Storytelling with Data

January 27, 2016

If you're waiting, please download Excel data for class here:

http://tinyurl.com/Storytellingwithdata

Hi!
My name is
Carey Anne

Founder & CEO of Open Data Nation
opendatanation.com
info@opendatanation.com

TODAY'S OBJECTIVES

WHY VISUALIZATION?
EXPLORE MULTIPLE VISUALIZATIONS & FEATURES
TELL A STORY WITH DATA

Let’s get some data for today!

Obtain and understand the data

You work for the City of Baltimore's Human Resource Department and you are tasked with understanding where all of the city's money goes. Your boss has left the question open ended. How will you answer this question?

http://tinyurl.com/Storytellingwithdata

https://data.baltimorecity.gov/City-Government/Baltimore-City-Employee-Salaries-FY2014/2j28-xzd7

With a partner…

Write out one or two questions that you would like to explore. What do you want to know about how and what employees are paid in Baltimore?

For example, you might look at annual wages to see if there are a few employees that are paid a lot.

Why do we care to visualize data?

Why visualize?

Compress data

Aid in comprehension

Summarize information

Synthesize results

Reveal hidden trends

Categorize information

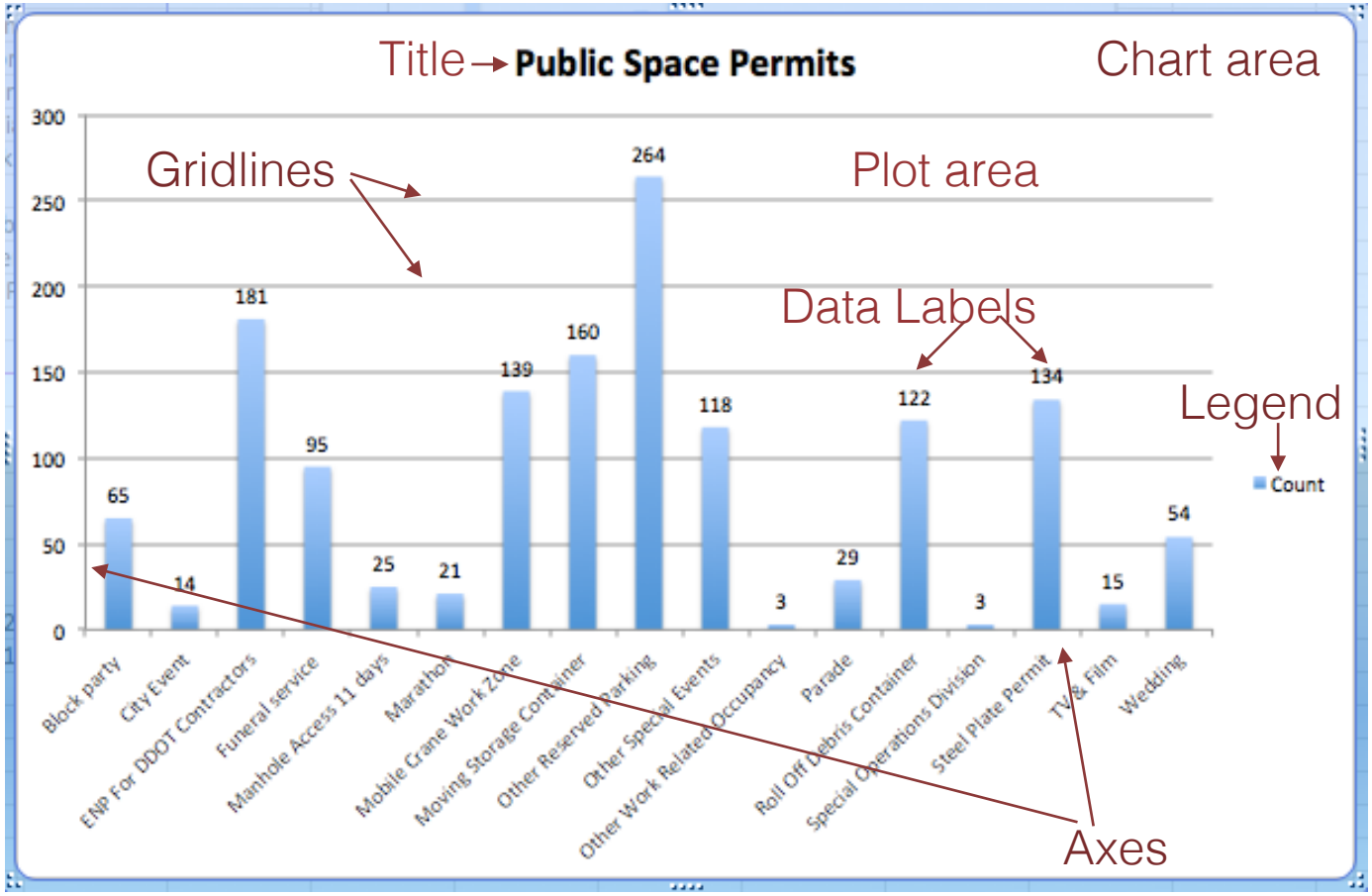
Display for presentation

Identify opportunities for future analysis

Let’s visualize and explore features!

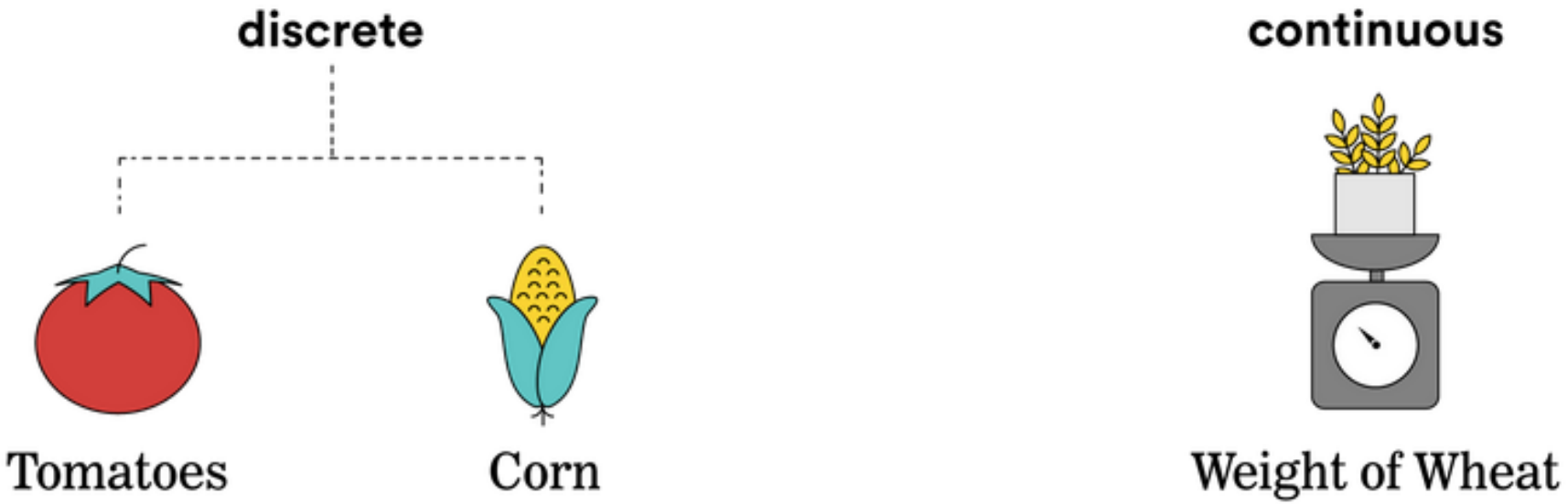
Chart anatomy 101
Bar chart and charting best practices
Histograms and using ‘Select Data’
Pie chart … just don’t.
Line graph and adding data labels
Scatterplot and statistics features

