DSC 201: Data Analysis & Visualization

Exploratory Data Analysis

Dr. David Koop
What is Exploratory Data Analysis?

• "Detective work" to summarize and explore datasets

• Includes:
  - Data acquisition and input
  - Data cleaning and wrangling ("tidying")
  - Data transformation and summarization
  - Data visualization
Exploratory Data Analysis

- John W. Tukey
  - Born in New Bedford
  - 1977: Highly influential book

- Emphasis on value of visualization in discovering trends, relationships

- From a review of the book:
  “Tukey favors analysis of data with little more than pencil and paper. Specifically, there is no need for a calculator, a computer, or a lettering guide to do the analyses he proposes” [R.M. Church, 1979]
“Reflective thought requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. The process takes time.”

– Donald A. Norman
Comparison with Older Statistical Methods

• Older method:
  - Take data and a probable model and learn model parameters
  - Good models are useful, help understand phenomena
  - What happens when we pick the wrong model, or don't know which one to pick?

• EDA:
  - Postpone the model assumptions, let the data speak first
  - Usually involves graphical techniques
  - Tukey used pen-and-paper approaches
  - Today, we can do much of this via computer, but insight may still take time
Types of EDA

- Univariate vs. multivariate
- Non-graphical vs. graphical
Univariate Non-graphical EDA

- Categorical Data:
  - Frequency counts, proportions
  - Groupings

- Quantitative Data:
  - Distribution
  - Summary statistics: mean, median, mode, variance, standard deviation, quantiles
Univariate Graphical EDA

• Histograms
  - Aggregation of data
  - Choose number of bins
  - Bin width makes a difference!

• Stem-and-leaf plots:
  - 1|710340
  - 2|06
  - 3|030223459
  - 4|028907
  - 5|00798273487
  - 6|128
  - 7|7897
  - 8|345
  - 9|1

Figure 4.3: Histograms of EDA2.dat with different bin widths.
Univariate Graphical EDA

- **Boxplots**
  - Show distribution
  - Multiple summary statistics can be read from the chart
  - Also provides a general shape of the data
  - Best for **unimodal** data
Multivariate non-graphical EDA

- Crosstabs and Pivot Tables
  - What is in the data? Count

- Correlation and covariance
  - How related are different columns?
  - How do they change together?
Multivariate Graphical EDA

- Scatterplots: look for correlation
  - Usually put outcome on y-axis
  - Can encode other variables
- Side-by-side boxplots
- Parallel coordinates
- Grouped bar charts
Assignment 5

- [http://www.cis.umassd.edu/~dkoop/dsc201/assignment5.html](http://www.cis.umassd.edu/~dkoop/dsc201/assignment5.html)
- Aggregation, resampling, and visualization of time series data
- Part 3: Multi-level index and groupbys
Final Exam

- [http://www.cis.umassd.edu/~dkoop/dsc201/final.html](http://www.cis.umassd.edu/~dkoop/dsc201/final.html)
- Similar format to midterm: multiple choice and short answer
- Should be familiar with **all** material
- Focuses more on the second half of the course
- Sample questions:
  - Given the following data frame, which columns are categorical, ordinal, and quantitative?
  - Given the following visualization, what are the marks and channels? What is the encoding?
  - Given the following data frame, suggest how a visualization can encode all the given data
  - What colormap would be most appropriate for this data frame?
Final Exam

• Sample Questions (continued)
  - If we do data analysis without considering the source of the data, what problems might occur?
  - Describe the split, apply, and combine steps involved in determining the average number of people who enter each NYC subway station each month.
  - Given the NYC subway data, what methods would you suggest to graphically analyze the relationship between the number of entries and the number of exits?
  - Given the following pandas code, add comments that describe what each block is doing.
  - Why are choropleth maps a poor choice for visualizing election data?
Final Exam

