Contents

1 Bertin's theory of data visualization

1.1 Semiology of graphics

retinal variables

A system for producing graphics

1. invariant (thing that stays same across all relationships in the graphic, title for graphic, topic - e.g. sales by county)
2. the components (they vary, represented by legend e.g. sales). Each component has 3 properties to analyze:
    - order
    - length - no of divisions
    - level - nominal, ordered, or quantitative concept

2 Grammar of Graphics

Chapter 2

3 R

element wise operations

- Package installer for packages with install dependencies checked.
- or use if{}

<- assignment operator, read as gets %>% - input to

- You can read directly from web as well.
- You should name chunks so that you can locate the errors later.

4 EDA

Ask questions about data.

- What kind of variation occurs within variables?
- What kind of covariation occurs between variables?

4.1 Variation

4.1.1 Bar chart

(`geom_bar`)to visualize variation in categorical variables.

4.1.2 Histogram

(`geom_hist` and `freq_poly`) to visualize variation in continuous variables.

4.1.3 Categorical and continuous together using color parameter for categorical variable.

4.1.4 Typical values

Set bin_width to a very low value to find most common and rare values.

4.1.5 Unusual values

To make it easy to see the unusual values, zoom in to small values of the y-axis with `coord_cartesian(ylim=c(0,upper_limit))`.

Note: ggplot's xlim and ylim thro away the data whereas this doesn't.

4.1.6 Missing Values

2 options to deal with unusual data:

1. Drop the entire row with the strange values:

```
diamonds2 <- diamonds %>%
filter(between(y, 3, 20))
```

2. Replace unusual data with missing values. (Better)

```
diamonds2 <- diamonds %>%
mutate(y = ifelse(y < 3 | y > 20, NA, y))
```

To suppress missing data removed warning in ggplot,

```
ggplot(....) +
geom_point(na.rm=TRUE)
```

4.2 Covariation

If variation describes the behavior within a variable, covariation describes the behavior between variables. Covariation is the tendency for the values of two or more variables to vary together in a related way.

4.2.1 A categorical and a continuous

- freq_poly with `y = ..density..` to change y-axis instead of count to compare different categories (which have different distribution) fairly.
- box_plot

4.2.2 2 categorical variables

geom_tile

4.2.3 2 continuous variables

- Scatter_plots
- For huge data, add transparency using alpha for better visualization.
- Other options are Hex_bin (`geom_hex`) and `geom_bin2d`
- Treat one continuous variable as categorical by binning it and then use previous techniques (1 categorical and 1 continuous).

4.3 Patterns and models

Patterns

Patterns in your data provide clues about relationships. If a systematic relationship exists between two variables it will appear as a pattern in the data.

Patterns reveal covariation. If variation is a phenomenon that creates uncertainty, covariation is a phenomenon that reduces it. If two variables covary, you can use the values of one variable to make better predictions about the values of the second. If the covariation is due to a causal relationship (a special case), then you can use the value of one variable to control the value of the second.

Models

Models are a tool for extracting patterns out of data

```
library(modelr)
mod <- lm(log(price) ~ log(carat), data = diamonds)
```

5 Wrangle

5.1 Import -> Tidy -> transform

Use unix tools to create data

- Pipe in unix means output of command on left is the input to command on right
- tsv better than csv since the data itself may contain commas.

```
uniq -c
sort -nr
```

5.1.1 Converting csv to tsv

```
grep "ctrl-v<tab>" data.csv
perl -pe 's/,/ctrl-v<tab>/g' < data.csv > data.tsv
```

```
import pandas as pd
def fun():
  return a
```

$$\vec{a}$$

< - input > - output

-p print every line -e the next command is a perl program

CSV Kit - to handle common irregularities.

csv is not a good format but used most frequently and hence a common problem in data science.

Figure 1: image