Chart anatomy 101

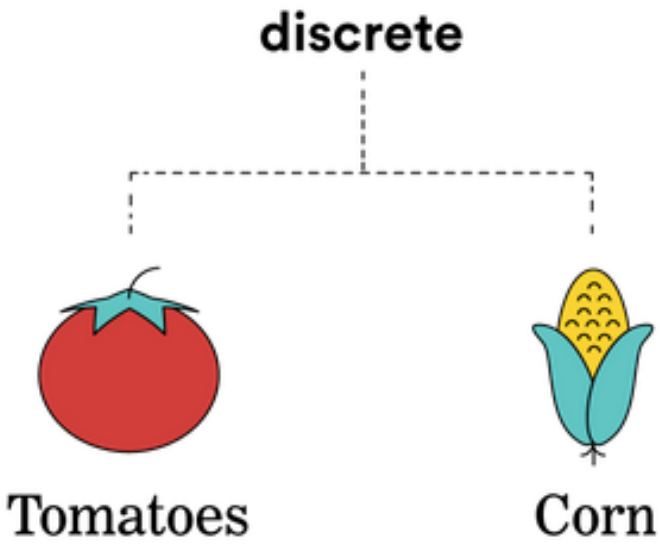


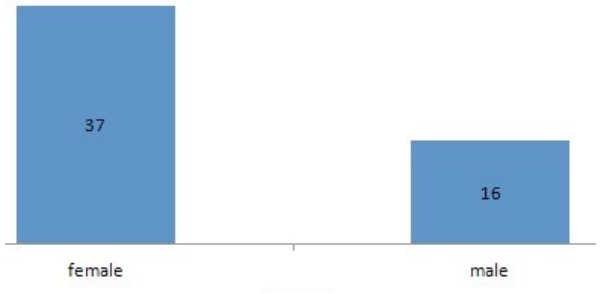
Bar charts

Discrete and continuous



Discrete and continuous





Bar chart

- Gaps between columns

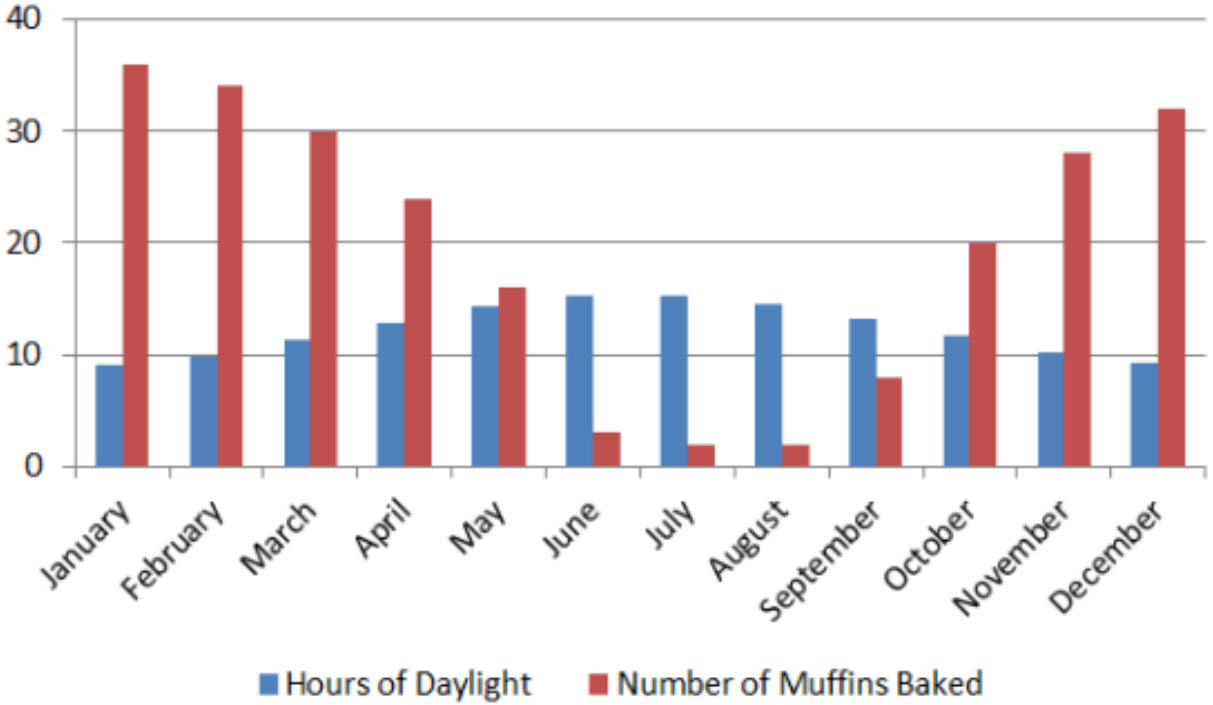
  Because data is categorical

- Y Axis is Count

  Because bars contain discrete observations of one value

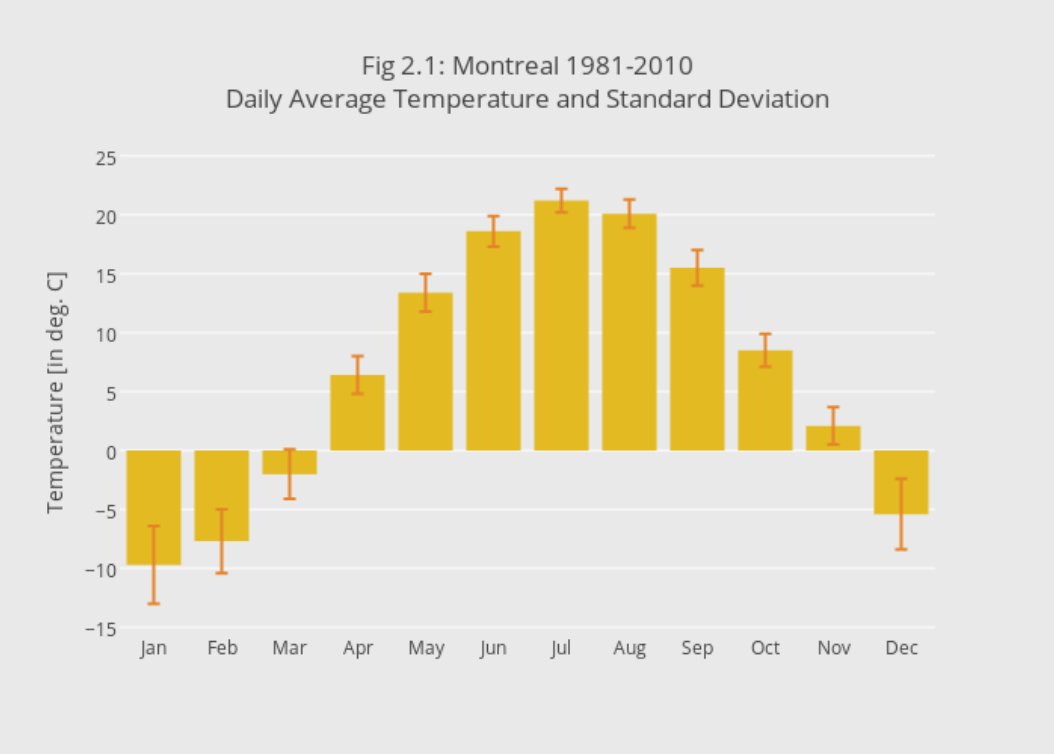
Bar chart examples

Grouped bar chart example



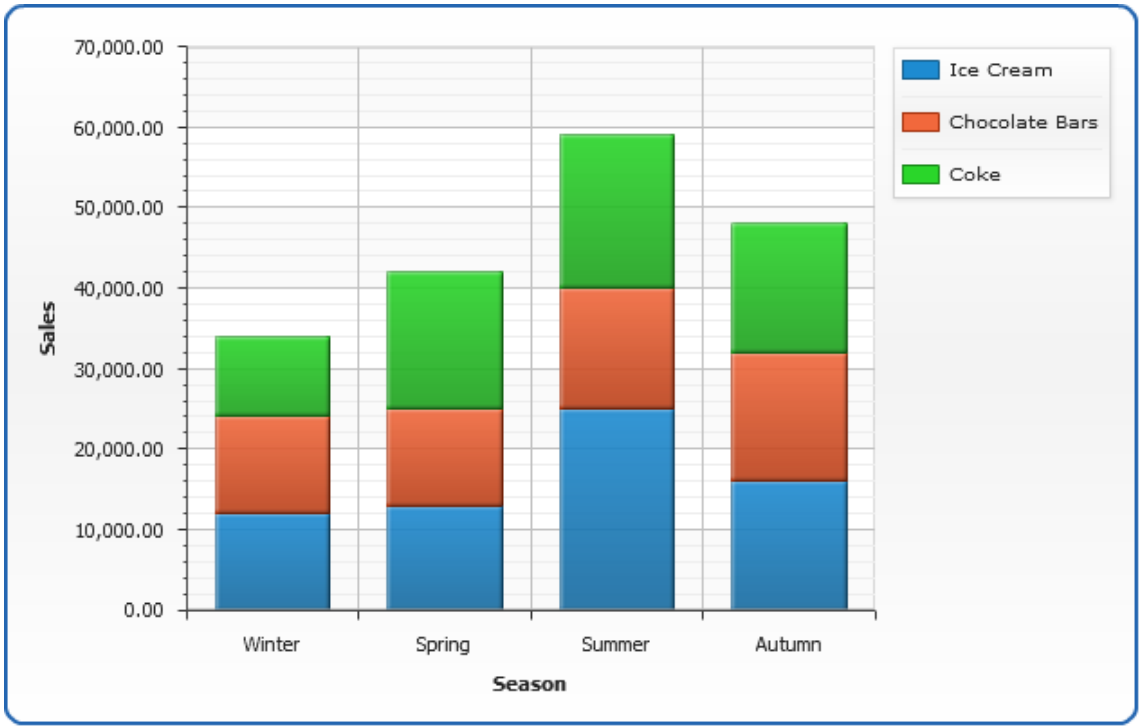
Source: https://lovestats.wordpress.com/

Deviation bar chart example



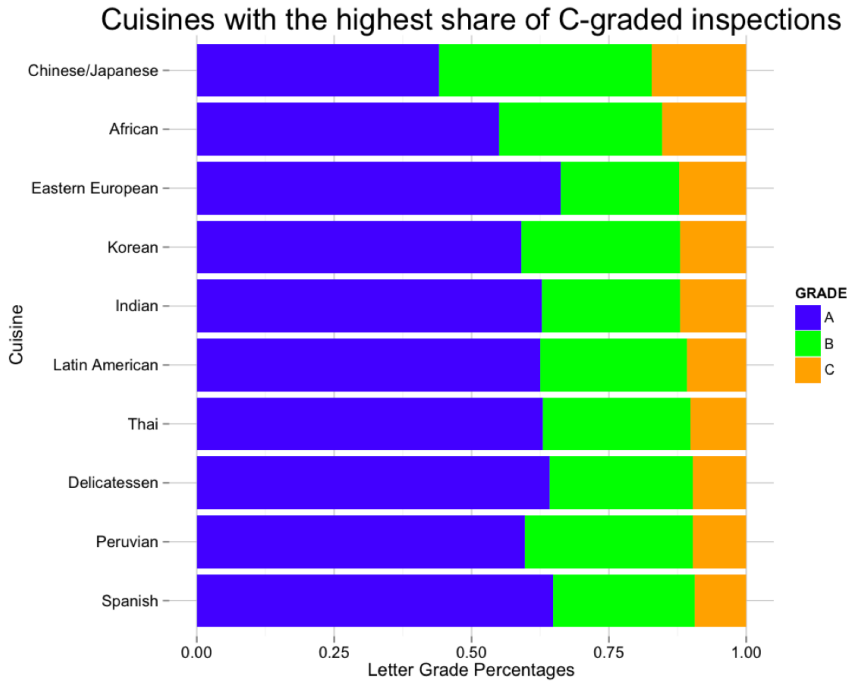
Source: https://plot.ly/python/bar-charts-tutorial/

Stacked bar chart example



Source: http://6.anychart.com

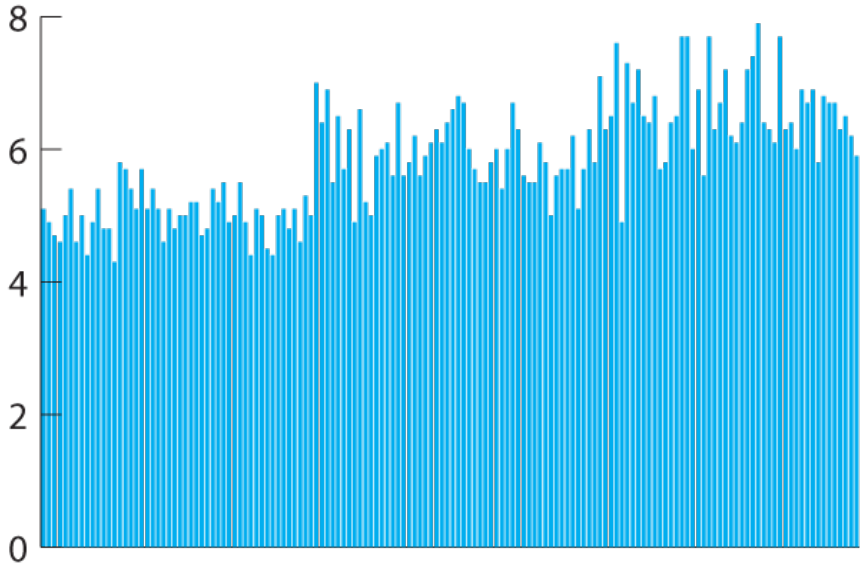
Stacked horizontal percentage bar chart



Best practices in bar charts

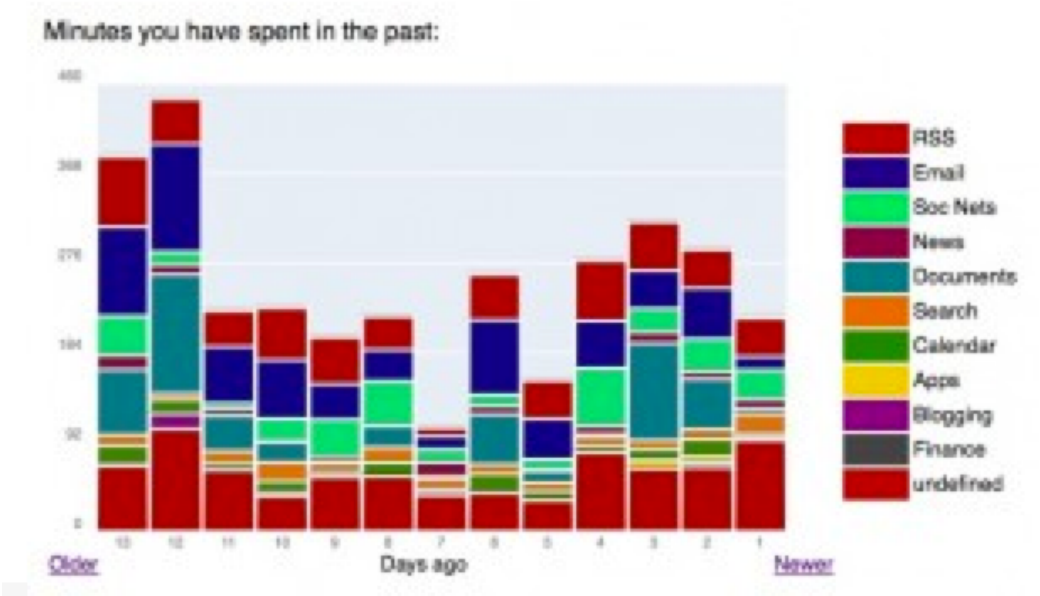
Bar charts best practices

- Limit number of bars to 10-12
- Too many categories make a chart hard to read:



Bar charts best practices

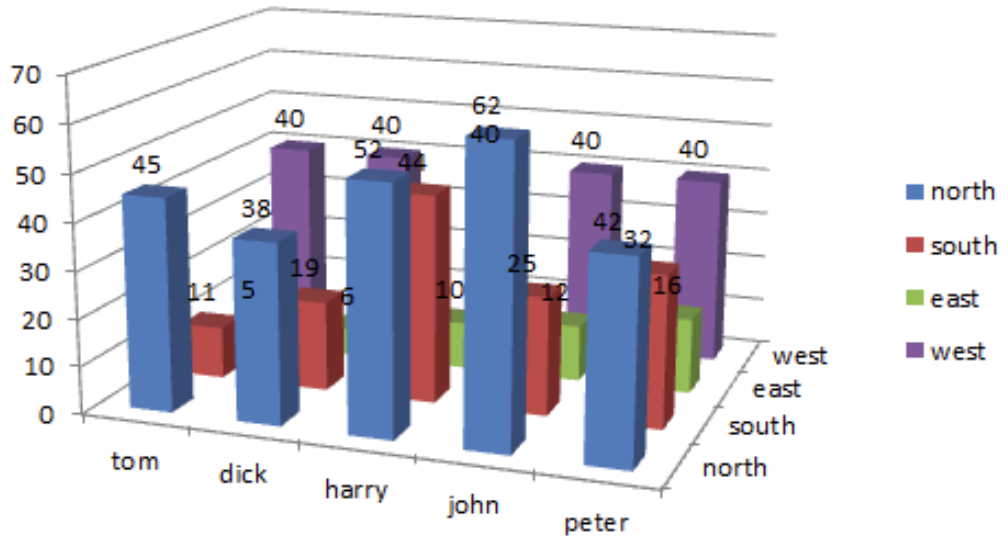
- Limit number of stacked categories to five or six
- Don’t be this guy:



Source: http://www.excelcharts.com/

Bar charts best practices

- Get rid of the 3-D effects. This isn't 1998.



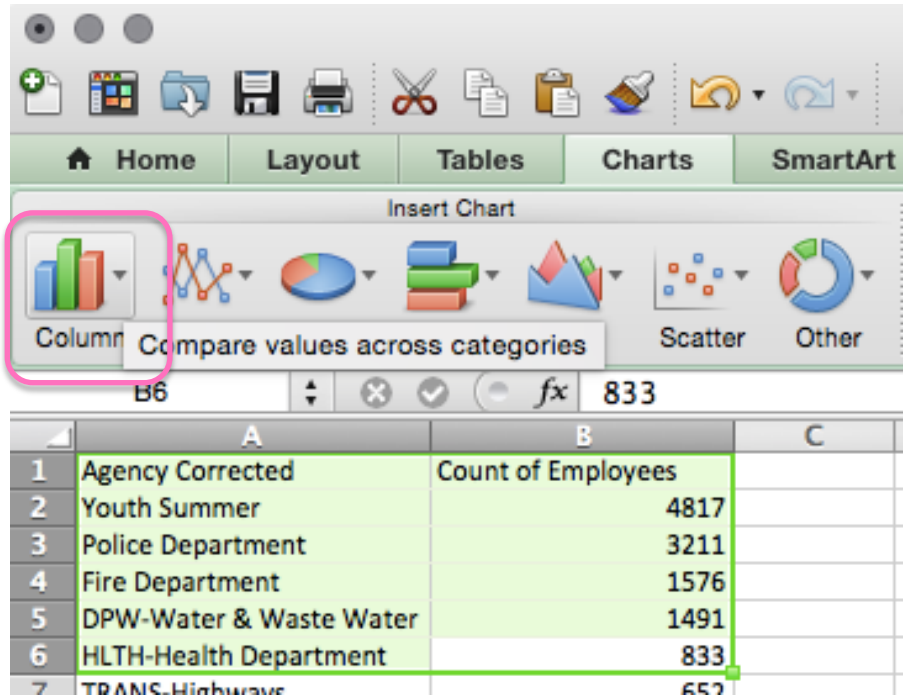
Source: http://www.forbes.com/sites/naomirobbins/2012/05/30/winner-of-the-bad-graph-contest-announced-2/

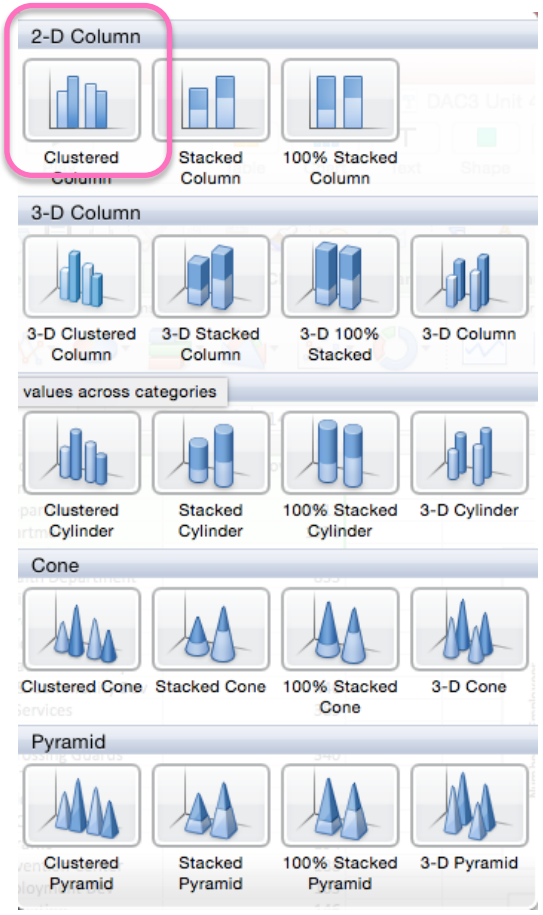
Bar Charts - Let’s practice.

