IMGD 2905

Presenting Data

Chapter 2

Outline

• Types of Charts
• Guidelines for Charts
• Common Mistakes

(next)
“Right” Chart Depends on **Variable Type**

- **Qualitative** (Categorical) variables
  - Can have states or subclasses
    - e.g., position: [striker, goalie, midfield]
  - Can be ordered or unordered
    - e.g., bronze, silver, gold → *ordered*
    - e.g., support, warrior, specialist → *unordered*

- **Quantitative** (Numeric) variables
  - Numeric levels
  - **Discrete or continuous**
    - e.g., goals in season, speed in meters
    - e.g., takedowns, win percentage

---

**Categorical: Bar Chart**

- Chart containing rectangles ("bars") where length represents count, amount, or percent
- Better than table for comparing numbers

Note: bars could be sideways, too

Demo: [imgdops.xlsx](http://www.cs.wpi.edu/~claypool/mqp/paywall/)
Categorical: Pareto Chart

- Bar chart, arranged most to least frequent
- Line showing cumulative percent
- Helps identify most common

Demo: imgdpops.xlsx

Sort by column D
New column E for percent \(=[D2/\text{SUM(D}$2:D$11)]\)
Note: $ “locks” value in (e.g., D$2 versus D2)
New column F for running \(=[\text{SUM(E}$2:E2)]\)
Select 2:11 in B, E and F. Insert combo plot

Categorical: Pie Chart

- Wedge-shaped areas (“pie slices”) – represent count, amount or percent of each category from whole
- Compare relative amounts at a glance
- Best if few slices since quantifying “size” of pie difficult
- Comparing pies also difficult

Demo: imgdpops.xlsx

“The Effects of Latency and Jitter on a First Person Shooter: Team Fortress 2”
http://www.cs.wpi.edu/~claypool/iqp/tf2/
Categorical: Cross-Classification Table

- Multi-column table with count or percent for 2+ categorical variables
  - Good for comparison across multi-categorical data

### Cross-tabulation

<table>
<thead>
<tr>
<th>Class rank</th>
<th>Do you live on campus?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Off-campus</td>
<td>On-campus</td>
</tr>
<tr>
<td>Freshman</td>
<td>37</td>
<td>169</td>
</tr>
<tr>
<td>Sophomore</td>
<td>42</td>
<td>48</td>
</tr>
<tr>
<td>Junior</td>
<td>66</td>
<td>3</td>
</tr>
<tr>
<td>Senior</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>231</td>
<td>167</td>
</tr>
</tbody>
</table>

Demo: [grades.xlsx](#)

Insert Pivot Chart
Select Major through Grade
Drag Majors to Axis
Drag Grade to Axis
Drag Grade to Values

---

Numeric: Frequency Distribution

- Groups of numeric values and frequency
- e.g., Survey of Champion “skins” bought with RP
  - 1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0
  - Cluster into groups
  - Report frequency per group

<table>
<thead>
<tr>
<th>Skins</th>
<th>Freq.</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>20%</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>30%</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>25%</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>15%</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>10%</td>
</tr>
</tbody>
</table>

- May include percentage
- Typically equal size
  - Sometimes ends are open (for extremes)
- Bin size/number variable
  - Too many and not readable
  - **Guide:** given data points
    - 100 or fewer 7-10
    - 101-200 11-15
    - 200+ 13-20

---
Cumulative Distribution

- Cumulative amount of data with value or less
- Easy to see min, max, median
- Compare shapes of distributions

Demo: lol-patches.xlsx

Select column R (Bug Fixes)
Sort low to high
New column S for percent [=ROW()/164]
Select column → paste down all
Select both column R and S
Insert → Scatter plot with lines

"Nerfs, Buffs and Bugs - Analysis of the Impact of Patching on League of Legends"
http://www.cs.wpi.edu/~claypool/papers/lol-crawler/

Histogram

- Bar chart for grouped numerical data
  - No (or small) gaps between adjacent bars

Demo: grades.xlsx

Select GPA data
Insert → Statistics Chart → Histogram
Can adjust bins, overflow/underflow

"Ages of professional League players"
http://www.leaguemath.com/early-vs-late-game-champions/
Stem and Leaf Display

- “Histogram-lite” for analysis w/out software
  - e.g., exam scores: 34, 81, 75, 51, 82, 96, 55, 66, 95, 87, 82, 88, 99, 50, 85, 72

9 | 6 5 9
8 | 1 2 7 2 8 5
7 | 5 2
6 | 6
5 | 1 5 0
4 |
3 | 4

Time Series Plot

- Associate data with date
- Line graph with dates (proportionally spaced!)

