Today: Using the Internet
Grammar of Graphics
1-D Categorical
Friday: ggplot2, 1-D Categorical

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

January 20, 2016

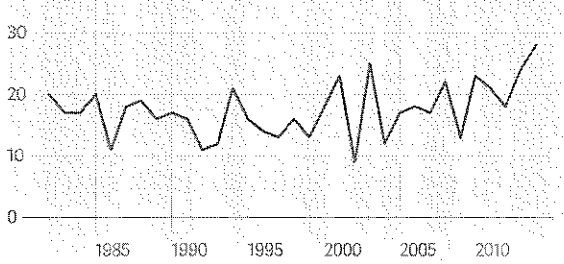
Via Quartz: Who Oscar Winners Thank

The Academy

It's standard procedure to thank the hosts, of course, so the Academy itself is always the big winner in mentions:

Mentions of "Academy" in all Oscar speeches since 1982

40 total mentions

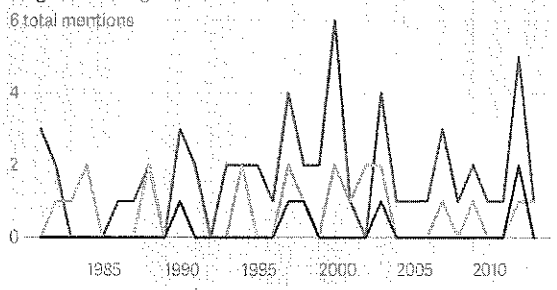


Via Quartz: Who Oscar Winners Thank

Management

No one really appreciates their publicist. At least not publicly:

Mentions of management in all Oscar acceptance speeches since 1982

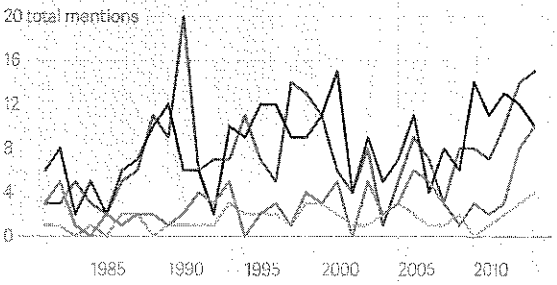


Via Quartz: Who Oscar Winners Thank

Family

Husbands apparently get more love than wives, and mentions of "children" collectively are pretty low as well:

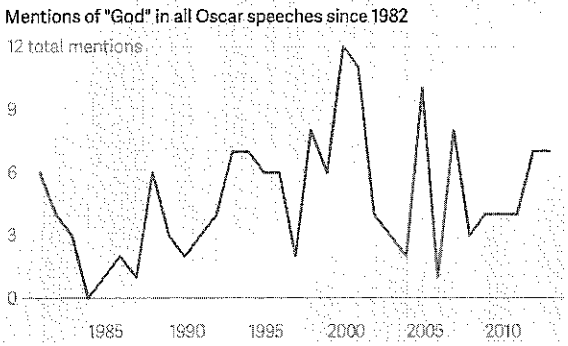
Mentions of family in all Oscar acceptance speeches since 1982



Via Quartz: Who Oscar Winners Thank

God

Tearful award winners thanking God from the podium have become something of a Hollywood cliché, but it actually happens relatively rarely. This is better thought of as a chart of winners saying "Oh my God!"



"Decorating" / Data-Ink

Graphics should not draw the viewer's attention away from the data.
Extras get in the way.

Note: Decoration does not refer to appropriate graph labeling.
Labels should always be clear, detailed, and thorough.
Label key parts of the data. Add text explanations if necessary.

Data Ink should primarily present information about the data:
the non-erasable, non-redundant core of a graphic

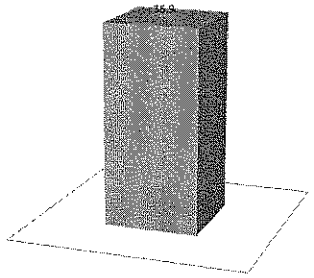
Tufte suggests using the *data-ink ratio*:

"Decorating" / Data-Ink

Two ways to increase the proportion of data-ink:

Remove non-data-ink:

Remove redundant data-ink:



R Package ggplot2 – Hadley Wickham

Based on "The Grammar of Graphics" by Leland Wilkinson, 2005

ggplot() # grammar of graphics plot

Each plot can be broken down into core components. Wilkinson defines the core components. Wickham puts them into practice in R.

