Exploratory Data Analysis

Looking carefully at your data is important:

- to identify mistakes in collection/processing
- to find violations of statistical assumptions
- to observe patterns in the data
- to make hypothesis.

Feeding unvisualized data to a machine learning algorithm is asking for trouble.
Why Data Visualization?

- Exploratory data analysis: what does your data really look like?
- Error detection: did you do something stupid?
- Presenting what you have learned to others.

A large fraction of the graphs and charts I see are terrible: visualization is harder than it looks.
Ascombe’s Quartet

All four data sets have exactly the same mean, variance, correlation, and regression line:

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>19.0</td>
<td>12.50</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
</tr>
</tbody>
</table>

mean: 9.0, 7.5, 9.0, 7.5, 9.0, 7.5, 9.0, 7.5
var.: 10.0, 3.75, 10.0, 3.75, 10.0, 3.75, 10.0, 3.75
corr.: 0.816, 0.816, 0.816, 0.816
Plotting Ascombe’s Quartet
Appreciating Art: Which is Better?

Sensible appreciation of art requires developing a particular visual aesthetic.
Tufte’s Visualization Aesthetic

Distinguishing good/bad visualizations requires a design aesthetic, and a vocabulary to talk about data representations:

- Maximize data ink-ratio
- Minimize lie factor
- Minimize chartjunk
- Use proper scales and clear labeling
Maximize Data-Ink Ratio

\[
\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}
\]
The Lie Factor: Dimensionality

(size of effect in graphic) / (size of effect in data)

The fixing a two- or three-dimensional representation by a single parameter yields a lie, because area or volume increase non-proportionally to length.
Graphical Integrity: Scale Distortion

Always start bar graphs at zero.

Always properly label your axes.

Use continuous scales: linear or labelled!
Aspect Ratios and Lie Factors

The steepness of apparent cliffs is a function of aspect ratio. Aim for 45° lines or Golden ratio as most interpretable.
Can this be the Same Data?

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.
Reduce Chartjunk

Extraneous visual elements distract from the message the data is trying to tell.

- Extra dimensionality
- Uninformative coloring
- Excessive grids and figurative decoration

In an exciting graphic, the data tells the story, not the chartjunk.
Chartjunk / Graphical Ducks
Can you Simplify this Plot?
Can You Further Simplify?
Better, but can you Further Simplify?
Anything Else that Can Go?
“Less is More”
MatPlotLib Supports Nice Plots

http://nbviewer.jupyter.org/5357268
Which Chart to Use?

Chart Suggestions—A Thought-Starter

- **Comparison**
  - What would you like to show?

- **Relationship**
  - Two Variables
  - Among Items

- **Composition**
  - Changing Over Time
  - New Periods, Many Periods

- **Distribution**
  - Single Variable
  - Few Data Points

- **Variable Width Column Chart**
  - Many Categories

- **Table or Table with Embedded Charts**
  - Many Items

- **Bar Chart**
  - Few Items

- **Column Chart**
  - Cyclic Data

- **Circular Area Chart**
  - Non-Cyclic Data

- **Line Chart**
  - Single or Few Categories

- **Columns Chart**
  - Few Periods

- **Bubble Chart**
  - Three Variables

- **Scatter Chart**
  - Two Variables

- **Stacked 100% Column Chart**
  - Only Relative Differences Matter

- **Stacked Column Chart**
  - Relative and Absolute Differences Matter

- **Stacked 100% Area Chart**
  - Only Relative Differences Matter

- **Stacked Area Chart**
  - Relative and Absolute Differences Matter

- **Pie Chart**
  - Simple Rate or Total

- **Waterfall Chart**
  - Accumulation or Subtraction to Total

- **3D Area Chart**
  - Components of Components

- **Sum of Subcomponents**

© 2006 A. Abdel — a.abdel@gmail.com
Tabular Data

Tables can have advantages over plots:

- Representation of numerical precision
- Understandable multivariate visualization: each column is a different dimension.
- Representation of heterogeneous data
- Compactness for small numbers of points.
Can this Table be Improved?

