CSE 442 - Data Visualization

Exploratory Data Analysis

Jeffrey Heer  University of Washington
What was the first data visualization?
~6200 BC Town Map of Catal Hyük, Konya Plain, Turkey
~950 AD Position of Sun, Moon and Planets
MACVLAE IN SOLE APPARENTES, OBSERVATAE
anno 1611. ad latitudinem in grad. 48. min. 40.

hora 9 et 10.
antemerdianam.

a c, horizon; b, arcus solis diurnus. Solsiens ex parte a, maculas exhibet quas vides, occasis vero c, eadem ratione primi motus nonnil inuerti. Et hanc matutinam vestrinam, mutationem, omnes macula quotidie subeunt. Quod semel exhibeas et mensuras, sufficiat.

0 BC

Sunspots over time, Scheiner 1626
Longitudinal distance between Toledo and Rome, van Langren 1644
The Rate of Water Evaporation, Lambert 1765
The Rate of Water Evaporation, Lambert 1765
The **Golden Age** of Data Visualization

1786 - 1900
Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.

BALANCE in FAVOUR of ENGLAND.

BALANCE AGAINST

The Commercial and Political Atlas, William Playfair 1786
Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.
1786  1826(?) Illiteracy in France, Pierre Charles Dupin
“to affect thro’ the Eyes what we fail to convey to the public through their word-proof ears”
1786 - 1864 British Coal Exports, Charles Minard
Consommations approximatives de la Houille dans la Grande Bretagne de 1850 à 1864.

Les abscisses représentent les années et les ordonnées les quantités annuelles de houille consommée. Les couleurs indiquent les espèces de consommations. Les longueurs d'ordonnées comprises dans une couleur sont les quantités de houille consommées à raison de deux millièmes pour un million de tonnes.

Données admises pour former le Tableau ci-contre.

Consommations — Sources des Renseignements.


District de Londres — id. — page 213

Produits de la Fonte — id — page 212 et pour les années avant 1865 calculé à raison de 3° de houille pour 1° de fonte, en admettant les quantités annuelles de fonte du Coal question page 192.

Production du fer — Mineral statistics — page 212 et pour les années avant 1865 calculé à raison de 3°35 de houille pour 1 tonne de fonte convertie en fer, et admettant 10° de la fonte produite convertie en fer.

Foyers domestiques — En y comprenant les petites manufactures. On l'estimait on 1848 à 19 millions de tonnes, (A) qu'on peut réduire à 18 millions à pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.

Eclairage au Gaz — Consommation estimée généralement de 3° au 5° de la production totale.

Exploitation des Chemins de Fer — En supposant pour consommation totale 10° par Kilomètre parcouru par les trains d'après les renseignements parlementaires.

Navigation à vapeur — Calculé à raison de 5° houille par cheval-vapeur et par heure, le nombre de chevaux étant celui des Steam Vessels pour 1864, et les steamers étant supposés marcher la moitié de l'année.

Avant 1864 j'ai supposé les consommations proportionnelles aux tonnages annuels des steamers du statistical abstract et du Board of trade.

(A) Voir l'excellent article houille de M. Lamé Fleury, Dictionnaire du Commerce, Page 311.
1786

1884 Rail Passengers and Freight from Paris
The Rise of Statistics
Rise of **formal statistical methods** in the physical and social sciences

**Little innovation** in graphical methods

A period of **application and popularization**

Graphical methods enter textbooks, curricula, and **mainstream use**
Four major influences act on data analysis today:
1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and larger bodies of data.
4. The emphasis on quantification in a wider variety of disciplines.
The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.
While some of the influences of statistical theory on data analysis have been helpful, others have not.
Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.
Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the flexibility of the informed human mind.

Accordingly, both approaches and techniques need to be structured so as to facilitate human involvement and intervention.
<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Set D</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.1</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.11</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

**Summary Statistics**

- \( u_X = 9.0 \)  \( \sigma_X = 3.317 \)
- \( u_Y = 7.5 \)  \( \sigma_Y = 2.03 \)

**Linear Regression**

- \( Y = 3 + 0.5 \times X \)
- \( R^2 = 0.67 \)

[Anscombe 1973]
Set A

Set B

Set C

Set D

[Anscombe 1973]
Topics

Exploratory Data Analysis
Data Wrangling
Exploratory Analysis Examples
Tableau / Polaris
Data Wrangling
I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any “analysis” at all.

