Intro to R

Hadley Wickham & Heike Hofmann

Goals

- A gentle introduction to R:
  learn how to view data and produce graphics
- Practice your question generation skills
- Learn which plots are best for answering which questions
- Revise reading plots
- Explore a large data set with graphics

Learning a new language is hard!

Diamonds data

- ~54,000 round diamonds from http://www.diamondse.info/
- Carat, colour, clarity, cut
- Total depth, table, depth, width, height
- Price

Getting started

-
  ```
  install.packages(”ggplot2”)
  # once per computer
  ```

-
  ```
  library(ggplot2)
  # every time you open R
  ```

-
  ```
  head(diamonds)
  str(diamonds)
  # Two ways to inspect a data set
  ```

-
  ```
  # Make sure you type things exactly;
  # R is very fussy
  ```

What can we learn from this data?

- Inspect the data
- Figure out what the variables are from http://www.diamondse.info/ and wikipedia
- **Write down** questions that you could answer with this data
- 4 minutes by yourself, then pair up for another 3 minutes, and we'll write ideas on the board

Answers

- Explore how one (or more) variables are distributed - barchart or histogram
- Explore how two variables are related - scatterplot, boxplot, tile plot
- Explore how two variables are related, conditioned on other variables - facetting

Scatterplot

- Two continuous variables
- `qplot(carat, price, data=diamonds)`
- `qplot(log(carat), log(price), data=diamonds)`
- `qplot(carat, price/carat, data=diamonds)`

Revision:
Interpreting a scatterplot

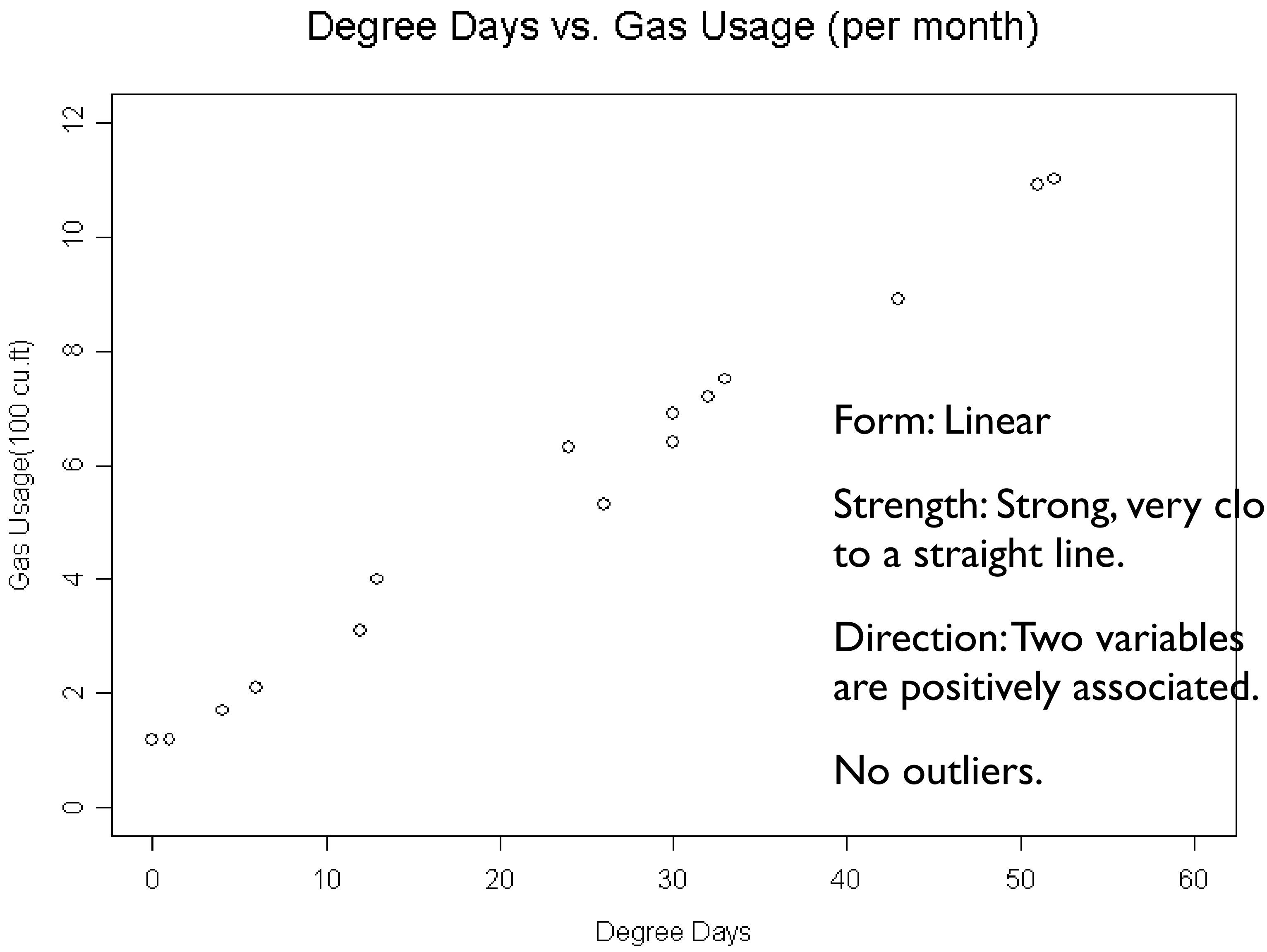
- Big patterns
  - Form and direction
  - Strength
- Small patterns
- Deviations from the pattern
  - Outliers

Interpreting Scatterplots

- **Form**
  - Is the plot linear? Is the plot curved? Is there a distinct pattern in the plot? Are there multiple groups?
- **Strength**
  - Does the plot follow the form very closely? Or is there a lot of variation?

Interpreting Scatterplots

- **Direction**
  - Is the pattern increasing? Is the plot decreasing?
    - **Positively**: Above (below) average in one variable tends to be associated with above (below) average in another variable.
    - **Negatively**: Above (below) average in one variable tends to be associated with below (above) average in another variable.

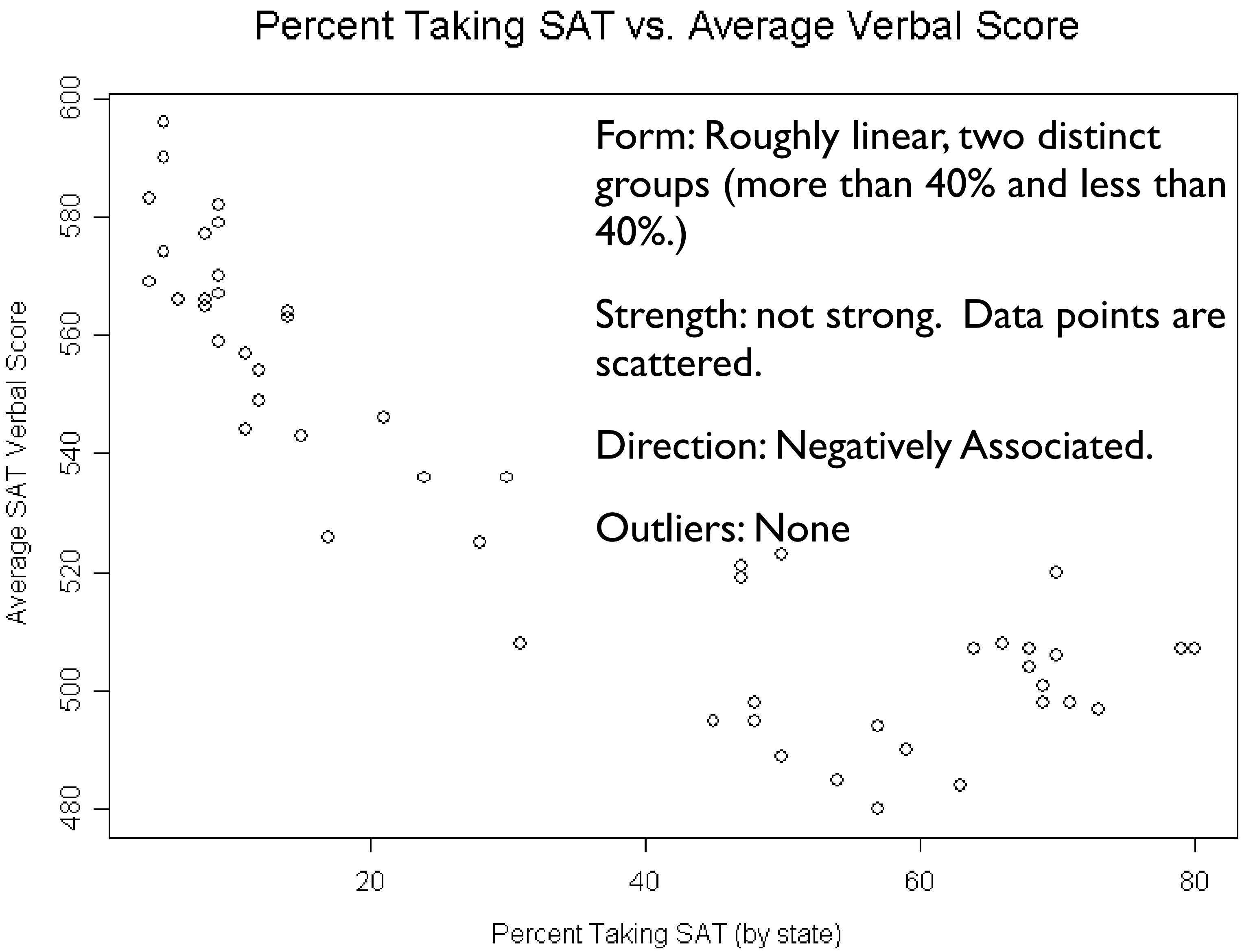


Form: Linear

Strength: Strong, very close to a straight line.

Direction: Two variables are positively associated.

No outliers.



Form: Roughly linear, two distinct groups (more than 40% and less than 40%.)

Strength: not strong. Data points are scattered.

Direction: Negatively Associated.

Outliers: None

Aesthetics

- Can map other variables to size or colour

-
  ```
  qplot(carat, price, data=diamonds,
  colour=color)
  ```

-
  ```
  qplot(carat, price, data=diamonds,
  size=carat)
  ```

-
  ```
  qplot(carat, price, data=diamonds,
  shape=cut)
  ```

Facetting

- Can **facet** to display plots for different subsets
- Row variables ~ column variables (. for none)
- `qplot(price, carat, data=diamonds, facets = . ~ color)`
- `qplot(price, carat, data=diamonds, facets = color ~ clarity)`

