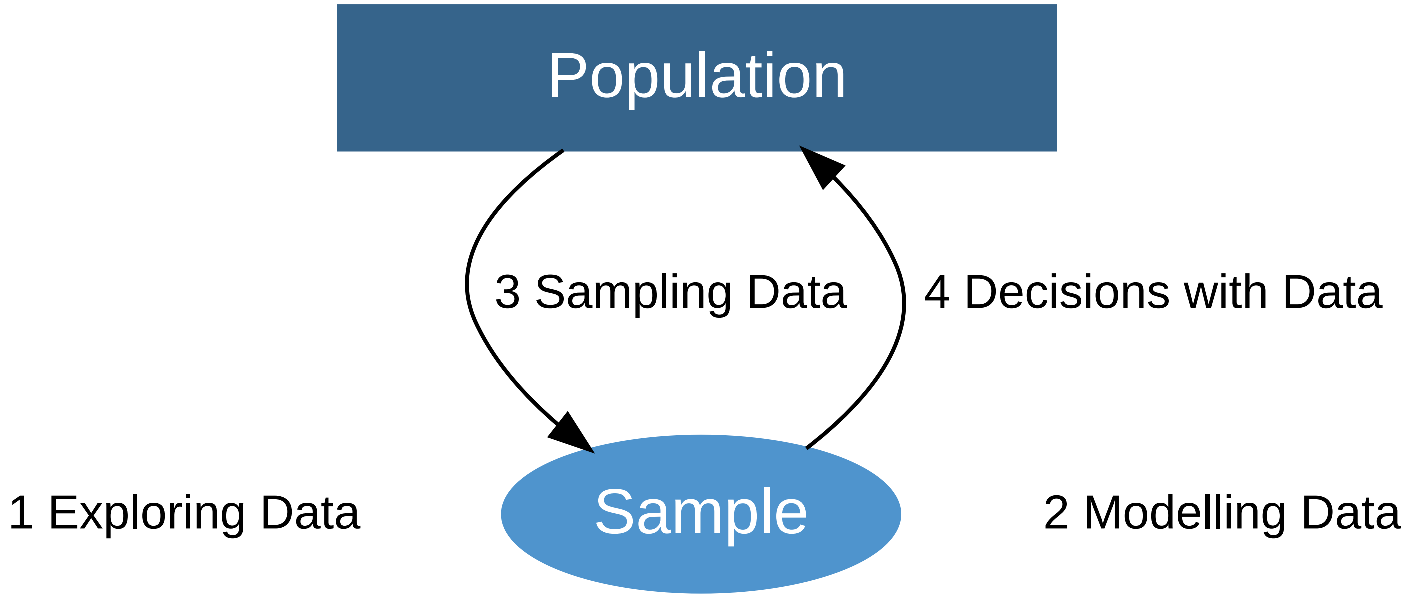
Data & Qualitative Data

Exploring Data | Data & Graphical Summaries

Unit Overview



Population

3 Sampling Data

4 Decisions with Data

1 Exploring Data

Sample

2 Modelling Data

Module1 Exploring Data

Design of Experiments

Where did the data come from & can we make reliable conclusions?

Data & Graphical Summaries

What type of data do we have & how can we visualise it?

Numerical Summaries

What are the main features of the data?

Data & Qualitative Data

Data Story | What causes Australian Road Fatalities?

Initial Data Analysis

Structure of the Data

Graphical Summaries

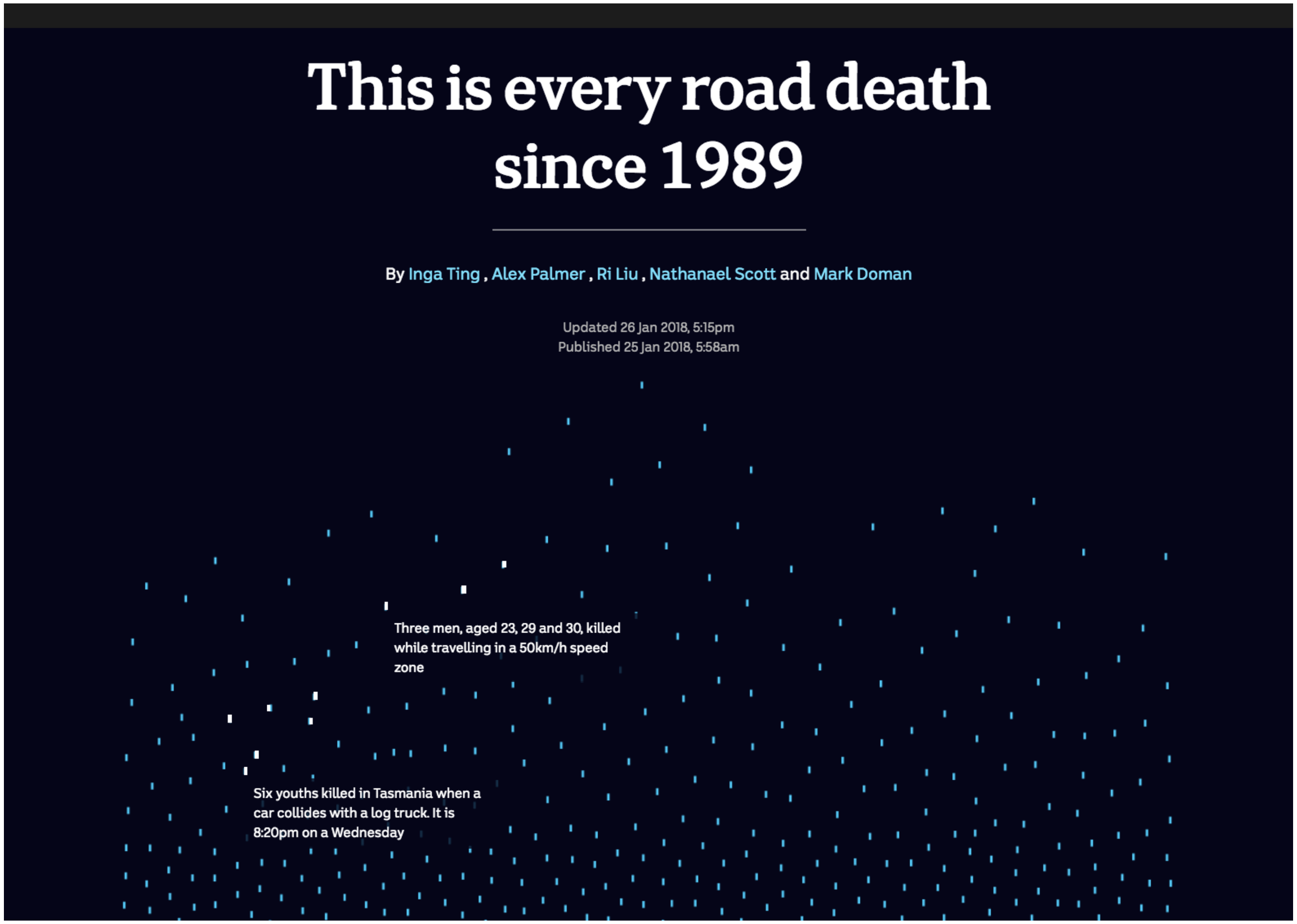
Qualitative Data

Extra Example

Summary

Data Story

What causes Australian Road Fatalities?



