CSE 412 - Intro to Data Visualization

Exploratory Data Analysis

Jane Hoffswell  University of Washington
Analysis Example:
Motion Pictures Data
Motion Pictures Data

Title String (N)
IMDB Rating Number (Q)
Rotten Tomatoes Rating Number (Q)
MPAA Rating String (O)
Release Date Date (T)
Lesson: Exercise Skepticism

Check **data quality** and your **assumptions**.

Start with **univariate summaries**, then start to consider **relationships among variables**.

Avoid premature fixation!
Analysis Example: Antibiotic Effectiveness
Data Set: Antibiotic Effectiveness

Genus of Bacteria String (N)
Species of Bacteria String (N)
Antibiotic Applied String (N)
Gram-Staining? Pos / Neg (N)
Min. Inhibitory Concent. (g) Number (Q)

Collected prior to 1951.
What questions might we ask?

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Antibiotic</th>
<th>Gram Staining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Penicillin</td>
<td>Streptomycin</td>
</tr>
<tr>
<td>Aerobacter aerogenes</td>
<td>870</td>
<td>1</td>
</tr>
<tr>
<td>Brucella abortus</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Brucella anthracis</td>
<td>0.001</td>
<td>0.01</td>
</tr>
<tr>
<td>Diplococcus pneumoniae</td>
<td>0.005</td>
<td>11</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>100</td>
<td>0.4</td>
</tr>
<tr>
<td>Klebsiella pneumoniae</td>
<td>850</td>
<td>1.2</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>800</td>
<td>5</td>
</tr>
<tr>
<td>Proteus vulgaris</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>850</td>
<td>2</td>
</tr>
<tr>
<td>Salmonella (Eberthella) typhosa</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>Salmonella schottmuelleri</td>
<td>10</td>
<td>0.8</td>
</tr>
<tr>
<td>Staphylococcus albus</td>
<td>0.007</td>
<td>0.1</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Streptococcus fecalis</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Streptococcus hemolyticus</td>
<td>0.001</td>
<td>14</td>
</tr>
<tr>
<td>Streptococcus viridans</td>
<td>0.005</td>
<td>10</td>
</tr>
</tbody>
</table>
How do the drugs compare?

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Penicillin</th>
<th>Antibiotic Streptomycin</th>
<th>Neomycin</th>
<th>Gram stain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerobacter aerogenes</td>
<td>870</td>
<td>1</td>
<td>1.6</td>
<td>–</td>
</tr>
<tr>
<td>Brucella abortus</td>
<td>1</td>
<td>2</td>
<td>0.02</td>
<td>–</td>
</tr>
<tr>
<td>Bacillus anthracis</td>
<td>0.001</td>
<td>0.01</td>
<td>0.007</td>
<td>+</td>
</tr>
<tr>
<td>Diplococcus pneumoniae</td>
<td>0.005</td>
<td>11</td>
<td>10</td>
<td>+</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>100</td>
<td>0.4</td>
<td>0.1</td>
<td>–</td>
</tr>
<tr>
<td>Klebsiella pneumoniae</td>
<td>850</td>
<td>1.2</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>800</td>
<td>5</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Proteus vulgaris</td>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>–</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>850</td>
<td>2</td>
<td>0.4</td>
<td>–</td>
</tr>
<tr>
<td>Salmonella (Eberthella) typhosa</td>
<td>1</td>
<td>0.4</td>
<td>0.008</td>
<td>–</td>
</tr>
<tr>
<td>Salmonella schottmuelleri</td>
<td>10</td>
<td>0.8</td>
<td>0.09</td>
<td>–</td>
</tr>
<tr>
<td>Staphylococcus albus</td>
<td>0.007</td>
<td>0.1</td>
<td>0.001</td>
<td>+</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>0.03</td>
<td>0.03</td>
<td>0.001</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus fecalis</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus hemolyticus</td>
<td>0.001</td>
<td>14</td>
<td>10</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus viridans</td>
<td>0.005</td>
<td>10</td>
<td>40</td>
<td>+</td>
</tr>
</tbody>
</table>

Original graphic by Will Burtin, 1951
How do the drugs compare?

Radius: $1 / \log(\text{MIC})$
Bar Color: Antibiotic
Background Color: Gram Staining
How do the drugs compare?
How do the drugs compare?

**X-axis:** Antibiotic | log(MIC)
**Y-axis:** Gram-Staining | Species
**Color:** Most-Effective?
Penicillin and Neomycin are more efficient broad-spectrum antibiotics than Penicillin.

Streptomycin and Neomycin are more efficient against gram-negative bacteria than Penicillin.

Gram-negative bacteria only

Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.

Gram-positive bacteria only

Penicillin is more efficient than either Streptomycin or Neomycin if the bacteria is known to be gram-positive.
Which antibiotic should one use?
Do the bacteria group by antibiotic resistance?
Do the bacteria group by antibiotic resistance?

Wainer & Lysen
American Scientist, 2009
Do the bacteria group by antibiotic resistance?

Wainer & Lysen
American Scientist, 2009
Do the bacteria group by antibiotic resistance?

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009
Do the bacteria group by antibiotic resistance?

Not a streptococcus! (realized ~30 yrs later)

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009
Do the bacteria group by resistance?
Do different drugs correlate?
Do the bacteria group by resistance?
Do different drugs correlate?

