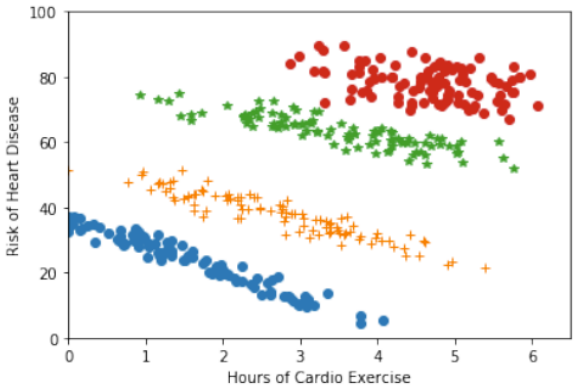
# Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

## A benefit of clustering

One advantage for downstream applications is, there might be patterns that are more obvious if you only look at related segment of the data. For example:




If we take the data as a whole (**no** clustering), we would draw a **positive** regression: it seems that exercise and heart disease increase **together**. That doesn't make sense!

But, if we divide it into **clusters**, based on age, we see a **negative** relationship: each individual group experiences **benefits** from exercise.

This particular issue is called **Simpson's Paradox**.

**Definition 1**

**Simpson's Paradox** is when a **trend** that appears in groups of data either **vanishes** or **reverses** when we look at all the data **together**.

It shows that sometimes, **patterns** that we see may reflect how we're **looking** at the data.

Rest assured, you don't need to know this paradox by **name**! But it's important to **understand** possible problems like it: it'll make you more **responsible** in the future!