Demo: majors.xlsx

Sel. year and majors
Insert ➔ Line Chart ➔ More Line Charts
Scatter Plot

- Two numerical variables, one on each axis
- Reveal patterns in relationship
- Setup “right” models (later)

[Image of a scatter plot with a graph and data points]

“Intelligent Simulation of Worldwide Application Distribution for OnLive's Server Network”
http://www.cs.wpi.edu/~claypool/mqp/onlive/

Demo: lol-rates.xlsx
Select two of (win, pick, ban)
Insert → scatter plot

Radar Plot

- Also called “star charts” or “kiviat plots”
- Good for quick visual comparison, especially when axes unequal

[Image of radar plots]

Gold compared to average, LoL NA teams, by role

Demo: lol-rates.xlsx
Select top line {win, pick, ban} + 1 row
num
Insert → Other → Radar scatter plot
Many More Charts!

https://en.wikipedia.org/wiki/Chart

- Bubble
- Waterfall
- Tree
- Gap
- Polar
- Violin
- Candlestick
- Kagi
- Gantt
- Nolan
- Pert
- Smith
- Skyline
- Vowel
- Nomogram
- Natal

- If common chart effective for message, use
- Otherwise, learn/use other charts as needed
- But remember – may need to explain how to read

Game Analytics Charts

http://dl.acm.org/citation.cfm?id=2812792

- Player choices (e.g., build units)
- Density of activities (e.g., where spend time on map)
- Movement through levels
Player Choices – Pie-Chart

Figure 1. Pie-charts show which types of towers have been built on the different building lots. The radius of the pie-chart is proportional to the number of towers built (Kayali, et al., 2014).

(Custom game, comparative study)

Player Location – Heat Map (1 of 2)

Figure 2. (a) Heatmap of death locations on the Team Fortress 2 map Goldrush. (b) Heatmap showing locations where players of a tower defense game collected coins dropped by defeated enemies (Kayali, et al., 2014).
Player Location – Heat Map (2 of 2)

Assassin’s Creed
Where play testers failed
Result: Make red areas easier


Note, Heat Map for Tables, Too!

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2014</td>
<td>2015</td>
<td>2016</td>
</tr>
<tr>
<td>2</td>
<td>January</td>
<td>600</td>
<td>708</td>
<td>594</td>
</tr>
<tr>
<td>3</td>
<td>February</td>
<td>607</td>
<td>984</td>
<td>743</td>
</tr>
<tr>
<td>4</td>
<td>March</td>
<td>901</td>
<td>886</td>
<td>908</td>
</tr>
<tr>
<td>5</td>
<td>April</td>
<td>608</td>
<td>615</td>
<td>835</td>
</tr>
<tr>
<td>6</td>
<td>May</td>
<td>715</td>
<td>833</td>
<td>734</td>
</tr>
<tr>
<td>7</td>
<td>June</td>
<td>520</td>
<td>663</td>
<td>618</td>
</tr>
<tr>
<td>8</td>
<td>July</td>
<td>731</td>
<td>521</td>
<td>950</td>
</tr>
<tr>
<td>9</td>
<td>August</td>
<td>709</td>
<td>663</td>
<td>987</td>
</tr>
<tr>
<td>10</td>
<td>September</td>
<td>633</td>
<td>863</td>
<td>579</td>
</tr>
<tr>
<td>11</td>
<td>October</td>
<td>533</td>
<td>651</td>
<td>841</td>
</tr>
<tr>
<td>12</td>
<td>November</td>
<td>996</td>
<td>958</td>
<td>749</td>
</tr>
<tr>
<td>13</td>
<td>December</td>
<td>792</td>
<td>717</td>
<td>875</td>
</tr>
</tbody>
</table>