• Sample Questions (continued):
  - How would you set up a pivot table to analyze the tipping data by sex, day of the week, and time of day? Why?
  - How do we know that length is a better channel than area? What does this mean for the effectiveness of a (1D) bubble chart versus a bar chart?
Review

• Python
  - Notebooks
  - Types
  - Variables
  - Functions
  - Lists & Tuples (also mutable vs. immutable)
  - Dictionaries & Sets
  - Classes
Review

- NumPy
  - Arrays
  - Why?
- Pandas
  - Series
  - Index
  - Data Frames
- Data
  - Formats: CSV, TSV, JSON, XML
  - Reading data from files
Review

- Data Wrangling:
  - Cleaning
  - Transforming
  - Reshaping
  - Merging
  - Tidy Data
Review

- Visualization:
  - Why?
  - Tools: matplotlib, Tableau, and more
  - Types of data: categorical, ordinal, quantitative
  - Encoding data: marks & channels
  - Expressiveness and Effectiveness
  - Visual encodings: bar charts, scatterplots, line charts, etc.
  - Color and colormaps
  - Interaction and multiple views
  - Maps
“Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.”

— T. Munzner
### Categorial, Ordinal, and Quantitative

#### Overview
- **Quantitative**: Data that can be measured or counted.
- **Ordinal**: Data that can be ranked or ordered, but the distances between the ranks may not be equal.
- **Categorical**: Data that falls into distinct categories or groups.

