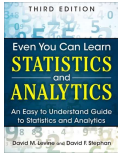


IMGD 2905

Presenting Data

Chapter 2

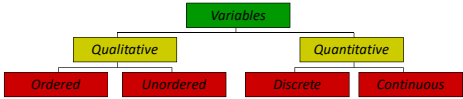


Outline

- Types of Charts (next)
- Guidelines for Charts
- Common Mistakes

“Right” Chart Depends on Variable Type

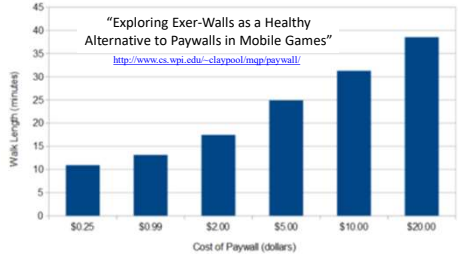
- **Qualitative** (Categorical) variables
 - Can have states or subclasses
 - e.g., position: [striker, goalie, midfielder]
 - Can be ordered or unordered
 - e.g., bronze, silver, gold → ordered
 - e.g., support, tank, jungler → unordered
- **Quantitative** (Numeric) variables
 - Numeric levels
 - Discrete or continuous
 - e.g., goals in season, speed in meters
 - e.g., kills / deaths ratio, win percentage



3

Categorical: Bar Chart

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

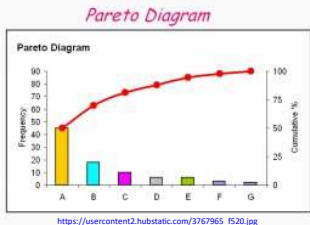


Note: bars could be sideways, too

Demo: [imgdpops.xlsx](#)

Categorical: Pareto Chart

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

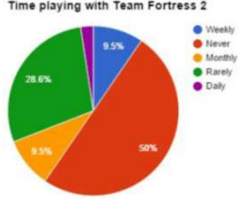


Sort by column D.
 New column E for percent [=D2/SUM(D\$2:D\$12)]
 New column F for running [=SUM(E\$2:E2)]
 Note: \$ “locks” value in (e.g., D\$12 versus D12)
 Select 2:11 in B, E and F. Insert combo plot

Demo: [imgdpops.xlsx](#)

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



“The Effects of Latency and Jitter on a First Person Shooter: Team Fortress 2”
<http://www.cs.wpi.edu/~claypool/jgpt12/>

Demo: [imgdpops.xlsx](#)

Categorical: Cross-Classification Table

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

Class rank * Do you live on campus? Crosstabulation

Count		Do you live on campus?		Total
		Off-campus	On-campus	
Class rank	Freshman	37	100	137
	Sophomore	42	48	90
	Junior	90	8	98
	Senior	62	1	63
Total		231	157	388

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
 - May include percentage and frequency
 - 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
- 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

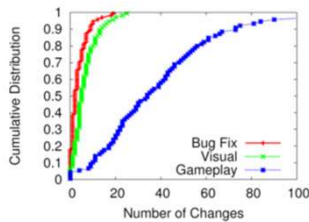
Skins	Freq.	Percent
0	4	20%
1	6	30%
2	5	25%
3	3	15%
4	2	10%

Cumulative Distribution

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

Demo: [lol-patches.xlsx](#)

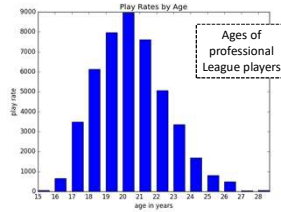
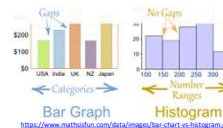
Select column R (Bug Fixes)
 Sort low to high
 New column S for percent [=ROW()/164]
 Select column → paste down all
 Select both column R and S
 Insert → Scatter plot with lines



"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 bwn adjacent bars



Demo: [grades.xlsx](#)

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

Stem and Leaf Display

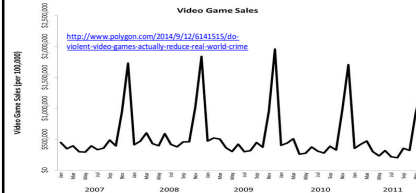
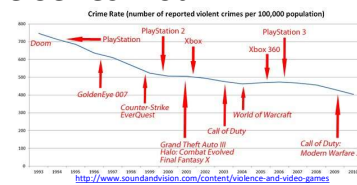
- "Histogram-lite" for analysis w/out software
 - e.g., exam scores: 34, 81, 75, 51, 82, 96, 55, 66, 95, 87, 82, 88, 99, 50, 85, 72

```

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)

Hours of study vs. Test scores

"Intelligent Simulation of Worldwide Application Distribution for OnLive's Server Network"
<http://www.cs.wpi.edu/~claypool/map/online/>

Demo: [lol-rates.xlsx](#)

Select two of {win, pick, ban}
 Insert → scatter plot

Radar Plot

Gold compared to average, LoL NA teams, by role

- Also called "star charts" or "kiviak plots"
- Good for quick visual comparison, especially when axes unequal

Demo: [lol-rates.xlsx](#)

Select top line {win, pick, ban} + 1 row num
 Insert → Other → Radar scatter plot

<http://www.thecoreports.com/lol/news/2561-using-gold-distribution-to-understand-team-dynamic-global-na-lcs-and-s>

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
- Learn/use other charts as needed

Game Analytics Charts

Gunter Wallner and Simone Kriglstein. "An Introduction to Gameplay Data Visualization", *Game Research Methods*, pages 231-250, ETC Press, ISBN: 978-1-312-88473-1, 2015.
<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

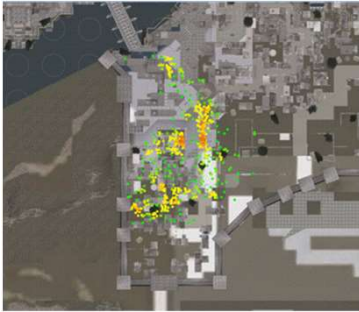
Player Choices – Pie-Chart

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

http://www.gamasutra.com/blog/jonathanDunkoff/20140320/213624/Game_Telemetry_with_DNA_Tracking_on_Assassins_Creed.php

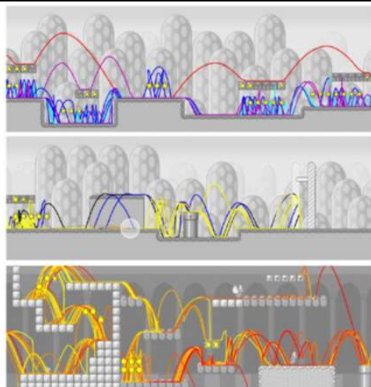
Note, Heat Map for Tables, Too!

	A	B	C	D
1		2014	2015	2016
2	January	600	708	594
3	February	607	984	749
4	March	901	886	908
5	April	608	615	835
6	May	715	833	734
7	June	520	663	618
8	July	731	521	950
9	August	709	663	987
10	September	633	863	979
11	October	533	651	841
12	November	996	958	749
13	December	792	717	875

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.*)

Figure 4. Examples of path visualizations coupled with color coding to communicate additional information. Top color coding reflects the reported expertise of players obtained through a pregame survey. Middle color depicts the area in which the player's character currently resides in. Bottom the color gradient reflects physiological data measured in the form of galvanic skin response (Mitra-Babus, et al., 2016)

Movement (2 of 2)

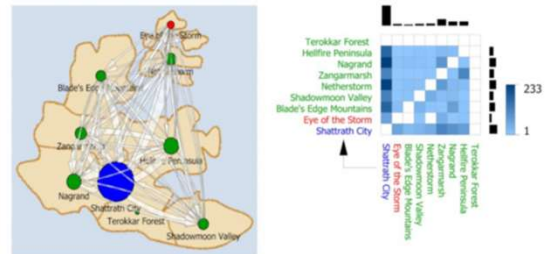


Figure 5. Left: Player movement between regions, cities, and battlegrounds on the World of Warcraft continent Outland. Right: Corresponding matrix view with cells colored according to the number of players moving from one area to another.

Player Behavior - Node-link

Game: DOGeometry - build road to veterinary house
Shows exploration, where stuck

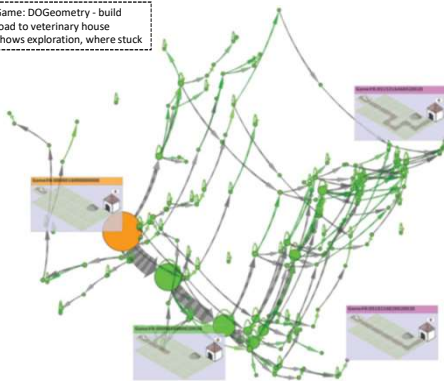
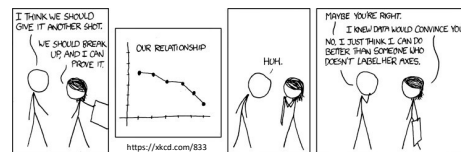


Figure 6. Node-link visualization of player behavior in an abstract puzzle game (Wallner and Krigstein, 2011).

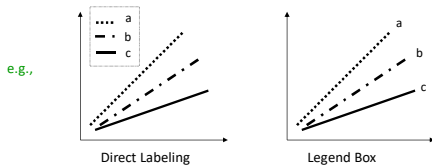
Outline

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



Guidelines for Good Charts (1 of 5)

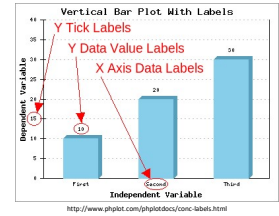
- **Require minimum effort from reader**
 - Perhaps *most* important metric
 - Given two, can pick one that takes less reader effort



25

Guidelines for Good Charts (2 of 5)

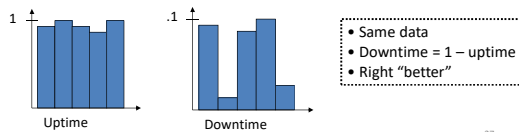
- **Maximize information**
 - Make self-sufficient
 - Key words in place of symbols
 - e.g., “Gold IV” and not “Player A”