Red means sales are low

Excel tutorial at: https://trumpexcel.com/heat-map-excel/
Movement (1 of 2)

(game: *Infinite Mario*, clone of *Super Mario Bros.*)

Movement (2 of 2)

Figure 5. Left: Player movement between regions, cities, and battlegrounds on the World of Warcraft continent Outland. Right: Corresponding matrix view with cells colored according to the number of players moving from one area to another.
Game: DOGeometry - build road to veterinary house
Shows exploration, where stuck

Outline

• Types of Charts (done)
• Guidelines for Charts (next)
  – Again, “art” not “rules”. Learn with experience.
    Recognize good/bad when see it.
• Common Mistakes

![Diagram showing player behavior and exploration](image-url)
Guidelines for Good Charts (1 of 5)

• **Require minimum effort from reader**
  – Perhaps *most* important metric
  – Given two, can pick one that takes less reader effort

![Direct Labeling vs Legend Box](image)

Guidelines for Good Charts (2 of 5)

• **Maximize information**
  – Make self-sufficient
  – Key words in place of symbols
    • e.g., “Gold IV” and not “Player A”
    • e.g., “Daily Games Played” not “Games Played”
  – Axis labels as informative as possible
    • e.g., “Game Time (seconds)” not “Game Time”
  – Help by using captions (or title, if stand-alone)
    • e.g., “Game time in seconds versus player skill in total hours played”

[Link to Phplot documentation](http://www.phplot.com/phplotdocs/conc-labels.html)
Guidelines for Good Charts (3 of 5)

• Minimize ink (1 of 2)
  – Maximize information-to-ink ratio
  – Too much unnecessary ink makes chart cluttered, hard to read
    • e.g., no gridlines unless needed to help read
  – Chart that gives easier-to-read for same data is preferred

Guidelines for Good Charts (3 of 5)

• Minimize ink (2 of 2)

https://www.slideshare.net/NicoleMarinsek/darkhorse-line-chart
Guidelines for Good Charts (4 of 5)

• Use commonly accepted practices
  − Present what people expect
  − e.g., origin at (0,0)
  − e.g., independent (cause) on x-axis, dependent (effect) on y-axis
  − e.g., x-axis scale is linear
  − e.g., increase left to right, bottom to top
  − e.g., scale divisions equal
• Departures are permitted, but require extra effort from reader → so use sparingly!

Guidelines for Good Charts (5 of 5)

• Avoid ambiguity
  − Show coordinate axes
    • at right angles
  − Show origin
    • usually at (0,0)
  − Identify individual curves and bars
    • With key/legend or label
  − Do not plot multiple variables on same chart
    • Single y-axis
Checklist for Good Charts

- **Axes**
  - Are both axes labeled?
  - Are the axis labels self-explanatory and concise?
  - Are the scale and divisions shown on both axes?
  - Are the min and max ranges appropriate?
  - Are the units indicated?

- **Lines/Curves/Points**
  - Is the number of lines/curves reasonably small?
  - Are curves labeled?
  - Are all symbols clearly distinguishable?
  - Is a concise, clear legend provided?
  - Does the legend obscure any data?

- **Information**
  - If the y-axis is variable, is an indication of spread (error bars) shown?
  - Are grid lines required to read data (if not, then remove)?

- **Scale**
  - Are units increasing left to right (x-axis) and bottom to top (y-axis)?
  - Do all charts use the same scale?
  - Are the scales contiguous?
  - Is bar chart order systematic?
  - Are bars appropriate width, spacing?

- **Overall**
  - Does the whole chart add information to reader?
  - Are there no curves/symbols/text that can be removed and still have the same information?
  - Does the chart have a title or caption (not both)?
  - Is the chart self-explanatory and concise?
  - Do the variables plotted give more information than alternatives?
  - Is chart referenced and discussed in any accompanying report?

---

Describing Chart in Report & Presentation

- **“Formula”**
  - Describe all axes
    - E.g., “The x-axis is time since game began, in seconds”
  - Describe data sets/trendlines
    - E.g., “The blue dots are the average maze completion time”
  - Then provide message
    - E.g., “Notice how the red bar is higher than the blue, indicating that ...”

