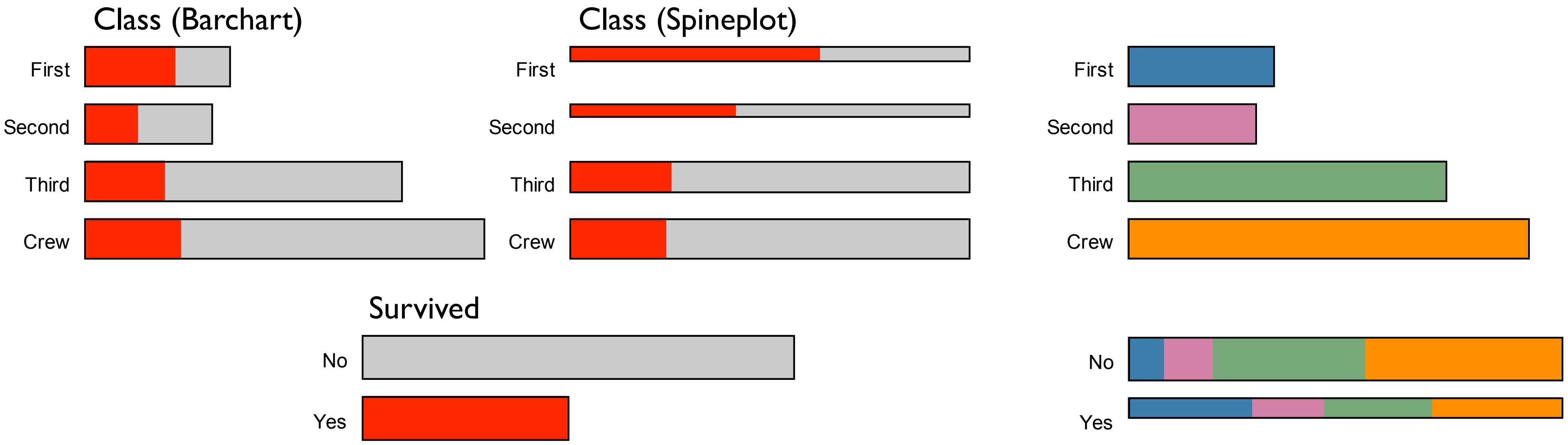
Chapter 3

Interactions between Two Variables

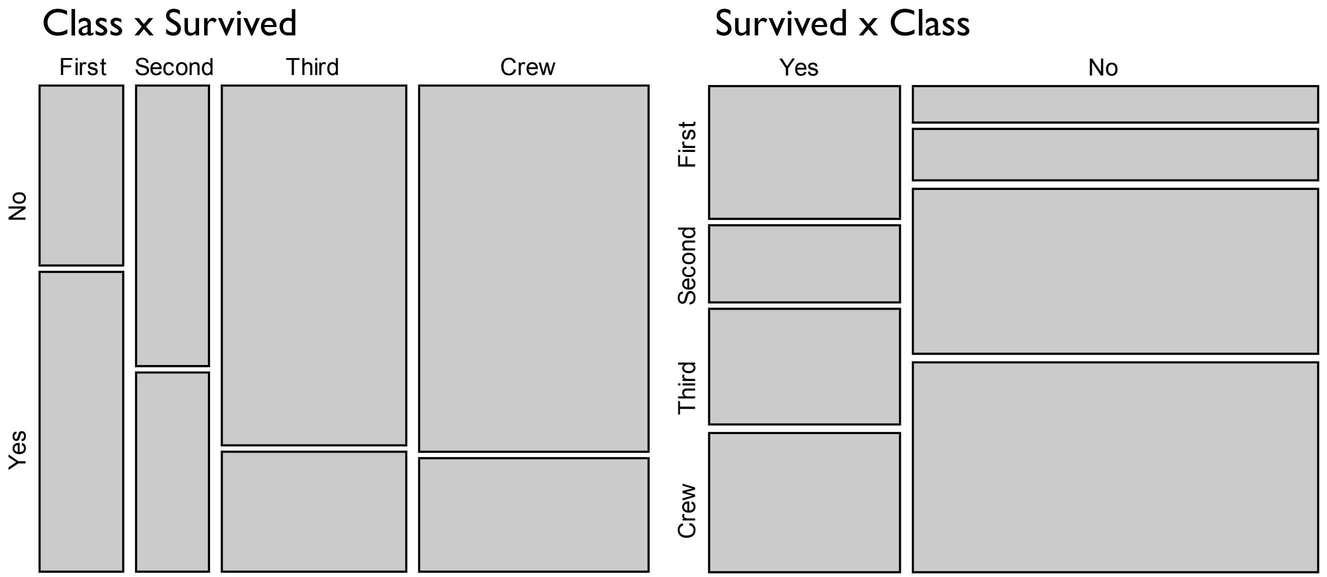
Two Categorical Variables

- The association between two categorical variables can be most easily examined via linked barcharts, i.e., condition a spineplot by selecting one or more category in a barchart
- If we want to examine more than one category at a time, we can also use color-brushing on the conditioning barchart and observe the classes in the spineplot of the dependent variable