<table>
<thead>
<tr>
<th>Country</th>
<th>Area</th>
<th>Density</th>
<th>Birthrate</th>
<th>Population</th>
<th>Mortality</th>
<th>GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russia</td>
<td>17075200</td>
<td>8.37</td>
<td>99.6</td>
<td>142893540</td>
<td>15.39</td>
<td>8900.0</td>
</tr>
<tr>
<td>Mexico</td>
<td>1972550</td>
<td>54.47</td>
<td>92.2</td>
<td>107449525</td>
<td>20.91</td>
<td>9000.0</td>
</tr>
<tr>
<td>Japan</td>
<td>377835</td>
<td>337.35</td>
<td>99.0</td>
<td>127463611</td>
<td>3.26</td>
<td>28200.0</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>244820</td>
<td>247.57</td>
<td>99.0</td>
<td>60609153</td>
<td>5.16</td>
<td>27700.0</td>
</tr>
<tr>
<td>New Zealand</td>
<td>268680</td>
<td>15.17</td>
<td>99.0</td>
<td>4076140</td>
<td>5.85</td>
<td>21600.0</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>647500</td>
<td>47.96</td>
<td>36.0</td>
<td>31056997</td>
<td>163.07</td>
<td>700.0</td>
</tr>
<tr>
<td>Israel</td>
<td>20770</td>
<td>305.83</td>
<td>95.4</td>
<td>6352117</td>
<td>7.03</td>
<td>19800.0</td>
</tr>
<tr>
<td>United States</td>
<td>9631420</td>
<td>30.99</td>
<td>97.0</td>
<td>298444215</td>
<td>6.5</td>
<td>37800.0</td>
</tr>
<tr>
<td>China</td>
<td>9596960</td>
<td>136.92</td>
<td>90.9</td>
<td>1313973713</td>
<td>24.18</td>
<td>5000.0</td>
</tr>
<tr>
<td>Tajikistan</td>
<td>143100</td>
<td>51.16</td>
<td>99.4</td>
<td>7320815</td>
<td>110.76</td>
<td>1000.0</td>
</tr>
<tr>
<td>Burma</td>
<td>678500</td>
<td>69.83</td>
<td>85.3</td>
<td>47382633</td>
<td>67.24</td>
<td>1800.0</td>
</tr>
<tr>
<td>Tanzania</td>
<td>945087</td>
<td>39.62</td>
<td>78.2</td>
<td>37445392</td>
<td>98.54</td>
<td>600.0</td>
</tr>
<tr>
<td>Tonga</td>
<td>748</td>
<td>153.33</td>
<td>98.5</td>
<td>114689</td>
<td>12.62</td>
<td>2200.0</td>
</tr>
<tr>
<td>Germany</td>
<td>357021</td>
<td>230.86</td>
<td>99.0</td>
<td>82422299</td>
<td>4.16</td>
<td>27600.0</td>
</tr>
<tr>
<td>Australia</td>
<td>7686850</td>
<td>2.64</td>
<td>100.0</td>
<td>20264082</td>
<td>4.69</td>
<td>29000.0</td>
</tr>
</tbody>
</table>
Dimensions for Improvement

- Order rows to invite comparisons.
- Order rows to highlight importance or pairwise relationships.
- Right justify uniform-precision numbers.
- Use emphasis, font, or color to highlight important entries.
- Avoid excessive-length column descriptors.
### Improved Tabular Presentation

