http://www.cis.umassd.edu/~dkoop/dsc201
Chicago Food Inspections Exploration

• Based on David Beazley's PyData Chicago talk
• YouTube video: https://www.youtube.com/watch?v=j6VSAsKAj98
• Our in-class exploration:
  - Don't focus on the syntax
  - Focus on:
    • What is information is available
    • **Questions** are interesting about this dataset
    • How to decide on good follow-up questions
    • What the computations mean
Data

• What is data?
  - Types
  - Semantics
• How is data structured?
  - Tables (Data Frames)
  - Databases
  - Data Cubes
• What formats is data stored in?
• Raw versus derived data
Data Types and Semantics

<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/AFS</th>
<th>JOINT RR TKT</th>
<th>7-D UNL</th>
<th>30-D UNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R011 42ND STREET &amp; 8TH AVENUE</td>
<td>00228985</td>
<td>00008471</td>
<td>00000441</td>
<td>0001455</td>
<td>0000134</td>
<td>00033341</td>
<td>00071255</td>
</tr>
<tr>
<td>2</td>
<td>R170 14TH STREET-UNION SQUARE</td>
<td>00224603</td>
<td>00011051</td>
<td>00000827</td>
<td>0003026</td>
<td>00000660</td>
<td>00089367</td>
<td>00199841</td>
</tr>
<tr>
<td>3</td>
<td>R046 42ND STREET &amp; GRAND CENTRAL</td>
<td>00207758</td>
<td>00007908</td>
<td>00000323</td>
<td>0001183</td>
<td>0003001</td>
<td>00040759</td>
<td>00096613</td>
</tr>
<tr>
<td>4</td>
<td>R012 34TH STREET &amp; 8TH AVENUE</td>
<td>00188311</td>
<td>00006490</td>
<td>00000498</td>
<td>0001279</td>
<td>0003622</td>
<td>00035527</td>
<td>00067483</td>
</tr>
<tr>
<td>5</td>
<td>R293 34TH STREET – PENN STATION</td>
<td>00168768</td>
<td>00006155</td>
<td>00000523</td>
<td>0001065</td>
<td>0005031</td>
<td>00030645</td>
<td>00054376</td>
</tr>
<tr>
<td>6</td>
<td>R033 42ND STREET/TIMES SQUARE</td>
<td>00159382</td>
<td>00005945</td>
<td>00000378</td>
<td>0001205</td>
<td>0000690</td>
<td>00058931</td>
<td>00078644</td>
</tr>
<tr>
<td>7</td>
<td>R022 34TH STREET &amp; 6TH AVENUE</td>
<td>00156008</td>
<td>00006276</td>
<td>00000487</td>
<td>0001543</td>
<td>0000712</td>
<td>00058910</td>
<td>00110466</td>
</tr>
<tr>
<td>8</td>
<td>R084 59TH STREET/COLUMBUS CIRCLE</td>
<td>00155262</td>
<td>00009484</td>
<td>00000589</td>
<td>0002071</td>
<td>0000542</td>
<td>00053397</td>
<td>00113966</td>
</tr>
<tr>
<td>9</td>
<td>R020 47-50 STREETS/ROCKEFELLER</td>
<td>00143500</td>
<td>00006402</td>
<td>00000384</td>
<td>0001159</td>
<td>0000723</td>
<td>00037978</td>
<td>00090745</td>
</tr>
<tr>
<td>10</td>
<td>R179 86TH STREET-LEXINGTON AVE</td>
<td>00142169</td>
<td>00010367</td>
<td>00000470</td>
<td>0001839</td>
<td>0000271</td>
<td>00050328</td>
<td>00125250</td>
</tr>
<tr>
<td>11</td>
<td>R023 34TH STREET &amp; 6TH AVENUE</td>
<td>00134052</td>
<td>00005005</td>
<td>00000348</td>
<td>0001112</td>
<td>0000649</td>
<td>00031531</td>
<td>00075040</td>
</tr>
<tr>
<td>12</td>
<td>R029 PARK PLACE</td>
<td>00121614</td>
<td>00004311</td>
<td>00000287</td>
<td>0000931</td>
<td>0000792</td>
<td>00025404</td>
<td>00065362</td>
</tr>
<tr>
<td>13</td>
<td>R047 42ND STREET &amp; GRAND CENTRAL</td>
<td>00100742</td>
<td>00004273</td>
<td>00000185</td>
<td>0000704</td>
<td>0001241</td>
<td>00022808</td>
<td>00068216</td>
</tr>
<tr>
<td>14</td>
<td>R031 34TH STREET &amp; 7TH AVENUE</td>
<td>00095076</td>
<td>00003990</td>
<td>00000232</td>
<td>0001459</td>
<td>0002428</td>
<td>00038671</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>R017 LEXINGTON AVENUE</td>
<td>00094655</td>
<td>00004688</td>
<td>00000190</td>
<td>0000833</td>
<td>0000754</td>
<td>00020018</td>
<td>00055066</td>
</tr>
<tr>
<td>16</td>
<td>R175 8TH AVENUE-14TH STREET</td>
<td>00094313</td>
<td>00003907</td>
<td>00000286</td>
<td>0001144</td>
<td>0000256</td>
<td>00038272</td>
<td>00074661</td>
</tr>
<tr>
<td>17</td>
<td>R057 BARCLAYS CENTER</td>
<td>00093804</td>
<td>00004204</td>
<td>00000454</td>
<td>0001386</td>
<td>0001491</td>
<td>00039113</td>
<td>00068119</td>
</tr>
<tr>
<td>18</td>
<td>R138 WEST 4TH ST-WASHINGTON SQ</td>
<td>00093562</td>
<td>00004677</td>
<td>00000251</td>
<td>0000965</td>
<td>0000127</td>
<td>00031628</td>
<td>00074458</td>
</tr>
</tbody>
</table>
Attribute Types

