Scalable Data Analysis (CIS 602-02)

Statistics

Dr. David Koop
### Visualization

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
</tr>
</tbody>
</table>

[F. J. Anscombe]
## Visualization

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean of x: 9
Variance of x: 11
Mean of y: 7.50
Variance of y: 4.122
Correlation: 0.816

[F. J. Anscombe]
Visualization

[F. J. Anscombe]
### MTA Fare Data Exploration

<table>
<thead>
<tr>
<th>REMOTE</th>
<th>STATION</th>
<th>FF</th>
<th>SEN/DIS</th>
<th>7-D AFAS UNL</th>
<th>D AFAS/RMF</th>
<th>JOINT RR TKT</th>
<th>7-D UNL</th>
<th>30-D UNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>R011</td>
<td>42ND STREET &amp; 8TH AVENUE</td>
<td>00228985</td>
<td>00008471</td>
<td>00000441</td>
<td>00001455</td>
<td>00000134</td>
<td>00033341</td>
<td>00071255</td>
</tr>
<tr>
<td>R170</td>
<td>14TH STREET-UNION SQUARE</td>
<td>00224603</td>
<td>00011051</td>
<td>00000827</td>
<td>00003026</td>
<td>00000660</td>
<td>00089367</td>
<td>00199841</td>
</tr>
<tr>
<td>R046</td>
<td>42ND STREET &amp; GRAND CENTRAL</td>
<td>00207758</td>
<td>00007908</td>
<td>00000323</td>
<td>00001183</td>
<td>00003001</td>
<td>00040759</td>
<td>00096613</td>
</tr>
<tr>
<td>R012</td>
<td>34TH STREET &amp; 8TH AVENUE</td>
<td>00188311</td>
<td>00006490</td>
<td>00000498</td>
<td>00001279</td>
<td>00003622</td>
<td>00035527</td>
<td>00067483</td>
</tr>
<tr>
<td>R293</td>
<td>34TH STREET - PENN STATION</td>
<td>00168768</td>
<td>00006155</td>
<td>00000523</td>
<td>00001065</td>
<td>00005031</td>
<td>00030645</td>
<td>00054376</td>
</tr>
<tr>
<td>R033</td>
<td>42ND STREET/TIMES SQUARE</td>
<td>00159382</td>
<td>00005945</td>
<td>00000378</td>
<td>00001205</td>
<td>00000690</td>
<td>00058931</td>
<td>00078644</td>
</tr>
<tr>
<td>R022</td>
<td>34TH STREET &amp; 6TH AVENUE</td>
<td>00156008</td>
<td>00006276</td>
<td>00000487</td>
<td>00001543</td>
<td>00000712</td>
<td>00058910</td>
<td>00110466</td>
</tr>
<tr>
<td>R084</td>
<td>59TH STREET/COLUMBUS CIRCLE</td>
<td>00155262</td>
<td>00009484</td>
<td>00000589</td>
<td>00002071</td>
<td>00000542</td>
<td>00053397</td>
<td>00113966</td>
</tr>
<tr>
<td>R020</td>
<td>47-50 STREETS/ROCKEFELLER</td>
<td>00143500</td>
<td>00006402</td>
<td>00000384</td>
<td>00001159</td>
<td>00000723</td>
<td>00037978</td>
<td>00090745</td>
</tr>
<tr>
<td>R179</td>
<td>86TH STREET-LEXINGTON AVE</td>
<td>00142169</td>
<td>00010367</td>
<td>00000470</td>
<td>00001839</td>
<td>00000271</td>
<td>00050328</td>
<td>00125250</td>
</tr>
<tr>
<td>R023</td>
<td>34TH STREET &amp; 6TH AVENUE</td>
<td>00134052</td>
<td>00005005</td>
<td>00000348</td>
<td>00001112</td>
<td>00000649</td>
<td>00031531</td>
<td>00075040</td>
</tr>
<tr>
<td>R029</td>
<td>PARK PLACE</td>
<td>00121614</td>
<td>00004311</td>
<td>00000287</td>
<td>00000931</td>
<td>00000792</td>
<td>00025404</td>
<td>00065362</td>
</tr>
<tr>
<td>R047</td>
<td>42ND STREET &amp; GRAND CENTRAL</td>
<td>00100742</td>
<td>00004273</td>
<td>00000185</td>
<td>00000704</td>
<td>00001241</td>
<td>00022808</td>
<td>00068216</td>
</tr>
<tr>
<td>R031</td>
<td>34TH STREET &amp; 7TH AVENUE</td>
<td>00095076</td>
<td>00003990</td>
<td>00000232</td>
<td>00001459</td>
<td>00002484</td>
<td>00038671</td>
<td></td>
</tr>
<tr>
<td>R017</td>
<td>LEXINGTON AVENUE</td>
<td>00094655</td>
<td>00004688</td>
<td>00000190</td>
<td>00000833</td>
<td>00000754</td>
<td>00020018</td>
<td>00055066</td>
</tr>
<tr>
<td>R175</td>
<td>8TH AVENUE-14TH STREET</td>
<td>00094313</td>
<td>00003907</td>
<td>00000286</td>
<td>00001144</td>
<td>00000256</td>
<td>00038722</td>
<td>00074661</td>
</tr>
<tr>
<td>R057</td>
<td>BARCLAYS CENTER</td>
<td>00093804</td>
<td>00004204</td>
<td>00000454</td>
<td>00001386</td>
<td>00001491</td>
<td>00039113</td>
<td>00068119</td>
</tr>
<tr>
<td>R138</td>
<td>WEST 4TH ST-WASHINGTON SQ</td>
<td>00093562</td>
<td>00004677</td>
<td>00000251</td>
<td>00000965</td>
<td>00000127</td>
<td>0031628</td>
<td>00074458</td>
</tr>
</tbody>
</table>