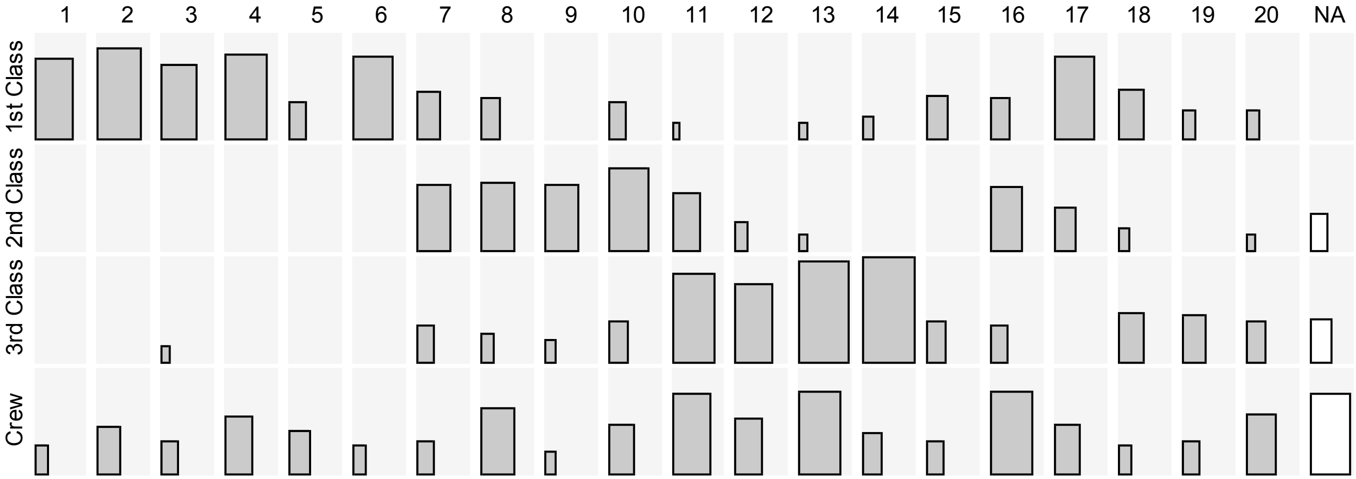
Mosaic Pots

- Although linked barcharts with color-brushing show the exact same information as a mosaic plot, a mosaic plot is more flexible in selecting subgroups and displaying highlighting within these groups
- Even with only two variables the order makes a big difference
- Classical mosaic plots are good to judge and compare associations
- With very many categories and/ or many empty or small cells, variations of mosaic plots are often far more efficient