Wainer & Lysen
American Scientist, 2009
Lesson: Iterative Exploration

Exploratory Process
1. Construct graphics to address questions
2. Inspect “answer” and assess new questions
3. Repeat…

Transform data appropriately (e.g., invert, log)

Show data variation, not design variation [Tufte]
Administrivia
A2: Exploratory Data Analysis

Use visualization software to form & answer questions

First steps:
Step 1: Pick domain & data
Step 2: Pose questions
Step 3: Profile the data
Iterate as needed

Create visualizations
Interact with data
Refine your questions

Author a report
Screenshots of most insightful views (8+)
Include titles and captions for each view

Due by 11:59pm Monday, Jan 25
Tableau / Polaris
Polaris [Stolte et al.]
Tableau

Data Display

Data Model

Encodings
Tableau / Polaris Approach

Insight: can simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Can also suggest encodings upon request
Tableau Demo

The dataset:
Every Congressional Candidate from 1996 to 2002
4 Election Cycles
9216 Candidacies
Dataset Schema

Year (Qi)
Candidate Code (N)
Candidate Name (N)
Incumbent / Challenger / Open-Seat (N)
Party Code (N) [1=Dem, 2=Rep, 3=Other]
Party Name (N)
Total Receipts (Qr)
State (N)
District (N)

This is a subset of the larger data set available from the FEC.
Hypotheses?

What might we learn from this data?
Hypotheses?

What might we learn from this data?
Correlation between receipts and winners?
Do receipts increase over time?
Which states spend the most?
Which party spends the most?
Margin of victory vs. amount spent?
Amount spent between competitors?
Tableau Demo
Specifying Table Configurations

Operands are the database fields
Each operand interpreted as a set {...}
Quantitative and Ordinal fields treated differently

Three operators:
concatenation (+)
cross product (x)
nest (/)
<table>
<thead>
<tr>
<th>Region</th>
<th>Technology</th>
<th>Office Supplies</th>
<th>Furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central</td>
<td>$150,000</td>
<td>$120,000</td>
<td>$100,000</td>
</tr>
<tr>
<td>East</td>
<td>$200,000</td>
<td>$180,000</td>
<td>$150,000</td>
</tr>
<tr>
<td>South</td>
<td>$250,000</td>
<td>$220,000</td>
<td>$180,000</td>
</tr>
<tr>
<td>West</td>
<td>$300,000</td>
<td>$280,000</td>
<td>$220,000</td>
</tr>
</tbody>
</table>

SUM(Sales): $2,297,201
Table Algebra

The operators (+, x, /) and operands (O, Q) provide an algebra for tabular visualization.

Algebraic statements are then mapped to:

- **Visualizations** - trellis plot partitions, visual encodings
- **Queries** - selection, projection, group-by aggregation

In Tableau, users make statements via drag-and-drop.

Note that this specifies operands *NOT* operators! Operators are inferred by data type (O, Q).
Table Algebra: Operands

**Ordinal fields**: interpret domain as a set that partitions table into rows and columns.
Quarter = {\{(Qtr1),(Qtr2),(Qtr3),(Qtr4)\}} ->

<table>
<thead>
<tr>
<th>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>95892</td>
<td>101760</td>
<td>105282</td>
<td>98225</td>
</tr>
</tbody>
</table>

**Quantitative fields**: treat domain as single element set and encode spatially as axes.
Profit = {\{(Profit[-410,650])\}} ->

![Profit scale graph]
Concatenation (+) Operator

Ordered union of set interpretations

Quarter + Product Type
= \{(Qtr1),(Qtr2),(Qtr3),(Qtr4)\} + \{(Coffee), (Espresso)\}
= \{(Qtr1),(Qtr2),(Qtr3),(Qtr4),(Coffee),(Espresso)\}

Profit + Sales = \{(Profit[-310,620]),(Sales[0,1000])\}

<table>
<thead>
<tr>
<th>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
<th>Coffee</th>
<th>Espresso</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>59</td>
<td>57</td>
<td>53</td>
<td>151</td>
<td>21</td>
</tr>
</tbody>
</table>
Cross (x) Operator

Cross-product of set interpretations

Quarter x Product Type =
{(Qtr1, Coffee), (Qtr1, Tea), (Qtr2, Coffee), (Qtr2, Tea), (Qtr3, Coffee), (Qtr3, Tea), (Qtr4, Coffee), (Qtr4, Tea)}

Product Type x Profit =

<table>
<thead>
<tr>
<th>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee</td>
<td>Espresso</td>
<td>Coffee</td>
<td>Espresso</td>
</tr>
<tr>
<td>131</td>
<td>19</td>
<td>160</td>
<td>20</td>
</tr>
</tbody>
</table>

![Chart showing profit for Coffee and Espresso products across quarters]
Nest (/) Operator

Cross-product filtered by existing records

Quarter x Month ->
creates twelve entries for each quarter. i.e.,
(Qtr1, December)

Quarter / Month ->
creates three entries per quarter based on
tuples in database (not semantics)
<table>
<thead>
<tr>
<th>State</th>
<th>Coffee</th>
<th>Espresso</th>
<th>Herbal Tea</th>
<th>Tea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorado</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connecticut</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Florida</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illinois</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iowa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Louisiana</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Massachusetts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missouri</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nevada</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Hampshire</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Mexico</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New York</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ohio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oklahoma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oregon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utah</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Washington</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wisconsin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Quantitative-Quantitative
Ordinal-Quantitative
Querying the Database

1. Select records from the database, filtering by user-defined criteria.
2. Partition the records into layers and panes. The same record may appear in multiple partitions.
3. Group, sort, and aggregate the relations within each pane.
4. Render and compose layers.
Quiz Section: Tableau

Tomorrow, Thursday January 14th

Introduction and hands-on experience in Tableau
Come prepared with Tableau installed
See announcement on Ed for instructions

Up Next: Jane's Office Hour (link on Canvas)