D. Koop, CIS 602-02, Fall 2015
MTA Fare Data Exploration
MTA Fare Data Exploration
MTA Fare Data Exploration

East 161st Street and River Avenue

Full Fares Purchased

Date

08-02 08-09 08-16 08-23 08-30 09-06 09-13 09-20 09-27 10-04 10-11 10-18 10-25 11-01
MTA Fare Data Exploration

East 161st Street and River Avenue

New York Yankees

2013 Regular Season Schedule

All games are Eastern time.
Visual Pop-out

Ranking Channels by Effectiveness

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

Identity Channels: Categorical Attributes
- Spatial region
- Color hue
- Motion
- Shape

[Munzner (ill. Maguire), 2014]
Voyager

Data Voyager

DATA

SHOWING DATA VARIATIONS FOR

[Variable Selection Panel]

[Visualization Gallery]

Abstract

— General visualization tools typically require manual specification of views: analysts must select data variables and then choose which transformations and visual encodings to apply. These decisions often involve both domain and visualization design expertise, and may impose a tedious specification process that impedes exploration. In this paper, we seek to complement manual chart construction with interactive navigation of a gallery of automatically-generated visualizations. We contribute Voyager, a mixed-initiative system that supports faceted browsing of recommended charts chosen according to statistical and perceptual measures. We describe Voyager's architecture, motivating design principles, and methods for generating and interacting with visualization recommendations. In a study comparing Voyager to a manual visualization specification tool, we find that Voyager facilitates exploration of previously unseen data and leads to increased data variable coverage. We then distill design implications for visualization tools, in particular the need to balance rapid exploration and targeted question-answering.

Index Terms — User interfaces, information visualization, exploratory analysis, visualization recommendation, mixed-initiative systems

Introduction

Exploratory visual analysis is highly iterative, involving both open-ended exploration and targeted question answering [16, 37]. Yet making visual encoding decisions while exploring unfamiliar data is non-trivial. Analysts may lack exposure to the shape and structure of their data, or begin with vague analysis goals. While analysts should typically examine each variable before investigating relationships between them [28], in practice they may fail to do so due to premature fixation on specific questions or the tedium of manual specification.

The primary interaction model of many popular visualization tools (e.g., [35, 44, 45]) is manual view specification. First, an analyst must select variables to examine. The analyst then may apply data transformations, for example binning or aggregation to summarize the data. Finally, she must design visual encodings for each resulting variable set. These actions may be expressed via code in a high-level language [44] or a graphical interface [35]. While existing tools are well suited to depth-first exploration strategies, the design of tools for breadth-oriented exploration remains an open problem. Here we focus on tools to assist breadth-oriented exploration, with the specific goal of promoting increased coverage of a data set.

To encourage broad exploration, visualization tools might automatically generate a diverse set of visualizations and have the user select.