- Example on Web page
  
  [http://web.cs.wpi.edu/~imgd2905/d19/samples/analysis-example.html](http://web.cs.wpi.edu/~imgd2905/d19/samples/analysis-example.html)
Guidelines for Good Charts (Summary)

- For each chart, go over “checklist”
- The more “yes” answers, the better
  - Remember, while guidelines, art and not science
  - So, may consciously decide not to follow these guidelines if better without them → but have *good* reason!
- In practice, takes several trials before arriving at “best” chart
- Want to present message the most: accurately, simply, concisely, logically
- Accompany with *description*! Text or verbal
  - Remember, audience/reader has *not* seen! – Make sure to introduce

Outline

- Types of Charts  (done)
- Guidelines for Charts  (done)
- Common Mistakes  (next)
Common Mistakes (1 of 6)

• Presenting too many alternatives on one chart
• Guidelines
  – More than 5 to 7 messages is too many
    • (Maybe related to the limit of human short-term memory?)
  – Line chart with 6+ curves
  – Column chart with 10+ bars
  – Pie chart with 8+ components
  – Each cell in histogram fewer than 5 values

Common Mistakes (2 of 6)

• Presenting many y-variables on single chart
  – Better to make separate graphs
  – Plotting many y-variables saves space, but better to requires reader to figure out relationship
  – Sometimes, space constraints (e.g., journal/conference papers),
    • So may “bend” but better to remove than “break”
Common Mistakes (3 of 6)

- **Using symbols in place of text**
- More difficult to read symbols than text
- Reader must flip through report to see symbol mapping to text
  - Even if “save” writers time, really “wastes” it since reader is likely to skip!

![Graph showing Y=1, Y=3, Y=5 and Beta vs Alpha]

Common Mistakes (4 of 6)

- **Placing extraneous information on chart**
  - Goal to convey message, so extra information distracting
  - e.g., Using gridlines only when exact values needed
  - e.g., Showing “per-user” data when only average user data needed

![Chart showing BREAD PRICES, 1848 and 2011]
Common Mistakes (5 of 6)

- Selecting scale ranges improperly
  - Most prepared by automatic rules
    • Give good first-guess
  - But
    • May include outlying data points, shrinking body
    • May have endpoints hard to read since on axis
    • May place too many (or too few) tics
  - In practice, (almost) always over-ride scale values

Common Mistakes (6 of 6)

- Using line chart instead of column chart
  - Lines joining successive points signify that they can be approximately interpolated
  - If don’t have meaning, should not use line chart

- No linear relationship between champion types
- Instead, use column chart
Misleading Charts

Non-Zero Origins to Emphasize (1 of 3)

- Normally, both axes meet at origin
- By moving and scaling, can magnify (or reduce!) difference

Which graph is better?
Non-Zero Origins to Emphasize
(2 of 3)

• Choose scale so that vertical height of highest point is at least \( \frac{3}{4} \) of the horizontal offset of right-most point
  – Three-quarters rule

• (And represent origin as 0,0)
Using Broken Scales in Column Charts

- By breaking scale in middle, can exaggerate differences
  - May be trivial, but then looks significant
  - Similar to “zero origin” problem

Using Double-Whammy Graph

- Two curves can have twice as much impact
  - But if two metrics are related, knowing one predicts other ... so use one!

(Card game “War”)
Plotting Quantities without Measure of Spread

• When random quantification, representing mean (or median) alone (or single data point!) not enough.

Pictograms Scaled by Height

• If scaling pictograms, do by area not height since eye drawn to area
  – e.g., “twice as good” → doubling height quadruples area.
Inappropriate Cell Size in Histogram

- Getting cell size “right” always takes more than one attempt
  - If too large, all points in same cell
  - If too small, lacks smoothness – hard to see shape

Same data. Left looks “normal” and right looks “exponential”

Pictorial Games (1 of 2)

- Can deceive as easily as can convey meaning
  - 1992 should be double 1989
  - Years not evenly spaced
Pictorial Games (2 of 2)

- Can deceive as easily as can convey meaning

- Missing x-axis (1997 too far over)
- Missing y-axis hard to compare (1950 height > 1970)
- Data points unclear