Anonymous Data Scientist

[Kandel et al. ’12]
In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.
Reported crime in Alabama

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>4525375</td>
<td>4029.3</td>
<td>987</td>
<td>2732.4</td>
<td>309.9</td>
</tr>
<tr>
<td>2005</td>
<td>4548327</td>
<td>3900</td>
<td>955.8</td>
<td>2656</td>
<td>289</td>
</tr>
<tr>
<td>2006</td>
<td>4599030</td>
<td>3937</td>
<td>968.9</td>
<td>2645.1</td>
<td>322.9</td>
</tr>
<tr>
<td>2007</td>
<td>4627851</td>
<td>3974.9</td>
<td>980.2</td>
<td>2687</td>
<td>307.7</td>
</tr>
<tr>
<td>2008</td>
<td>4661900</td>
<td>4081.9</td>
<td>1080.7</td>
<td>2712.6</td>
<td>288.6</td>
</tr>
</tbody>
</table>

Reported crime in Alaska

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>657755</td>
<td>3370.9</td>
<td>573.6</td>
<td>2456.7</td>
<td>340.6</td>
</tr>
<tr>
<td>2005</td>
<td>663253</td>
<td>3615</td>
<td>622.8</td>
<td>2601</td>
<td>391</td>
</tr>
<tr>
<td>2006</td>
<td>670053</td>
<td>3582</td>
<td>615.2</td>
<td>2588.5</td>
<td>378.3</td>
</tr>
<tr>
<td>2007</td>
<td>683478</td>
<td>3373.9</td>
<td>538.9</td>
<td>2480</td>
<td>355.1</td>
</tr>
<tr>
<td>2008</td>
<td>686293</td>
<td>2928.3</td>
<td>470.9</td>
<td>2219.9</td>
<td>237.5</td>
</tr>
</tbody>
</table>

Reported crime in Arizona

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>5739879</td>
<td>5073.3</td>
<td>991</td>
<td>3118.7</td>
<td>963.5</td>
</tr>
<tr>
<td>2005</td>
<td>5953007</td>
<td>4827</td>
<td>946.2</td>
<td>2958</td>
<td>922</td>
</tr>
<tr>
<td>2006</td>
<td>6166318</td>
<td>4741.6</td>
<td>953</td>
<td>2874.1</td>
<td>914.4</td>
</tr>
<tr>
<td>2007</td>
<td>6338755</td>
<td>4502.6</td>
<td>935.4</td>
<td>2780.5</td>
<td>786.7</td>
</tr>
<tr>
<td>2008</td>
<td>6500180</td>
<td>4087.3</td>
<td>894.2</td>
<td>2605.3</td>
<td>587.8</td>
</tr>
</tbody>
</table>

Reported crime in Arkansas

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>2750000</td>
<td>4033.1</td>
<td>1096.4</td>
<td>2699.7</td>
<td>237</td>
</tr>
<tr>
<td>2005</td>
<td>2775708</td>
<td>4068</td>
<td>1085.1</td>
<td>2720</td>
<td>262</td>
</tr>
<tr>
<td>2006</td>
<td>2810872</td>
<td>4021.6</td>
<td>1154.4</td>
<td>2596.7</td>
<td>270.4</td>
</tr>
<tr>
<td>2007</td>
<td>2834797</td>
<td>3945.5</td>
<td>1124.4</td>
<td>2574.6</td>
<td>246.5</td>
</tr>
<tr>
<td>2008</td>
<td>2855390</td>
<td>3843.7</td>
<td>1182.7</td>
<td>2433.4</td>
<td>227.6</td>
</tr>
</tbody>
</table>

Reported crime in California

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>35842038</td>
<td>3423.9</td>
<td>686.1</td>
<td>2033.1</td>
<td>704.8</td>
</tr>
<tr>
<td>2005</td>
<td>36154147</td>
<td>3321</td>
<td>692.9</td>
<td>1915</td>
<td>712</td>
</tr>
<tr>
<td>2006</td>
<td>36457549</td>
<td>3175.2</td>
<td>676.9</td>
<td>1831.5</td>
<td>666.8</td>
</tr>
<tr>
<td>2007</td>
<td>36553215</td>
<td>3032.6</td>
<td>648.4</td>
<td>1784.1</td>
<td>600.2</td>
</tr>
<tr>
<td>2008</td>
<td>36756666</td>
<td>2940.3</td>
<td>646.8</td>
<td>1769.8</td>
<td>523.8</td>
</tr>
</tbody>
</table>

Reported crime in Colorado

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>4601821</td>
<td>3918.5</td>
<td>717.3</td>
<td>2679.5</td>
<td>521.6</td>
</tr>
</tbody>
</table>
DataWrangler

Wrangler: Interactive Visual Specification of Data Transformation Scripts
Sean Kandel et al. CHI’11
Data Wrangling

One often needs to manipulate data prior to analysis. Tasks include reformatting, cleaning, quality assessment, and integration.