[Wongsuphasawat et al., 2015]
We now present a selected subset of the study results, focusing on verbalized their thought process in a think-aloud protocol. We did not had used Python/matplotlib and 9 had used R/ggplot.

Data variable coverage, bookmarking activity, user survey responses, events. Finally, we collected data from the exit survey and interview, recorded interaction logs, capturing all input device and application observed each analysis session and took notes. Audio was recorded to pixels. After completing two analysis sessions, participants completed end the session early if they were satisfied with their exploration.

Subjects 30 minutes to explore the dataset. Subjects were allowed to might bias them toward premature fixation on those questions. We gave explore the data, and specifically to "get a comprehensive sense of what briefly introduced subjects to the test dataset. We asked participants to trial, using a dataset distinct from those used for actual analysis. We then participants with a $15 gift certificate.

Study datasets before, nor had they used Voyager or PoleStar (though experience. All subjects had used visualization tools including Tableau, interest, are of similar complexity, and concern phenomena accessible.

Fig. 10. PoleStar, a visualization specification tool inspired by Tableau.

No subject had analyzed the which the fixed effect in question has been removed.

Of the 179 total visualizations bookmarked in Voyager, 124 (69%) views without manipulating them. For PoleStar, in both cases we include a data variable automatically added by the recommendation engine which the fixed effect in question has been removed.

We recruited 16 participants (6 female, 10 male), allParticipants had used Excel. Among other tools, 9 had used Tableau, 13 with both tools. When asked to rate their confidence in the comprehensive sense of what briefly introduced subjects to the test dataset. We asked participants to trial, using a dataset distinct from those used for actual analysis. We then participants with a $15 gift certificate.

![Image](image_url)
Voyager

• Show breadth-first data variation
  - Contrast with visualization where a user deals with fixed data and wants to examine it specifically
  - We often don't have a specific question: "Get a comprehensive sense of what the dataset contains and use the bookmark features to collect interesting patterns, trends, or other insights worth sharing with colleagues"

• Univariate plots and suggested paired plots (potentially with transformations)

• Ranking of Recommendations

• Hybrid of PoleStar and Voyager?

• Scalability?
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Student (Pos.)</th>
<th>Student (Neg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/29</td>
<td>Visualization</td>
<td>Chaitanya Chandurkar</td>
<td>Ramya Reddy Mara</td>
</tr>
<tr>
<td>10/1</td>
<td>Statistics</td>
<td>Pragnya Srinivasan</td>
<td>Shakti Bhattarai</td>
</tr>
<tr>
<td>10/6</td>
<td>Machine Learning</td>
<td>Sumukhi Kappa</td>
<td>Vishnu Vardhan Kumar Pallati</td>
</tr>
<tr>
<td>10/8</td>
<td>Clustering</td>
<td>Akeim Findlay</td>
<td>Richard de Groof</td>
</tr>
<tr>
<td>10/15</td>
<td>Databases</td>
<td>Gursharanpreet Singh</td>
<td>Kalesha Nagineni</td>
</tr>
<tr>
<td>10/20</td>
<td>Databases</td>
<td>Priya Vishnudas Shanbhag</td>
<td>Shree Lekha Kakkerla</td>
</tr>
<tr>
<td>10/22</td>
<td>Data Cubes</td>
<td>Nilesh Bhadane</td>
<td>Tanmay Thakar</td>
</tr>
<tr>
<td>11/3</td>
<td>Natural Language Processing</td>
<td>Jayeshkumar Vijayaraghavalu</td>
<td>Sanjana Bhardwaj</td>
</tr>
<tr>
<td>11/5</td>
<td>Cloud Computing</td>
<td>Arsalan Aqeel Hafiz</td>
<td>Zennia Sandhu</td>
</tr>
<tr>
<td>11/10</td>
<td>Map Reduce</td>
<td>Arpit Parikh</td>
<td>Rutvi Dave</td>
</tr>
<tr>
<td>11/12</td>
<td>General Cluster Computing</td>
<td>Harshada Gorhe</td>
<td>Mehmet Duman</td>
</tr>
<tr>
<td>11/17</td>
<td>Streaming Data</td>
<td>Anurag Dhirendra Singh</td>
<td>Rishu Vaid</td>
</tr>
<tr>
<td>11/19</td>
<td>Out of Core Algorithms</td>
<td>Rameshta Reddy Kotha</td>
<td>Hari Bharti</td>
</tr>
<tr>
<td>11/24</td>
<td>Graph Algorithms</td>
<td>Dhvani Patel</td>
<td></td>
</tr>
<tr>
<td>12/1</td>
<td>Reproducibility</td>
<td>Xiaoochun Chen</td>
<td></td>
</tr>
</tbody>
</table>

