Collecting and cleaning data

Module 1

DATA
DATA
COLLECTION

Collecting Data

- The most critical issue in any data science or modeling project is finding the right data set.
- Who might actually have the data I need?
- Why might they decide to make it available to me?
- How can I get my hands on it?

Hunting

- Who has the data, and how can you get it?

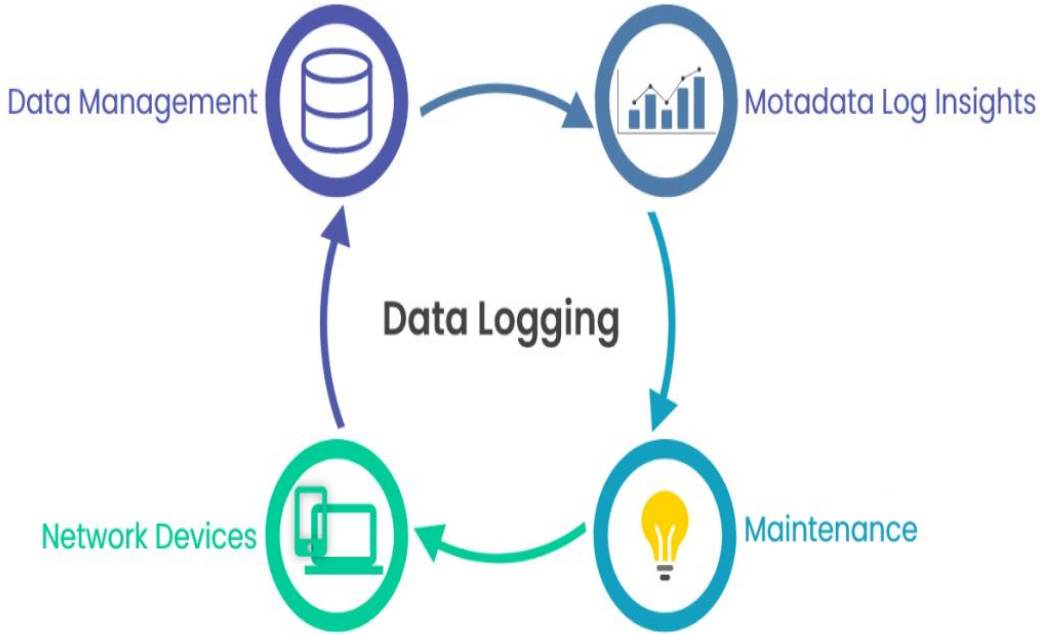
1. Companies and Proprietary Data Sources
   - x Business issues, and the fear of helping their competition.
   - x Privacy issues, and the fear of offending their customers.
2. Government Data Sources
   - ✔ Without compromising the national interest or violating privacy
   - ✔ RTI – Right to Information
4. Academic Data Sets
5. Sweat Equity

Scraping

- There are two distinct steps to make this happen, spidering and scraping:
- ***Spidering*** is the process of downloading the right set of pages for analysis.
  - ***web crawling***
- ***Scraping*** is the fine art of stripping this content from each page to prepare it for computational analysis.

Logging

- Logging is the practice of recording information about a program's execution in a systematic and organized way.
- The important considerations in designing any logging system are:
  - Build it to endure with limited maintenance. Set it and forget it, by provisioning it with enough storage for unlimited expansion, and a backup.
  - Store all fields of possible value, without going crazy.
  - Use a human-readable format or transactions database, so you can understand exactly what is in there when the time comes, months or years later, to sit down and analyze your data.



https://www.motadata.com/blog/what-is-data-logging/

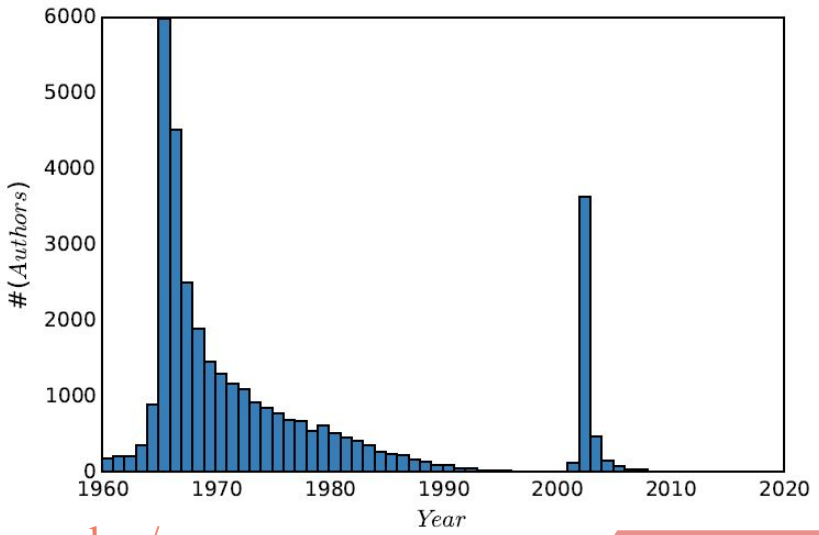
DATA CLEANING

Cleaning Data

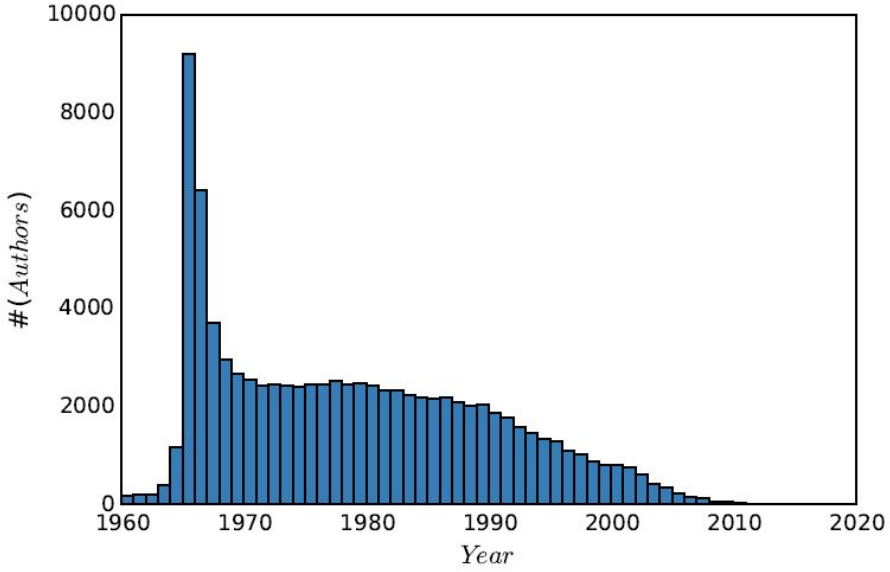
1. Errors vs. Artifacts
   1. Data **errors** represent information that is fundamentally lost in acquisition
   2. **Artifacts** are generally systematic problems arising from processing done to the raw information it was constructed from.

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q



https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/



- The cleaned data removes these artifacts, and the resulting distribution looks correct.

Data Compatibility

- Data compatibility refers to the ability of data sources to communicate and exchange information with each other and with a data management framework

Unit Conversions
Numerical Representation Conversions

- Quantifying observations in physical systems requires standard units of measurement.
- Step 1. Identify the unit you have. These are the Starting Units.
- Step 2. Identify the unit you want. These are the Desired Units.
- Step 3. Identify appropriate unit conversion factor(s). These are the Linking (or Ratio) Unit(s). Use EXACT conversion factors whenever available.
- Step 4. Cancel units and perform the math calculations (e.g., multiply, divide). Repeat the calculation (double check).
- Step 5. Evaluate the result. Does the answer make sense?

Name Unification

- Integrating records from two distinct data sets requires them to share a common key field

Time/Date Unification

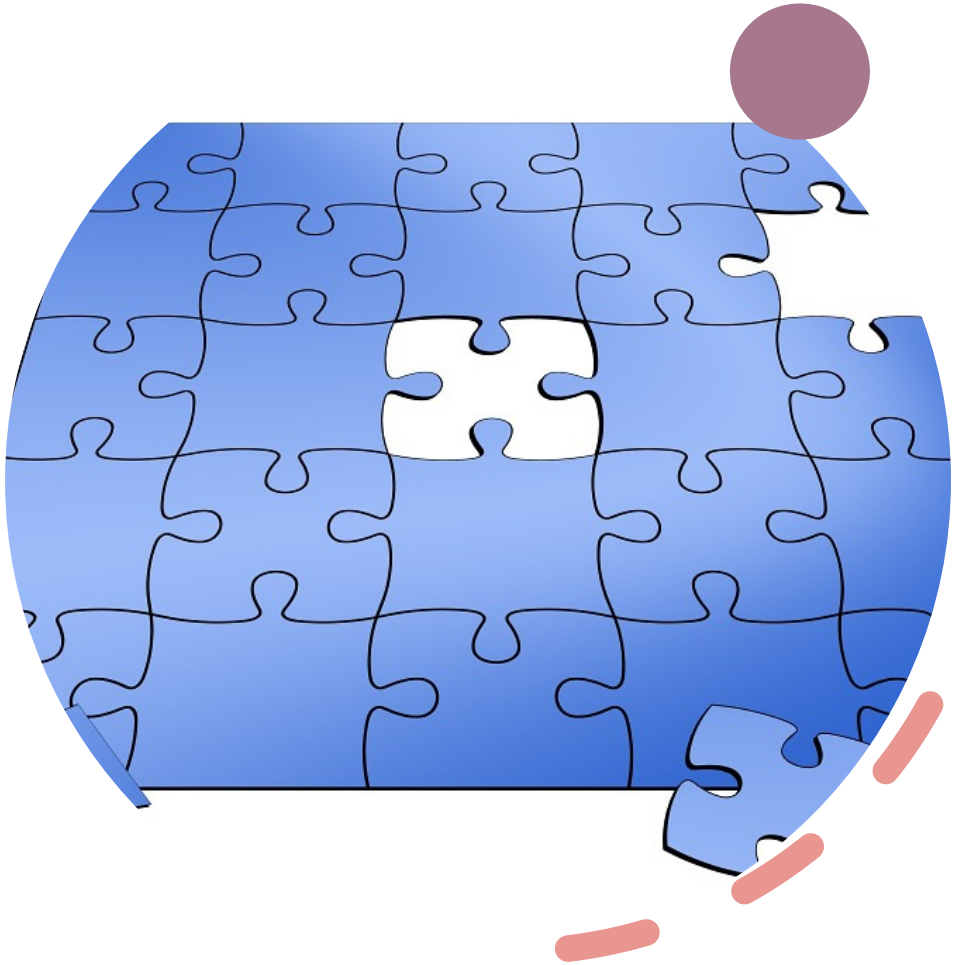
- Data/time stamps are used to infer the relative order of events, and group events by relative simultaneity.
- Integrating event data from multiple sources requires careful cleaning to ensure meaningful results.