You work for the City of Baltimore's Human Resource Department and you are tasked with understanding where all of the city's money goes.

In the bar chart worksheet, explore:
What are the top 5 employers by number of employees?

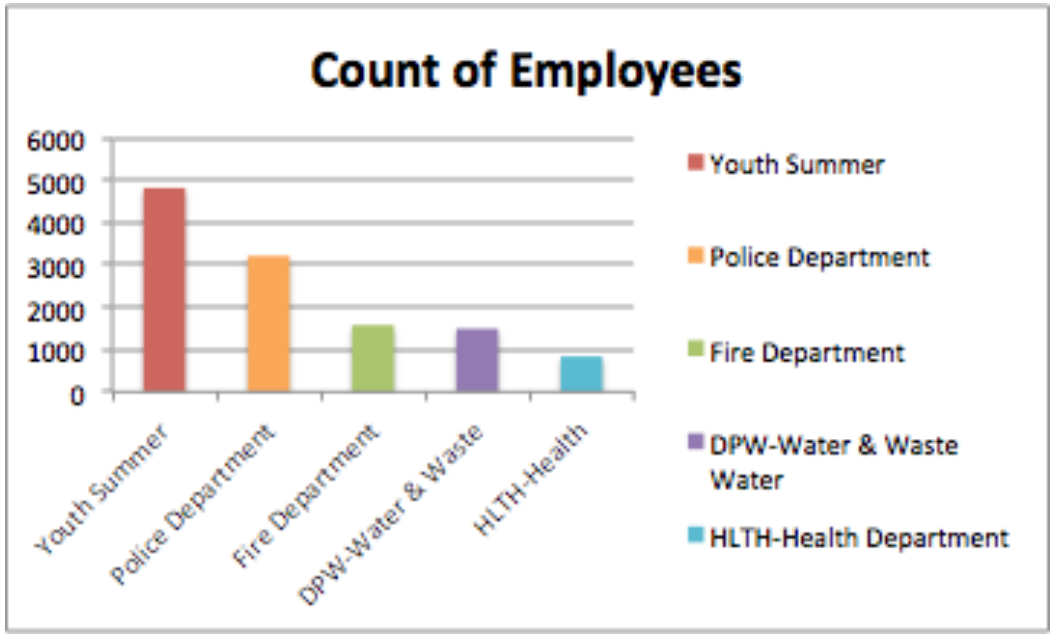
Let’s work together





Our results:

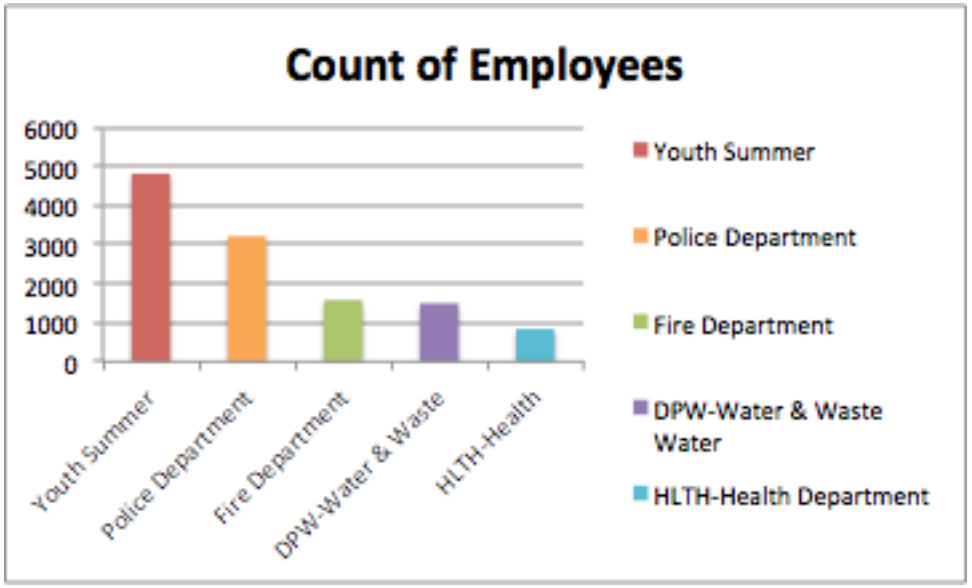
- Youth Summer employs the most, with nearly 5,000 employees.
- The Police Department employs the next most,
- The Police Department employs almost double the third most employer, the Fire Department.



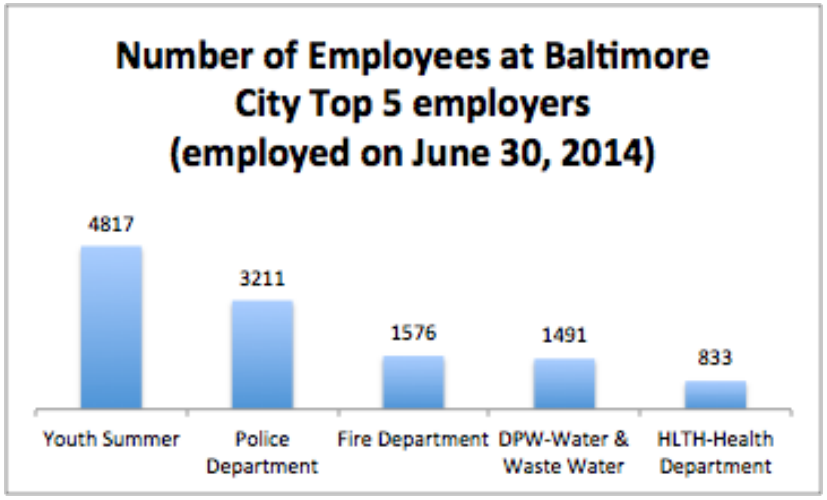
Design critique

Visualization best practices

1. Make the title descriptive
2. Label axes when necessary
3. Get rid of extra whitespace
4. Get rid of redundancies
5. Keep the colors distinct
6. Everything has a purpose!



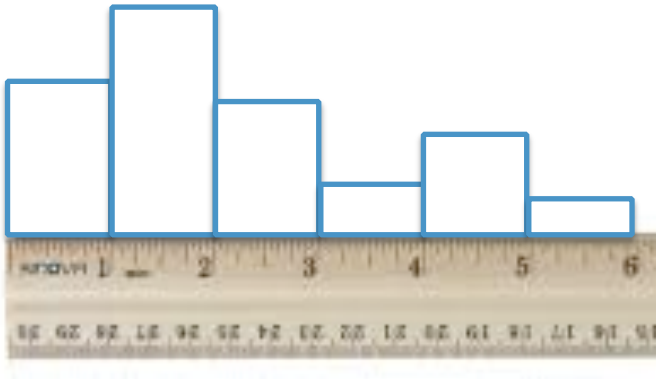
Visualizations incorporating best practices



Histograms

Discrete and continuous

Histogram



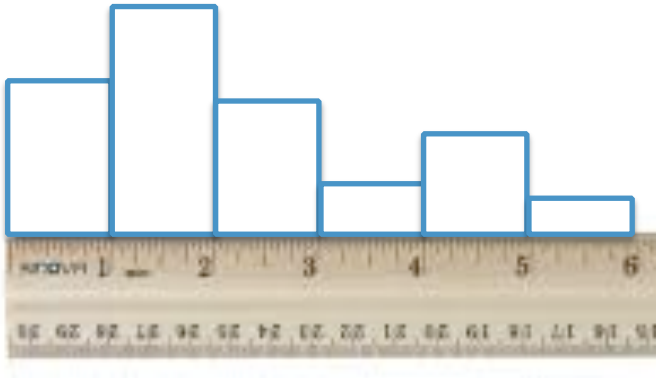
- No gaps between columns
  Because data is continuous
- Y Axis is Frequency
  Because bins contain a range of data

Bins divide the data into intervals

- All be the same size
- Include all the data, even outliers
- Ideally are whole numbers

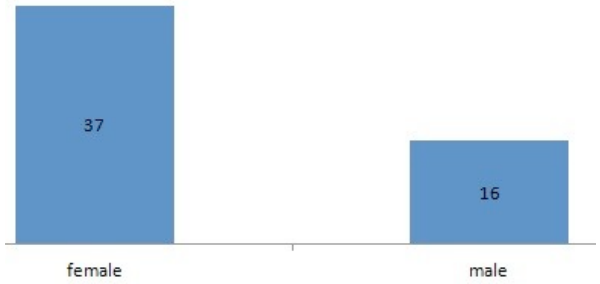
TRY THIS!

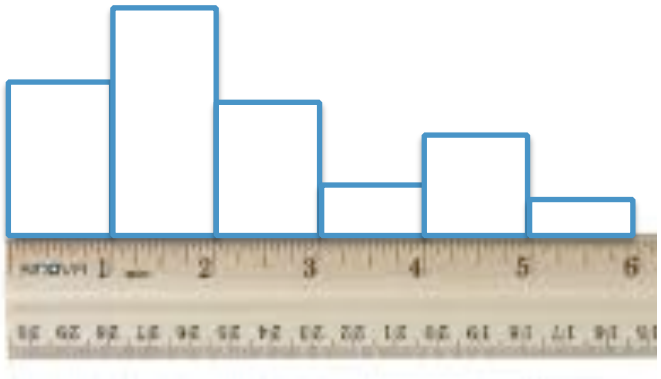
In which bin is:
1/2 inch?
4.1 inches?
3 inches?



BINS = 1 2 3 4 5 6

Discrete and continuous





Bar chart

- Gaps between columns
  Because data is categorical
- Y Axis is Count
  Because bars contain discrete observations of one value

Histogram

- No gaps between columns
  Because data is continuous
- Y Axis is Frequency
  Because bins contain a range of data

TRY THIS!

With Baltimore City Data, let's determine appropriate for the variable 'Annual Salary'

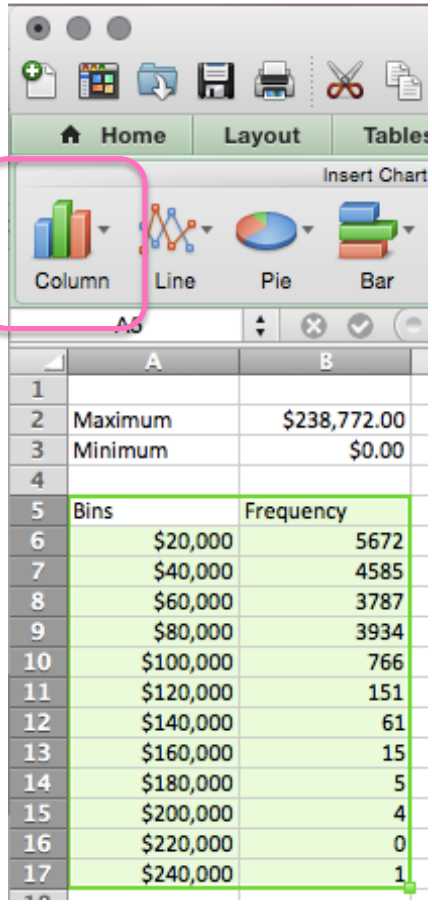
Histograms - Let’s try together

You work for the City of Baltimore's Human Resource Department and you are tasked with understanding where all of the city's money goes.

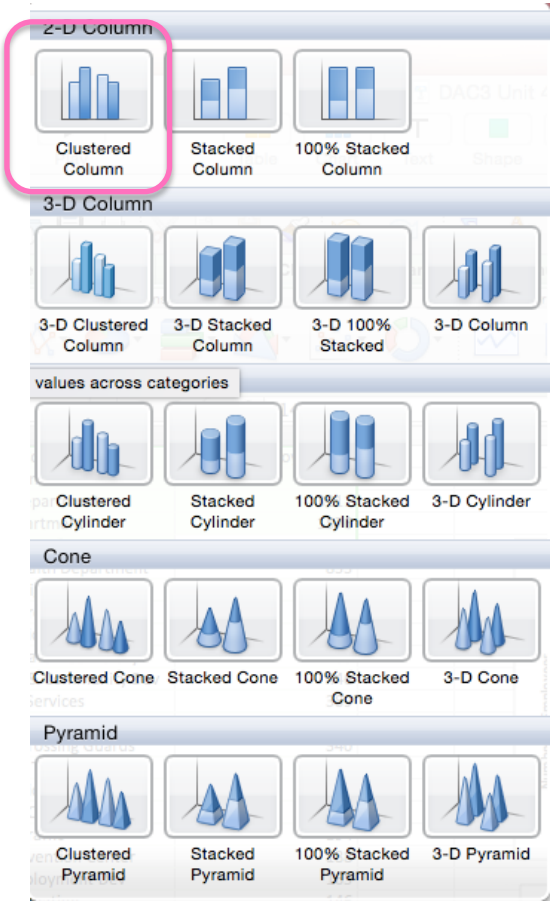
In a sheet of data labeled 'histogram,' explore:

1. What is the distribution of wages in Baltimore?
2. Are there many employees who are paid a lot of money?

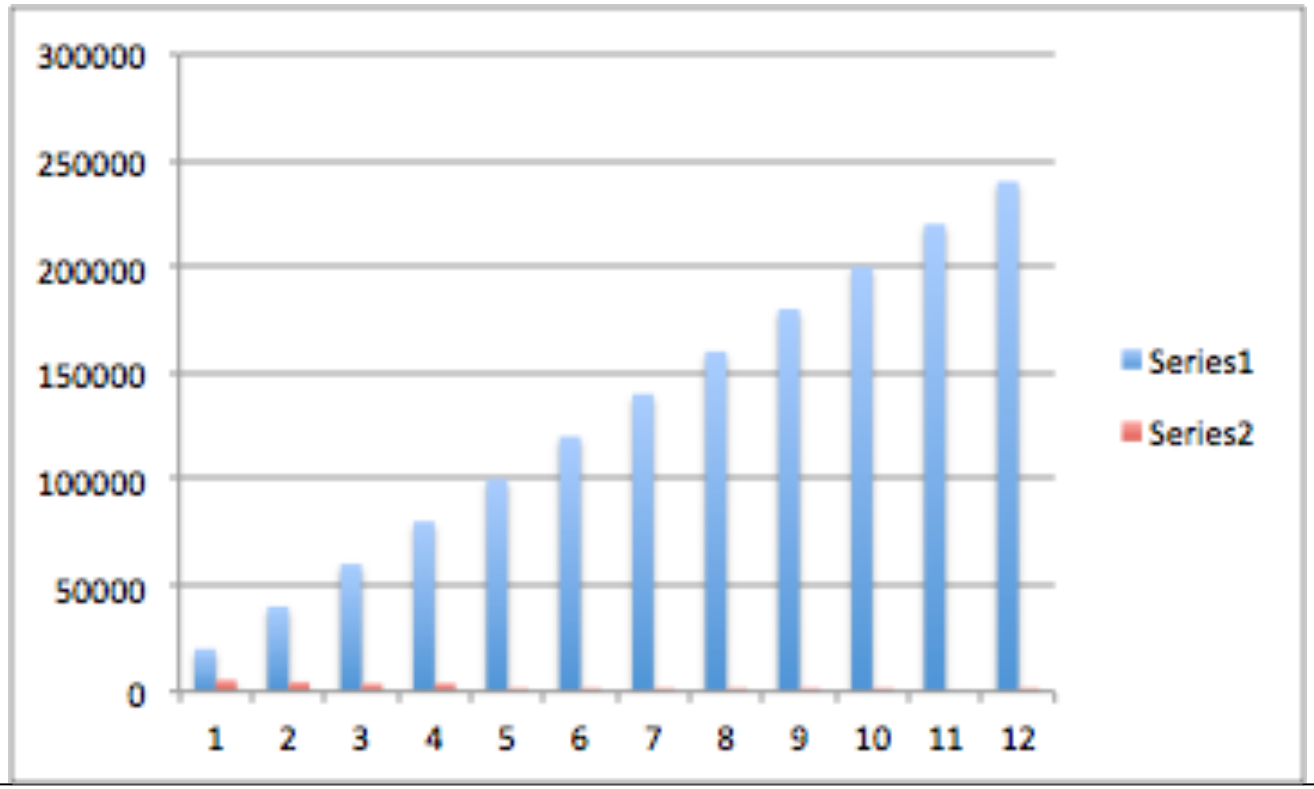
Let's work together



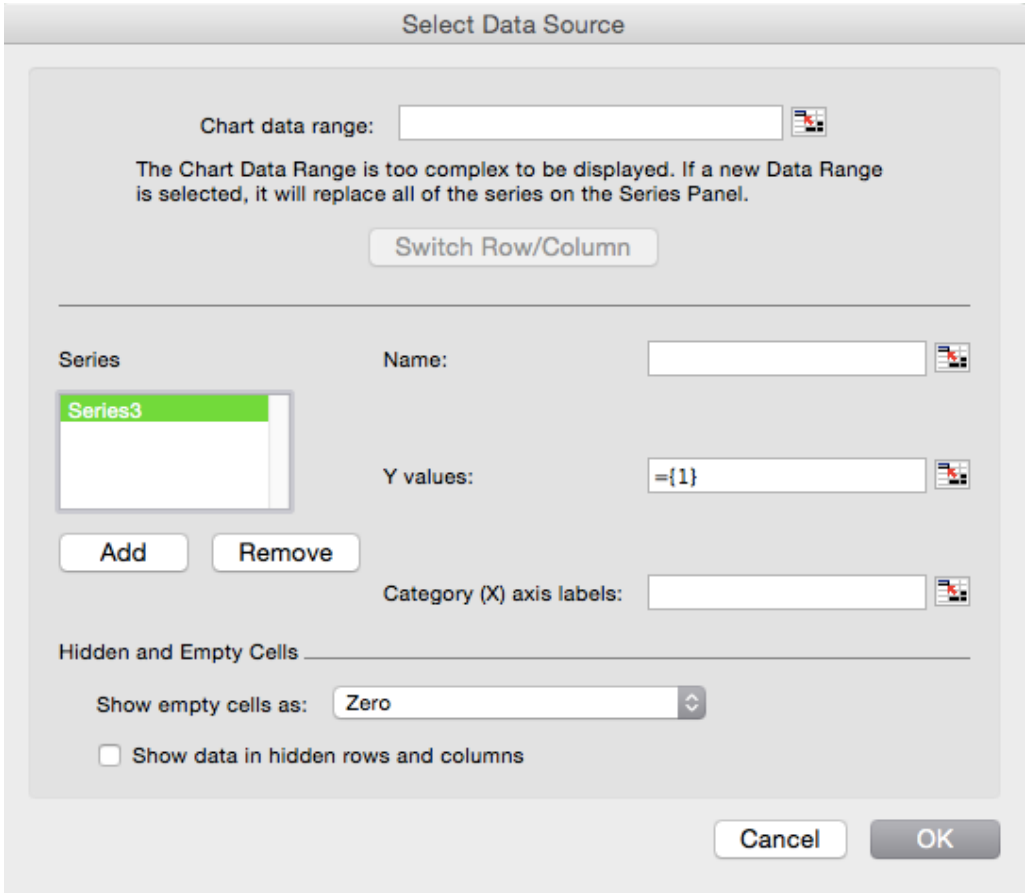
	A	B
1		
2	Maximum	$238,772.00
3	Minimum	$0.00
4		
5	Bins	Frequency
6	$20,000	5672
7	$40,000	4585
8	$60,000	3787
9	$80,000	3934
10	$100,000	766
11	$120,000	151
12	$140,000	61
13	$160,000	15
14	$180,000	5
15	$200,000	4
16	$220,000	0
17	$240,000	1



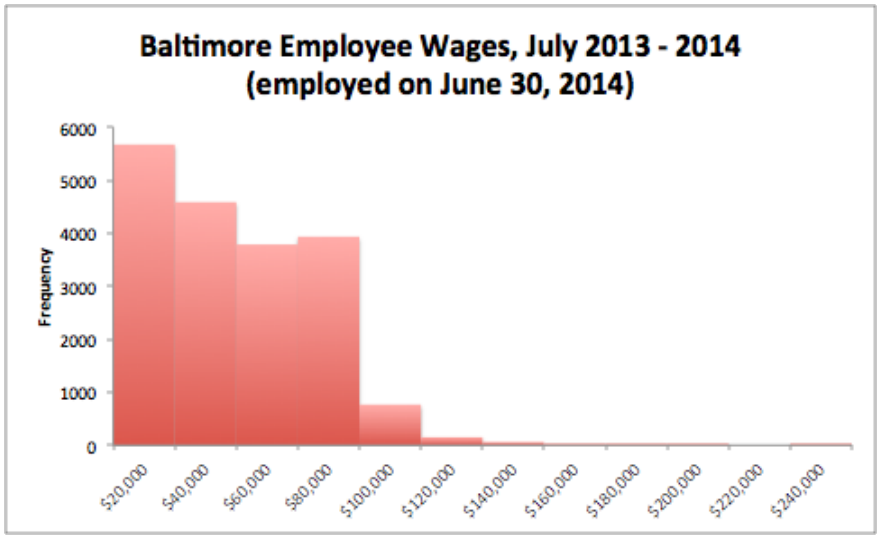
What in the $%@*!

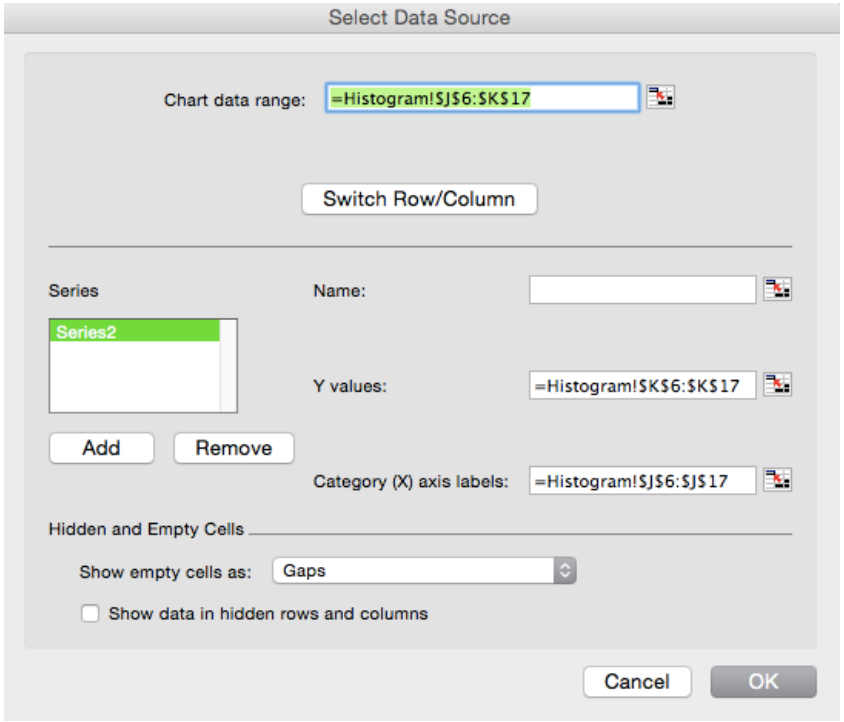

300000
250000
200000
150000
100000
50000
0
1
2
3
4
5
6
7
8
9
10
11
12
Series1
Series2

Select data window is your new best friend



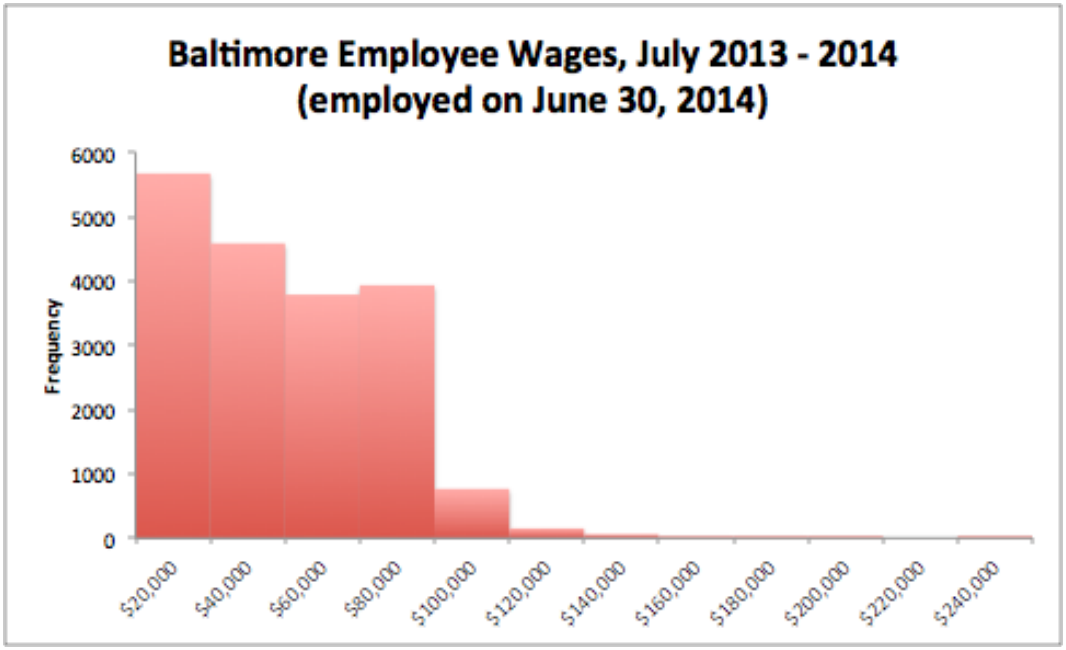
Use ‘Select Data’ to clean up histogram





A histogram shows the frequency of continuous values in bins or buckets

- Most commonly, Baltimore City employees earn less than $20,000.
- Most employees in Baltimore earn between $0 and $80,000
- The highest paid Baltimore employee earns over $238,000.



What would happen if we created a bar chart?

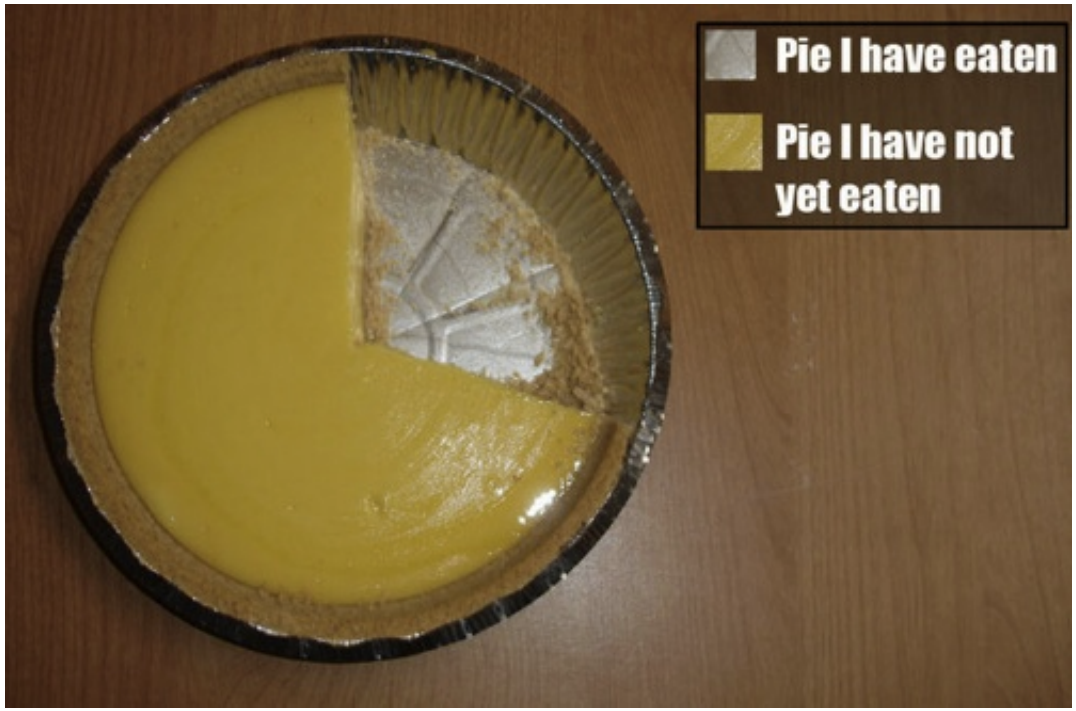
A **bar chart** shows the **count** of **discrete** observations **by category**

Pie charts

Pie chart

Shows the sum of a whole value

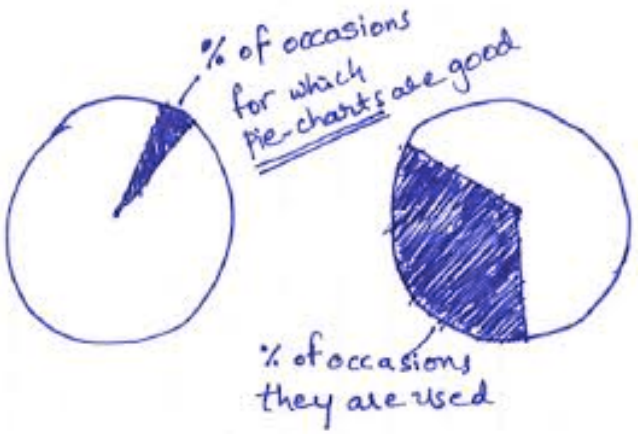
Typically not good practice because the human eye has a difficult time interpreting volume of a wedge or circle.