If you need to switch, coordinate with another student and email me to approve.
Projects

• Options:
  - Data analysis on some existing data: think about the questions you want to try to answer
  - Improve some technique for data analysis

• Data Sources:
  - Search the web for topics you're interested in
  - https://github.com/caesar0301/awesome-public-datasets
  - Local data

• If you are doing a research project in a particular area, let's try to work something out so that the course project relates
Statistics

• "Lies, D**ned lies, and statistics", Benjamin Disraeli (& Mark Twain)
• Example of Problematic Statistics:
  - Mean, Median, and Mode
  - Mean food truck rankings
  - Flipping a coin four times
Population and Sample

- Population: All humpback whales
- Sample: 50 Humpback whales found within 500 miles of Cape Cod between June 15 and September 15, 2013
Descriptive and Inferential

- Descriptive: summarizing and describing data
- Inferential: using samples to make an inference about the population
- Die Rolls:
  - 1234623524524111342354613
  - Can do descriptive statistics about the rolls (frequencies)
  - Is it a biased die? Use inferential statistics
Descriptive Statistics

• Mean, median, mode, range

• Standard deviation (variance) measures how far-flung data is (difference from mean)

\[ s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}. \]

• Interquartile Range: Inner 25% to 75% of values (by positions)
Boxplot

Experiment No.

Speed of light (km/s minus 299,000)

true speed

Experiment No.: 1, 2, 3, 4, 5
Random Variable

• A variable whose value is subject to chance
• Examples: coin flips to be performed, outcomes of an experiment
• Random variables have expected values and variances which are not the same as sample means of sample standard deviations
Law of Large Numbers

- With enough trials, the sample mean should be close to the expected value.
- Means that there is stability with enough observations.
Probability Distributions

- Distribution of a random variable is the possible outcomes and their probabilities
- Types: Discrete, continuous, and mixed

![Diagram of a density curve with probabilities as areas under the curve.](image-url)

Example 6.2: Probabilities of a continuous random variable

\[ P(x_1 < X < x_2) \]

[Source: J. Isotalo., Basics of Statistics]
A random variable $X$ following a normal distribution with a mean of $\mu$ and a standard deviation of $\sigma$ is denoted by $X \sim N(\mu, \sigma)$.

There are other symmetric bell-shaped density curves that are not normal. The normal density curves are specified by a particular equation. The height of the density curve at any point $x$ is given by the density function $f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2}$.

We will not make direct use of this fact, although it is the basis of mathematical work with normal distribution. Note that the density function is completely determined by $\mu$ and $\sigma$.

Example 6.4.

Normal Distribution

Values of $X$

$\mu - 3\sigma$ $\mu - 2\sigma$ $\mu - \sigma$ $\mu$ $\mu + \sigma$ $\mu + 2\sigma$ $\mu + 3\sigma$

Figure 11: Normal distribution.

Definition 6.8 (Standard normal distribution).

A continuous and random variable $Z$ is said to have a standard normal distribution if $Z$ is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$, i.e., $Z \sim N(0, 1)$.

Normal Distribution

[J. Isotalo., Basics of Statistics]
Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event. The probability mass function of the Poisson distribution is given by:

\[ P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \]

where \( \lambda \) is the average rate of occurrence of the event, and \( k \) is the number of events occurring in the given interval.

The graph shows the Poisson distribution with a mean value of \( \lambda \).

The x-axis represents the number of events (\( k \)), and the y-axis represents the probability of observing \( k \) events (\( P(X = k) \)).
Introduction to Bayesian Methods

Cam Davidson-Pilon

Presented by: Pragnya Srinivasan and Shakti Bhattarai
Bayesian Methods

- Inferential statistics
- Difference between Frequentist and Bayesian perspectives: both are useful
- Law of large numbers
- Prior and posterior
- Distributions
- Statistical models
Next Week

• Machine Learning
  - General approaches
  - Clustering
• Reading Responses
• Assignment 2
• Project Proposal