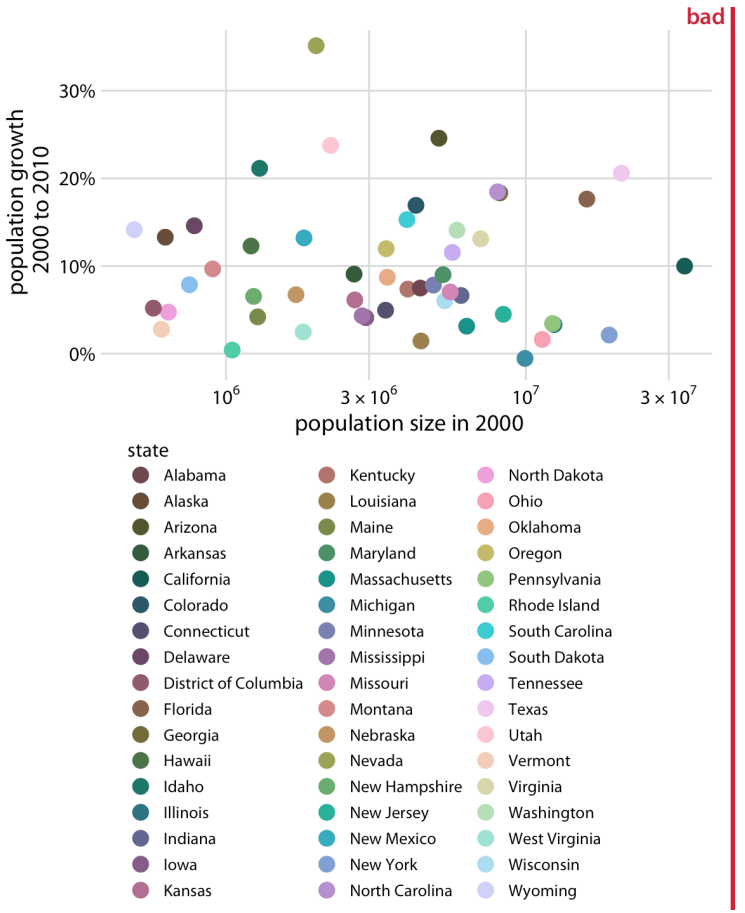
```
ggplot(data = test_dataset, aes(x = age, y = income, fill = age)) +
  geom_bar(stat = "identity") +
  xlab("Age") +
  ylab("Income")
```

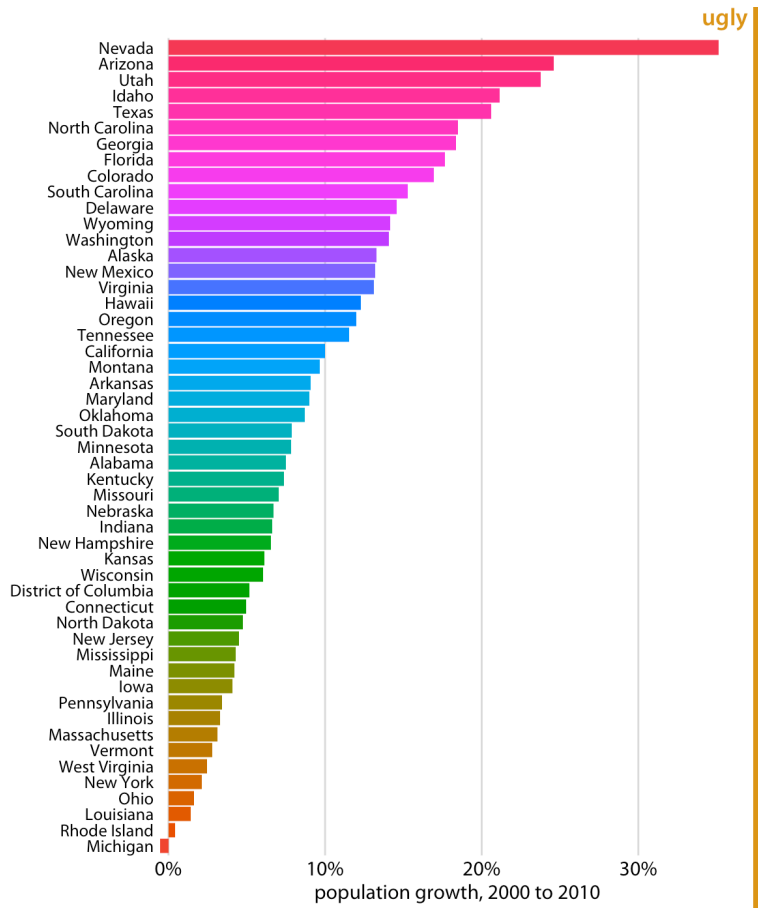


we can also:

- change the order of the elements in our legend
- change the name of the categories in the legend
- change the position of the legend…. and a lot more

What's bad about these visualizations? What can we do better?





via: https://clauswilke.com/dataviz/pitfalls_of_color_use_files/figure-html/popgrowth-vs-popsize-colored-1.png

via: https://clauswilke.com/dataviz/pitfalls_of_color_use_files/figure-html/popgrowth-US-rainbow-1.png

colors

- Fewer colors enable the brain to process the information more clearly
- There are three main types of color palettes used in the world of data visualizations:
  - **Qualitative palette —** each color is distinct from the others
  - **Sequential palette —** a single color in a variety of saturations or a gradient
  - **Diverging palette —** color variables sit on a spectrum, such as cold to hot

  🔄 What can we change in the graphics we saw before?

Info and Inspo:

colors need to be easy to distinguish - What about color-blindness?

1. deutanomaly: green-blind
2. protanomaly: red-blind
3. tritanomaly: blue-blind

⟹ check acessability via a website or via an R-package colorblindr

1. Revisit last weeks plots - how can you transform them with the principles we learned today?

Choose **one** of the two data sets and do the exercises regarding this set.

1. Create a new variable called `Avg_Rooms` that displays the average amount of rooms in each household in a certain neighborhood (block)

Hint: for one observation there might be 20 households, and 50 rooms in total - this leads to an average of 2.5 rooms per household

2. Plot Avg_Rooms against Median_Income - interpret the result
3. Eliminate values in Avg_Rooms that are greater then 10 rooms from your data set - then plot task 2 again
4. as before: Explain your decision of the type of visualization + Interpret the results

1. Create a new variable called `full_score` that is the mean score of all three testscores (Math, Reading, Writing)
2. Calculate grades from the `full_score` using the american system (hint here)
3. Plot the distribution of grades by number of study hours per week
   Hint: create a factor variable from WklyStudyHours

   Hint: You could do this in one plot with multiple lines or in more than one graphic where all the plots are printed in one frame (https://intro2r.com/mult_graphs.html)
4. as before: Explain your decision of the type of visualization + Interpret the results

Starting next week my colleague Victoria Hünewaldt will take over and start with statistical learning and machine learning.

We’ll meet again in January!