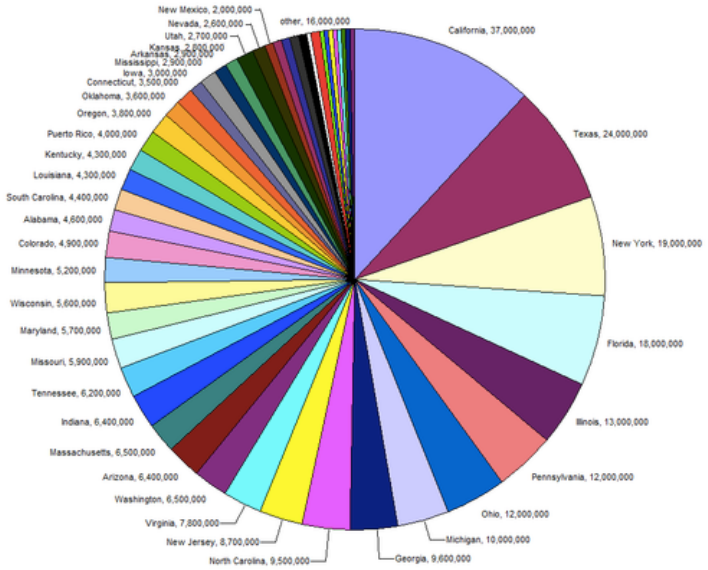


Best practices in pie charts

Best practices with pie charts

Don’t. Use a stacked column instead. If you don't head this warning then at least…



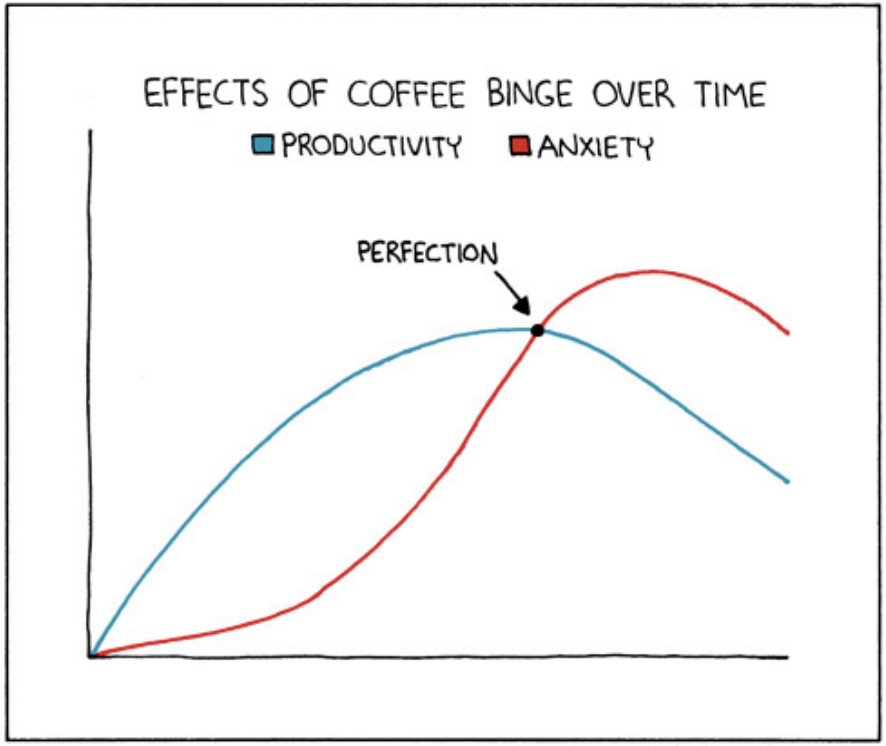


- Limit to 6 or fewer categories
- Make sure wedges sum to 100%

Line graphs

Line graphs

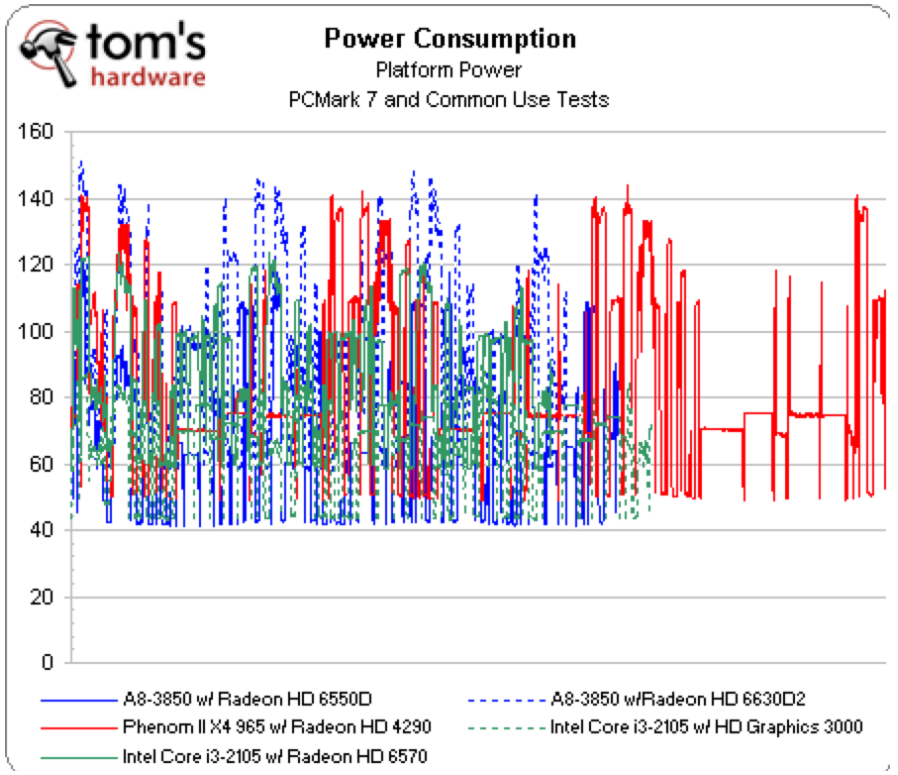
Often the horizontal axis represents time (a “time series” chart)



Best practices in line charts

Line graph best practices

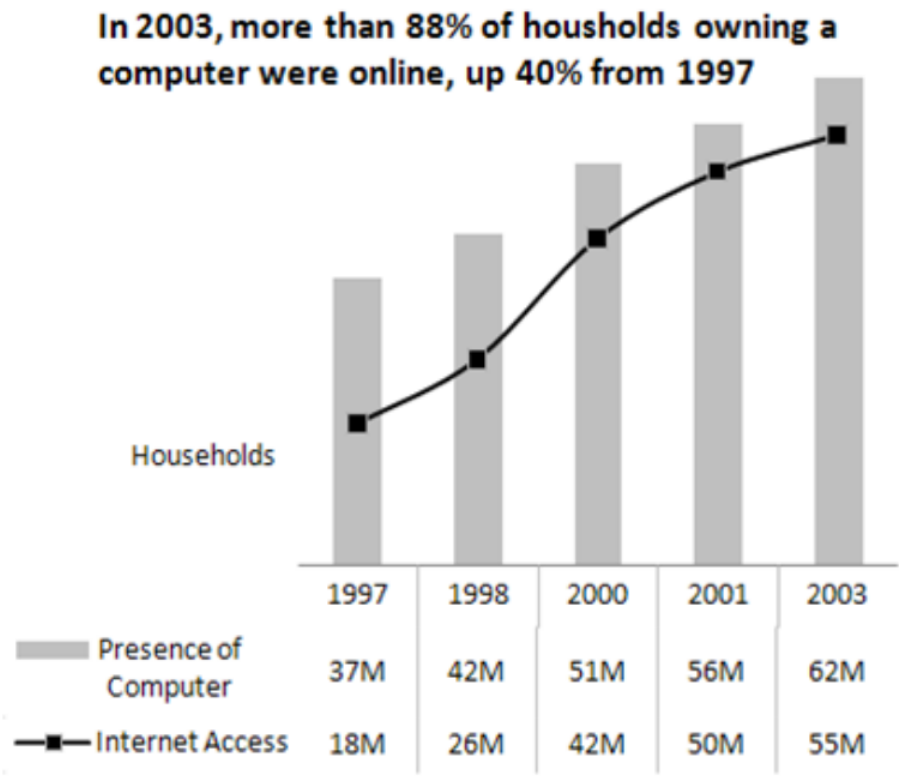
Avoid having too many lines!



Source: http://www.tomshardware.com/reviews/amd-a8-3850-llano,2975-18.html

Line graph best practices

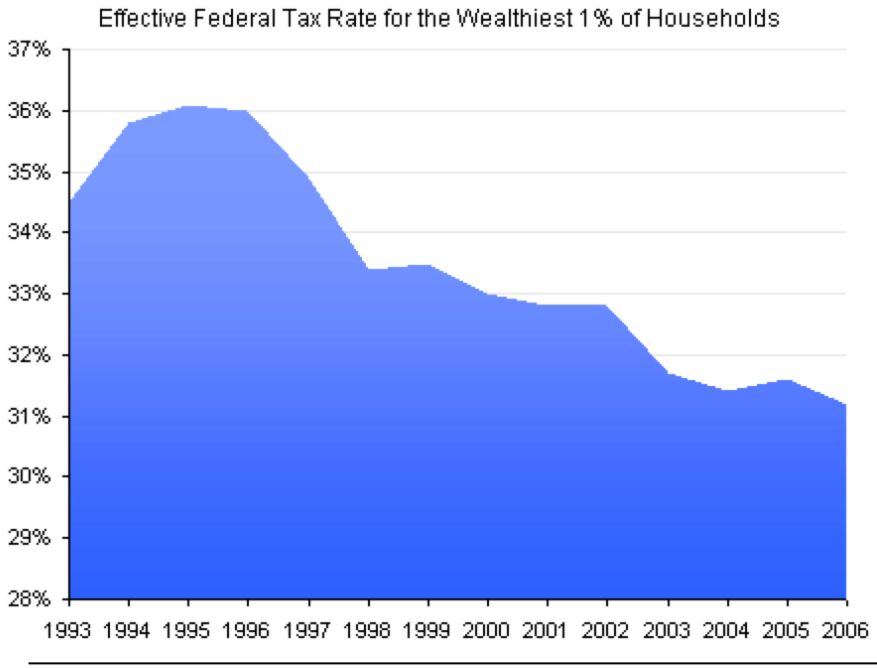
Make sure your axes are always labeled at consistent intervals

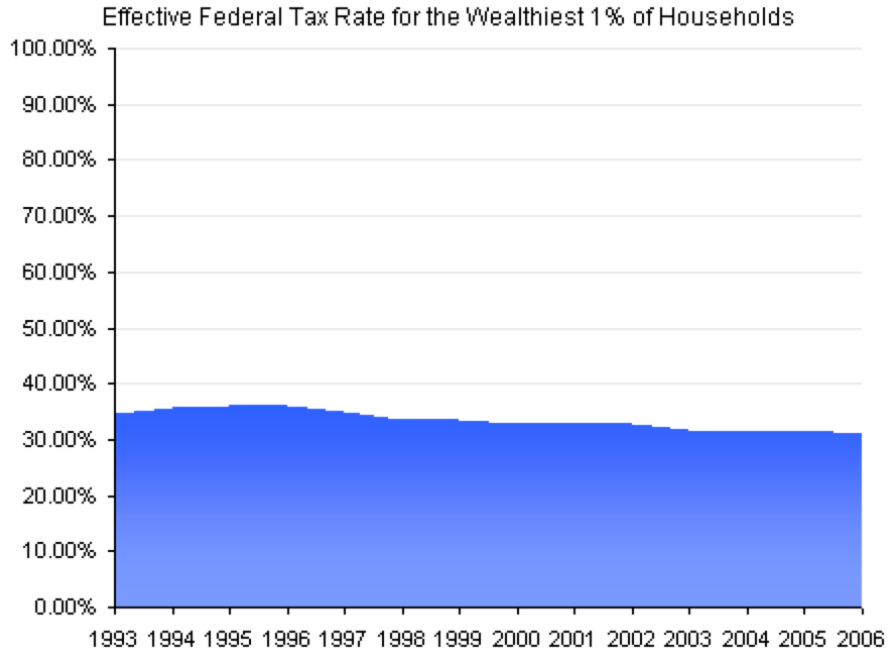


Source: http://www.perceptualedge.com/articles/visual_business_intelligence/line_graphs_and_irregular_intervals.pdf

Line graph best practices

Scale makes a big difference!





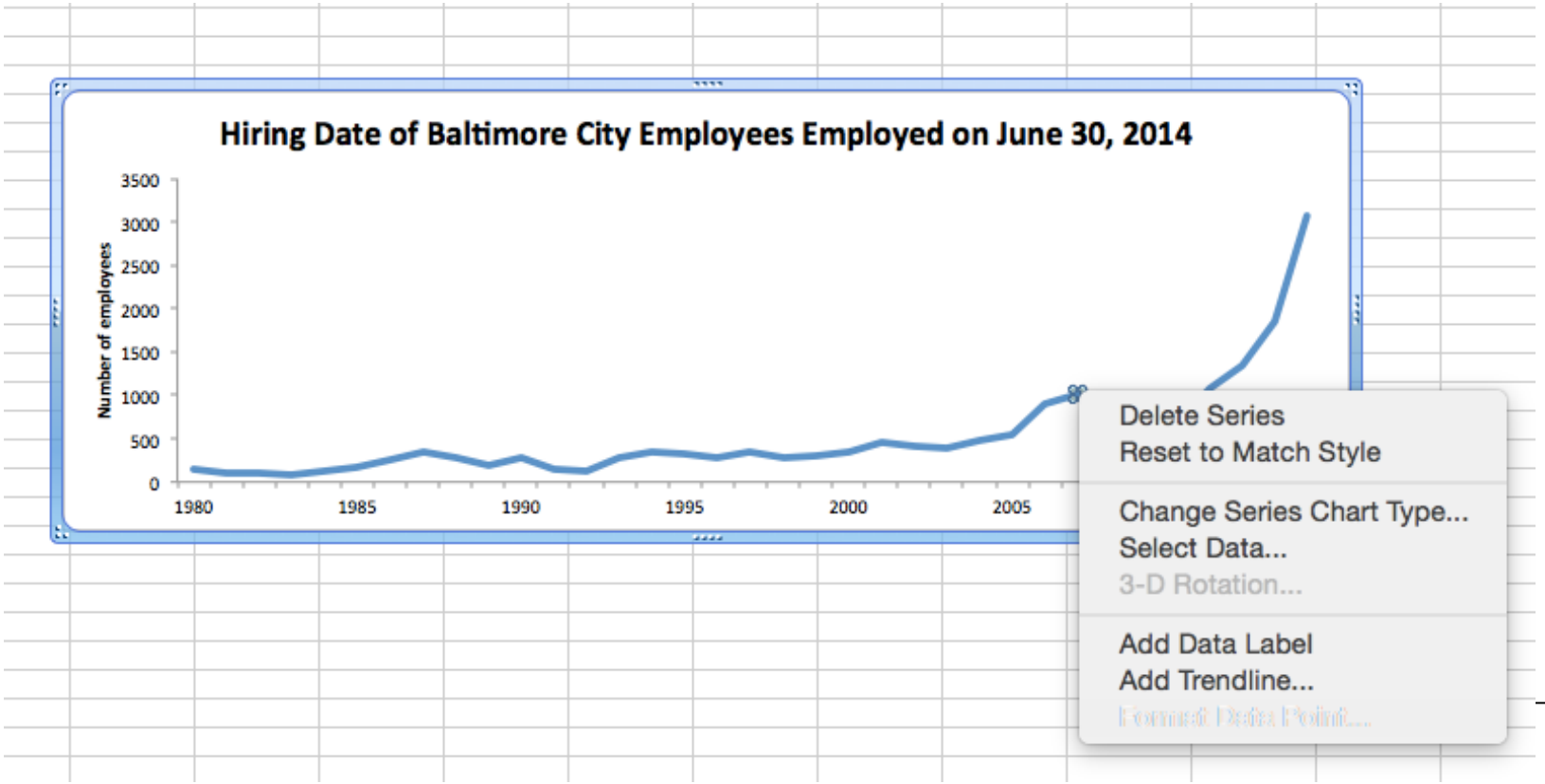
Source: http://peltiertech.com/tax-the-rich-or-deceptive-axis-scales/

Line charts - You do.

You work for the City of Baltimore's Human Resource Department and you are tasked with understanding where all of the city's money goes.

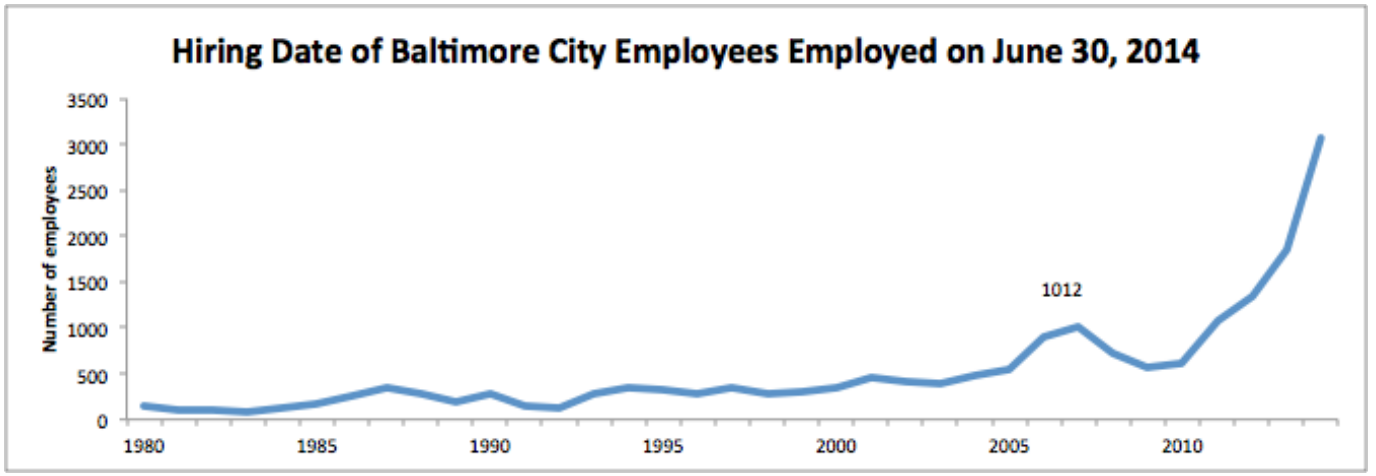
In a sheet of data labeled 'line chart,' explore:
1. When were current employees hired, starting in 2008?

Add a data label to highlight key points.



A line chart shows trends over time.

- Most employees that are employed in 2014 were hired after 2010.
- Employees who were hired in 2007 (1012) are more likely than 2008, 2009, and 2010 to still be employed.

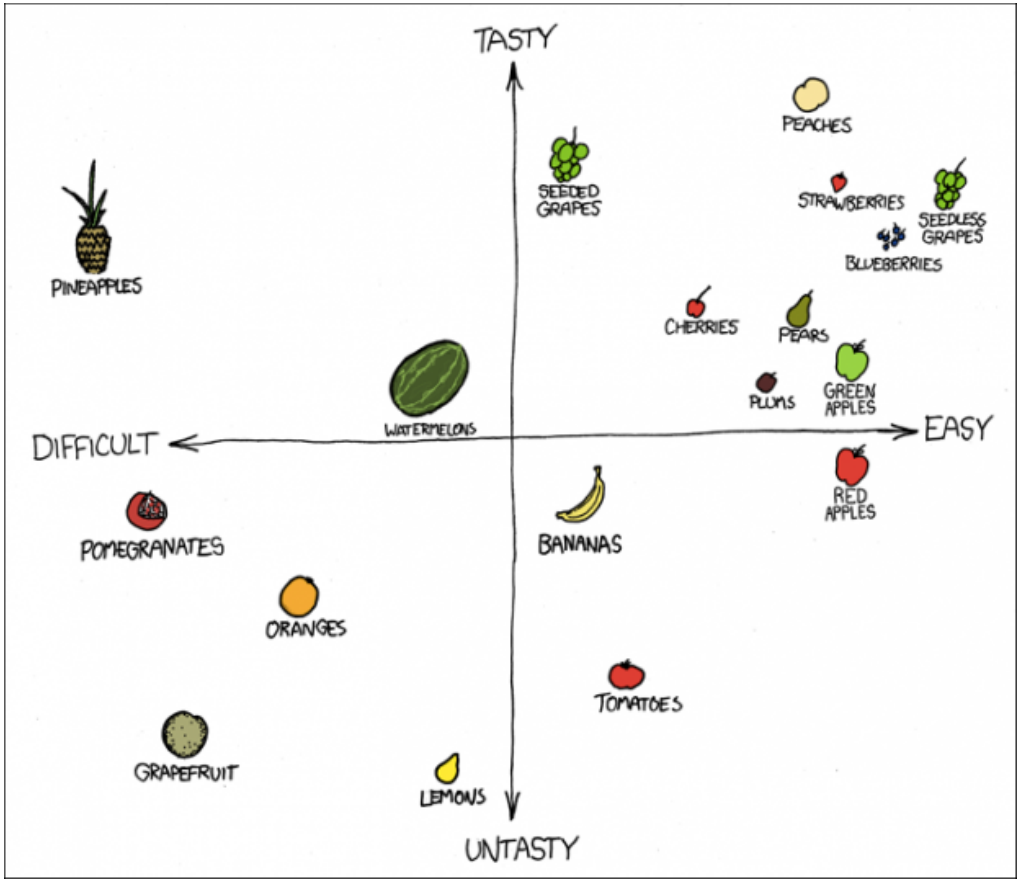


Scatterplots

What is a scatter plot?

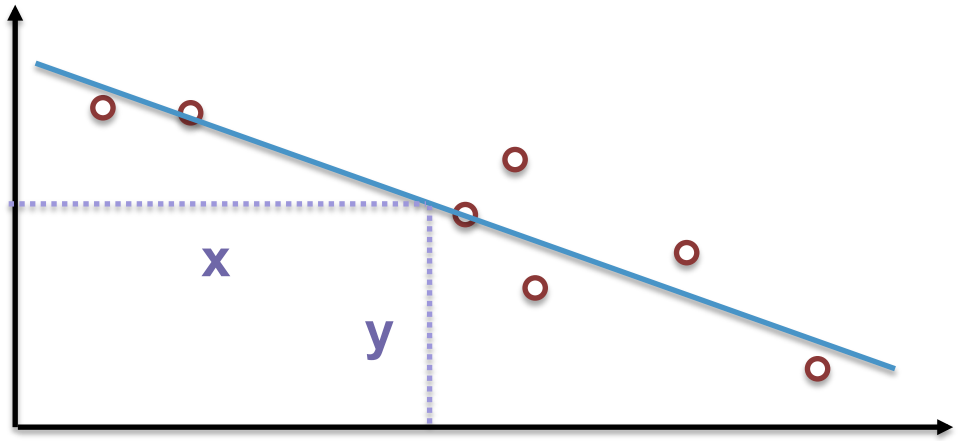
A scatter plot compares a set of numeric, continuous values.

Useful tool to identify outlier points



Interpreting a trend line

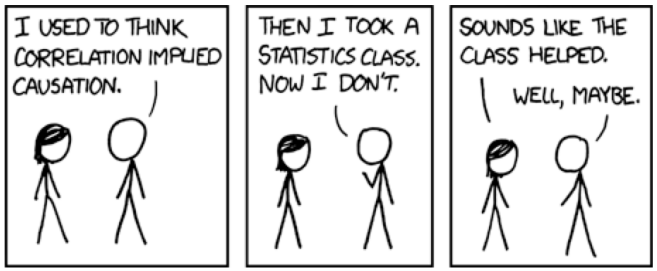
- A trend line is a visual representation of correlation between the two variables
- Interpretation: for every change of the X-axis variable by **x** units, we expect the Y-axis variable to change by **y** units

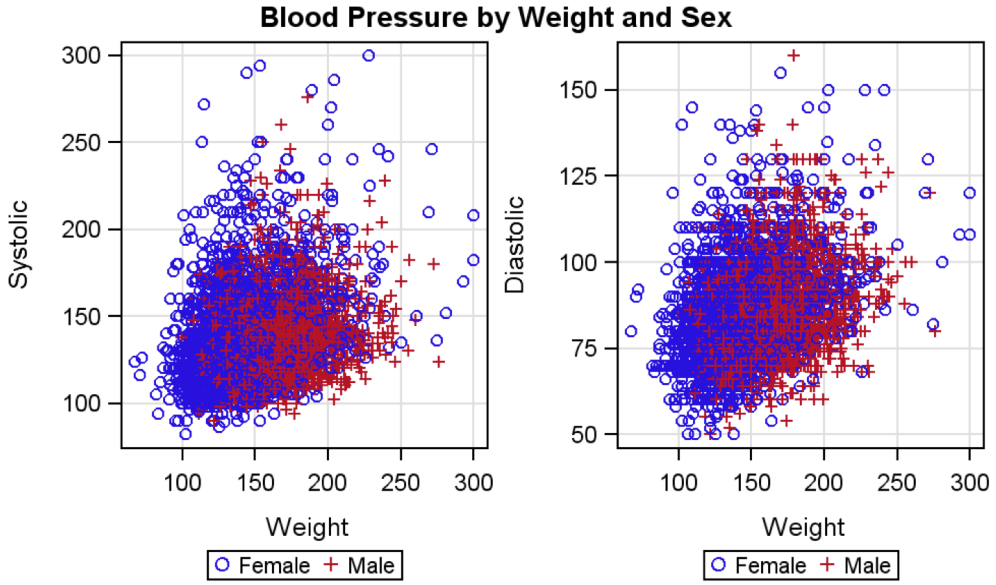


Best practices in scatter plots

Words of caution

- Limit the data size: too many data points will look like a blob
- If mapping colors, limit to 6 categories at a time
- Correlation does not imply causation!





Source: http://blogs.sas.com/content/graphicallyspeaking/2013/02/12/gtl-layouts/

Scatter Plots - Let’s do together.

You work for the City of Baltimore's Human Resource Department and you are tasked with understanding where all of the city's money goes.

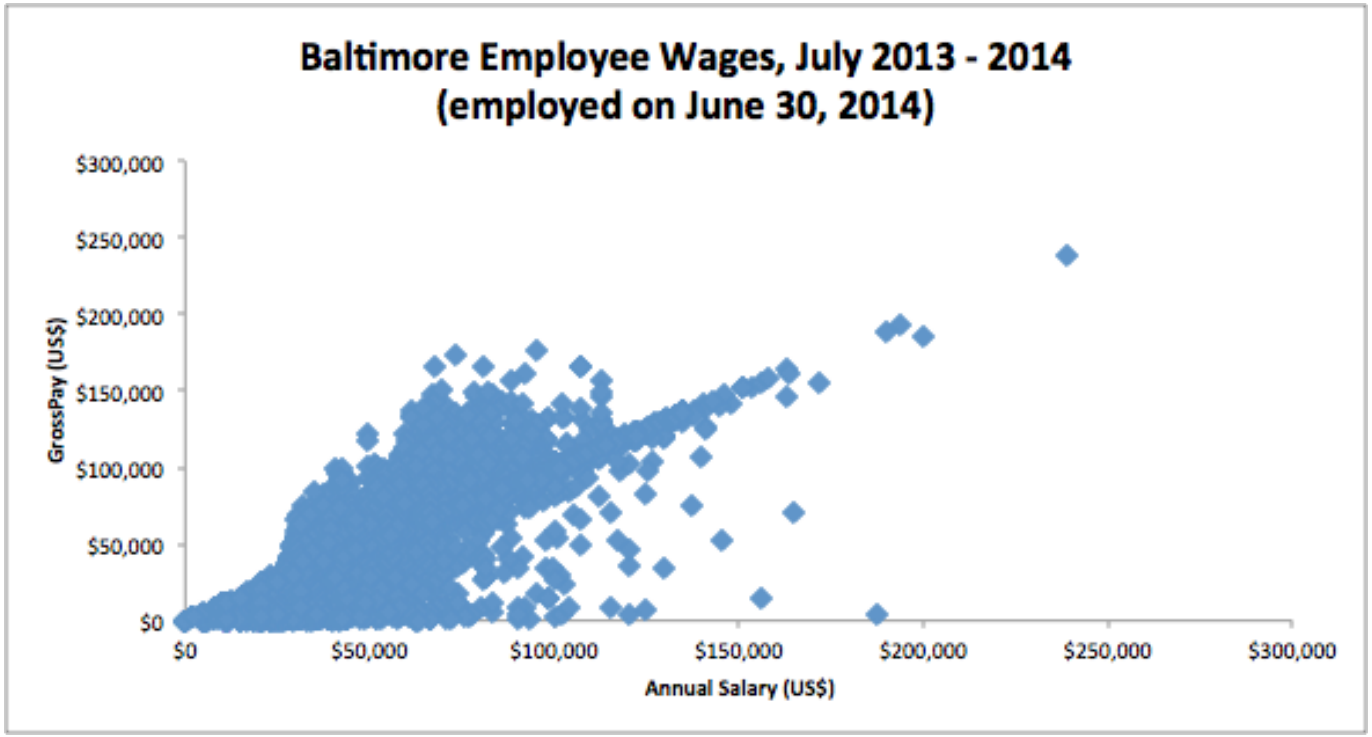
In a sheet of data labeled 'scatterplot,' explore:

1. Is there a relationship between gross pay and annual salary?
2. Does the Mayor's Office have more or less withholdings than others?

Source: http://peltiertech.com/tax-the-rich-or-deceptive-axis-scales/

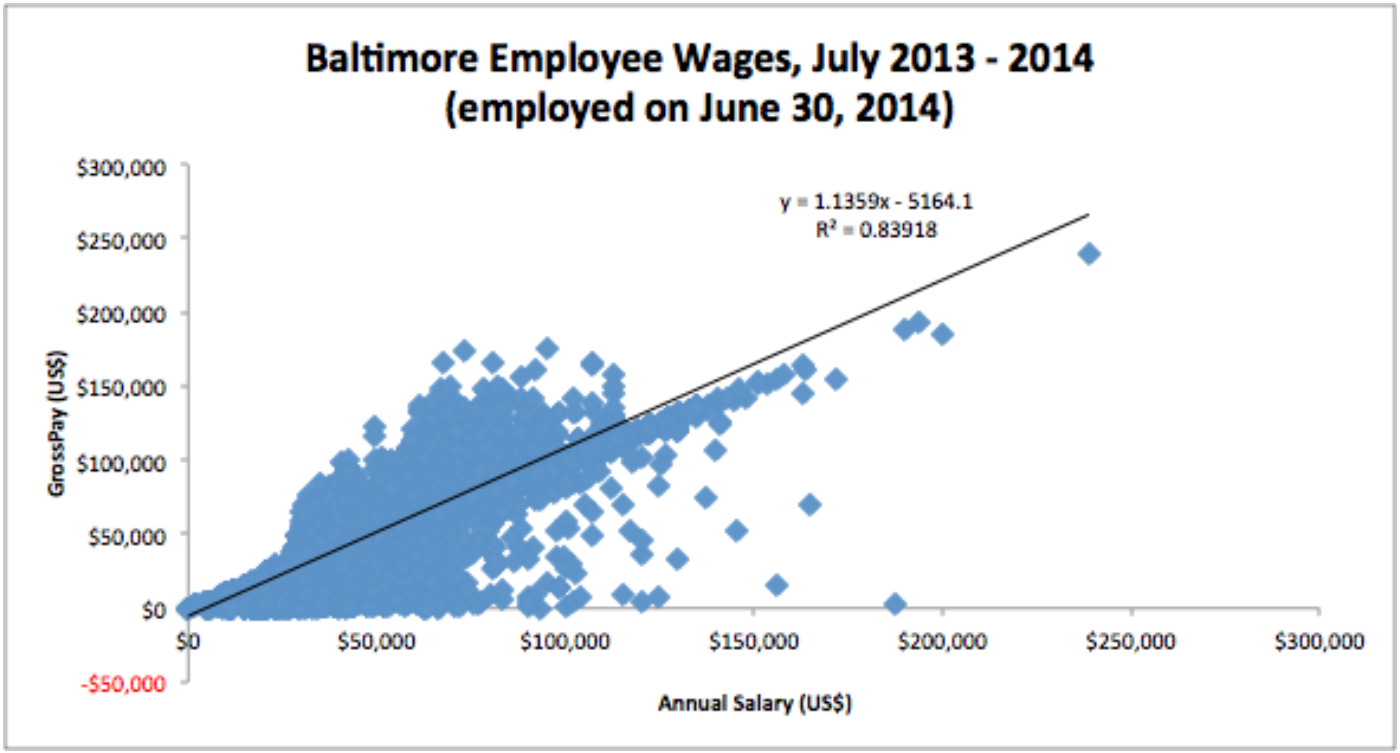
Scatter plot: showing a trend

How does the salary compare to gross pay of employees?



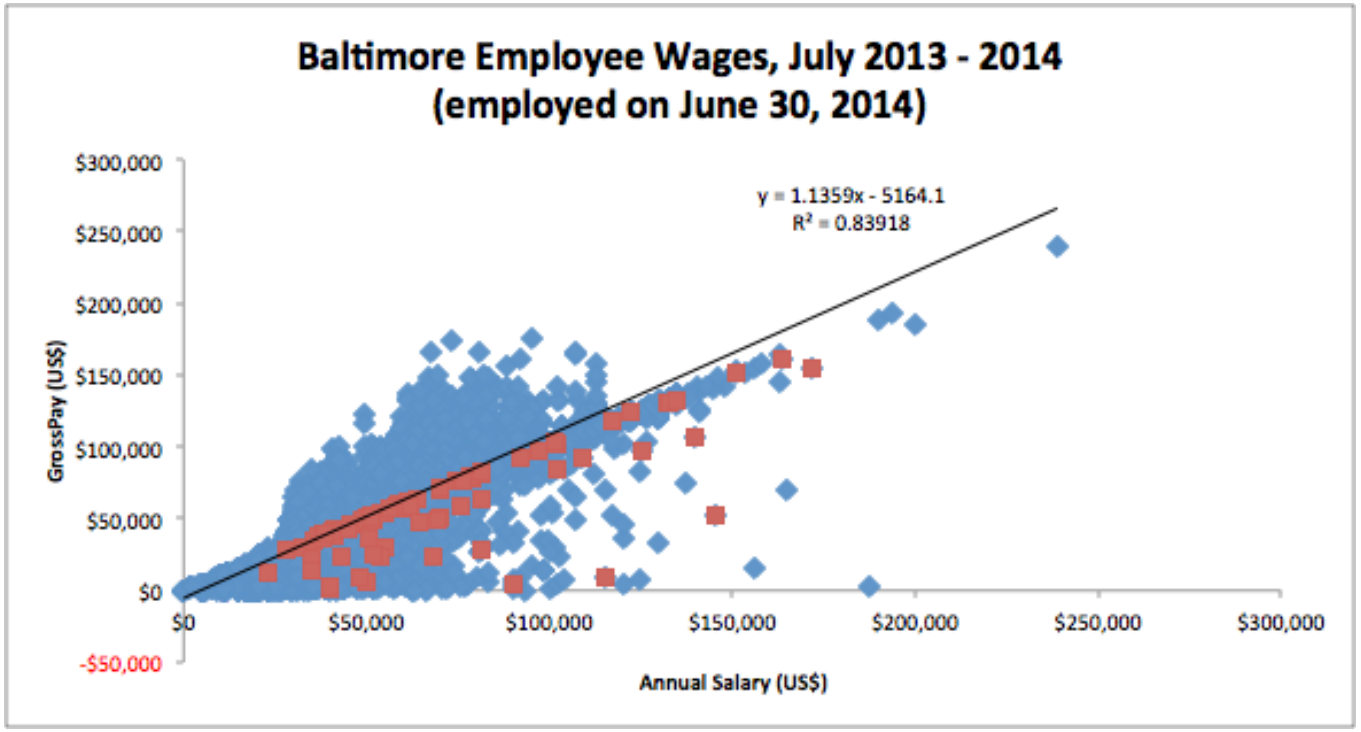
Scatter plot: showing a trend

Add a trend line, equation, & R2



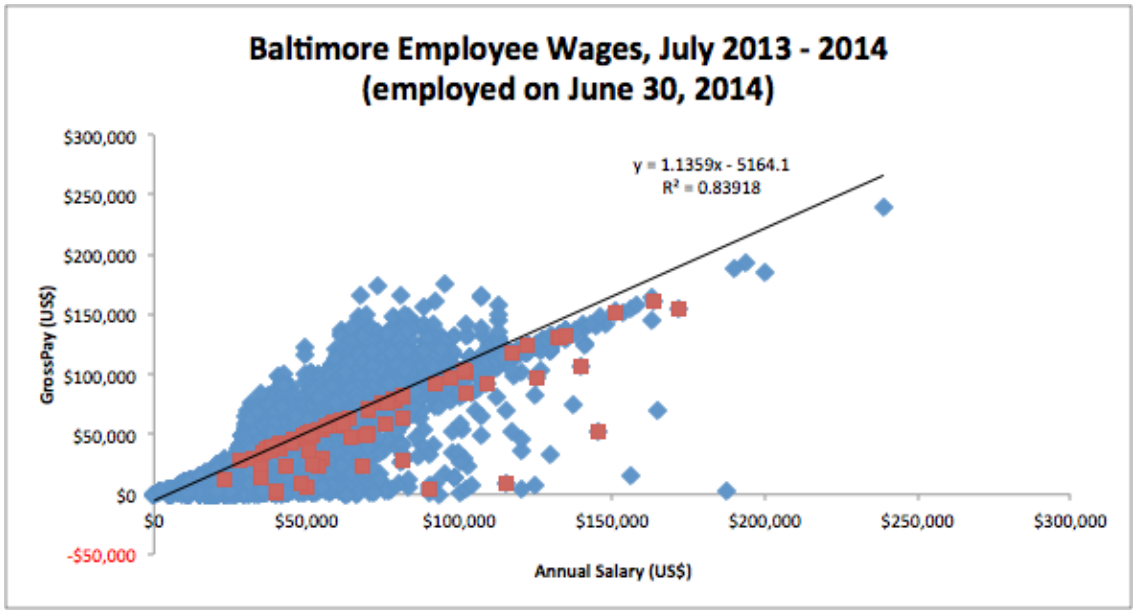
Scatter plot: showing a trend

Add a second series



Scatter plot: showing a trend

- As annual salary increases, gross pay increases
- For every unit change in annual salary, gross pay increases by 1.13
- The relationship is strong (R2 =.89)
- The Mayor's Office follows the same trend as all employees of the City



Lastly, let’s put the story together…

How do we ensure that the audience is able to retain, recall and retell our data-driven stories?

Telling stories with data is both a science and an art.

‘Storytelling with data’ is cross disciplinary:

- Visual designer
- Coder
- Statistician
- Storyteller

Let’s return to the question at hand

You work for the City of Baltimore’s Human Resource Department and you are tasked with understanding where all of the city’s money goes. Your boss has left the question open ended. How will you answer this question?

Know thy audience

Before you start: know thy audience

- Define your audience - client, manager, general public?
- Define your audience's needs
  - Why does your audience care about this?
  - What motivates your audience?
- Decide on the appropriate mode of communication

Before you start: example continued

- **Audience:** Senior management
- **Why do they care?** They want to know how the city is spending money to make budgetary decisions
- **What motivates them?**
  - Saving money
  - Not wasting money
- **Best medium:** presentation and/or report

Structure an argument

A data narrative is just like that argumentative essay you had to write in high school

ARGUMENTATIVE - ONE side only		
INTRO		• *general statement / hook* • *elaboration > > scope (can include a definition)* • *thesis statement clearly stating the position (one side) of the author*
BODY	ARGUMENT 1 FOR *OR* AGAINST	*topic sentence* *+ support*
	ARGUMENT 2 FOR *OR* AGAINST	*topic sentence* *+ support*
	ARGUMENT 3 FOR *OR* AGAINST	*topic sentence* *+ support*
CONCLUSION *summary of position & ideas / link to action*		• *restate thesis statement and opinion* • *summarise ideas* • *closing comments/final thoughts*

wiki-witch

INTRO	• ***general statement / hook*** • ***elaboration > > scope*** *(can include a definition)* • ***thesis statement*** *clearly stating the position (one side) of the author*

Creating the introduction: provide context

- **Hook**: Tell your audience why this is this the most important thing they should care about right now
- **Scope:** Summarize what you've been asked to do, and why
- **Thesis:** Summarize what you're going to argue
- **Background:** Describe the context of the problem if you think that the audience may not be familiar

Creating the introduction: provide context

- **Hook**: In a time of financial constraint, the City of Baltimore spends $754 million on employee wages, more than any other public expenditure.

- **Scope:** I will explore the wages of employees of Baltimore in FY2014 to determine
  - who is being paid the most
  - how much are most employees paid
  - how has this changed over time, and
  - how gross wages are related to annual salaries.

- **Thesis:** There are trends in Baltimore City employee wages that need to be reviewed to determine if the City can save money.

