

## 4. Exercises for today

### Preface

We'll be working with California Housing Data and Exam Data today and in next weeks session. The two data sets can be found in the package `sozoekds` (see last weeks exercise on how to download it or look into the [Read Me](#) ).

The data sets have added information about their content - like how the variables are measured, what the measure etc. Looking at the description is helpful since a variable in a data set could have a name like `Median_Age` (in `calhouse`) which I might think is the...

- Median age of the inhabitants of the house?
- The age of the person who did the observation?
- The median age of a house within a block?

`Median_Age` measures the last thing - it it measures it on a block-scale. This way it's already an aggregated value for multiple buildings. We can still calculate the mean or median over all observations.

### Todays and next weeks exercises

Now: Work in a Quarto Document with your group on the Jupyter Hub!

- Choose just **one** of the two data sets (`calhouse` or `examdata`) and do the responding exercises.
- While doing the tasks:
  - Please choose an accurate type of plot for the purpose of the question.
  - Give an explanation on why this type of visualization is accurate.
  - Also look at the scales, colors, headlines, type and size of lines/dots, legends...  
- and choose a style that fits.

### California Housing

#### *today*

1. Create a diagram that visualizes the distribution of the Median House Value
2. Afterwards create a diagram that puts Median House Value and the distance to the coast together
3. Add labels, color and scales in a way that suit the purpose of your graphic
4. Explain your decision of the type of visualization
5. Interpret the result

#### *next week*

1. Revisit last weeks plots - how can you re-style them with the principles we learned today?
2. New plot(s)
  - a. 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*
  - b. Plot Avg\_Rooms against Median\_Income - interpret the result
  - c. Eliminate values in Avg\_Rooms that are greater than 10 rooms from your data set - then plot task 2 again
  - d. as before: Explain your decision of the type of visualization + Interpret the results

## Exam Scores

### *today*

1. Create a diagram that visualizes the distribution of the MathScore
2. Plot the distribution of the different test scores in regards to Gender
3. Add labels, color and scales in a way that suit the purpose of your graphic
4. Explain your decision of the type of visualization
5. Interpret the result

### *next week*

1. Revisit last weeks plots - how can you re-style them with the principles we learned today?
2. New plot(s)
  - a. Create a new variable called full\_score that is the mean score of all three testcores (Math, Reading, Writing)
  - b. Calculate grades from the full\_score using the american system ([hint here](#))
  - c. 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](https://intro2r.com/mult_graphs.html))

- d. as before: Explain your decision of the type of visualization + Interpret the results