This is every road death since 1989

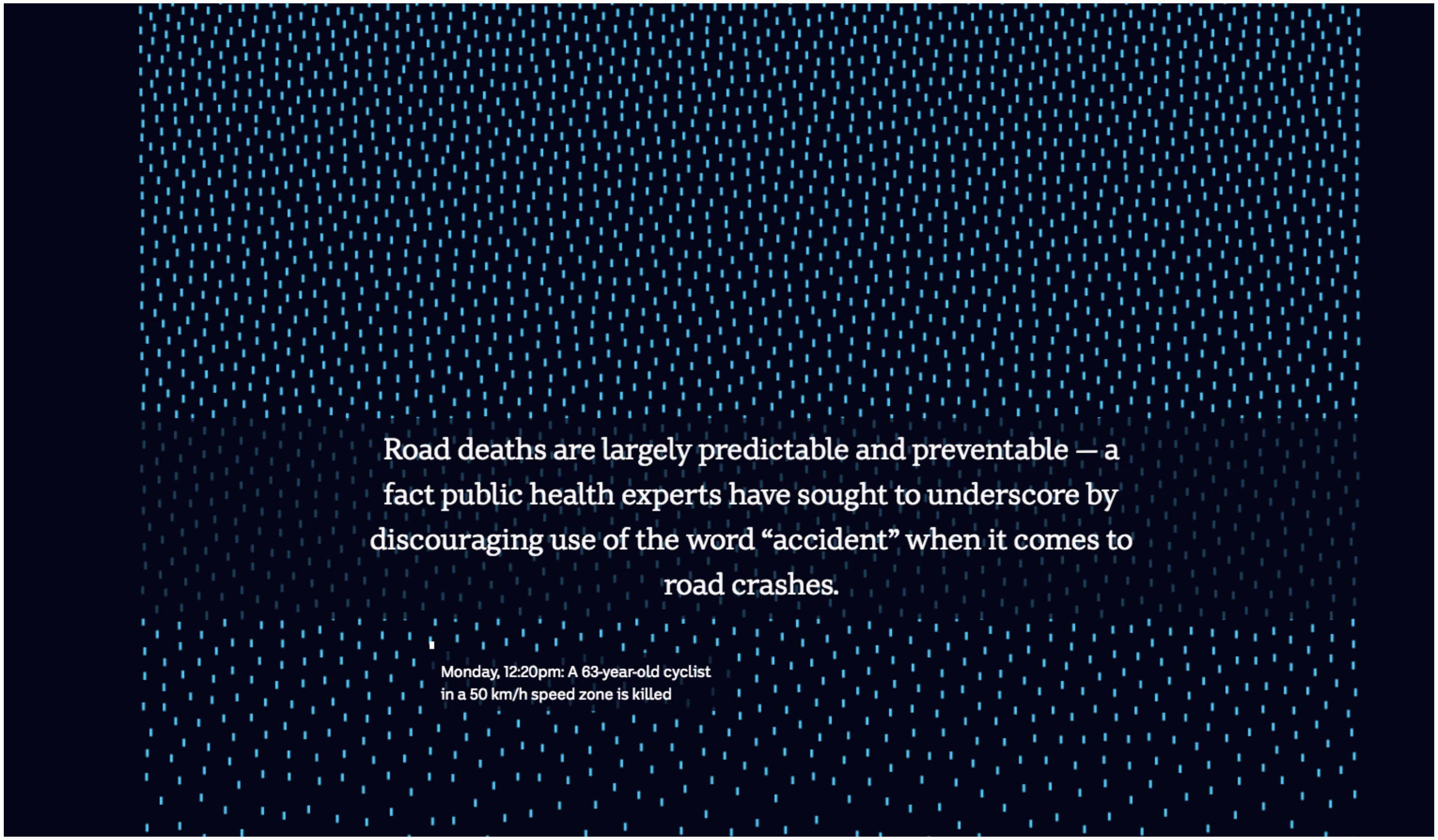
By Inga Ting , Alex Palmer , Ri Liu , Nathanael Scott and Mark Doman

Updated 26 Jan 2018, 5:15pm
Published 25 Jan 2018, 5:58am

Three men, aged 23, 29 and 30, killed while travelling in a 50km/h speed zone

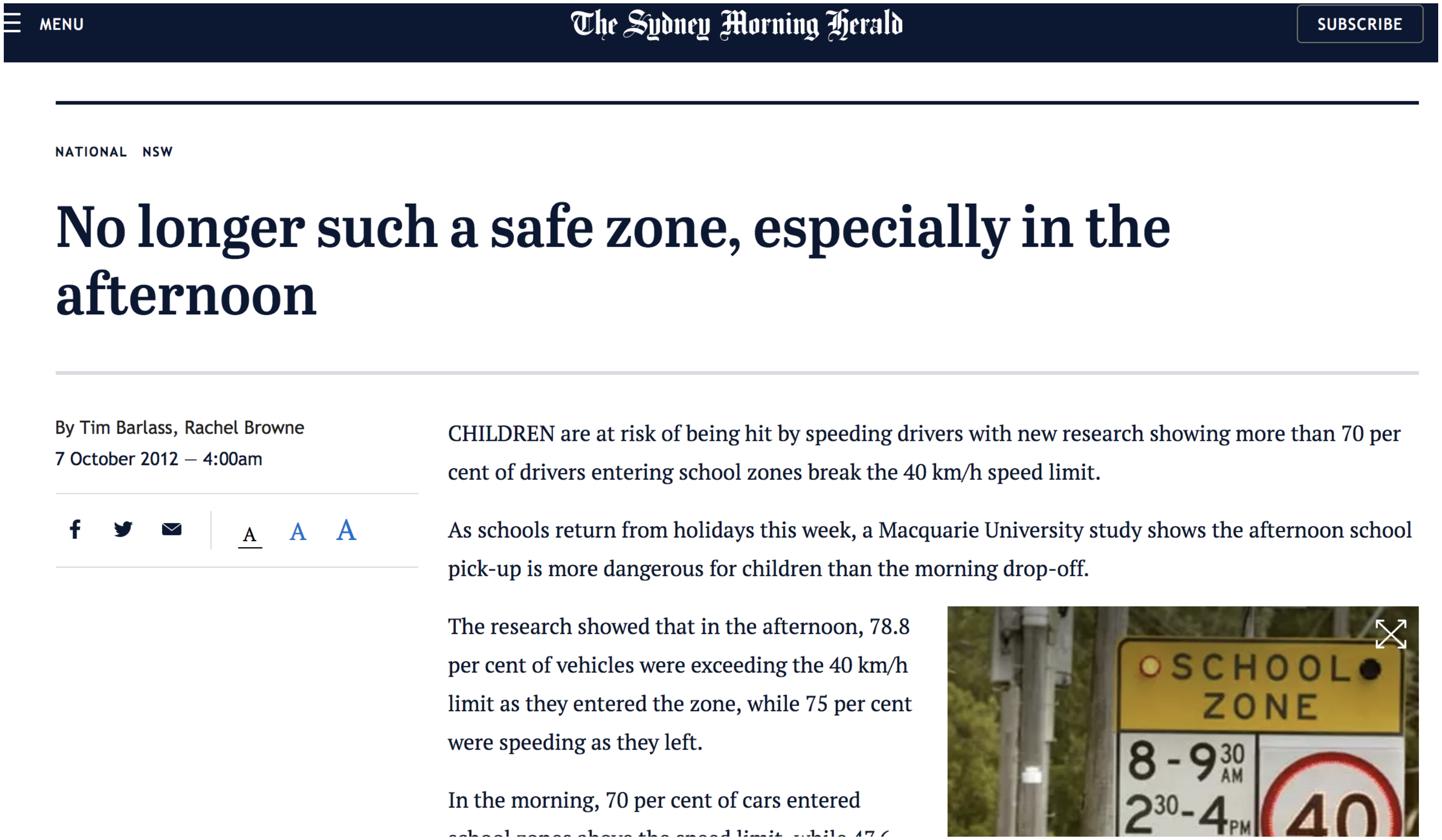
Six youths killed in Tasmania when a car collides with a log truck. It is 8:20pm on a Wednesday

🔗 ABC Animation



Road deaths are largely predictable and preventable — a fact public health experts have sought to underscore by discouraging use of the word “accident” when it comes to road crashes.

Monday, 12:20pm: A 63-year-old cyclist in a 50 km/h speed zone is killed



The Sydney Morning Herald

NATIONAL NSW

No longer such a safe zone, especially in the afternoon

By Tim Barlass, Rachel Browne
7 October 2012 — 4:00am

CHILDREN are at risk of being hit by speeding drivers with new research showing more than 70 per cent of drivers entering school zones break the 40 km/h speed limit.

As schools return from holidays this week, a Macquarie University study shows the afternoon school pick-up is more dangerous for children than the morning drop-off.

The research showed that in the afternoon, 78.8 per cent of vehicles were exceeding the 40 km/h limit as they entered the zone, while 75 per cent were speeding as they left.

In the morning, 70 per cent of cars entered school zones above the speed limit, while 47.6

SMH

Australian Road Fatality Data

- Despite preventative measures such as compulsory seat belts (from 1970) and school zones (2001), the number of road fatalities in Australia continues to rise.
  - In 2015, 1,209 died in road fatalities.
  - Why?
- We are going to investigate data from the Australian Bureau of Statistics (ABS) for January-April 2016.

```
# Read in data
data = read.csv("data/2016Fatalities.csv",header=T)
# Names of Variables
names(data)
```

```
##  [1] "Crash.ID"                        "State"
##  [3] "Date"                            "Day"
##  [5] "Month"                           "Year"
##  [7] "Dayweek"                         "Time"
##  [9] "Hour"                            "Min"
## [11] "Crash.Type"                      "BusInvolvement"
## [13] "RigidTruck..Involvement"         "Articulated.Truck..Involvement."
## [15] "SpeedLimit"                      "RoadUser"
## [17] "Gender"                          "Age"
```

Find Data Dictionary

Statistical Thinking

What questions do you have?

- How many road fatalities have there been so far this year, and how does it compare to last year?
- What is the most common day and time for a crash?
- Does biological sex affect the type of road fatality?
- What is the chance that a motorcycle rider is involved in a road fatality?
- How many people wear seatbelts?

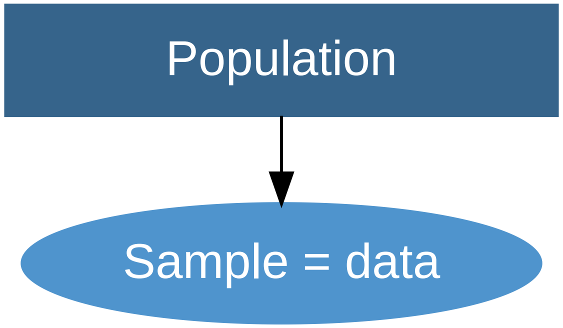
Initial Data Analysis

What is data?

Data

Data is **information** about the set of **subjects** being studied (like road fatalities).

- Most commonly, data refers to the sample, not the population.



Different types of data

There are many different types of data, in different formats.

For example:

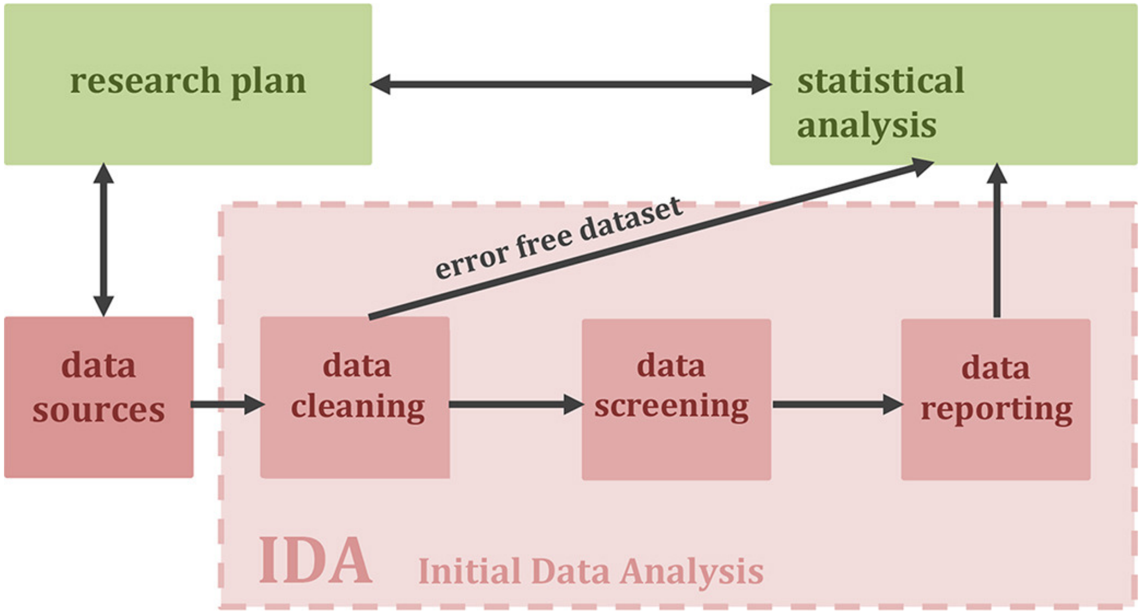
- survey data
- spreadsheet type data
- MRI image data

Initial Data Analysis (IDA)

Initial Data Analysis (IDA)

Initial Data Analysis is a first general look at the data, without formally answering the research questions.

- IDA helps you to see whether the data can answer your research questions.
- IDA may pose other research questions.
- IDA can
  - identify the data's main qualities;
  - suggest the population from which a sample derives.



“The purposes of IDA are to ensure that the later statistical analysis can be performed efficiently and to minimize the risk of incorrect or misleading results.”

Huebner et al. 2016

What’s involved in IDA?

Initial Data Analysis commonly involves:

- data background: checking the quality and integrity of the data
- data structure: what information has been collected?
- data wrangling: scraping, cleaning, tidying, reshaping, splitting, combining
- data summaries: graphical and numerical

Here we focus on **structure** & **graphical summaries** for qualitative and quantitative data.

Structure of the Data

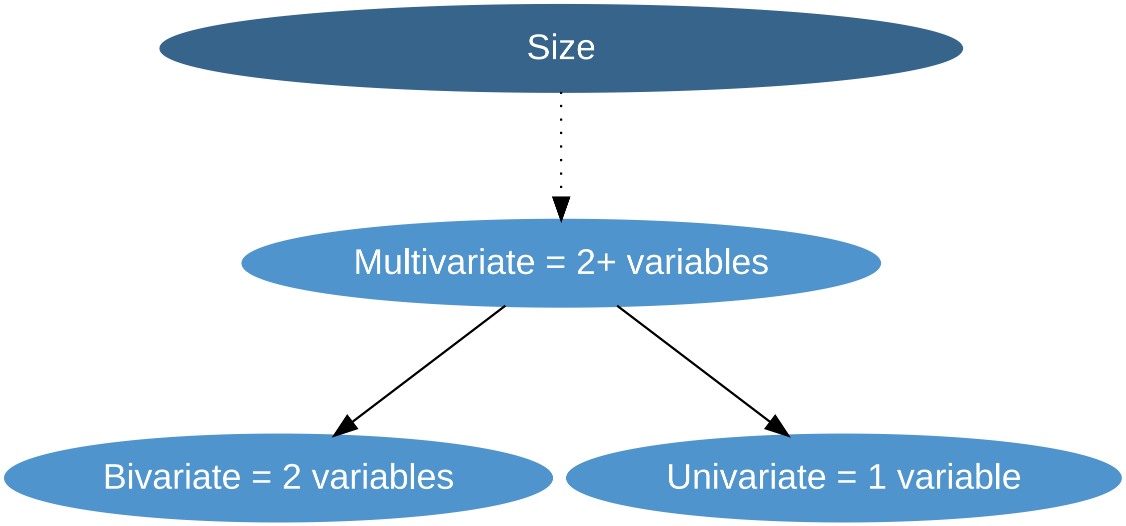
Variables

Variable

A **variable** measures or describes some attribute of the subjects.

- Data with p variables is said to have **dimension** p.

Number of variables



Statistical Thinking

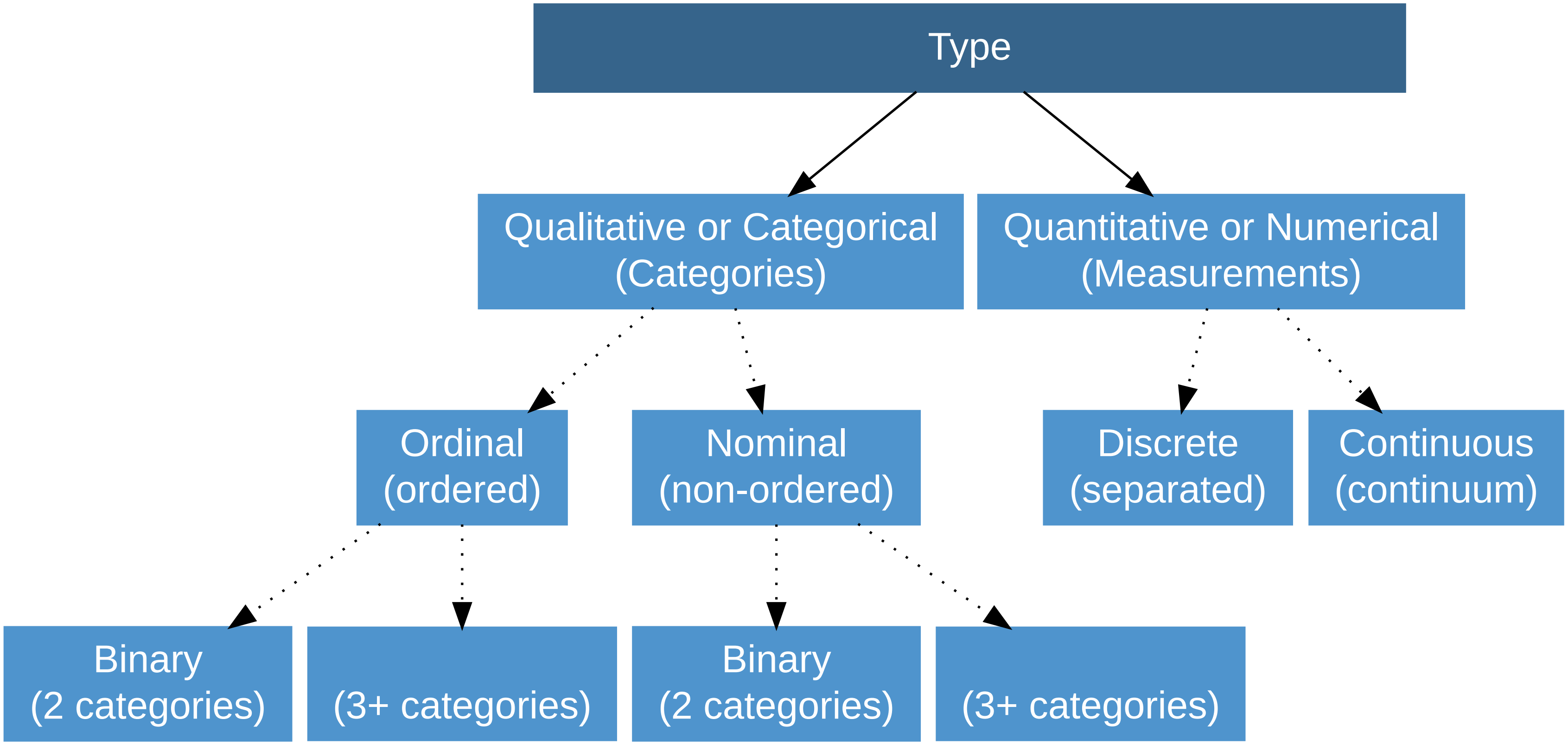
How many variables does the road fatality data have?

- The road fatality data has dimension $p = 17$, as the CrashID serves as an anonymous identifier.

```
# Size of Data
dim(data)
```

```
## [1] 442  18
```

Types of variables



Statistical Thinking

Classify the variable `Age` in the Road Fatality Data.

- Technically Age is a quantitative, continuous variable, but here the ages have been reported as discrete ‘integer’ (by rounding down to the nearest year).
- Age may be also be recorded as a qualitative variable in a survey, as respondents may be more willing to give their age category. However, it is optimal to record quantitative data if possible.

Suggest a similar variable.

- Income.

```r
# Structure of Data
str(data)
```

```
## 'data.frame':    442 obs. of  18 variables:
##  $ Crash.ID                       : num  3.2e+12 3.2e+12 2.2e+12 3.2e+12 1.2e+12 ...
##  $ State                          : Factor w/ 8 levels "ACT","NSW","NT",..: 4 4 7 4 2 8 2 7 8 5 ...
##  $ Date                           : Factor w/ 113 levels "1-Apr-16","1-Feb-16",..: 29 62 73 57 102 106 70 58 103 32 ...
##  $ Day                            : int  16 24 26 22 7 8 26 23 7 17 ...
##  $ Month                          : Factor w/ 4 levels "April","February",..: 2 1 4 4 1 1 1 1 2 1 ...
##  $ Year                           : int  2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...
##  $ Dayweek                        : Factor w/ 7 levels "Friday","Monday",..: 6 4 3 6 5 1 6 3 4 4 ...
##  $ Time                           : Factor w/ 225 levels "0:00","0:12",..: 53 187 90 73 219 106 108 67 44 46 ...
##  $ Hour                           : int  13 5 16 15 9 17 17 14 12 12 ...
##  $ Min                            : int  0 0 0 0 20 10 19 32 26 35 ...
##  $ Crash.Type                     : Factor w/ 3 levels "Multiple vehicle",..: 1 3 2 2 2 1 2 3 1 1 ...
##  $ BusInvolvement                 : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RigidTruck..Involvement        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Articulated.Truck..Involvement.: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SpeedLimit                     : int  100 70 100 50 50 50 50 -9 60 110 ...
##  $ RoadUser                       : Factor w/ 6 levels "Bicyclist (includes pillion passengers)",..: 2 2 6 6 6 6 6 5 5 5 ...
##  $ Gender                         : Factor w/ 2 levels "Female","Male": 1 2 1 2 2 2 2 1 1 1 ...
##  $ Age                            : int  -9 -9 0 1 1 1 2 4 5 5 ...
```

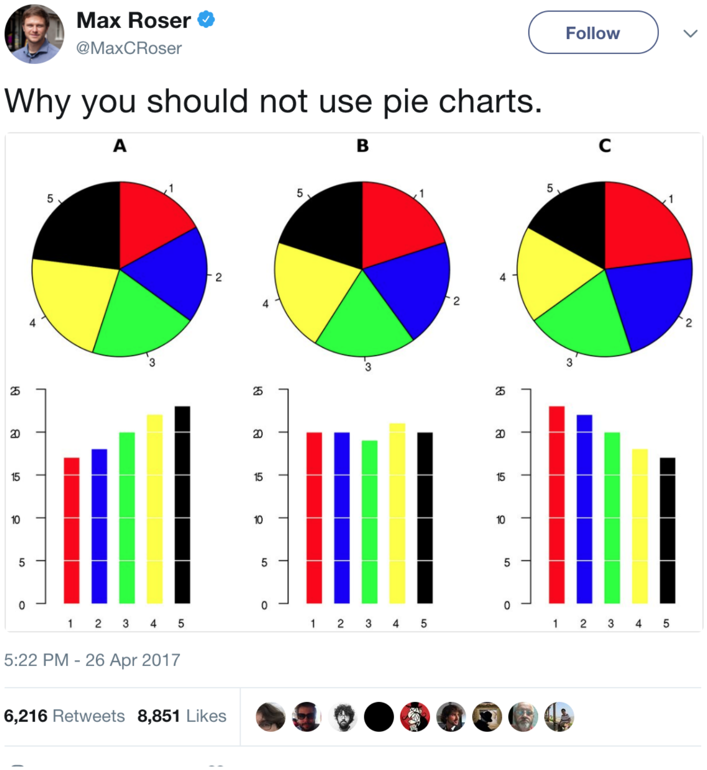
Graphical Summaries

Choosing a graphical summary

The aim of a graphical summary is to best highlight features of this data.

- To some extent we use trial and error.
- While the pie chart may be popular, it is usually not informative.
- The relatively new ‘Shiny Apps’ present data in accessible ways. See examples.

Twitter debate about pie charts



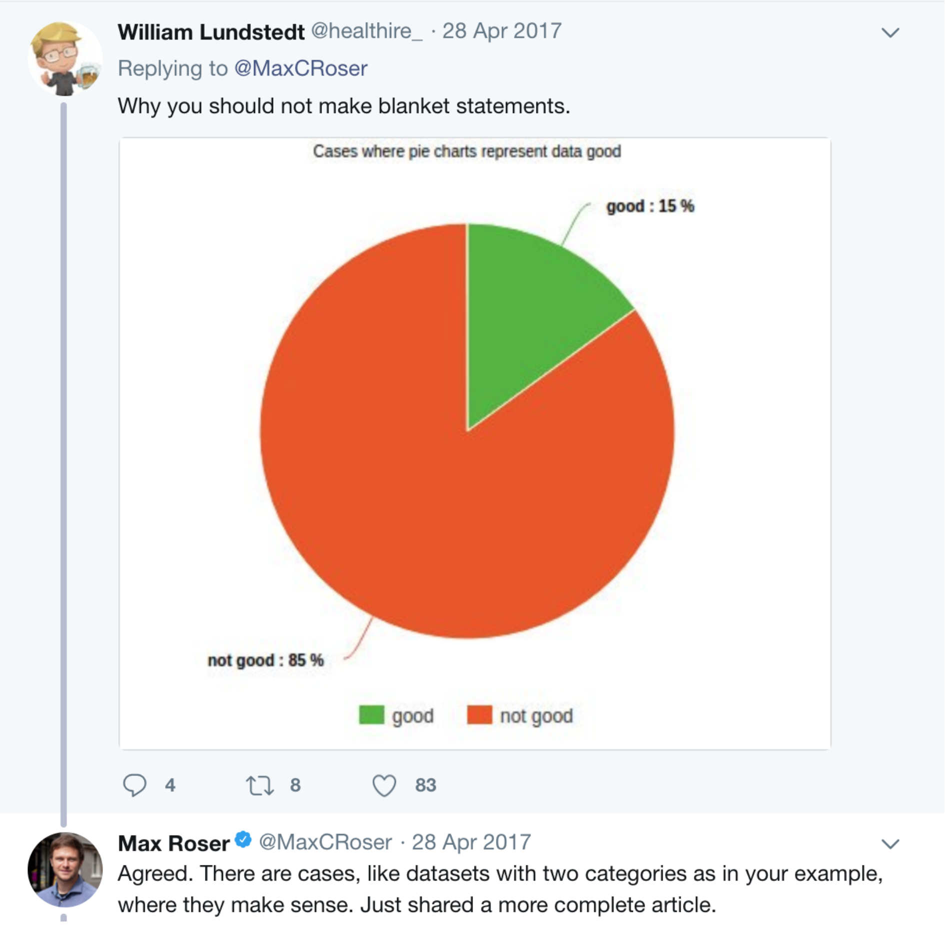
Max Roser
@MaxCRoser

Why you should not use pie charts.

5:22 PM - 26 Apr 2017

6,216 Retweets 8,851 Likes

Twitter



William Lundstedt @healthire_ · 28 Apr 2017

Replying to @MaxCRoser

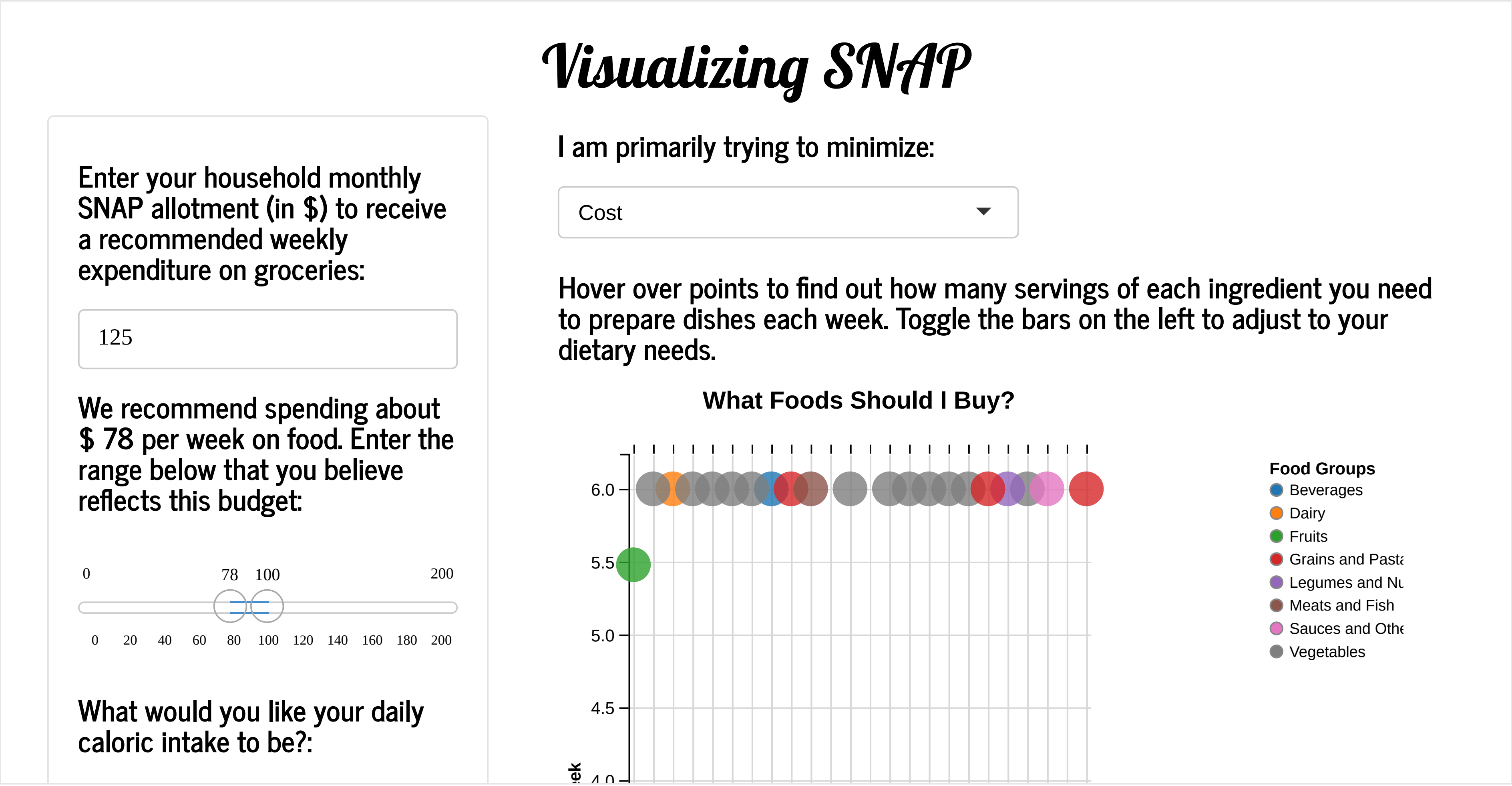
Why you should not make blanket statements.

Max Roser @MaxCRoser · 28 Apr 2017

Agreed. There are cases, like datasets with two categories as in your example, where they make sense. Just shared a more complete article.

Twitter

Shiny Apps



Visualizing SNAP

Enter your household monthly SNAP allotment (in $) to receive a recommended weekly expenditure on groceries:

125

We recommend spending about $ 78 per week on food. Enter the range below that you believe reflects this budget:

What would you like your daily caloric intake to be?:

I am primarily trying to minimize:

Cost

Hover over points to find out how many servings of each ingredient you need to prepare dishes each week. Toggle the bars on the left to adjust to your dietary needs.

What Foods Should I Buy?

Qualitative Data

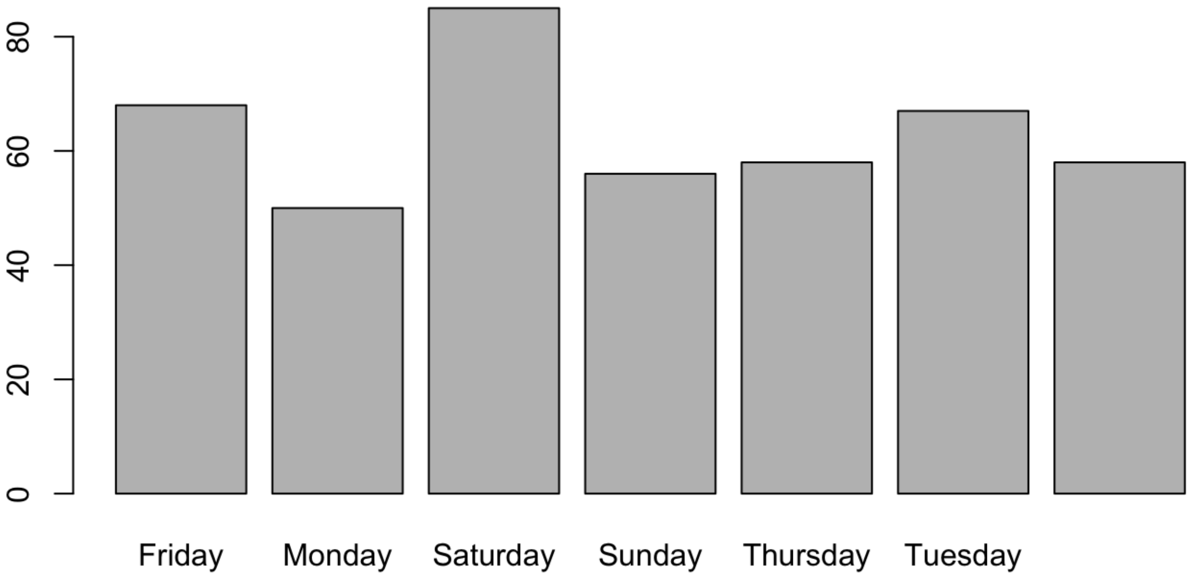
Simple bar plot

The bar plot (or bar chart or bar graph) is a simple summary of qualitative data.

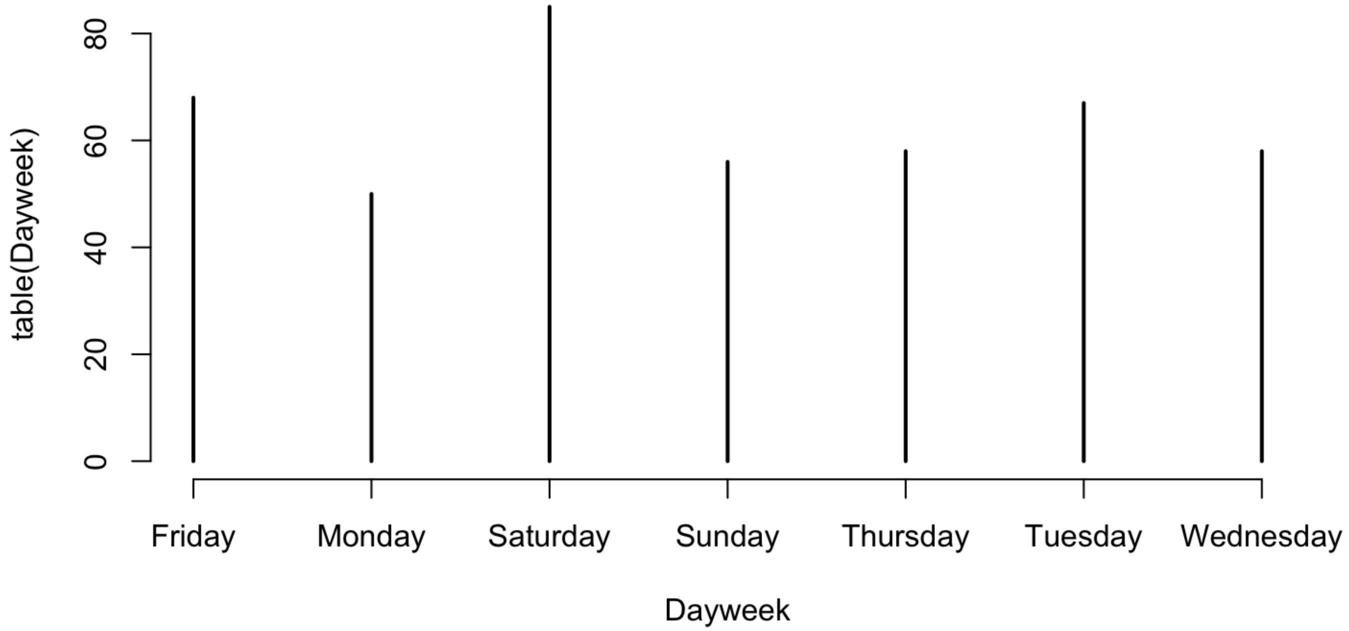
```
# Select the DayWeek variable from the whole data frame
Dayweek = data$Dayweek

# Produce a frequency table of fatalities per day of the week
table(Dayweek)
```

```
## Dayweek
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday 
##        68        50        85        56        58        67        58 
```

```
# Produce a bar chart
barplot(table(Dayweek))
```



```
# A simpler version
plot(table(Dayweek))
```



Statistical Thinking

What was the most common day of road fatality?

- Saturday

Why might that be the case?

- More volume of cars on the road, or people driving faster?

What data would you need to check your hypotheses?

- Data on volume and speed of cars on the road each day.

Double bar plot

Things get more interesting when we consider 2 qualitative variables.

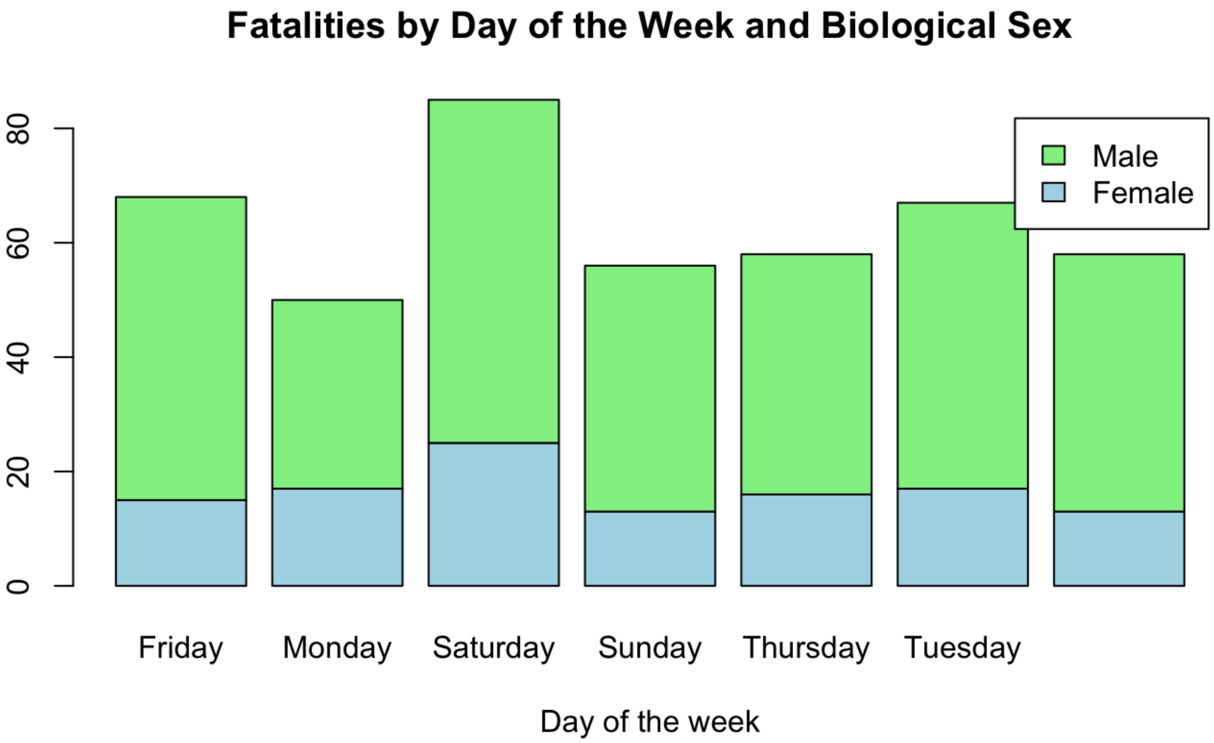
```
# Select DayWeek and Gender (Biological Sex) variables
Dayweek = data$Dayweek
Gender = data$Gender

# Produce a double frequency table (contingency table)
data1 = table(Gender, Dayweek)
data1
```

```
##         Dayweek
## Gender   Friday Monday Saturday Sunday Thursday Tuesday Wednesday
##   Female     15     17       25     13       16      17        13
##   Male       53     33       60     43       42      50        45
```

Note: Here we have called the variables by the names in the data set. Here “Gender” refers to biological sex, as it was historically recorded in this dataset. Read more.