- **Background:** Baltimore is in a financial crisis.

B O D Y	**ARGUMENT 1** **FOR *OR*** **AGAINST**	***topic sentence*** *+ support*
	ARGUMENT 2 **FOR *OR*** **AGAINST**	***topic sentence*** *+ support*
	ARGUMENT 3 **FOR *OR*** **AGAINST**	***topic sentence*** *+ support*

Body of presentation: build up your argument

1. State each supporting argument
2. Provide evidence from the data

Body of presentation: best practices

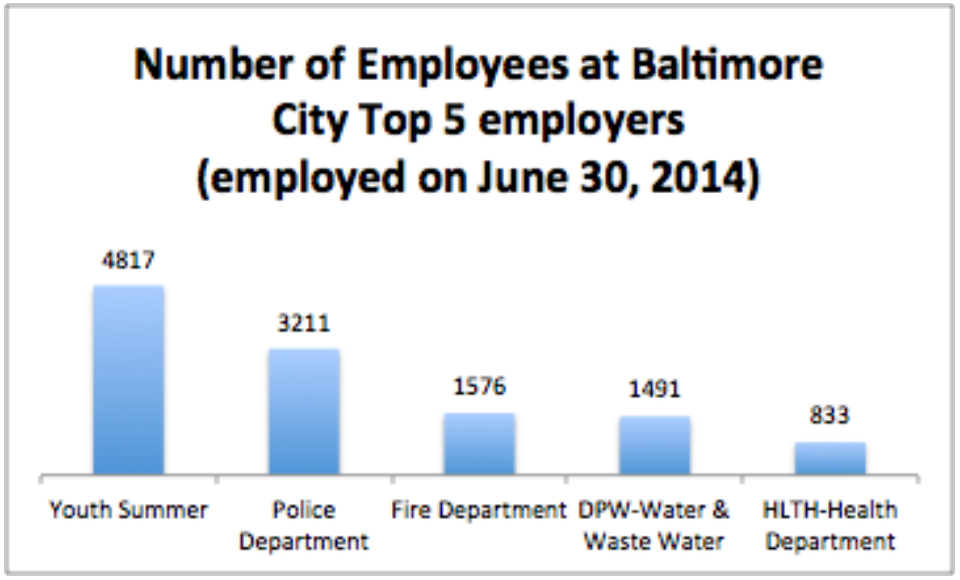
1. Cut out distractions
    - Everything you show should build up to or support your argument
    - You can have more than one point/slide/graph to support each argument

2. Make headers informative
    - Tell the audience what they should focus on, not just what they're looking at

3. Avoid clutter
    - Many concise slides is better than one busy slide

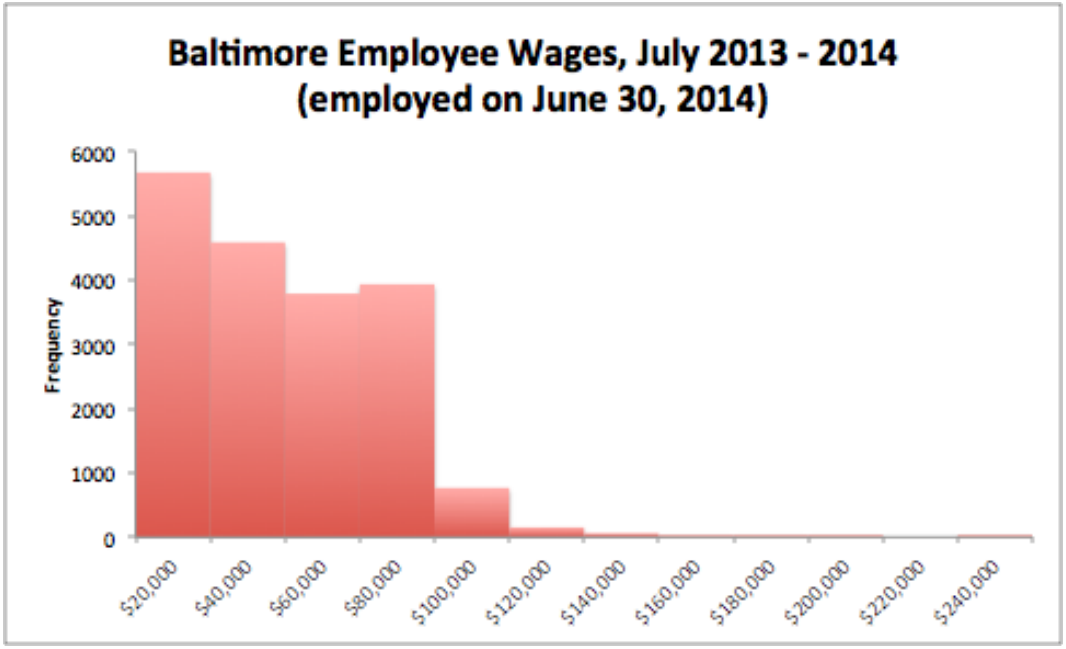
Youth Summer employs the most, with nearly 5,000 employees.

- The Police Department employs the next most,
- The Police Department employs almost double the third most employer, the Fire Department.



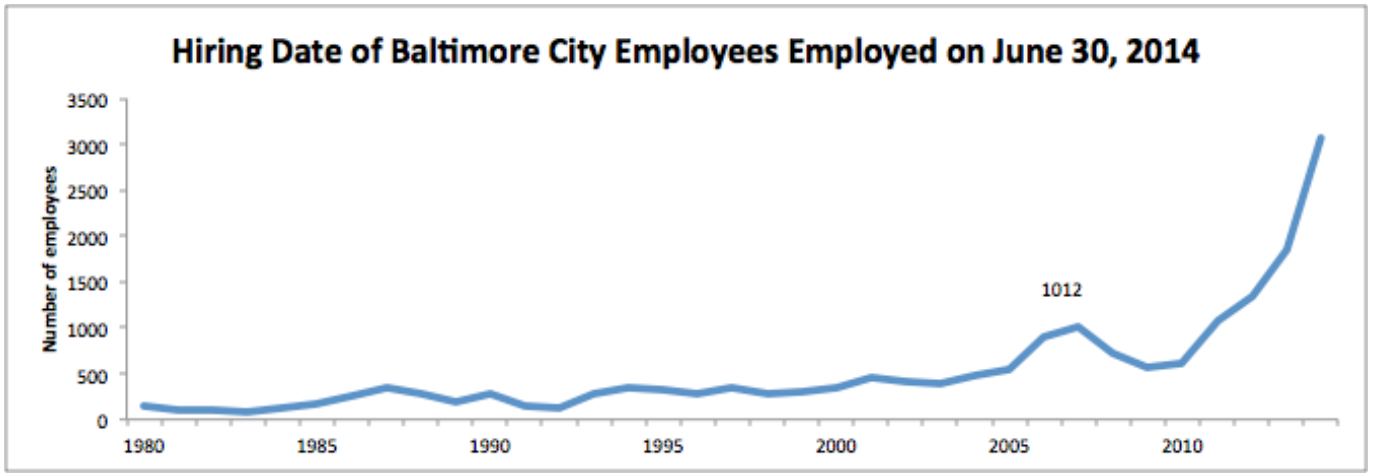
Most commonly, Baltimore City employees earn less than $20,000.

- Most employees in Baltimore earn between $0 and $80,000
- The highest paid Baltimore employee earns over $238,000.



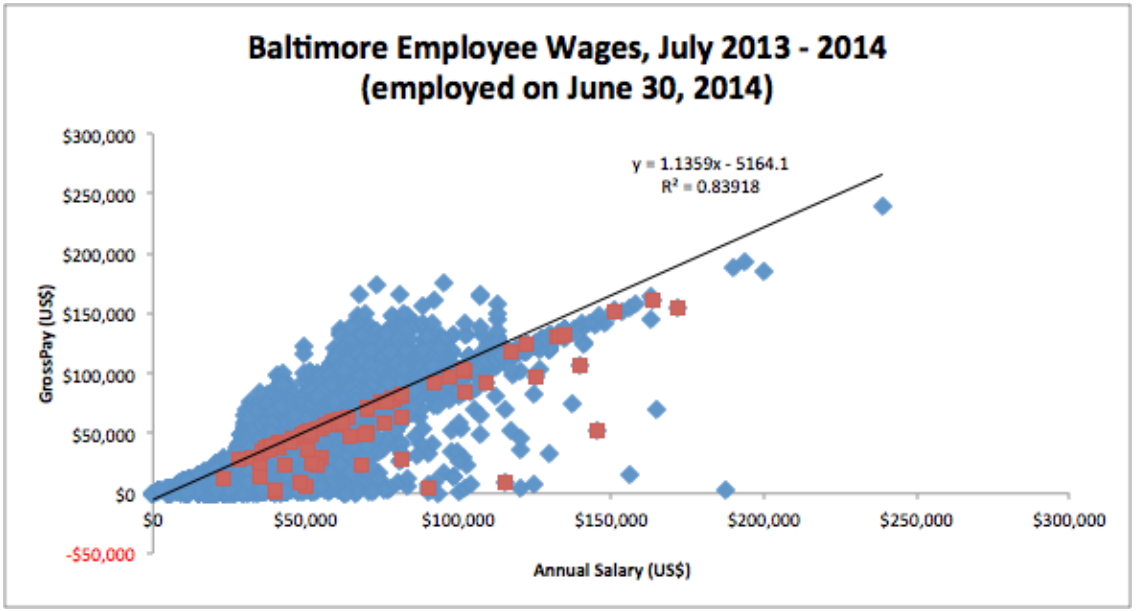
Most employees that are employed in 2014 were hired after 2010.

- Employees who were hired in 2007 are more likely than 2008, 2009, and 2010 to still be employed.



As annual salary increases, gross pay increases.

- For every unit change in annual salary, gross pay increases by 1.13
- The relationship is strong (R2 =.89)
- The Mayor's Office follows the same trend as all employees of the City



CONCLUSION *summary of position & ideas / link to action*	• ***restate thesis statement and opinion*** • ***summarise ideas*** • ***closing comments/final thoughts***

Conclusion

1. Restate your thesis/conclusion
2. Summarize how the evidence in the body of your narrative supported it
3. Mention caveats to your calculations and/or possible next steps

Example: Conclusion

Our analysis revealed that in Baltimore City:

- Youth Summer employs the most, with nearly 5,000 employees
- Most commonly, Baltimore City employees earn less than $20,000
- Most employees that are employed in 2014 were hired after 2010
- As annual salary increases, gross pay increases

Example: Next steps

Caveats:

- Wages are for all employees, including temporary or seasonal workers.

Next steps:

- Limit the data to explore wage characteristics for full-time employees
- Dive deeper to understand why employees hired in 2007 are still employed
- Explore whether other agencies have different wage characteristics

Final thoughts: Make it look good

1. Check for spelling/punctuation errors
2. Make sure all text is large enough to read
3. Format numbers for readability
4. Label all charts and tables (don’t forget to indicate date ranges where appropriate!)
5. Make it visually appealing

CLOSING DISCUSSION

WHY VISUALIZATION?
EXPLORE MULTIPLE VISUALIZATIONS & FEATURES
TELL A STORY WITH DATA

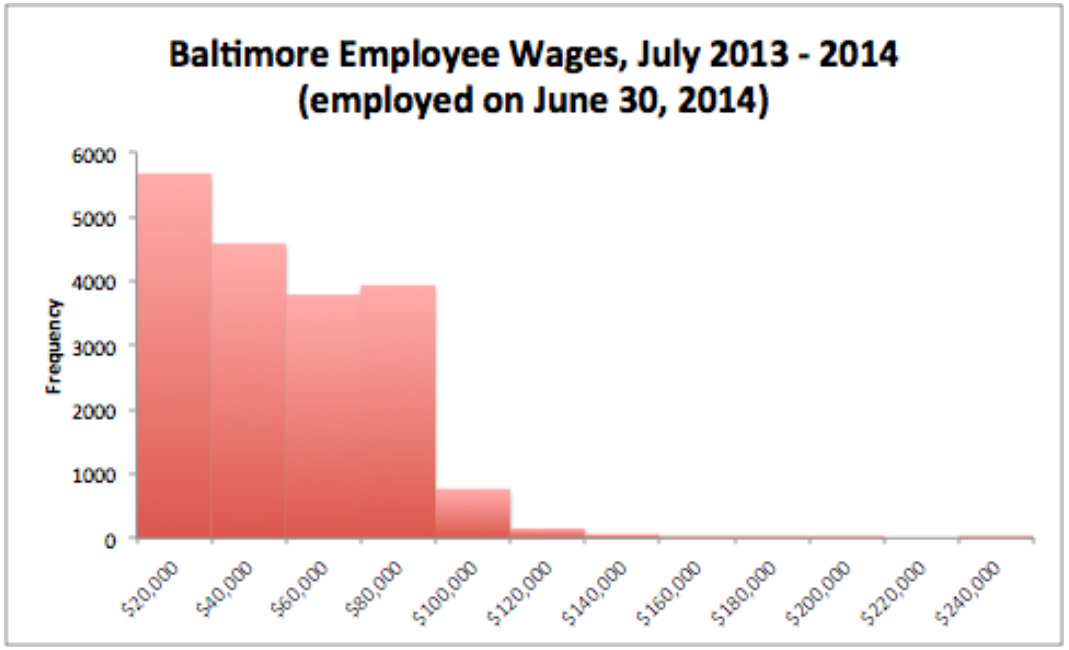
GOOD BYE FOR NOW!

- Carey Anne Nadeau, Founder & CEO of Open Data Nation
- info@opendatanation.com
- generalassemb.ly
- facebook.com/gnrlassembly
- @ga

Extra Material
How to create a histogram using the frequency formula

Histogram in three steps

1. Assign bins
2. Use FREQUENCY formula to create a frequency table
3. Modify bar chart to create a histogram



Use the FREQUENCY function

- In the column adjacent to the bins,
  highlight range of cells and enter frequency formula

=Frequency(DATA_ARRAY, BIN_ARRAY)

= Create a frequency table (look for the data here, look for the bins here)

- Don’t press enter. At the same time, press Control (or Command) + Shift + Enter

Clustered Column from a Clustered Column Chart

- Click Insert > Chart> Clustered Column Chart
- Select the data series
- Press COMMAND +1 to bring up ‘Format Data Series’ dialogue box
- OR right click and select ‘Format Data Series’
- Reduce the gap width to 0%