Facets vs aesthetics

- Will need to experiment as to which one answers your question/tells the story best
- Remember, just like with pivot tables we want comparisons of interest to be close together

Your turn

- Work through each of the example plots
- Try variations to answer your questions

Finished?

- Continue to polish your questions about the data
- Go to http://had.co.nz/ggplot2 and figure out how to make other plots that you know about

Histograms and barcharts

- Used to display the **distribution** of a variable
  - Continuous variable → histogram
  - Categorical variable → bar chart
- For the histogram, you should always vary the binwidth

Examples

```
qplot(cut, data=diamonds, geom=”bar”)

qplot(price, data=diamonds,
geom=”histogram”)

qplot(price, data=diamonds,
geom=”histogram”, binwidth=500)

qplot(price, data=diamonds,
geom=”histogram”, binwidth=100)

qplot(price, data=diamonds,
geom=”histogram”, binwidth=10)
```

Aesthetics & facetting

- Like for scatterplot, you can map `fill` to another variable, or use facetting to compare subsets
- Facetting is generally more useful, as it is easier to compare different groups

Your turn

- Explore the distribution of carat
- What can you see? What might explain that pattern?
- Make sure to experiment with bin width!
- Use facetting to explore the relationship between price and colour

Zooming

- `qplot(price, data=diamonds, geom=”histogram”, xlim=c(0, 5000))`