<table>
<thead>
<tr>
<th>Country</th>
<th>Population</th>
<th>Area</th>
<th>Density</th>
<th>Mortality</th>
<th>GDP</th>
<th>Birth Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>31,056,997</td>
<td>647,500</td>
<td>47.96</td>
<td>163.07</td>
<td>700</td>
<td>36.0</td>
</tr>
<tr>
<td>Australia</td>
<td>20,264,082</td>
<td>7,686,850</td>
<td>2.64</td>
<td>4.69</td>
<td>29,000</td>
<td>100.0</td>
</tr>
<tr>
<td>Burma</td>
<td>47,382,633</td>
<td>678,500</td>
<td>69.83</td>
<td>67.24</td>
<td>1,800</td>
<td>85.3</td>
</tr>
<tr>
<td>China</td>
<td>1,313,973,713</td>
<td>9,596,960</td>
<td>136.92</td>
<td>24.18</td>
<td>5,000</td>
<td>90.9</td>
</tr>
<tr>
<td>Germany</td>
<td>82,422,299</td>
<td>357,021</td>
<td>230.86</td>
<td>4.16</td>
<td>27,060</td>
<td>99.0</td>
</tr>
<tr>
<td>Israel</td>
<td>6,352,117</td>
<td>20,770</td>
<td>305.83</td>
<td>7.03</td>
<td>19,800</td>
<td>95.4</td>
</tr>
<tr>
<td>Japan</td>
<td>127,463,611</td>
<td>377,835</td>
<td>337.35</td>
<td>3.26</td>
<td>28,200</td>
<td>99.0</td>
</tr>
<tr>
<td>Mexico</td>
<td>107,449,525</td>
<td>1,972,550</td>
<td>54.47</td>
<td>20.91</td>
<td>9,000</td>
<td>92.2</td>
</tr>
<tr>
<td>New Zealand</td>
<td>4,076,140</td>
<td>268,680</td>
<td>15.17</td>
<td>5.85</td>
<td>21,600</td>
<td>99.0</td>
</tr>
<tr>
<td>Russia</td>
<td>142,893,540</td>
<td>17,075,200</td>
<td>8.37</td>
<td>15.39</td>
<td>8,900</td>
<td>99.6</td>
</tr>
<tr>
<td>Tajikistan</td>
<td>7,320,815</td>
<td>143,100</td>
<td>51.16</td>
<td>110.76</td>
<td>1,000</td>
<td>99.4</td>
</tr>
<tr>
<td>Tanzania</td>
<td>37,445,392</td>
<td>945,087</td>
<td>39.62</td>
<td>98.54</td>
<td>600</td>
<td>78.2</td>
</tr>
<tr>
<td>Tonga</td>
<td>114,689</td>
<td>748</td>
<td>153.33</td>
<td>12.62</td>
<td>2,200</td>
<td>98.5</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>60,609,153</td>
<td>244,820</td>
<td>247.57</td>
<td>5.16</td>
<td>27,700</td>
<td>99.0</td>
</tr>
<tr>
<td>United States</td>
<td>298,444,215</td>
<td>9,631,420</td>
<td>30.99</td>
<td>6.50</td>
<td>37,800</td>
<td>97.0</td>
</tr>
</tbody>
</table>
Line Charts

- Show data points, not just fits.
- Line segments show connections, so do not use in categorical data.
- Connecting points by lines is often chartjunk. Better is usually a trend line or fit with the data points.
Scatter plots show the values of each point, and are a great way to present 2D data sets. Higher dimensional datasets are often best projected to 2D, through self-organizing maps or principle component analysis, although can be represented through bubble plots.
Reduce Overplotting by Small Points
Heatmaps Reveal Finer Structure

Color points on the basis of frequency
Bubble Charts for Extra Dimensions

Using color, shape, size, and shading of “dots” enables dot plots to represent additional dimensions.
Bar Plots vs. Pie Charts

**Bar plots** show the frequency of proportion of categorical variables. **Pie charts** use more space and are harder to read and compare.

Partitioning each bar into pieces yields the stacked bar chart.

Pie charts are arguably better for showing percentages of totality, and people do seem to like them, so they may be harmless in small amounts.
Which Pie Chart is Better?

Left Pie Chart:
- Trump: 45%
- Cruz: 25%
- Kasich: 14%
- Rubio: 11%
- Others or uncommitted: 5%

Right Pie Chart:
- Trump: 50%
- Cruz: 20%
- Kasich: 15%
- Rubio: 10%
- Others or uncommitted: 5%
Histograms

Histograms (and CDFs) visualize distributions over continuous variables:

Histograms are better for displaying peaks, CDFs for showing tails.
Histograms: Bin Size / Count Matters

Histogram[d, 5], Histogram[d, 10], Histogram[d, 100]

In[62] := {Histogram[d, 5], Histogram[d, 10], Histogram[d, 100]}

Out[62]=

Histogram[d100, 5], Histogram[d100, 10], Histogram[d100, 100]

In[65] := {Histogram[d100, 5], Histogram[d100, 10], Histogram[d100, 100]}

Out[65]=

d100 = Take[d, 100];
Frequency vs. Density Histograms

Dividing counts by the total yields a probability density plot, which is more interpretable: