START OF DAY 10

Reading: None
Big Data Utah 2014

Marissa Saunders & Brad Stone
Air Quality Competition Report

- Domain
- Data
- Challenges
- Approach
- Results
Group Project Progress Report
3 Minute Synopsis

• What have you done?
• Preliminary results?
• What remains to be done?
• How will you go about it?
Data Visualization
Napoleon Invasion of Russia, 1812
Map representing the losses over time of French army troops during the Russian campaign, 1812-1813.
Conceived by Charles Joseph Minard, Inspector General of Public Works, retired.

Paris, 20 November 1869

The number of men present at any given time is represented by the width of the grey line; one mm. indicates ten thousand men. Figures are also written besides the lines. Grey designates men moving into Russia; black, for those leaving. Sources for the data are the works of messrs. Thiels, Segr, Fezensac, Chambry and the unpublished diary of Jacob, who became an Army Pharmacist on 28 October. In order to visualize the army's losses more clearly, I have drawn this as if the units under prince Jerome and Marshall Davoust (temporarily separated from the main body to go to Minsk and Miledow, which then joined up with the main army again), had stayed with the army throughout.

Figure 58. Minard's map of Napoleon's Russian campaign.
This graphic has been translated from French to English and modified to most effectively display the temperature data.
Snow's Cholera Map, 1855
Far East Asia at Night

Notice how dark it is!

North Korea

Seoul
South Korea
Role of Visualization

• Support interactive exploration
• Help in result presentation
• Disadvantage: requires human eyes
• Can be misleading
## Bad Visualization

<table>
<thead>
<tr>
<th>Year</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>2110</td>
</tr>
<tr>
<td>2000</td>
<td>2105</td>
</tr>
<tr>
<td>2001</td>
<td>2120</td>
</tr>
<tr>
<td>2002</td>
<td>2121</td>
</tr>
<tr>
<td>2003</td>
<td>2124</td>
</tr>
</tbody>
</table>

Y-Axis scale gives **WRONG** impression of big change
BETTER VISUALIZATION

<table>
<thead>
<tr>
<th>Year</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>2110</td>
</tr>
<tr>
<td>2000</td>
<td>2105</td>
</tr>
<tr>
<td>2001</td>
<td>2120</td>
</tr>
<tr>
<td>2002</td>
<td>2121</td>
</tr>
<tr>
<td>2003</td>
<td>2124</td>
</tr>
</tbody>
</table>

Axis from 0 to 2000 scale gives CORRECT impression of small change
Another Bad Visualization

Lie Factor

The degree to which the graphical representation of a particular effect matches the reality

\[
\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}
\]

For the fuel economy graph:

\[
\frac{(5.3 - 0.6)}{(27.5 - 18.0)} = \frac{7.833}{0.528} = 14.8
\]

Tufte’s requirement: 0.95 < Lie Factor < 1.05

Tufte’s Principles of Graphical Excellence

• Give the viewer
  – the greatest number of ideas
  – in the shortest time
  – with the least ink in the smallest space.

• Tell the truth about the data!

Visualization Methods

• Visualizing in 1-D, 2-D and 3-D
  – Well-known visualization methods
  – E.g., histogram, box plot, scatter plot

• Visualizing more dimensions
  – Scatter Plot Matrix
  – Parallel Coordinates
  – Chernoff Faces
  – Stick Figures
Scatter Plot Matrix

Represent each possible pair of variables in their own 2-D scatterplot (car data)

**Q:** Useful for what?
A: linear correlations (e.g. horsepower & weight)

**Q:** Misses what?
A: multivariate effects
Parallel Coordinates

- Encode variables along a horizontal line
- Vertical line specifies values

Dataset in a Cartesian coordinates

Same dataset in parallel coordinates

Invented by
Alfred Inselberg
while at IBM, 1985
### Example: Iris (I)

<table>
<thead>
<tr>
<th>sepal length</th>
<th>sepal width</th>
<th>petal length</th>
<th>petal width</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

---

Diagram:
- Sepal Length: 5.1
- Sepal Width: 3.5
- Petal Length: 1.4
- Petal Width: 0.2
Example: Iris (II)

Each data point is a line.

Similar points correspond to similar lines.

Lines crossing over correspond to negatively correlated attributes.

Problems:
  - order of axes up to ~20 dimensions
Chernoff Faces

Encode different variables’ values in characteristics of human face

Fun applets:
- http://www.cs.uchicago.edu/~wiseman/chernoff/

Chernoff-Faces [Che 73, Tuf 83]
Stick Figures

Two variables mapped to X, Y axes
Other variables mapped to limb lengths and angles
Census data showing age, income, sex, education, etc.

Closed figures correspond to women and we can see more of them on the left.

Note also a young woman with high income.
Summary

• Many methods
• Visualization is possible in more than 3-D
• Aim for graphical excellence (and truth!)
• Free and open-source software
  – Ggobi
  – Xmdv
  – Others (see www.kdnuggets.com/software/visualization.html)
Homework: Association Rule Mining

END OF DAY 10