- **Attribute Types**
  - Categorical
  - Ordered
    - Ordinal
    - Quantitative

- **Ordering Direction**
  - Sequential
  - Diverging
  - Cyclic

[Munzner (ill. Maguire), 2014]
Derived Data

- Often, data in its original form isn't as useful as we would like
- Examples: Data about a basketball team's games
- Example 1: 1stHalfPoints, 2ndHalfPoints
  - More useful to know total number of points
  - Points = 1stHalfPoints + 2ndHalfPoints
- Example 2: Points, OpponentPoints
  - Want to have a column indicating win/loss
  - Win = True if (Points > OpponentPoints) else False
- Example 3: Points
  - Want to have a column indicating how that point total ranks
  - Rank = index in sorted list of all Point values
Databases

• Usually more than one table: keys and foreign keys
• **Database Management System:** software to work with databases
• Optimized storage, security, schema design, administration
• Lots of engineering to process transactions **reliably**
  - ACID (atomicity, consistency, isolation, durability)
• Create new tables, insert new rows, update/delete old rows, query tables for results
• Often coupled with the Structured Query Language (SQL)
  - `SELECT population FROM state_pop
    WHERE state = 'MA' and year = '2016';`
• Other database technologies: in-memory stores, column-based stores, NoSQL, graph databases, XML databases
Data Cubes

- More **dimensions**: not only rows/cols
- aka OLAP cube (online analytical processing)
- Often a **hierarchy** for each dimension
- Hierarchy coupled with **aggregation**
- Example: sales data with a time dimension
  - Examine years, months, or days
  - Find the year with maximum sales
  - Find the average sales per month
  - Find the day with the minimum sales
Data Formats

• Start with files (not databases or data cubes)

• **File format** is a set of rules a file's structure should obey

• Common file formats for data:
  - Comma-separated values (CSV)
  - Tab-separated values (TSV) (or other delimiters)
  - Fixed-width format: old school
  - JavaScript Object Notation (JSON)
  - eXtensible Markup Language (XML)
  - Excel
Data Formats

• Start with files (not databases or data cubes)

• **File format** is a set of rules a file's structure should obey

• Common file formats for data:
  - Comma-separated values (CSV)
  - Tab-separated values (TSV) (or other delimiters)
  - Fixed-width format: old school
  - JavaScript Object Notation (JSON)
  - eXtensible Markup Language (XML)
  - Excel
Comma-separated values (CSV) Format

• Comma is a field separator, newlines denote records
  - a,b,c,d,message
  1,2,3,4,hello
  5,6,7,8,world
  9,10,11,12,foo