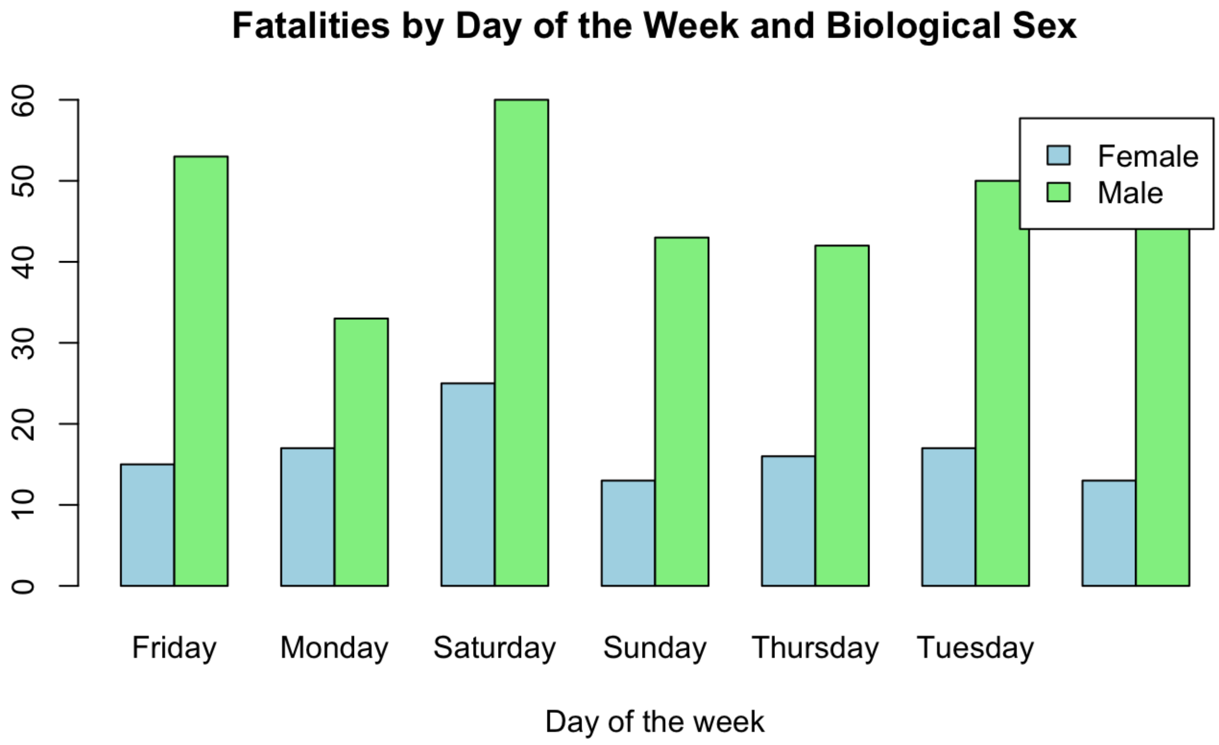
Stacked Bar Plot

```
barplot(data1, main="Fatalities by Day of the Week and Biological Sex",
  xlab="Day of the week", col=c("lightblue","lightgreen"),
    legend = rownames(data1))
```



Side-by-Side Bar Plot

```
barplot(data1, main="Fatalities by Day of the Week and Biological Sex",
  xlab="Day of the week", col=c("lightblue","lightgreen"),
    legend = rownames(data1), beside=TRUE)
```



Statistical Thinking

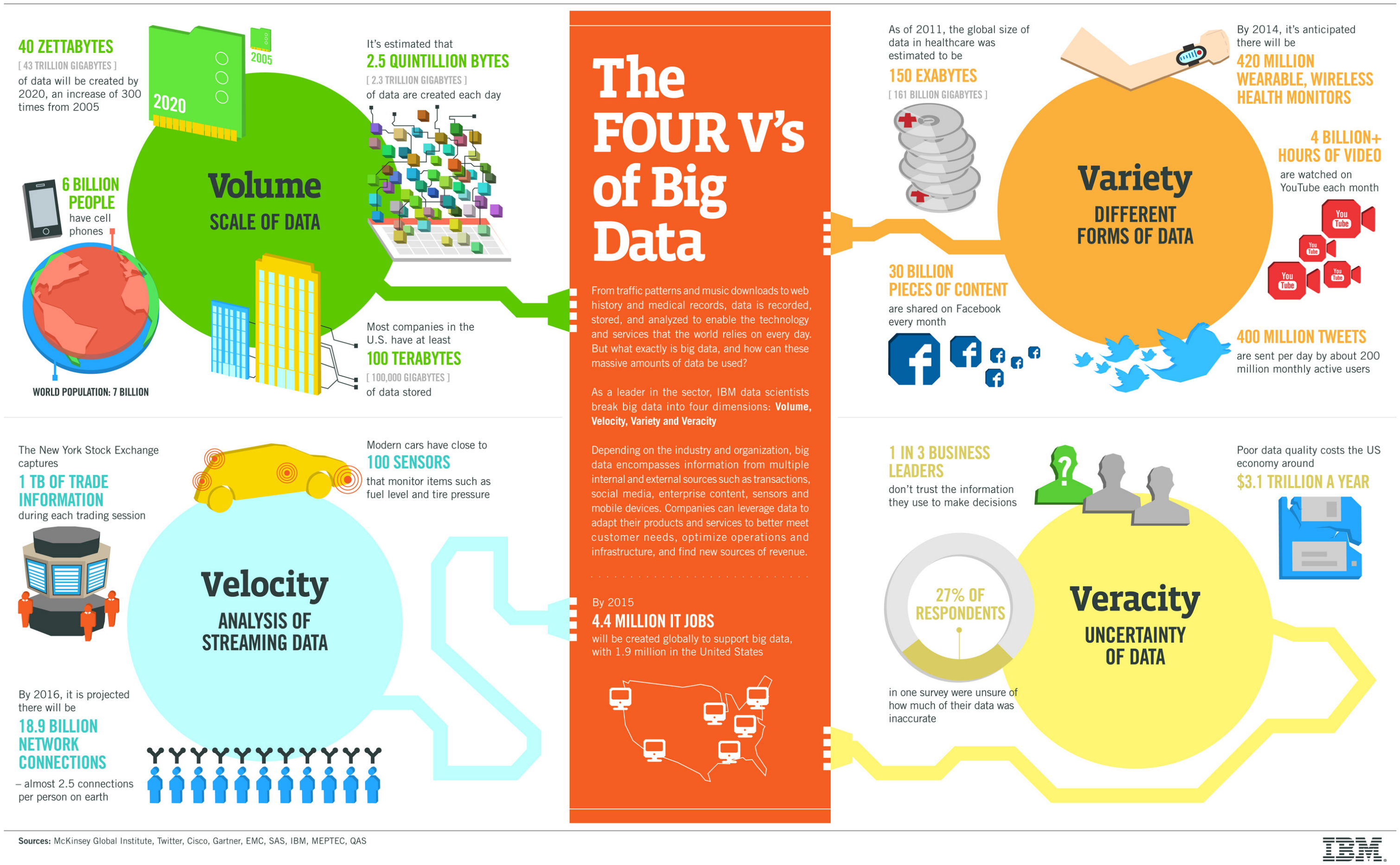
Do you have a prefence for stacked or side-by-side? Why?