#### Table Example

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td><strong>B</strong></td>
<td><strong>C</strong></td>
<td><strong>S</strong></td>
<td><strong>T</strong></td>
<td><strong>U</strong></td>
</tr>
<tr>
<td><strong>Order ID</strong></td>
<td><strong>Order Date</strong></td>
<td><strong>Order Priority</strong></td>
<td><strong>Product Container</strong></td>
<td><strong>Product Base Margin</strong></td>
<td><strong>Ship Date</strong></td>
</tr>
<tr>
<td>3</td>
<td>10/14/06</td>
<td>5-Low</td>
<td>Large Box</td>
<td>0.8</td>
<td>10/21/06</td>
</tr>
<tr>
<td>6</td>
<td>2/21/08</td>
<td>4-Not Specified</td>
<td>Small Pack</td>
<td>0.55</td>
<td>2/22/08</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Small Pack</td>
<td>0.79</td>
<td>7/17/07</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Jumbo Box</td>
<td>0.72</td>
<td>7/17/07</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.6</td>
<td>7/18/07</td>
</tr>
<tr>
<td>35</td>
<td>10/23/07</td>
<td>4-Not Specified</td>
<td>Wrap Bag</td>
<td>0.52</td>
<td>10/24/07</td>
</tr>
<tr>
<td>35</td>
<td>10/23/07</td>
<td>4-Not Specified</td>
<td>Small Box</td>
<td>0.58</td>
<td>10/25/07</td>
</tr>
<tr>
<td>36</td>
<td>11/3/07</td>
<td>1-Urgent</td>
<td>Small Box</td>
<td>0.55</td>
<td>11/3/07</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.65</td>
<td>7/18/07</td>
</tr>
<tr>
<td>35</td>
<td>10/23/07</td>
<td>4-Not Specified</td>
<td>Wrap Bag</td>
<td>0.52</td>
<td>10/24/07</td>
</tr>
<tr>
<td>35</td>
<td>10/23/07</td>
<td>4-Not Specified</td>
<td>Small Box</td>
<td>0.58</td>
<td>10/25/07</td>
</tr>
<tr>
<td>36</td>
<td>11/3/07</td>
<td>1-Urgent</td>
<td>Small Box</td>
<td>0.55</td>
<td>11/3/07</td>
</tr>
<tr>
<td>36</td>
<td>11/3/07</td>
<td>1-Urgent</td>
<td>Small Pack</td>
<td>0.49</td>
<td>3/19/07</td>
</tr>
<tr>
<td>36</td>
<td>11/3/07</td>
<td>1-Urgent</td>
<td>Small Pack</td>
<td>0.49</td>
<td>3/19/07</td>
</tr>
<tr>
<td>66</td>
<td>12/18/06</td>
<td>5-Low</td>
<td>Wrap Bag</td>
<td>0.56</td>
<td>1/20/05</td>
</tr>
<tr>
<td>69</td>
<td>6/4/05</td>
<td>4-Not Specified</td>
<td>Small Pack</td>
<td>0.44</td>
<td>6/6/05</td>
</tr>
<tr>
<td>69</td>
<td>6/4/05</td>
<td>4-Not Specified</td>
<td>Small Pack</td>
<td>0.44</td>
<td>6/6/05</td>
</tr>
<tr>
<td>70</td>
<td>12/18/06</td>
<td>5-Low</td>
<td>Small Box</td>
<td>0.59</td>
<td>12/23/06</td>
</tr>
<tr>
<td>70</td>
<td>12/18/06</td>
<td>5-Low</td>
<td>Small Box</td>
<td>0.59</td>
<td>12/23/06</td>
</tr>
<tr>
<td>96</td>
<td>4/17/05</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.55</td>
<td>4/19/05</td>
</tr>
<tr>
<td>97</td>
<td>1/29/06</td>
<td>3-Medium</td>
<td>Medium Box</td>
<td>0.38</td>
<td>1/30/06</td>
</tr>
<tr>
<td>129</td>
<td>11/19/08</td>
<td>5-Low</td>
<td>Medium Box</td>
<td>0.37</td>
<td>11/28/08</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.37</td>
<td>5/9/08</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.38</td>
<td>5/10/08</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.38</td>
<td>5/10/08</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.38</td>
<td>5/10/08</td>
</tr>
<tr>
<td>132</td>
<td>6/11/06</td>
<td>3-Medium</td>
<td>Medium Box</td>
<td>0.6</td>
<td>6/12/06</td>
</tr>
<tr>
<td>132</td>
<td>6/11/06</td>
<td>3-Medium</td>
<td>Medium Box</td>
<td>0.6</td>
<td>6/12/06</td>
</tr>
<tr>
<td>132</td>
<td>6/11/06</td>
<td>3-Medium</td>
<td>Jumbo Box</td>
<td>0.69</td>
<td>6/14/06</td>
</tr>
<tr>
<td>134</td>
<td>5/1/08</td>
<td>4-Not Specified</td>
<td>Large Box</td>
<td>0.82</td>
<td>5/3/08</td>
</tr>
<tr>
<td>135</td>
<td>10/21/07</td>
<td>4-Not Specified</td>
<td>Small Pack</td>
<td>0.64</td>
<td>10/23/07</td>
</tr>
<tr>
<td>166</td>
<td>9/12/07</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.55</td>
<td>9/14/07</td>
</tr>
<tr>
<td>193</td>
<td>8/8/06</td>
<td>1-Urgent</td>
<td>Medium Box</td>
<td>0.57</td>
<td>8/10/06</td>
</tr>
<tr>
<td>194</td>
<td>4/5/08</td>
<td>3-Medium</td>
<td>Wrap Bag</td>
<td>0.42</td>
<td>4/7/08</td>
</tr>
</tbody>
</table>
Visual Encoding

• How do we encode data visually?
  - **Marks** are the basic graphical elements in a visualization
  - **Channels** are ways to control the appearance of the marks

• Marks classified by dimensionality:
  - Points
  - Lines
  - Areas
  - Also can have surfaces, volumes
  - Think of marks as a mathematical definition, or if familiar with tools like Adobe Illustrator or Inkscape, the path & point definitions
Visual 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

[Visual Channels by Effectiveness](Munzner (ill. Maguire), 2014)
Test perception of **area** differences

![Diagram of area perception with two shapes A and B](image_url)
Types of Visualizations: Scatterplots