Financial Unification

- Money makes the world go round, which is why so many data science projects revolve around financial time series.
- One issue here is currency conversion, representing international prices using a standardized financial unit.
- Currency exchange rates can vary by a few percent within a given day, so certain applications require time-sensitive conversions.
- Conversion rates are not truly standardized.
- the most meaningful way to represent price changes over time is probably not differences but returns, which normalize the difference by the initial price

Dealing with Missing Values

- Not all data sets are complete.
- An important aspect of data cleaning is identifying fields for which data isn't there, and then properly compensating for them:
- What is the year of death of a living person?
- What should you do with a survey question left blank, or filled with an obviously outlandish value?
- What is the relative frequency of events too rare to see in a limited-size sample?



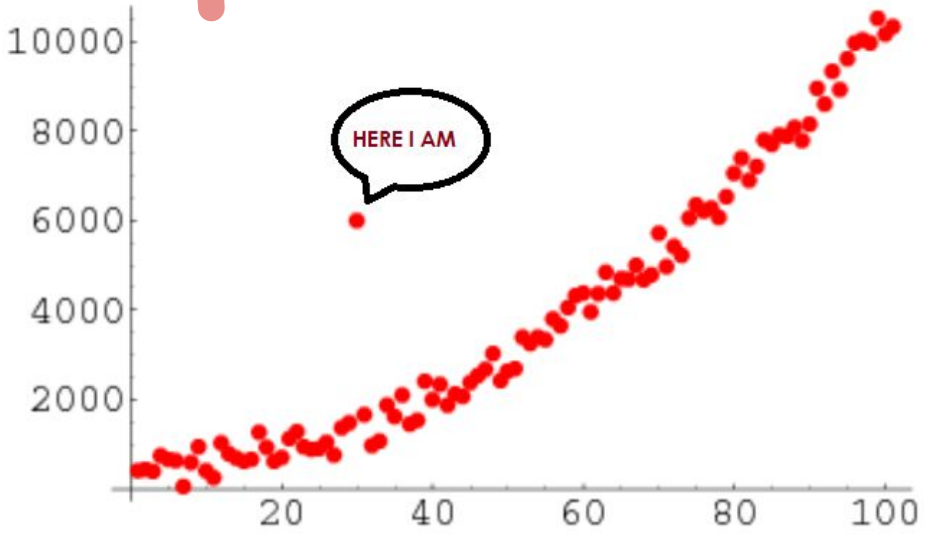
Dealing with Missing Values - So how should we deal with missing values?

- The simplest approach is to drop all records containing missing values.
- Estimate or impute missing values
  - Heuristic-based imputation
  - Mean value imputation
  - Random value imputation
  - Imputation by nearest neighbor
  - Imputation by interpolation

Take-Home Lesson: Separately maintain both the raw data and its cleaned version. The raw data is the ground truth, and must be preserved intact for future analysis. The cleaned data may be improved using imputation to fill in missing values. But keep raw data distinct from cleaned, so we can investigate different approaches to guessing.

Outlier Detection

- Mistakes in data collection can easily produce outliers that can interfere with proper analysis
- Outlier elements are often created by data entry mistakes



WHAT IS DATA VISUALIZATION?

Visualizing Data

- Effective data visualization is an important aspect of data science, for at least three distinct reasons:

1. Exploratory data analysis
2. Error detection
3. Communication

Exploratory Data Analysis

www.educba.com

Exploratory Data Analysis

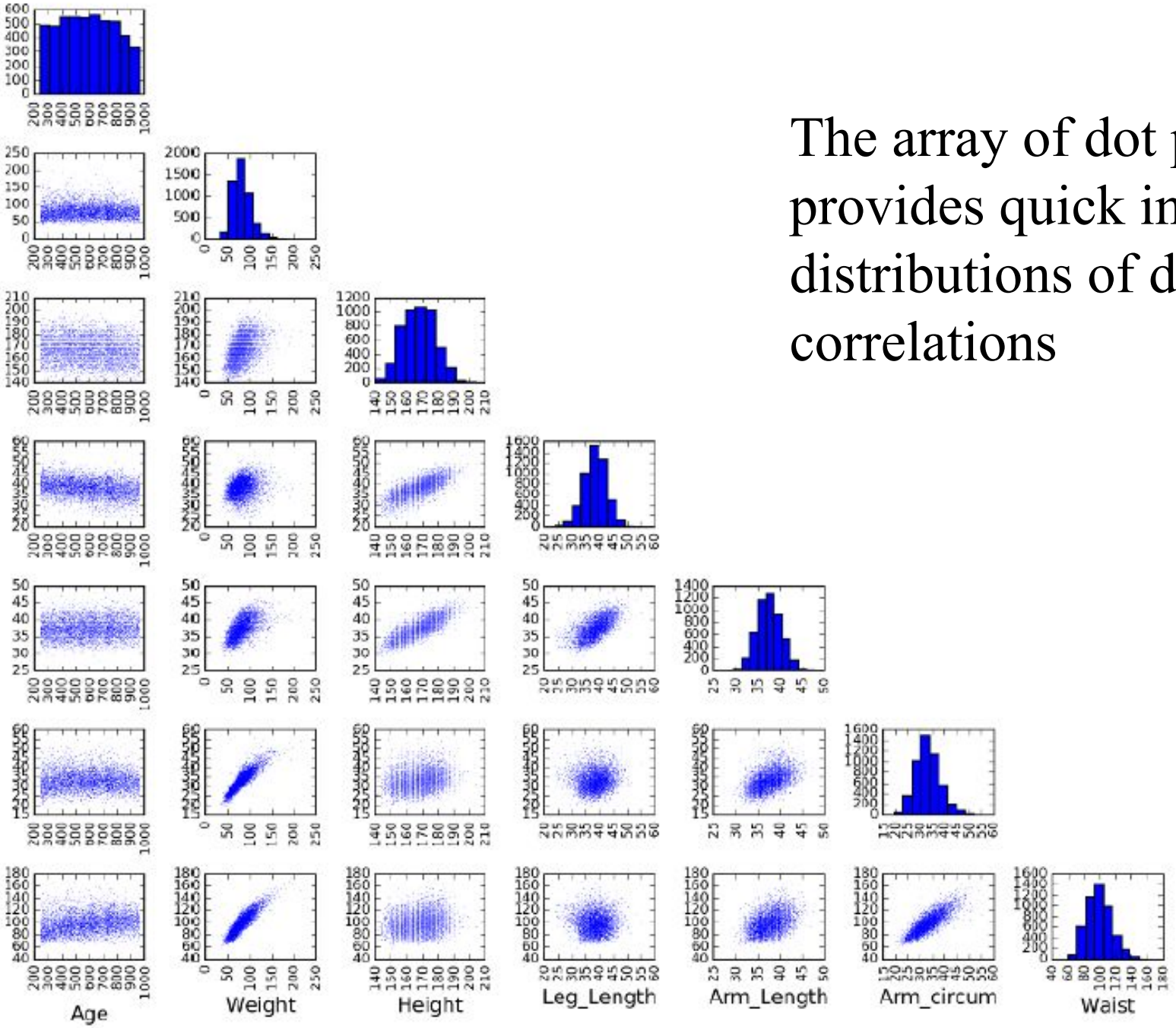
- Exploratory data analysis is the search for patterns and trends in a given data set
- Visualization techniques play an important part in this quest

Confronting a New Data Set

	Min	25%	Median	75%	Max
Age	241	418	584	748	959
Weight	32.4	67.2	78.8	92.6	218.2
Height	140	160	167	175	204
Leg Length	23.7	35.7	38.4	41	55.5
Arm Length	29.5	35.5	37.4	39.4	47.7
Arm Circumference	19.5	29.7	32.8	36.1	141.1
Waist	59.1	87.5	97.95	108.3	172

- Answer the basic questions
  - Who constructed this data set, when, and why‘
  - How big is it?
  - What do the fields mean?
- Look for familiar or interpretable records
- Summary statistics
- Pairwise correlations
- Class breakdowns
- Plots of distributions:

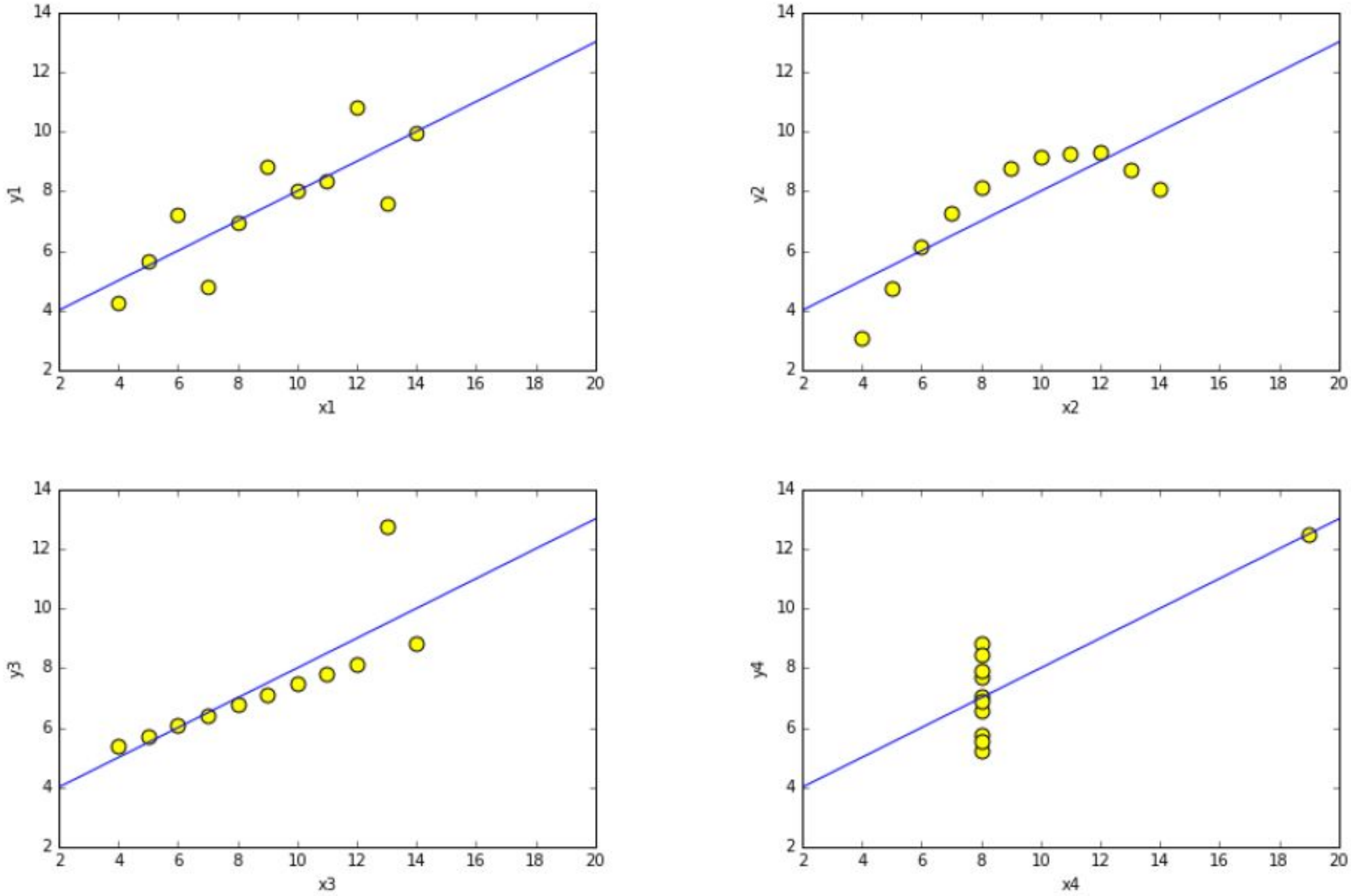
	Age	Weight	Height	Leg Length	Arm Length	Arm Circum	Waist
Age	1.000						
Weight	0.017	1.000					
Height	−0.105	0.443	1.000				
Leg_Len	−0.268	0.238	0.745	1.000			
Arm_Len	0.053	0.583	0.801	0.614	1.000		
Arm_Circ	0.007	0.890	0.226	0.088	0.444	1.000	
Waist	0.227	0.892	0.181	−0.029	0.402	0.820	1.000



The array of dot plots of variable pairs provides quick insight into the distributions of data values and their correlations

Summary Statistics and Anscombe's Quartet

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.31	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
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	

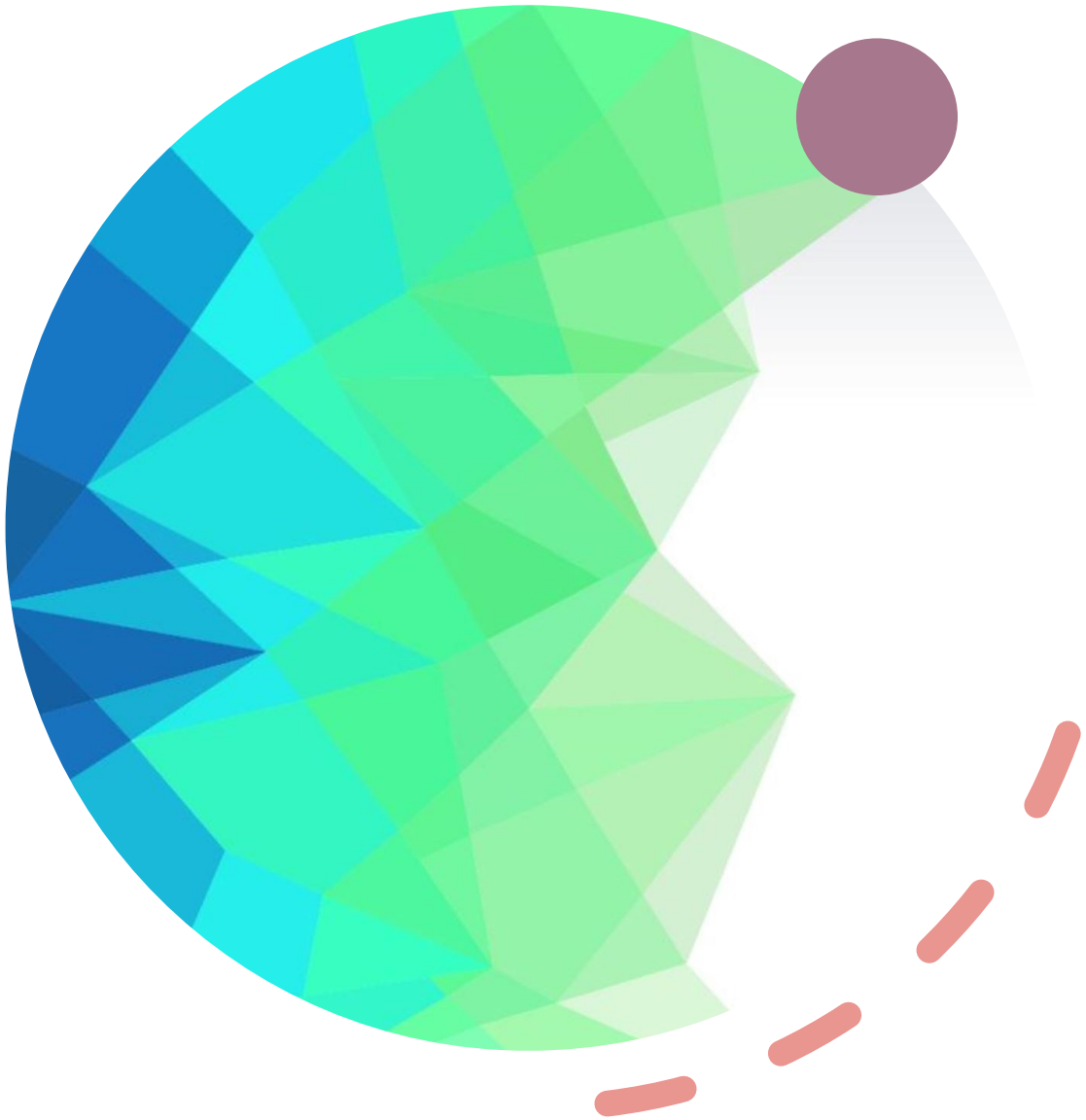

y1
x1
y2
x2
y3
x3
y4
x4

Visualization Tools

- Exploratory data analysis
  - Spreadsheet programs like Excel
  - Notebook-based programming environments like iPython, R, and Mathematica
- Publication/presentation quality charts
  - Plotting libraries like MatPlotLib or Gnuplot
  - Excel
  - R has a very extensive library of data visualizations
- Interactive visualization for external applications
  - Dashboards
  - Python
  - Tableau