Are these plots telling us anything useful? How could they be misread?

- There seems to be a similar proportion of fatalities across each day, for biological sex.
- We could posit that men are more likely to be involved in fatal accidents than women. However, perhaps there are more men on the road than women. More data is needed.

Graphical summaries of big data

- **Big data** is the massive amounts of data being collected in fields such as genomics, astrophysics, marketing and sociology.
- Big data is commonly **high dimensional**, which means that there are more variables p than subjects n.
    - For example, genomics data can have 3 billion variables, as a person's DNA sequence is 3 billion basepairs long.
- Big data can be described by many "V"s: high volume, high velocity, high variety, high variability, low veracity/validity, high vulnerabiity, high volatility and high value.
- Big data requires more complex visualisations.



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM.

Extra Example

Investigating Speed Limit

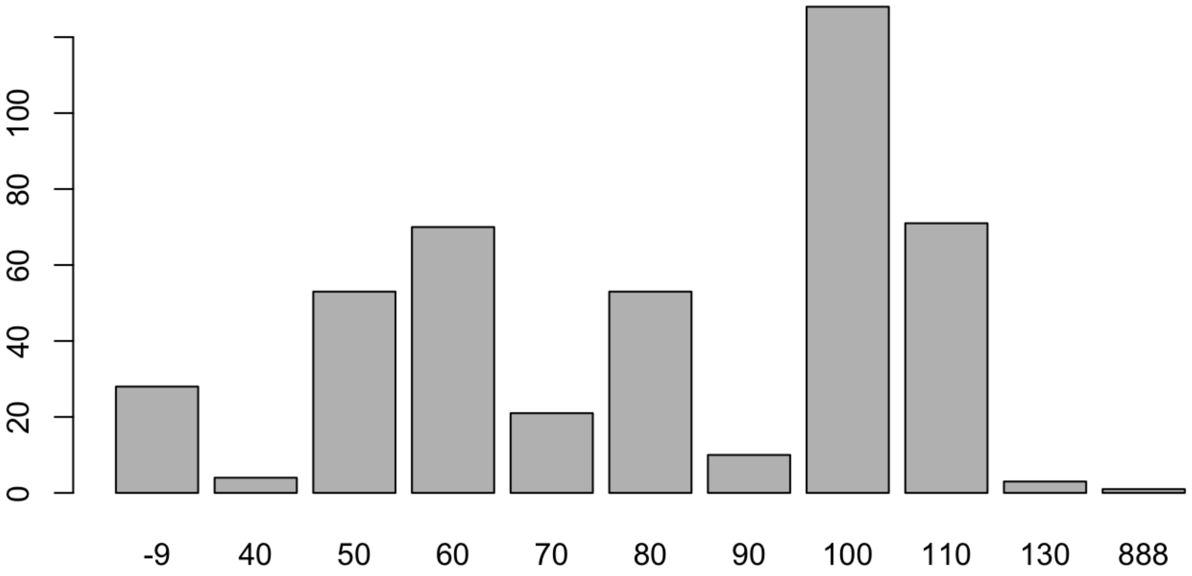
Speed Limit could be considered as a discrete, quantitative variable, but it is best classified as an (ordinal) qualitative variable. Why?

```
SpeedLimit = data$SpeedLimit
table(SpeedLimit)
```

```
## SpeedLimit
##  -9  40  50  60  70  80  90 100 110 130 888
##  28   4  53  70  21  53  10 128  71   3   1
```

Simple Bar Plot

```
barplot(table(SpeedLimit))
```



Statistical Thinking

What is curious about this data? Why?

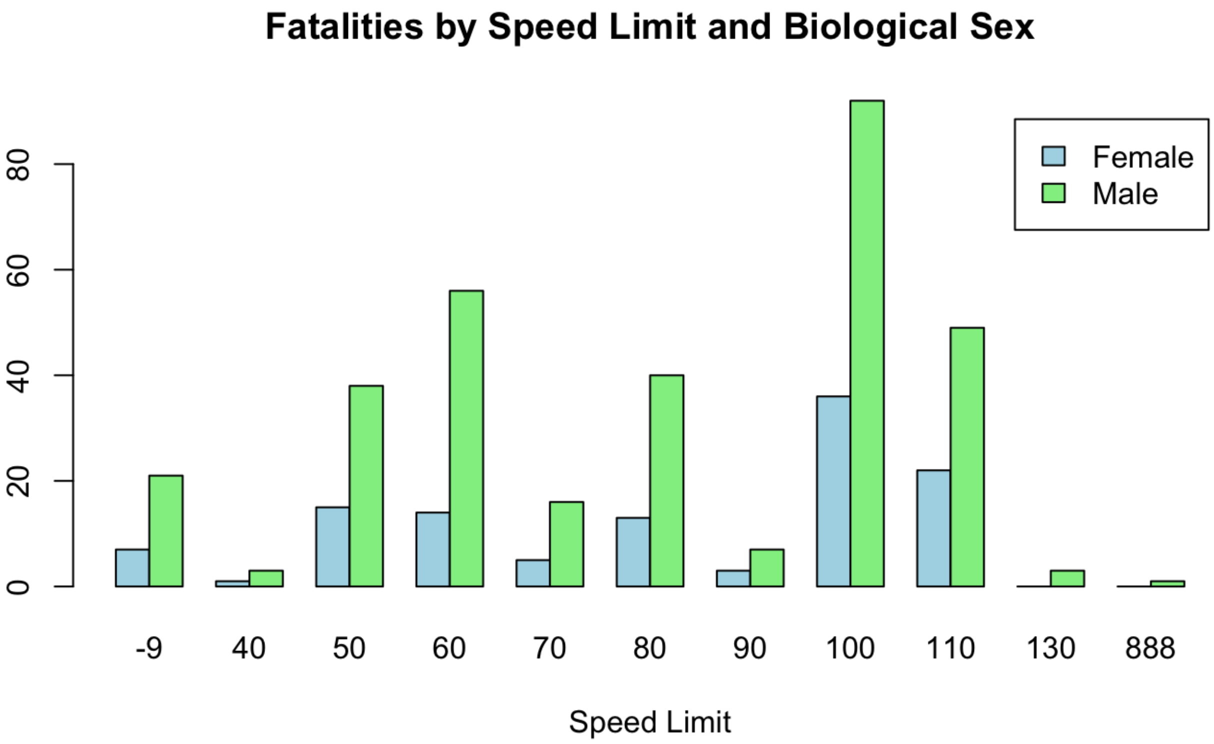
- -9 indicates a missing value. Why speed of 888? We could ‘clean’ the data.

What was the most common speed at which a road fatality occurred? How might this affect public policy?

- Notice how many fatalities happen at high speeds (100km/hr+).
- This might lead to change in speed limits or investigation of road conditions for high speeds.
- Can we assume that a vehicle travelling in that speed zone was travelling at that speed limit?

Double Bar Plot

```
#Gender = data$Gender
data2 = table(Gender,SpeedLimit)
barplot(data2, main="Fatalities by Speed Limit and Biological Sex",
  xlab="Speed Limit", col=c("lightblue","lightgreen"),
    legend = rownames(data1), beside=TRUE)
```



Statistical Thinking

Are there any interesting patterns?

Summary

The type of variables determines what type of graphical summary.

1 Qual	2 Qual	1 Quant	2 Quant	1 Quant + 1 Qual
Simple Bar Plot	Double Bar Plot	Histogram Box Plot	Scatter Plot	Comparative Box Plot

Key Words

data, subjects, population, sample, Initial Data Analysis, data background, data structure, data wrangling, data summaries, variable, dimension, multivariate, bivariate, univariate, qualitative/categorical, quantitative/numerical, ordinal, nominal, discrete, continous, binary, bar plot (chart)

Further Thinking

Big data in the humanitarian sector

Data is the new seatbelt

Data is the new seat belt | Yiem Sunbhanich | TEDxColumbus