• May have a header (a,b,c,d,message), but not required

• No type information: we do not know what the columns are (numbers, strings, floating point, etc.)
  - Default: just keep everything as a string
  - Type inference: Figure out what type to make each column based on what they look like

• What about commas in a value? → double quotes
Delimiter-separated Values

- Comma is a **delimiter**, specifies boundary between fields
- Could be a tab, pipe (|), or perhaps spaces instead
- All of these follow similar styles to CSV
Fixed-width Format

• Old school
• Each field gets a certain number of spots in the file
• Example:

<table>
<thead>
<tr>
<th>id</th>
<th>360.242940</th>
<th>149.910199</th>
<th>11950.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>id1594</td>
<td>444.953632</td>
<td>166.985655</td>
<td>11788.4</td>
</tr>
<tr>
<td>id1849</td>
<td>364.136849</td>
<td>183.628767</td>
<td>11806.2</td>
</tr>
<tr>
<td>id1230</td>
<td>413.836124</td>
<td>184.375703</td>
<td>11916.8</td>
</tr>
<tr>
<td>id1948</td>
<td>502.953953</td>
<td>173.237159</td>
<td>12468.3</td>
</tr>
</tbody>
</table>

• Specify exact character ranges for each field, e.g. 0-6 is the id
JavaScript Object Notation (JSON)

• A format for web data
• Associative arrays and lists
• Example:
  ```json
  { "name": "Wes", 
    "places_lived": ["United States", "Spain", "Germany"], 
    "pet": null, 
    "siblings": 
      [{ "name": "Scott", "age": 25, "pet": "Zuko"}, 
        { "name": "Katie", "age": 33, "pet": "Cisco"}]
  }
  ```
• Only contains literals (no variables) but allows null
• Values: strings, arrays, dictionaries, numbers, booleans, or null
  - Dictionary keys must be strings
  - Quotation marks help differentiate string or numeric values
eXtensible Markup Language (XML)