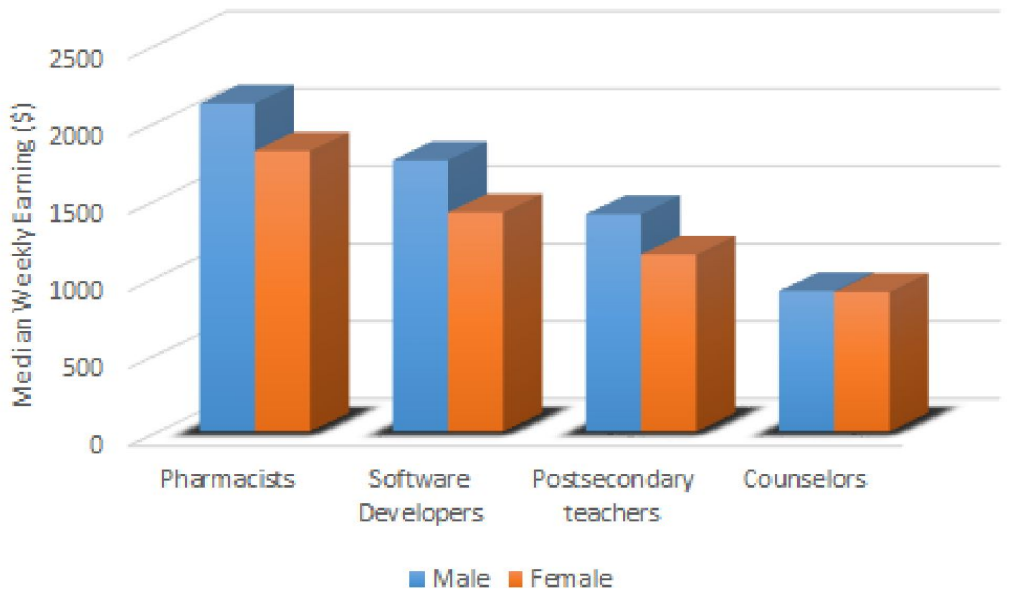
Developing a Visualization Aesthetic

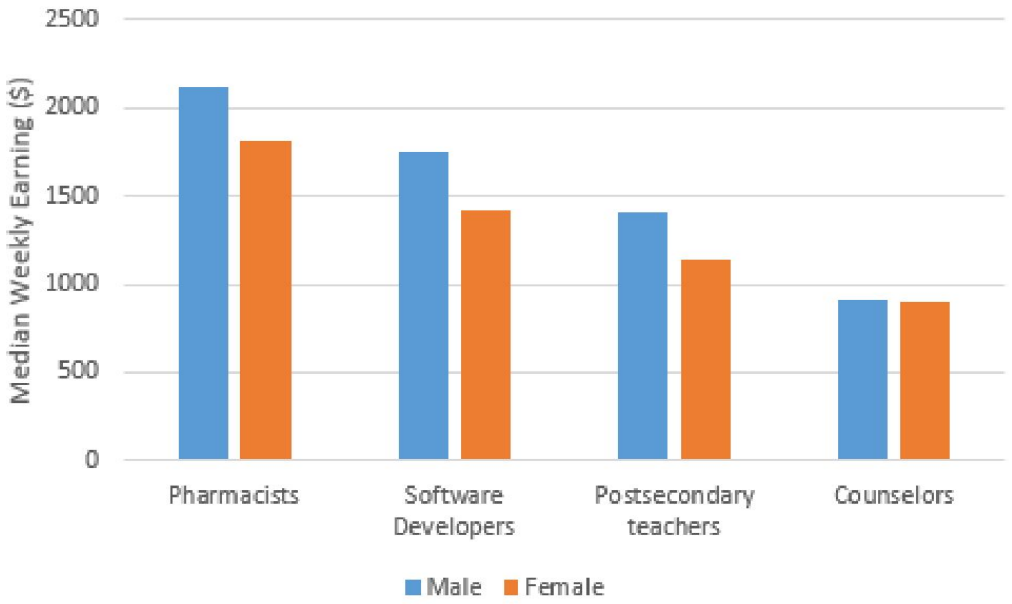
- Maximize data-ink ratio
- Minimize the lie factor
- Minimize chartjunk
- Use proper scales and clear labelling
- Make effective use of color
- Exploit the power of repetition



Maximizing Data-Ink Ratio

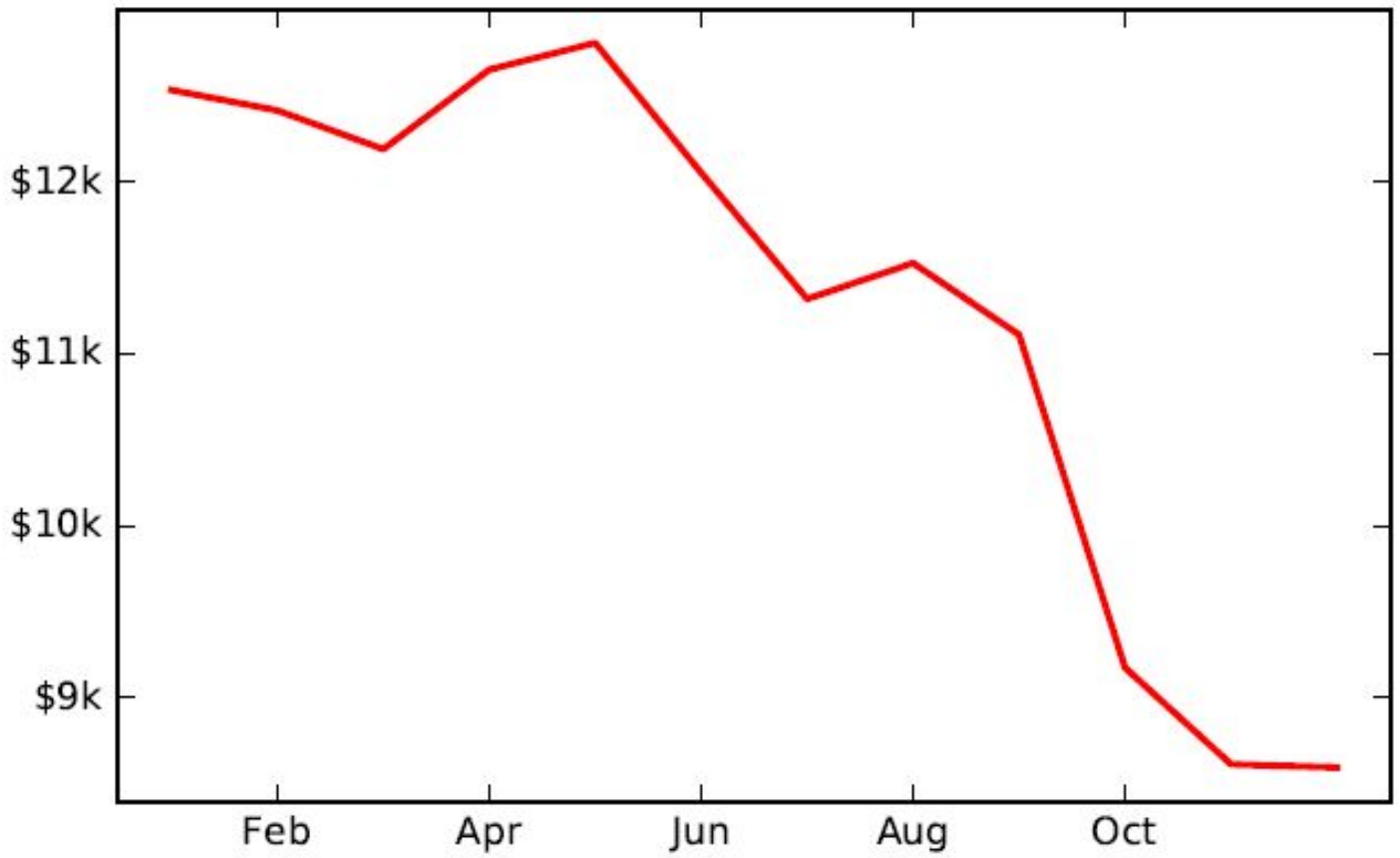
- $$\text{Data-Ink Ratio} = \frac{\text{Data-Ink}}{\text{Total ink used in graphic}}$$

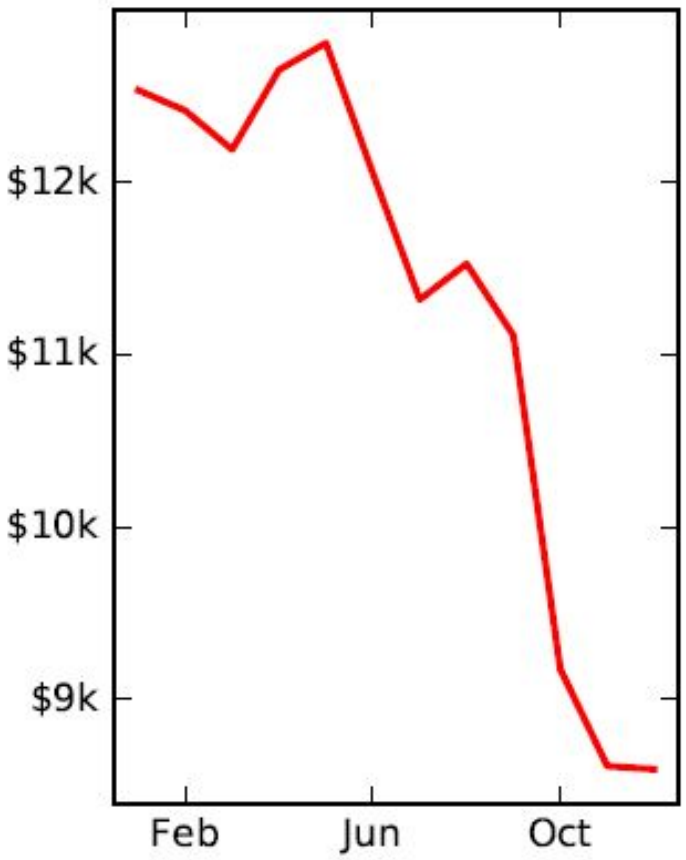


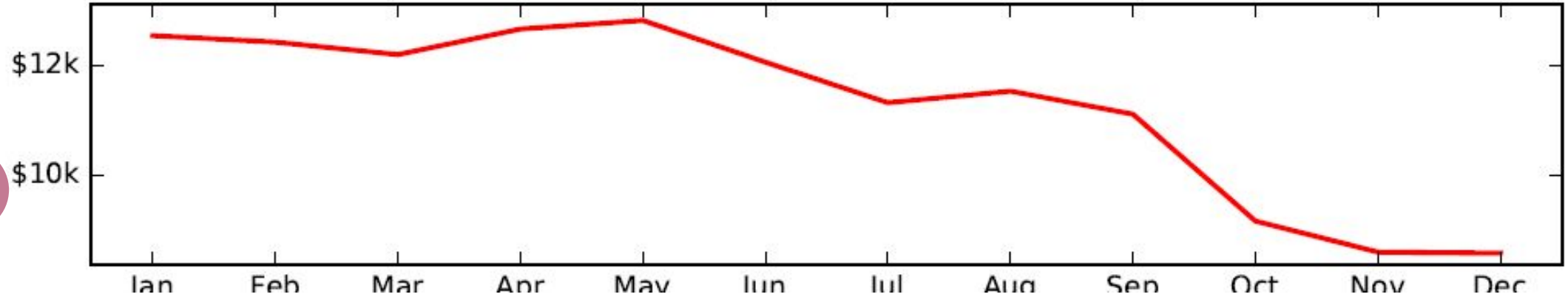


Minimizing the Lie Factor

- $lie\ factor = \frac{(size\ of\ an\ effect\ in\ the\ graphic)}{(size\ of\ the\ effect\ in\ the\ data)}$
- Bad practices include
  - Presenting means without variance
  - Presenting interpolations without the actual data
  - Distortions of scale
  - Eliminating tick labels from numerical axes
  - Hide the origin point from the plot

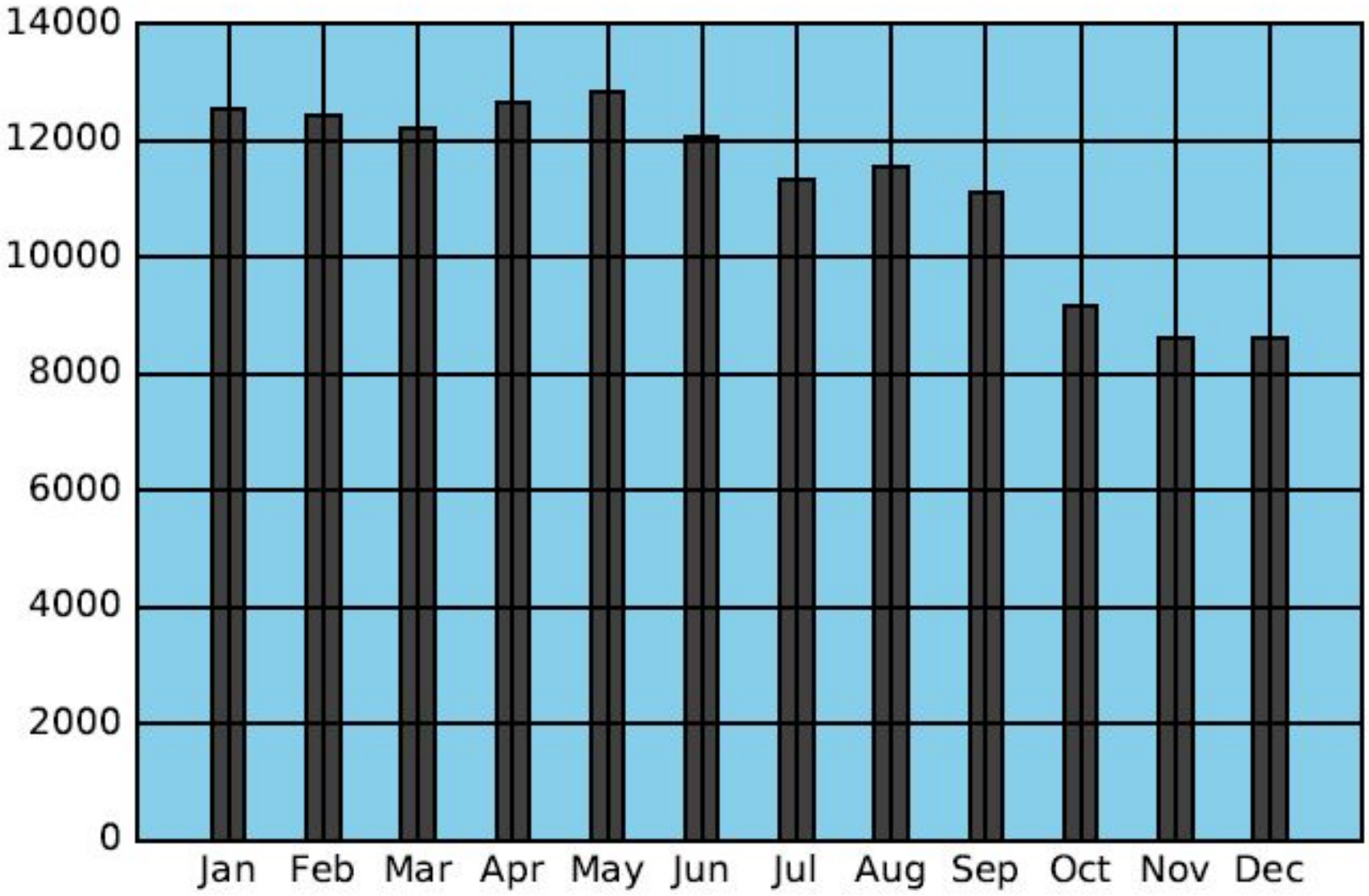

$12k
$11k
$10k
$9k
Feb
Apr
Jun
Aug
Oct


$12k
$11k
$10k
$9k
Feb
Jun
Oct


$12k
$10k
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

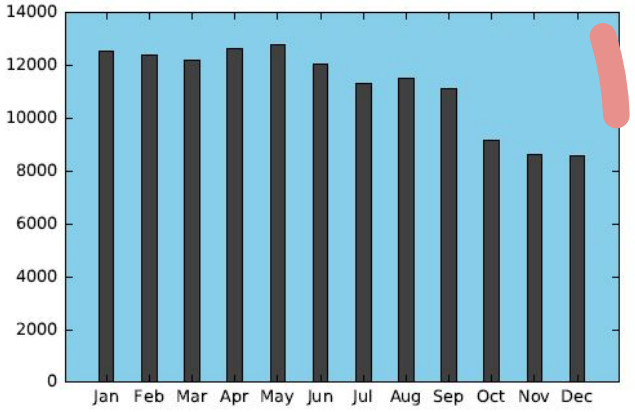
Org

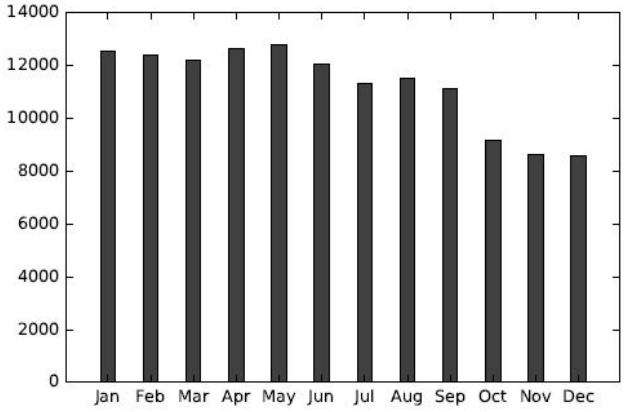
A monthly time series of sales.

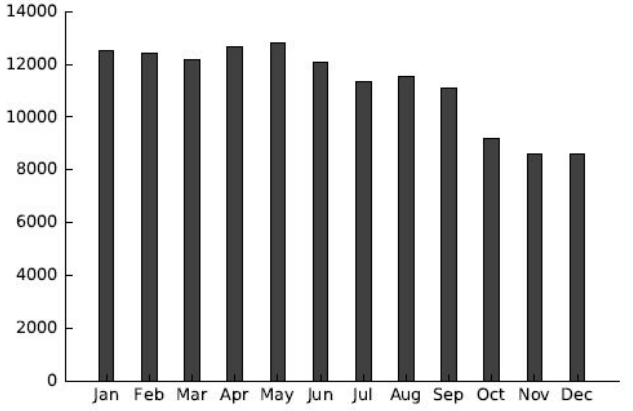


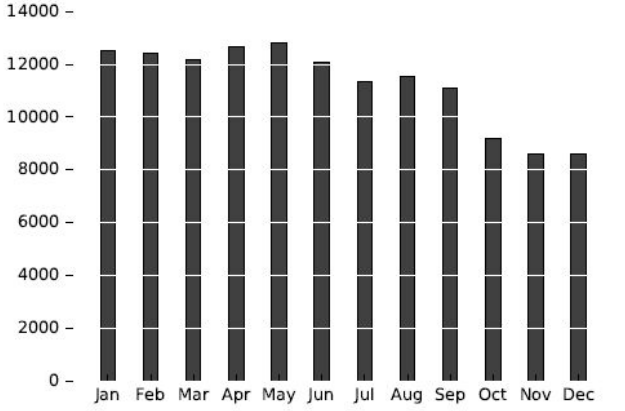
Minimizing Chartjunk

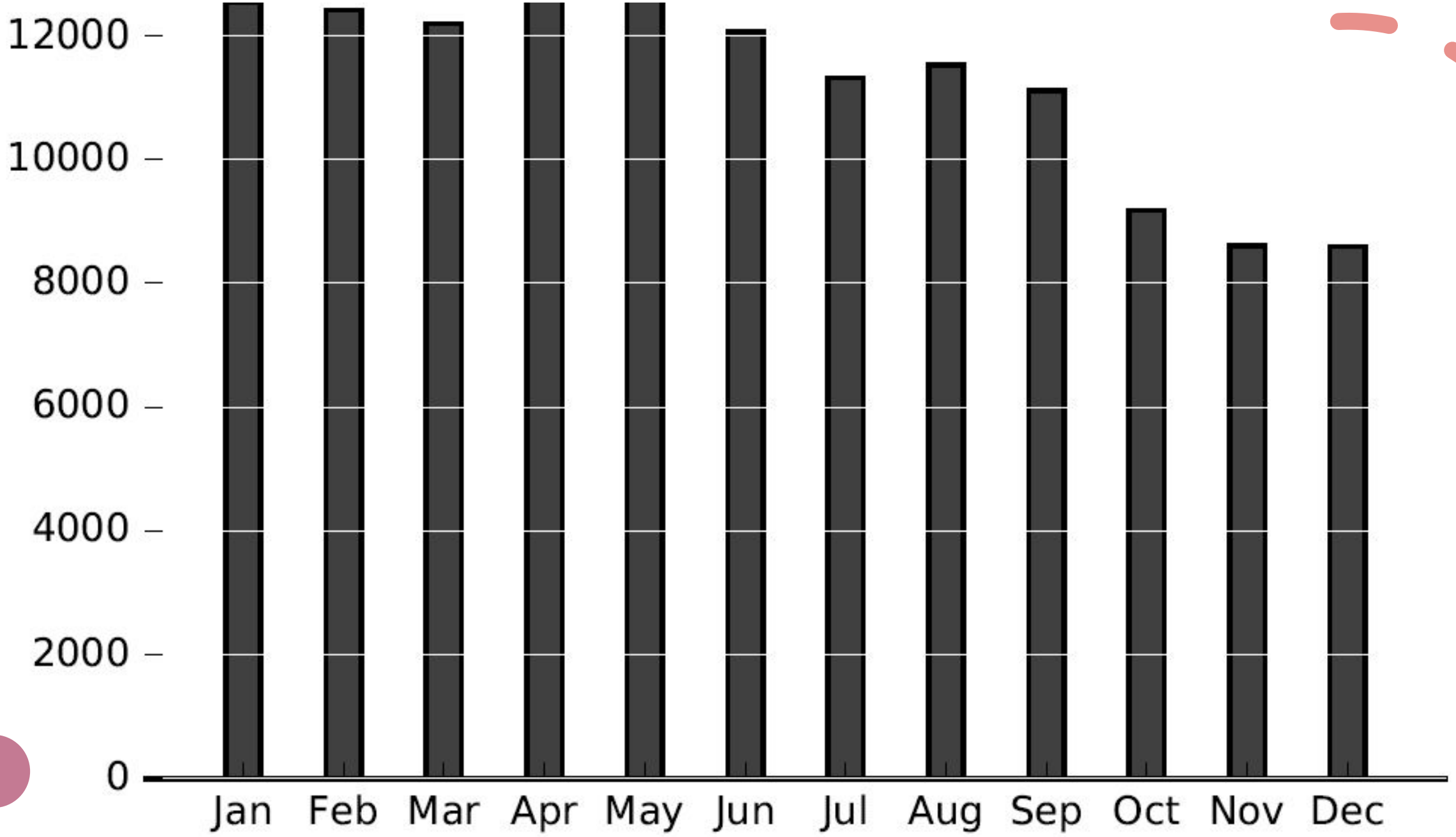
- Jailbreak your data (upper left)
- Stop throwing shade (upper right)
- Think outside the box (lower left)
- Make missing ink work for you (lower right)










12000
10000
8000
6000
4000
2000
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Proper Scaling and Labeling

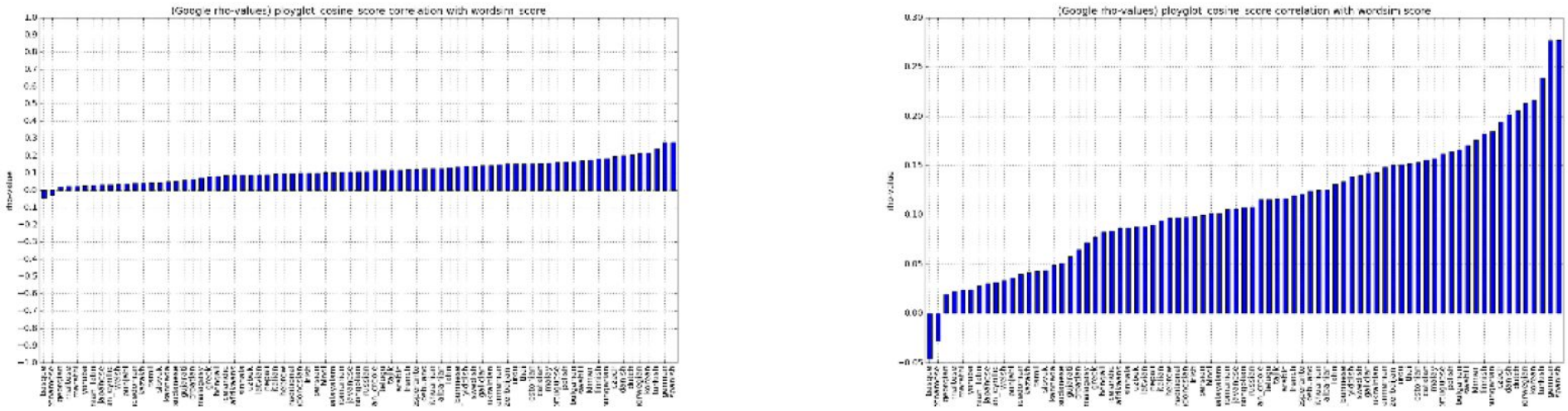


Figure 6.9: Scaling over the maximum possible range (left) is silly when all it shows is white space. Better scaling permits more meaningful comparisons (right).

Effective Use of Color and Shading

- Colors play two major roles in charts, namely marking class distinctions and encoding numerical values.
- Representing points of different types, clusters, or classes with different colors encodes another layer of information on a conventional dot plot.
- This is a great idea when we are trying to establish the extent of differences in the data distribution across classes.
- The most critical thing is that the classes be easily distinguishable from each other, by using bold primary colors.

The Power of Repetition

- Small multiple plots and tables are excellent ways to represent multivariate data.
- Time series plots enable us to compare the same quantities at different calendar points

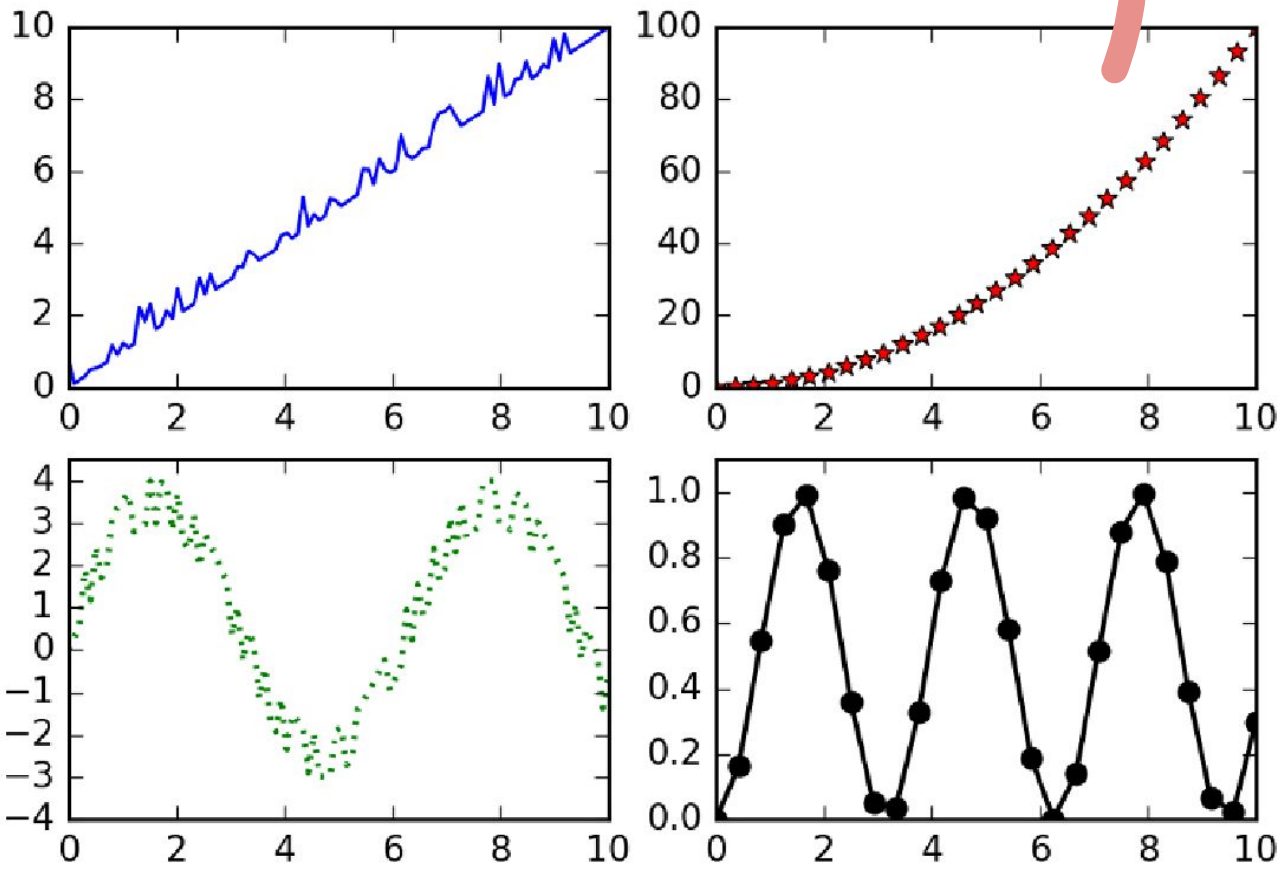
Chart Types

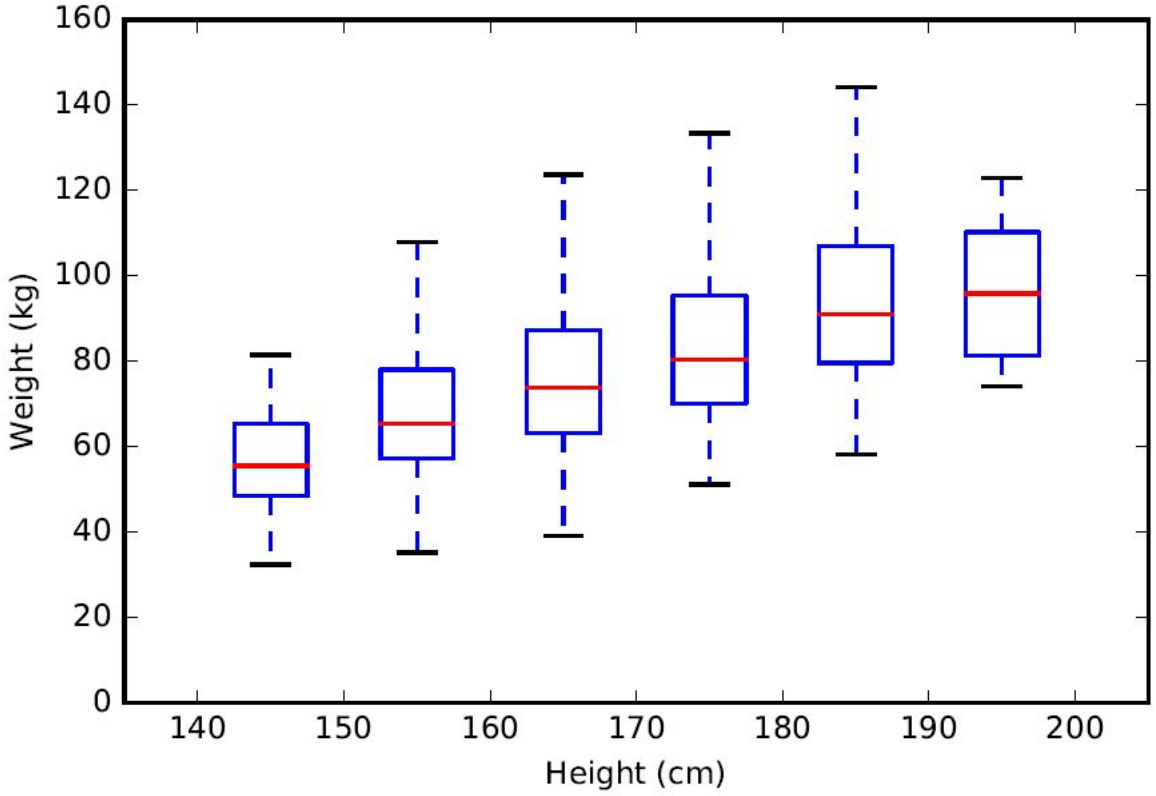
- Tabular Data
  - Representation of precision
  - Representation of scale
  - Multivariate visualization
  - Heterogeneous data
  - Compactness
- Best practices include
  - Order rows to invite comparison
  - Order columns to highlight importance, or pairwise relationships
  - Right-justify uniform-precision numbers
  - Use **emphasis**, font, or color to *highlight* important entries
  - Avoid excessive-length column descriptors

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

Dot and Line Plots

- Advantages of line charts include Interpolation and fitting





Box and whisker plots

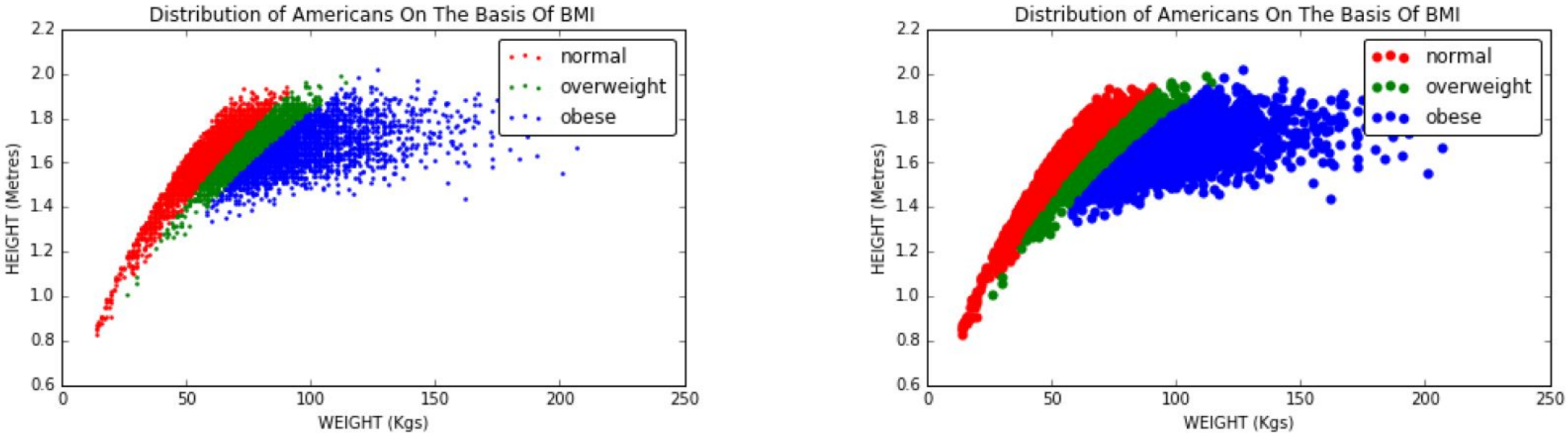
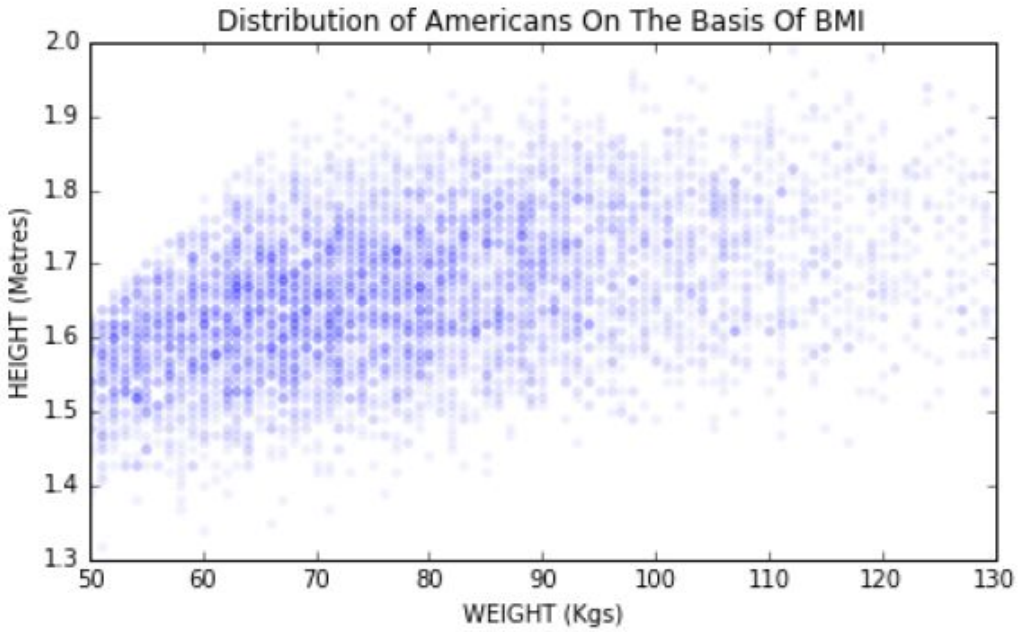


Figure 6.14: Smaller dots on scatter plots (left) reveal more detail than the default dot size (right).

Scatter Plots



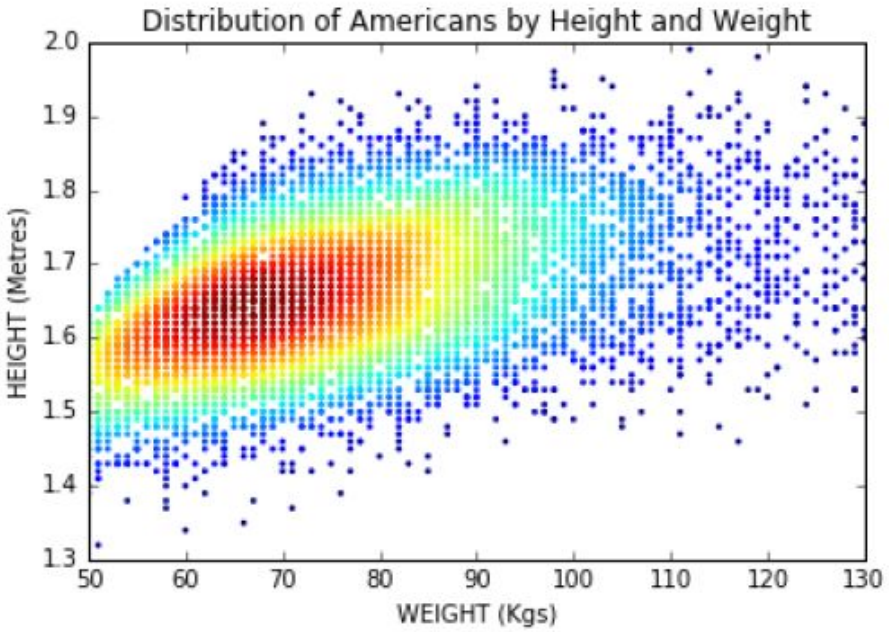


Figure 6.15: Overlapping dots can obscure scatter plots, particularly for large data sets. Reducing the opacity of the dots (left) shows some of the fine structure of the data (left). But a colored heatmap more dramatically reveals the distribution of the points (right).