 - e.g., “Daily Games Played” not “Games Played”
 - Axis labels as informative as possible
 - e.g., “Game Time (seconds)” not “Game Time”
 - Help by using captions (or title, if stand-alone)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



26

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



27

Guidelines for Good Charts (3 of 5)

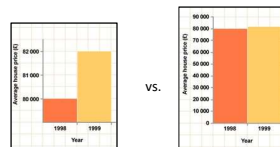
- **Minimize ink (2 of 2)**



<https://www.slideshare.net/NicoleMarinsek/darkhorse-line-chart>

Guidelines for Good Charts (4 of 5)

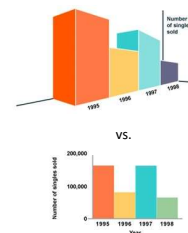
- **Use commonly accepted practices**
 - Present what people expect
 - e.g., origin at (0,0)
 - e.g., independent (cause) on x-axis, dependent (effect) on y-axis
 - e.g., x-axis scale is **linear**
 - e.g., increase **left to right**, **bottom to top**
 - e.g., scale divisions **equal**
- Departures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



29

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



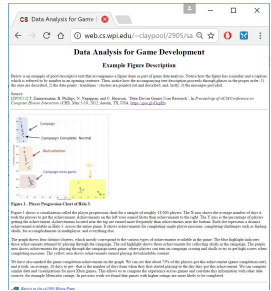
30

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?
- **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?
- **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)?

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/d17/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

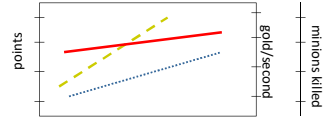
- Types of Charts (done)
- Guidelines for Charts (done)
- Common Mistakes (next)

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!

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

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) ticks
 - 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

Misleading Charts

There are three kinds of lies: lies, damned lies, and statistics. —Disraeli

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. —H. G. Wells

It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so. —Artemus Ward

Round numbers are always false. —Samuel Johnson

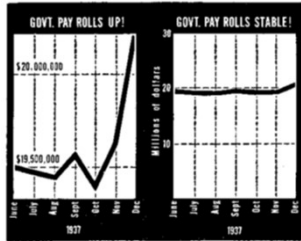
W. W. NORTON & COMPANY - INC. New York

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?

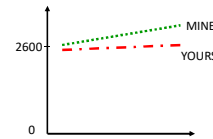
Non-Zero Origins to Emphasize (2 of 3)



Dun's Review, 1938

Non-Zero Origins to Emphasize (3 of 3)

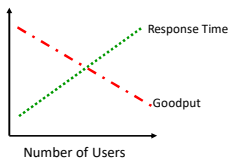
- Choose scale so that vertical height of highest point is at least $\frac{3}{4}$ of the horizontal offset of right-most point
 - Three-quarters rule
- (And represent origin as 0,0)



44

Using Double-Whammy Graph

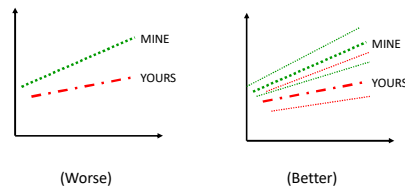
- Two curves can have twice as much impact
 - But if two metrics are related, knowing one predicts other ... so use one!



45

Plotting Quantities without Measure of Spread

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



46

Pictograms Scaled by Height

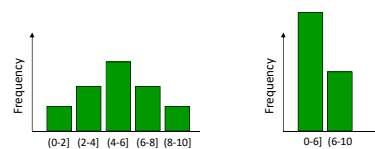
- If scaling pictograms, do by area not height since eye drawn to area
 - e.g., twice as good \rightarrow doubling height quadruples area



47

Using 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

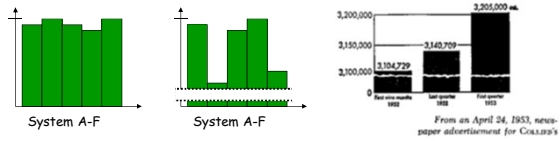


Same data. Left is "normal" and right is "exponential"

48

Using Broken Scales in Column Charts

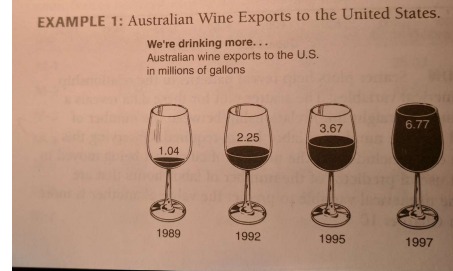
- By breaking scale in middle, can exaggerate differences
 - May be trivial, but then looks significant
 - Similar to “zero origin” problem



49

Pictorial Games (1 of 2)

- Can deceive as easily as can convey meaning



50

Pictorial Games (2 of 2)

- Can deceive as easily as can convey meaning