Highly recommend these workshop slides:
https://opr.princeton.edu/workshops/Downloads/2015Jan_ggplot2Koffman.pdf

R Package ggplot2 – Hadley Wickham

1. **data**: in ggplot2, data must be stored as an R data frame
2. **coordinate system**: describes 2-D space that data is projected onto
   e.g., Cartesian coordinates, polar coordinates, map projections, ...
3. **geoms**: describe type of geometric objects that represent data
   e.g., points, lines, polygons, ...
4. **aesthetics**: describe visual characteristics that represent data
   e.g., for example, position, size, color, shape, transparency, fill
5. **scales**: for each aesthetic, describe how visual characteristic is converted to display values
   e.g., log scales, color scales, size scales, shape scales, ...
6. **stats** : describe statistical transformations that help summarize data
   e.g., counts, means, medians, regression lines, ...
7. **facets**: describe how data is split into subsets and displayed as multiple small graphs

How Do I Learn ggplot?

The best way to learn how ggplot works is through examples!

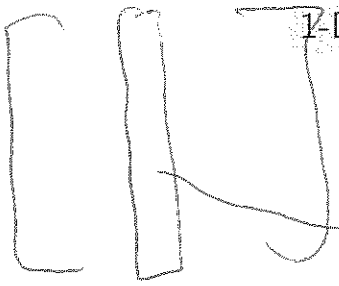
We'll go through several examples of this in Lab 02 and HW 02

Next Up: 1-D Categorical Data

Recall: Data can be **categorical** or **continuous**

Categorical data can be **ordered** or **unordered / nominal**

Data



1-D Categorical Data

Structure: vector of length n = # of rows of original dataset

How could we summarize this data? frequencies of each category
What information would you report? ↳ counts
↳ "frequentist probabilities"
percentages, proportions

- # of unique categories
- what are the unique categories?
- most / least frequent cat.?
- ordered or unordered?

2-D categorical → contingency table

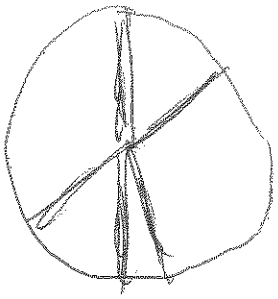
1-D Categorical Data

To show the differences among the categories, need to use *area plots*:

In graph of 1D categorical variable, we want to see differences in area of the graph corresponding to each category

Examples of area plots?

- Bar graph } variations in height of bar
- pie charts
- spine graph
- rose diagrams



1-D Categorical Data – Pie Charts

Polar coordinates

Pie Charts: circle divided up into sections ("pie slices") such that the area of each section is proportional to the number of observations with each unique categorical value.

θ = Theta ~~α~~ HW 3?

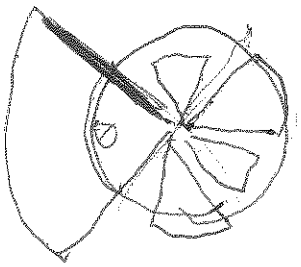
A = area $\propto$ frequency / proportion

r = radius $\propto$ nothing

↳ all are the same

Polar Coordinates

1-D Categorical Data – Rose Diagrams



Rose Diagrams: circle sections are created for each category. All sections have the same width/arc/angle. The radius is proportional to the square root of the category frequency. Sections are called "petals". Developed by Florence Nightingale (example will be posted to Blackboard).

r = radius $\propto$

$A \propto$ frequency / proportion

$A = \pi r^2 \Rightarrow A \propto r^2$

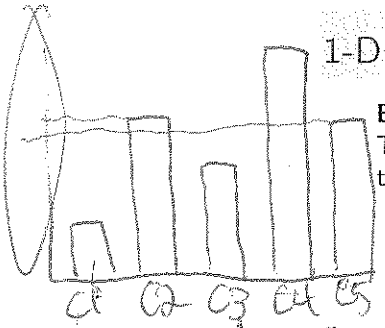
$\Rightarrow r \propto \sqrt{\text{freq/proportion}}$

θ = "theta" $\propto$ nothing

freq. or prop.

area of rectangle = width x height

1-D Categorical Data – Bar Charts



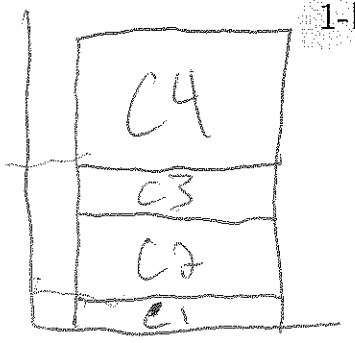
Bar Charts: rectangular bar is created for each unique categorical value. The area and height of the bar is proportional to % of observations with the categorical value. Bars usually have equal width.

width of bars ∝ nothing

heights ∝ frequency / proportion of observations that fall into that particular category

area ∝ same as height

1-D Categorical Data – Spine Charts



Spine Charts: rectangular bar is created for each unique categorical value. ~~The height of all bars is equal, and the width of the bar corresponds to the proportion in that~~ category.

harder to compare stacked heights as opposed to bar chart

width ∝ equal

heights ∝ frequency / proportion

area ∝

But: STACKED → hard to compare.