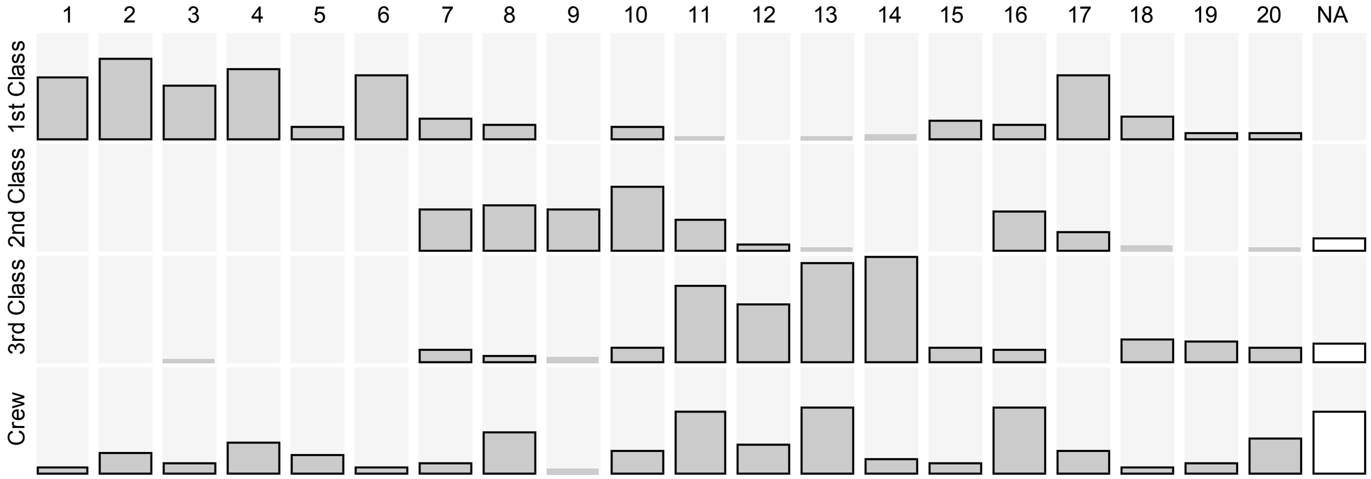
Fluctuation Diagram

- Fluctuation Diagrams (as all other variations of mosaic plots) assign the same area for all category crossings
- This allows to visualize the structure of the variables in cases where we have many empty and/or very small cells
- Due to the regular grid on which all cells are placed, this kind of plot is not particularly well readable for more than 2 variables
- The plot is best to read when the tiles are close to quadratic



Multiple Barcharts

- In contrast to fluctuation diagrams, multiple barcharts do only scale the tiles along one dimension
- Thus, we can plot (traditional) barcharts which are conditioned upon the levels of (an)other variable(s)
- If we consider horizontal and vertical layouts, the possibility to switch to spineplots and allow to modify the splitting direction at all splits, we get **very** many possible plots

