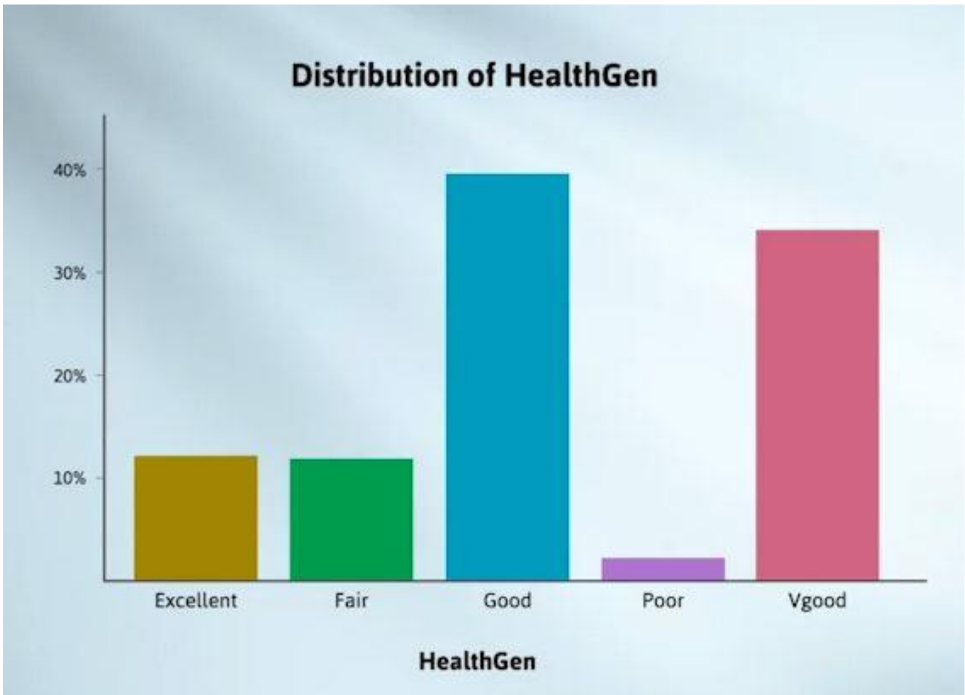
# DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

THE UNIVERSITY OF AUCKLAND

## WEEK 2

### 2.4 ORDERING CATEGORIES by Chris Wild

In this video, we're picking up our categorical variable story where we left off last time and adding a few small refinements.

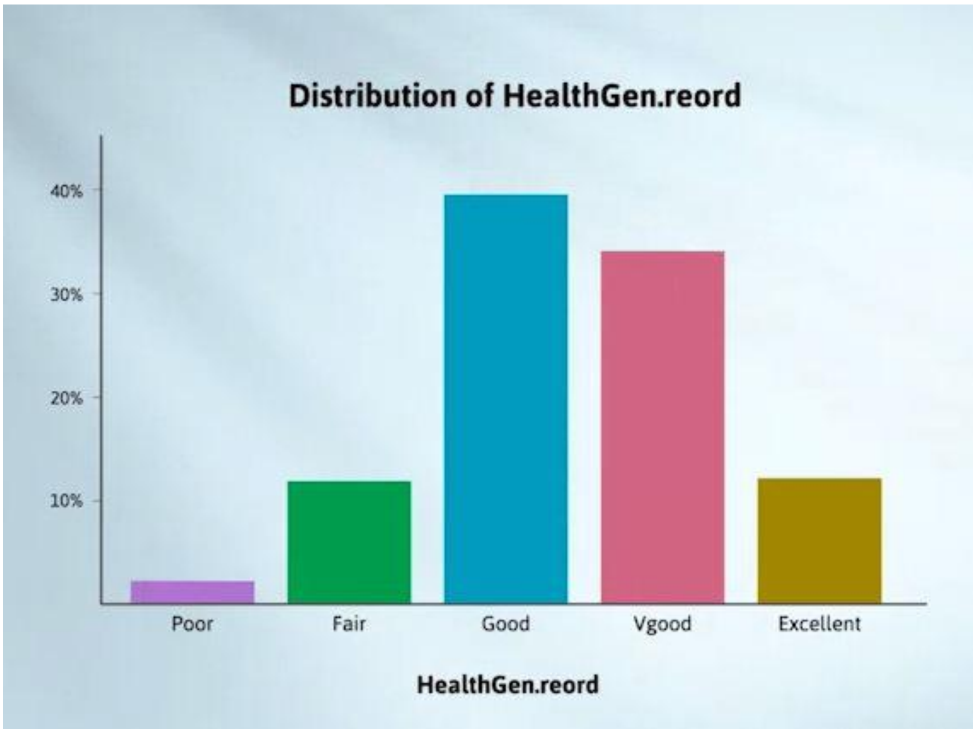


This is the NHANES variable HealthGen. It's a self-reporting of the person's general state of health.