- Self-describing format with nesting
- Each field has tags
- Example:
  ```xml
  - <INDICATOR>
    <INDICATOR_SEQ>373889</INDICATOR_SEQ>
    <PARENT_SEQ></PARENT_SEQ>
    <AGENCY_NAME>Metro-North Railroad</AGENCY_NAME>
    <INDICATOR_NAME>Escalator Avail.</INDICATOR_NAME>
    <PERIOD_YEAR>2011</PERIOD_YEAR>
    <PERIOD_MONTH>12</PERIOD_MONTH>
    <CATEGORY>Service Indicators</CATEGORY>
    <FREQUENCY>M</FREQUENCY>
    <YTD_TARGET>97.00</YTD_TARGET>
  </INDICATOR>
  
  - Top element is the root
Exploratory Data Analysis
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]
Exploratory Data Analysis

• EDA is about understanding data and checking assumptions

• Questions:
  - Is the data correct?
  - Does it match our previous expectations
  - Is there…
    • a pattern?
    • a relationship?
    • a correlation?
    • a trend?
    • …
Analysis is a cycle

Start with basic computations, analyze results, and ask new questions
Comparison between EDA & 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 (one attribute) vs. multivariate (2+ attributes)
- Non-graphical vs. graphical
  - Non-graphical ~ statistics
  - Graphical ~ visualizations
- All are important!
Univariate Non-graphical EDA

• Categorical Data:

• Quantitative Data:
Univariate Non-graphical EDA

• Categorical Data:
  - Frequency counts, proportions
  - Groupings

• Quantitative Data:
  - Distribution
  - Summary statistics: mean, median, mode, variance, standard deviation, quantiles
Word Counts

• "This is the house that Jack built. This is the malt that lay in the house that Jack built. This is the rat that ate the malt. That lay in the house that Jack built."

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>6</td>
<td>0.171</td>
</tr>
<tr>
<td>that</td>
<td>5</td>
<td>0.143</td>
</tr>
<tr>
<td>This</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>is</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>house</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>Jack</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>built.</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>malt</td>
<td>2</td>
<td>0.057</td>
</tr>
<tr>
<td>lay</td>
<td>2</td>
<td>0.057</td>
</tr>
<tr>
<td>in</td>
<td>2</td>
<td>0.057</td>
</tr>
</tbody>
</table>
Univariate Graphical EDA

• Categorical Data
Categorical Data: Group
Bar Chart
Univariate Graphical EDA

- Quantitative Data
Show Every Attribute Value

• Barcode Chart

• Strip Chart

• Jittered Strip Chart

• Bee Swarm Chart

[J. Cherdarchuk, Darkhorse Analytics]
Bar Code Charts

Note: Only individual events are shown; relays are excluded. All swim times are plotted relative to the 2016 Olympics Qualifying Time for respective events. Medals shown in non-Olympic years are for Pan Pacific Championships or World Championships. Data includes most major meets. Source: Swimrankings.net [New York Times, 2016]

2000

2004

2008

2012

2016

5% Faster …

Faster times

'01

'02

'03

'05

'06

'07

'09

'10

'11

'14

'15

'13

2016 Olympic Qualifying Times

Each line is Phelps's time in a single race

Phelps could do no wrong at the Beijing Olympics: he won gold in all of his events.

Phelps retired during 2013, and was suspended for a D.U.I. for much of 2015.

He is training for Rio with a regimen similar to the one that helped him dominate in Beijing.

5% Slower …

Slower times
Swimmers often swim slower when training for important events.
Bee Swarm Charts

Legal
Management
Healthcare Practitioners
Finance
Architecture & Engineering
Business Operations
Life & Social Science
Computer & Mathematical
Technicians
Protective Service
Entertainment & Media
Sales
Education & Library
Community & Social Services
Construction
Maintenance & Repair
Administrative Support
Extraction
Production
Transportation
Military
Cleaning & Maintenance
Healthcare Support
Food Preparation
Personal Care & Service
Farming & Forestry

$0k $25k $50k $75k $100k $125k $150k $175k >$200k
Bin the Data

- **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
Stem-and-leaf Plots
Histograms

Observed ranks of posts by subreddit

"The reddit Front Page is Not a Meritocracy", T. W. Schneider
Simple Distributions

**UNIFORM**

**NORMAL**
REPEATED MEASURES, VARIATION IN POPULATIONS, ETC.

**EXPONENTIAL**
DURATIONS BETWEEN EVENTS, ETC.

[W. Willett, 2016]
Skew in Distributions

SKewed Left

SKewed Right

[W. Willett, 2016]
Common Distributions

- Uniform
- Bernoulli
- Hypergeometric
- Binomial
- Poisson
- Geometric
- Exponential
- Negative Binomial
- Log Normal
- Normal (Gaussian)
- Chi-Squared
- Weibull
- Student’s t
- Gamma
- Beta

[Cloudera]
Boxplots

- Show distribution
- Multiple summary statistics can be read from the chart
- Also provides a general shape of the data
- Best for unimodal data

[N. Yao]
Boxplots

Annual earnings distributions, 10 years after starting school

How to read

10th percentile 25th percentile Median annual earnings 75th percentile 90th percentile

$0 $20K $40K $60K $80K $100K $120K $140K $160K $180K $200K $220K $240K

Harvard

UPenn

Princeton

Columbia

Cornell

Dartmouth

Yale

Brown

[Washington Post, 2015]
Multivariate non-graphical EDA
Multivariate non-graphical EDA

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

• Correlation and covariance
  - Correlation: how related are different attributes?
    • Positive correlation (related)
    • Negative correlation (related)
    • Zero (unrelated)
  - Covariance: how do two attributes change together?
Crosstabs & Pivot Tables

- Count groups and subgroups
- At least two different attributes
- Can subdivide vertically and horizontally for more subgroups
- Sometimes totals are useful

<table>
<thead>
<tr>
<th>region</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>206</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>51</td>
</tr>
</tbody>
</table>
Multivariate graphical EDA
Scatterplots and Correlation

Strong positive correlation

Moderate positive correlation

No correlation

Moderate negative correlation

Strong negative correlation

Curvilinear relationship
Next

• Visualization
  - Tableau
  - Design Rules
• Install Tableau
  - Students receive a free license
  - https://www.tableau.com/academic/students
• Watch Tableau tutorials:
  - https://www.tableau.com/learn/training
• Assignment 1: To be posted soon