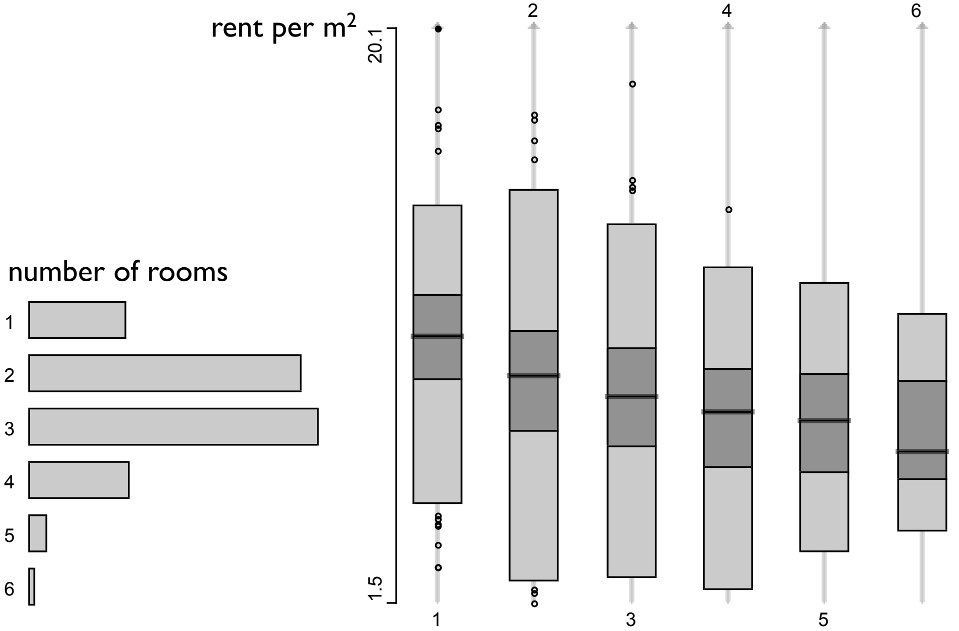


1 Categorical Variable and 1 Continuous Variable

- **Categorical → Continuous**
  The most efficient way to compare the distribution of a continuous variable, given the levels of a categorical variable is a boxplot y by x

  As boxplots don’t reflect the sizes of the underlying groups, a barchart should be used aside as an aid

  (there are modifications to boxplots which try to indicate the group sizes, but they are neither standard nor effective)

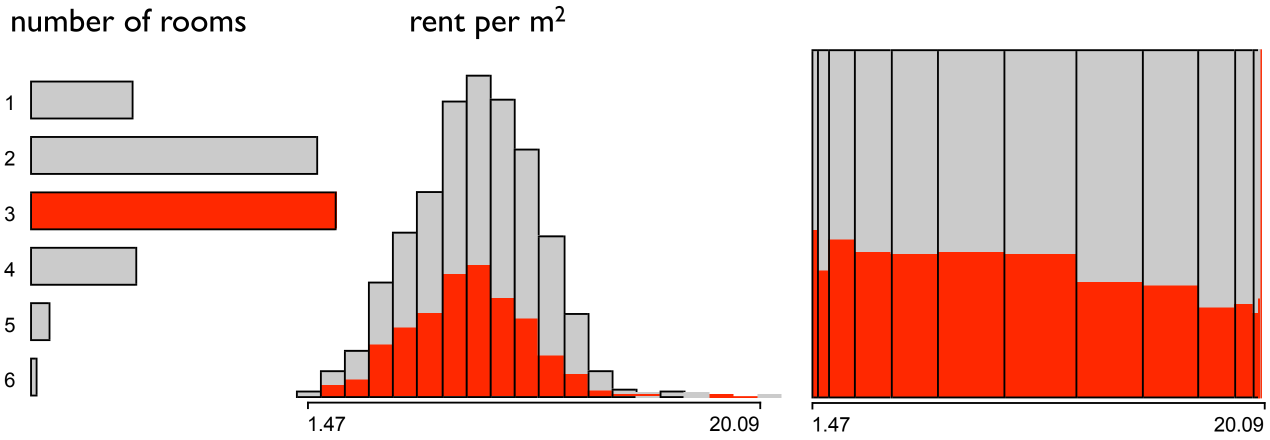


1 Categorical Variable and 1 Continuous Variable

- **Categorical → Continuous**
  If we want to examine the structure of the selected category in more detail, we can step through the different levels of the categorical variable while observing the highlighting in a histogram

  The highlighting we observe is:

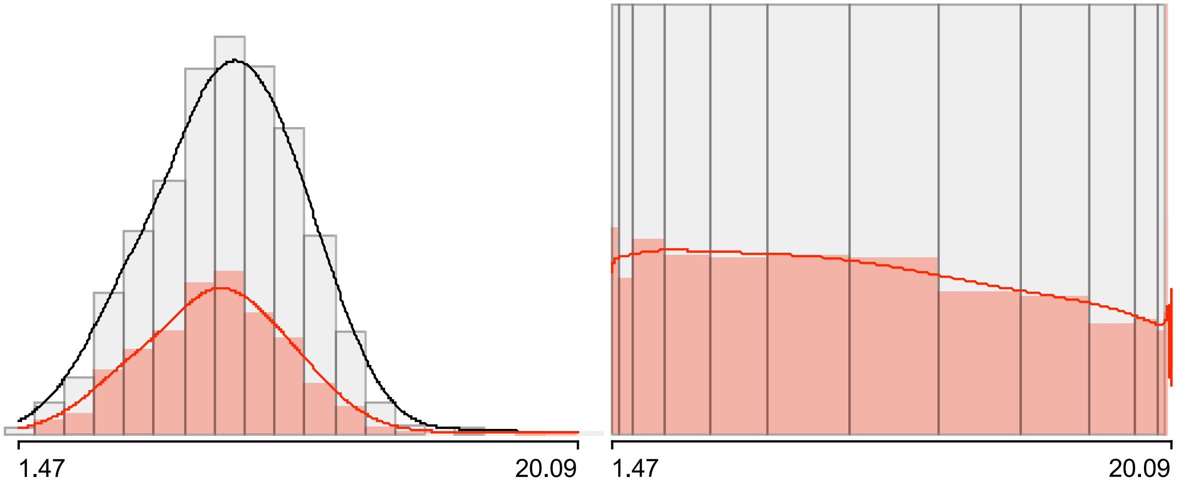
$$P(\text{“No. of Rooms”} = 3 \mid \text{“Price per m}^2\text{”} \in [8.00, 9.00)) = 41.03\%$$



1 Categorical Variable and 1 Continuous Variable

- **Categorical → Continuous**
  Using density estimates we get a smoother, overall trend of the distribution of the selected level with the same plot set-up

  Note:
  The density estimate is defined to integrate to 1, such that it must be scaled with the proportion of highlighted cases in order to match the selected cases in the histogram/spinogram

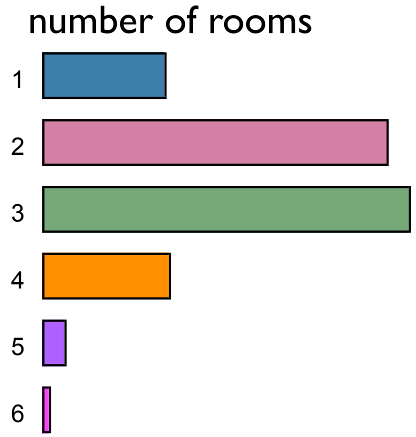


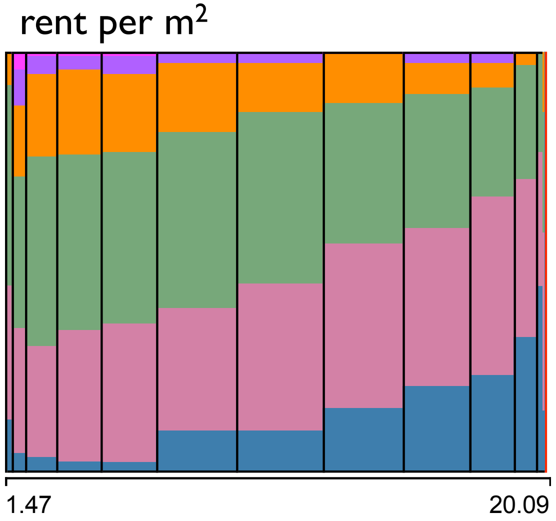
1 Categorical Variable and 1 Continuous Variable

- **Categorical → Continuous**
  Using color-brushing we may observe the distribution of several levels at a time

  With not too many levels, the stacked proportions may still be readable quite accurately

  Using this kind of color-coding is completely useless in standard histograms


number of rooms


rent per m²