Approaches include:
Manual manipulation in spreadsheets
Code: arquero (JS), dplyr (R), pandas (Python)
Trifacta Wrangler  http://www.trifacta.com/products/wrangler/
Open Refine   http://openrefine.org/
Tidy Data  [Wickham 2014]

How do rows, columns, and tables match up with observations, variables, and types? In “tidy” data:
1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

The advantage is that this provides a flexible starting point for analysis, transformation, and visualization.

Our pivoted table variant was not “tidy”!

(This is a variant of normalized forms in DB theory)
Data Quality

“The first sign that a visualization is good is that it shows you a problem in your data...

...every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

Martin Wattenberg
Violent Infants!

Marauding Centenarians!

Query Result: 4792 out of 4792 (100%)
Graph Viewer

Roll-up by:
- All

Visualization:
- Matrix

Sort by:
- None

Edge centrality filters:
Visualize Friends by School?

Berkeley
Cornell
Harvard
Harvard University
Stanford
Stanford University
UC Berkeley
UC Davis
University of California at Berkeley
University of California, Berkeley
University of California, Davis
Data Quality Hurdles

Missing Data  no measurements, redacted, …?
Erroneous Values  misspelling, outliers, …?
Type Conversion  e.g., zip code to lat-lon
Entity Resolution  diff. values for the same thing?
Data Integration  effort/errors when combining data

LESSON: Anticipate problems with your data. Many research problems around these issues!
Administrivia
A2: Deceptive Visualization

Design **two** static visualizations for a dataset:
1. An *earnest* visualization that faithfully conveys the data
2. A *deceptive* visualization that tries to mislead viewers

Your two visualizations may address different questions. Try to design a deceptive visualization that appears to be earnest: *can you trick your classmates and course staff?*

You are free to choose your own dataset, but we have also provided some preselected datasets for you.

Submit two images and a brief write-up on Canvas.

Due by **Wed 1/26 11:59pm.**
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!**
Tableau / Polaris
Polaris [Stolte et al.]
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
The dataset:
Federal Elections Commission Receipts
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
Dimensionality Reduction
Dimensionality Reduction (DR)

Project nD data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

Different DR methods make different trade-offs: for example to **preserve global structure** (e.g., PCA) or **emphasize local structure** (e.g., nearest-neighbor approaches, including t-SNE and UMAP).
Principal Components Analysis

1. Mean-center the data.
2. Find basis vectors that maximize the data variance.
3. Plot the data using the top vectors.
Principal Components Analysis

Linear transform: scale and rotate original space.

Lines (vectors) project to lines.

Preserves global distances.
PCA of Genomes [Demiralp et al. ‘13]
Reduction Techniques

**LINEAR - PRESERVE GLOBAL STRUCTURE**

Principal Components Analysis (PCA)
Linear transformation of basis vectors, ordered by amount of data variance they explain.

**NON-LINEAR - PRESERVE LOCAL TOPOLOGY**

t-Dist. Stochastic Neighbor Embedding (t-SNE)
Probabilistically model distance, optimize positions.

Uniform Manifold Approx. & Projection (UMAP)
Identify local manifolds, then stitch them together.
Non-Linear Techniques

Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods:

- **t-SNE** - probabilistic interpretation of distance
- **UMAP** - tries to balance local/global trade-off
Visualizing t-SNE [Wattenberg et al. ’16]

Results can be highly sensitive to the algorithm parameters!
Are you seeing real structures, or algorithmic hallucinations?
How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.
t-SNE projection of latent space of language translation model.

The stratosphere extends from about 10km to about 50km in altitude.

- English: The stratosphere extends from about 10km to about 50km in altitude.
- Korean: 성층권은 고도 약 10km부터 약 50km까지 확장됩니다.
- Japanese: 成層圏は、高度10kmから50kmの範囲にあります.
Time Curves [Bach et al. ‘16]

Timeline:
1 2 3 4 5 6 7
Time difference
Circles are data cases with a time stamp. Similar colors indicate similar data cases.

Folding:
1 2 3 4 5 6 7
(a) Folding time

Time curve:
1 2 3 4 5 6 7
Similarity
The temporal ordering of data cases is preserved. Spatial proximity now indicates similarity.

Wikipedia “Chocolate” Article

U.S. Precipitation over 1 Year