For most data analysis programmes, the default behaviour is to represent categories in alpha-numeric order. This means that letters sort alphabetically, and numbers sort in front of letters.

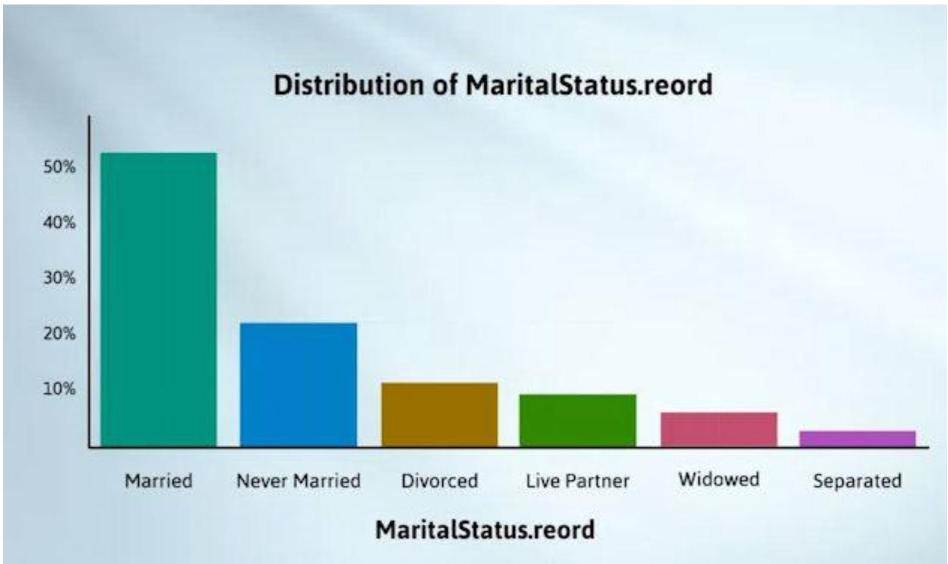
Sometimes, by dumb luck, this works just fine. But at other times, well, the ordering of the groups here, Excellent, going to Fair, then Good, then Poor, then Very Good is just plain dumb.

There is a natural ordering to the HealthGen categories, and it should be used as we've done here.

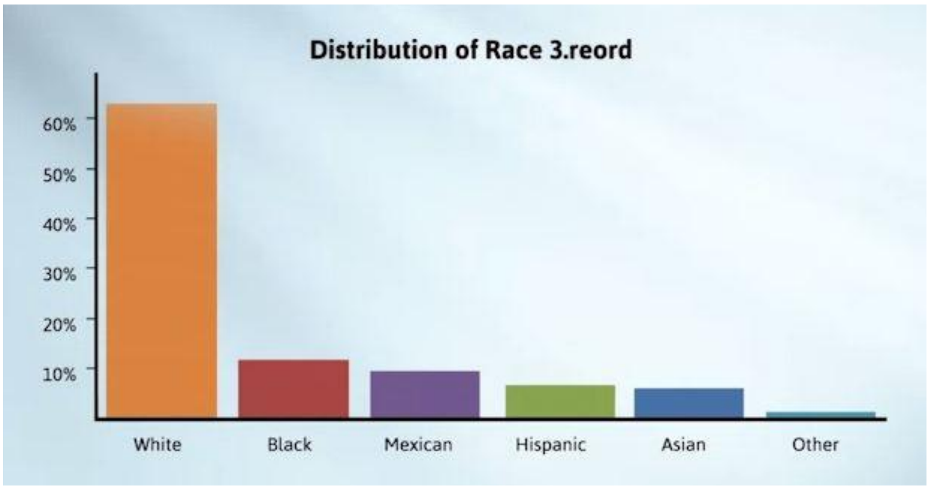


We'll pause for a little bit of name-calling. If the categories have a natural ordering, the variable is said to be ordinal. If there's no natural ordering, the variable is said to be nominal.

If there's no particular reason for setting categories in some other order, ordering them by height order is usually the most informative. We call this ordering by frequency. Frequency is just another name for a count.



An exception is a catch-all category like "other."



People usually use other to sweep up a set of entities that have nothing much in common except that they don't belong to any of the other groups. It's done to avoid having too many very small categories. For presentation graphics, I prefer to put such a category out to the far right-hand side regardless of its popularity because it's largely a meaningless category.

To wrap up, we have the following ways of ordering.

Alphabetic order helps us to locate a particular item in a list just as we do with a name in a phone book.

Frequency order is best for letting us see relative popularity. But if there's a natural order, we'd normally want to use it.

And now I'll leave you with these questions to remind you of the ideas we've just been covering.

QUESTIONS

- How does most software order categories by default?
- What other orderings should we consider and what are their advantages?