WS #12 - Feature Engineering

Monday, October 21, 2024

Your Name: ____

Names of people you worked with: ____

Do you get enough exercise? What do you do?

Task:

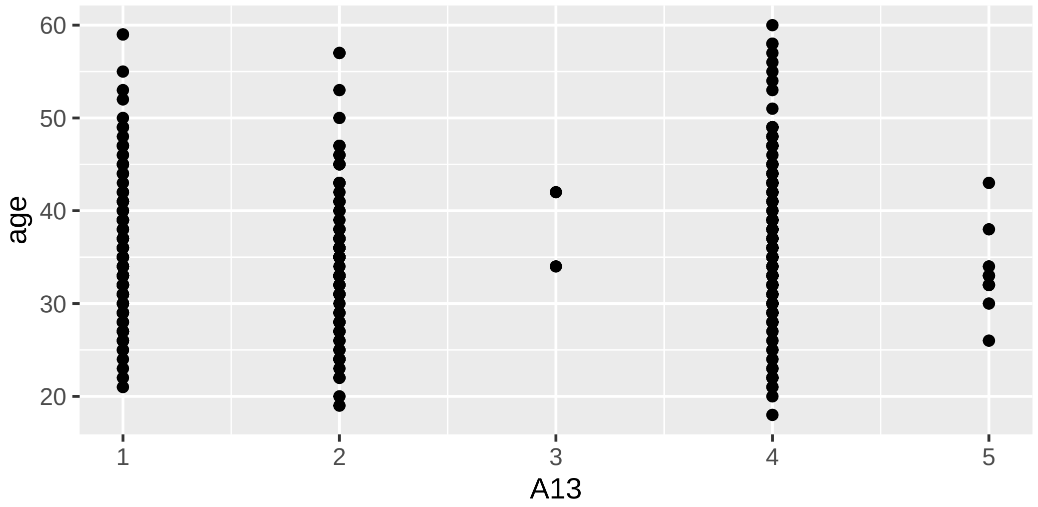
Consider the following scenarios where the idea is to build a linear model. For each:

- Why can't we use the X variable as is?
- How should X be adjusted?

Scenario 1:

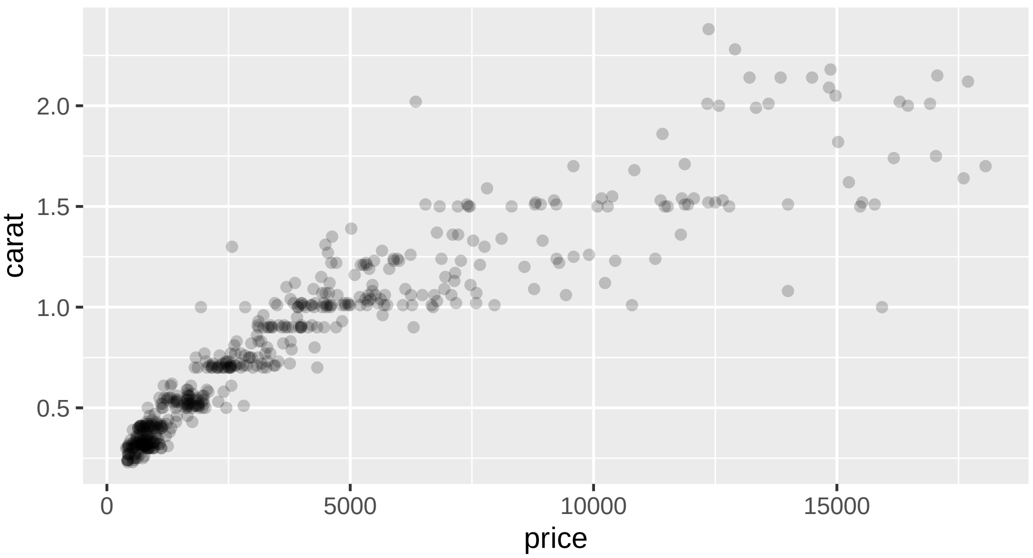
Data from the Health Evaluation and Linkage to Primary Care study. Y is `age`; X is:

> `A13`: Usual employment pattern in last 6 months (1=Full time, 2=Part time, 3=Student, 4=Unemployed, 5=Control envir)



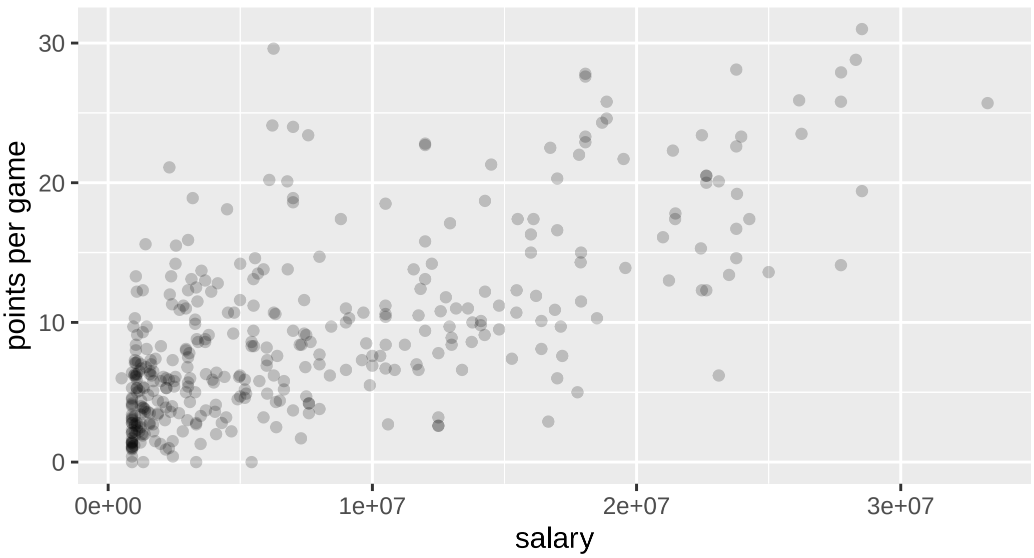
Scenario 2a:

500 randomly selected diamonds, Y is `carat` (roughly, the size of the dimond), X is `price`.



Scenario 2b:

X is NBA `salary` for 2017-18; Y is `points per game`.



Solution

Scenario 1

- Categorical variables should not be used as linear numbers. There is not a “one unit increase” between each value of the category.
- To fix the problem, we use each category as its own binary (yes/no) variable.

Scenarios 2

- In both images, there are a large number of observations in the bottom left corner. The extreme observations (top right) are likely to swamp any linear model built on the raw variables. Additionally, the relationships really don’t seem linear.
- Consider transforming the X variable (and/or the Y variable), possibly with a natural log transformation.