Intro to graphics in R

Heike Hofmann, Hadley Wickham

Outline

- Histograms
- Relationships of
  - Continuous vs Categorical Data
  - Categorical vs Categorical Data

Getting Started

- Start R
- load ggplot2 into your workspace:
  library(ggplot2)
- Bring up a summary of the diamonds data:
  summary(diamonds)

Histograms

- Divide data into bins
- Count number of observations in each bin

Always experiment with the bin width!

If you use the default bins in your assignments,
I will not give you any marks

Histograms

- 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=50)

Interpreting Histograms

- Big Pattern: Shape of the data
  - peaked vs flat,
  - skew vs symmetric

- Small Pattern:
  - location/number of modes
  - gaps (or areas of low density)

Interpreting Histograms

- Stability
  - Reaction to changes in binwidth/anchor points

Your turn

- Create histograms for carat, height, width and depth
- Experiment with bin size
- For height, width and depth, you may want to set xlim=c(0,5) as well (why?)

Investigating relationships

- Used scatterplots for two continuous variables
- What about a continuous and a categorical?
- What about two categorical?

Cont vs cat

- If we use a scatterplot, there is a lot of overplotting
- Some solutions:
  - jitter points randomly so they don't overlap
  - summarise the distribution using boxplots or histograms

Cont vs cat

- qplot(color, price/carat, data=diamonds)
- qplot(color, price/carat, data=diamonds, geom="jitter")
- qplot(color, price/carat, data=diamonds, geom="jitter", position=position_jitter(xjitter=2))
- qplot(color, price/carat, data=diamonds, geom="boxplot")

Cont vs cat

- qplot(price/carat, data=diamonds, facets= color ~ ., type="histogram")
- qplot(price/carat, data=diamonds, facets= color ~ ., type="histogram", binwidth=100)

Your turn

- Explore relationships between clarity, color and cut, with price, carats or price/carat
- Are there any interesting patterns?

- Advanced: Try to work out what alpha blending is, and figure out, how it works

Cat vs cat

- Use fluctuation diagram
- ggfluctuation(table(diamonds$cut, diamonds $color))
- Heatmaps not as useful: ggfluctuation(table(diamonds$cut, diamonds $color), type="colour")
- Very important categorical variables in sensible order
- May want to standardise

Experimentation

- diamonds$cut <- factor(diamonds$cut, c("Fair","Good","Very Good", "Premium","Ideal"))
- diamonds$clarity <- factor(diamonds $clarity, c("IF", "VVS1", "VVS2", "VS1", "VS2", "SI1", "SI2", "I1"))
- cut_color <- table(diamonds$cut, diamonds $color)
- names(dimnames(cut_color)) <- c("cut", "color")

Experimentation

- ggfluctuation(cut_color)
- ggfluctuation(cut_color / rowSums(cut_color))
- ggfluctuation(cut_color / colSums(cut_color))

Experimentation

- melt(cut_color)
- qplot(cut, value, data=melt(cut_color))
- qplot(cut, value, data=melt(cut_color), geom="line", id=color)
- qplot(cut, value, data=melt(cut_color / rowSums(cut_color)), type="line", id=color, colour=color)

Your turn

- Investigate other relationships between categorical variables