Bar Plots and Pie Charts

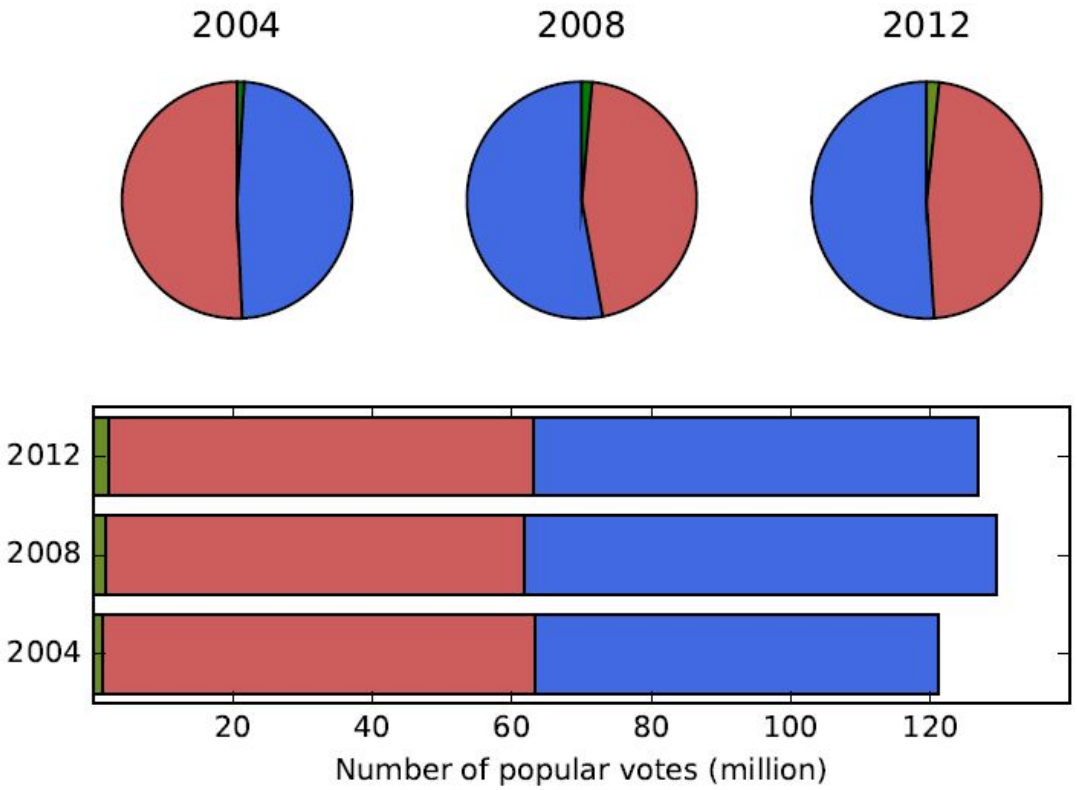


Figure 6.17: Voter data from three U.S. presidential elections. Bar plots and pie charts display the frequency of proportion of categorical variables. Relative magnitudes in a time series can be displayed by modulating the area of the line or circle.

Histograms

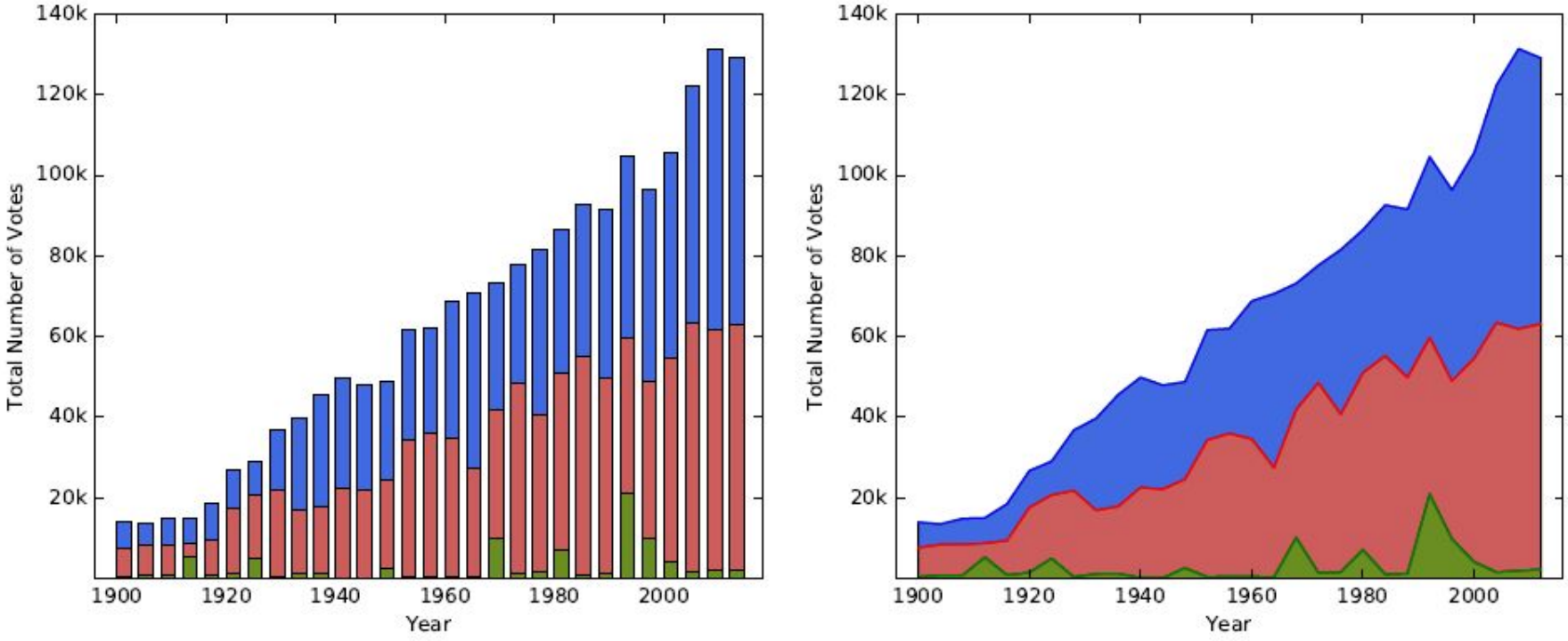


Figure 6.21: Time series of vote totals in U.S. presidential elections by party enable us to see the changes in magnitude and distribution. Democrats are shown in blue, and Republicans in red. It is hard to visualize changes, particularly in the middle layers of the stack.

Data Maps

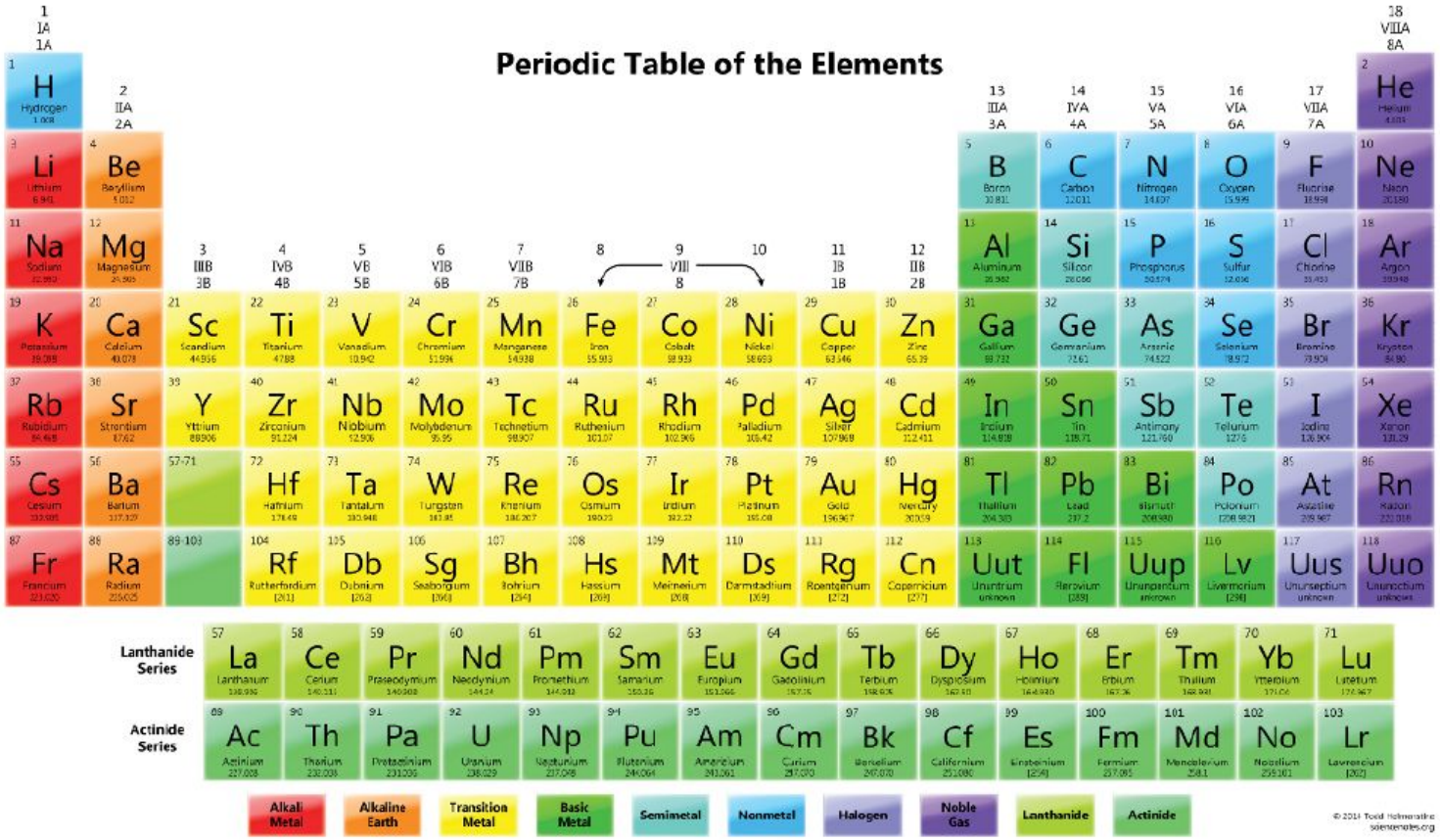


Figure 6.26: The periodic table maps the elements into logical groupings, reflecting chemical properties. *Source*: http://sciencenotes.org