- Data: two **quantitative** values
- Task: find trends, clusters, outliers
- How: marks at spatial position in horizontal and vertical directions
- Correlation: dependence between two attributes
  - Positive and negative correlation
  - Indicated by lines
- Coordinate system (axes) and labels are important!
Color and Colormaps

- A colormap specifies a mapping between colors and data values
- Channels: hue, saturation, and luminance
- Luminance perception is non-linear...
- Issues with rainbow colormaps
- Segmented vs. continuous
- Univariate and bivariate

- **Binary**
  - y
  - n

- **Diverging**
  - -1
  - 0
  - +1

- **Categorical**
  - T
  - F
  - A

- **Sequential**
  - 3
  - 2
  - 1
Choropleth Map

[M. Ericson, New York Times]
Review

• Data Aggregation
  - Why?
  - Split-Apply-Combine
  - Pivot Tables & Crosstabs
  - Visualizations and Aggregation

• Time Series
  - Representing date and time
  - Frequencies and ranges
  - Time zones
  - Resampling: downsampling and up sampling
  - Window functions
Aggregation of time series data, a special use case of `groupby`, is referred to as **resampling** in this book and will receive separate treatment in Chapter 10.

**GroupBy Mechanics**

Hadley Wickham, an author of many popular packages for the R programming language, coined the term **split-apply-combine** for talking about group operations, and I think that’s a good description of the process. In the first stage of the process, data contained in a pandas object, whether a Series, DataFrame, or otherwise, is **split** into groups based on one or more **keys** that you provide. The splitting is performed on a particular axis of an object. For example, a DataFrame can be grouped on its rows (`axis=0`) or its columns (`axis=1`). Once this is done, a function is **applied** to each group, producing a new value. Finally, the results of all those function applications are **combined** into a result object. The form of the resulting object will usually depend on what’s being done to the data. See Figure 9-1 for a mockup of a simple group aggregation.

**Figure 9-1. Illustration of a group aggregation**

Each grouping key can take many forms, and the keys do not have to be all of the same type:

- A list or array of values that is the same length as the axis being grouped
- A value indicating a column name in a DataFrame

[W. McKinney, Python for Data Analysis]
Pivot Tables and Crosstabs

- `tips.pivot_table(index=['sex', 'smoker'])`

```
<table>
<thead>
<tr>
<th>sex</th>
<th>smoker</th>
<th>size</th>
<th>tip</th>
<th>tip_pct</th>
<th>total_bill</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Female</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>2.592593</td>
<td>2.773519</td>
<td>0.156921</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>2.242424</td>
<td>2.931515</td>
<td>0.182150</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>2.711340</td>
<td>3.113402</td>
<td>0.160669</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>2.500000</td>
<td>3.051167</td>
<td>0.152771</td>
</tr>
</tbody>
</table>
```

- `pd.crosstab([tips.time, tips.day], tips.smoker, margins=True)`

```
<table>
<thead>
<tr>
<th>time</th>
<th>smoker</th>
<th>No</th>
<th>Yes</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>day</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dinner</td>
<td>Fri</td>
<td>3</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Sat</td>
<td>45</td>
<td>42</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>Sun</td>
<td>57</td>
<td>19</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Thur</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Lunch</td>
<td>Fri</td>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Thur</td>
<td>44</td>
<td>17</td>
<td>61</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>151</td>
<td>93</td>
<td>244</td>
</tr>
</tbody>
</table>
```
Visualization and Aggregation: Histograms

- Very similar to bar charts
- Often shown without space between (continuity)
- Choice of number of bins
  - Important!
  - Viewers may infer different trends based on the layout

[Munzner (ill. Maguire), 2014]
Date and Time

• Many different representations: 2016-12-01, 1 Dec 2016, 13:15, 1:15pm
• Time zones: UTC and localizing time
• Time series: data indexed by timestamps
  - Operations: shifting, differences
  - Aggregation
  - Visualizations
  - Window functions
Review

- Exploratory Data Analysis
  - Why?
  - Univariate and multivariate
  - Non-graphical and graphical