

Exercise sheet 2: Car Insurance Dataset

Background

The dataset contains information for one year from a German car insurance company on 52,903 policies. The data were collected in 1996.

Information in the dataset

Variable	Description
<i>Birthyr</i>	policyholder's year of birth
<i>Gender</i>	policyholder's gender
<i>Bundesland</i>	German Bundesland
<i>Country</i>	policyholder's nationality (?)
<i>Regyr</i>	year car registered
<i>Areareg</i>	area of registration
<i>kw</i>	engine power (kW)
<i>Job</i>	farmer, civil servant or other
<i>Claim</i>	type of claim: material damage, personal injury, none
<i>Payment</i>	claim amount paid out
<i>NoClaim</i>	no claims bonus class

Some questions

1. Which variables might be most useful? Draw appropriate plots of them to see if there are any interesting patterns. If necessary, adjust the scales to make them more interpretable and comparable.
2. How would you describe the distribution of year of birth. Draw a histogram *Birthyr*. What values would you suggest for the anchorpoint and the class width in the histogram?
3. Is the age of the car, *Regyr*, related to the age of the driver, *Birthyr*?
4. Do women drive more or less powerful cars than men? Draw barcharts of *Gender* and *kw* and use linking.
5. Are *Claim* and *Job* related?
6. Are *Country* and *Bundesland* related?

Of course, there may be many other interesting questions you could consider!