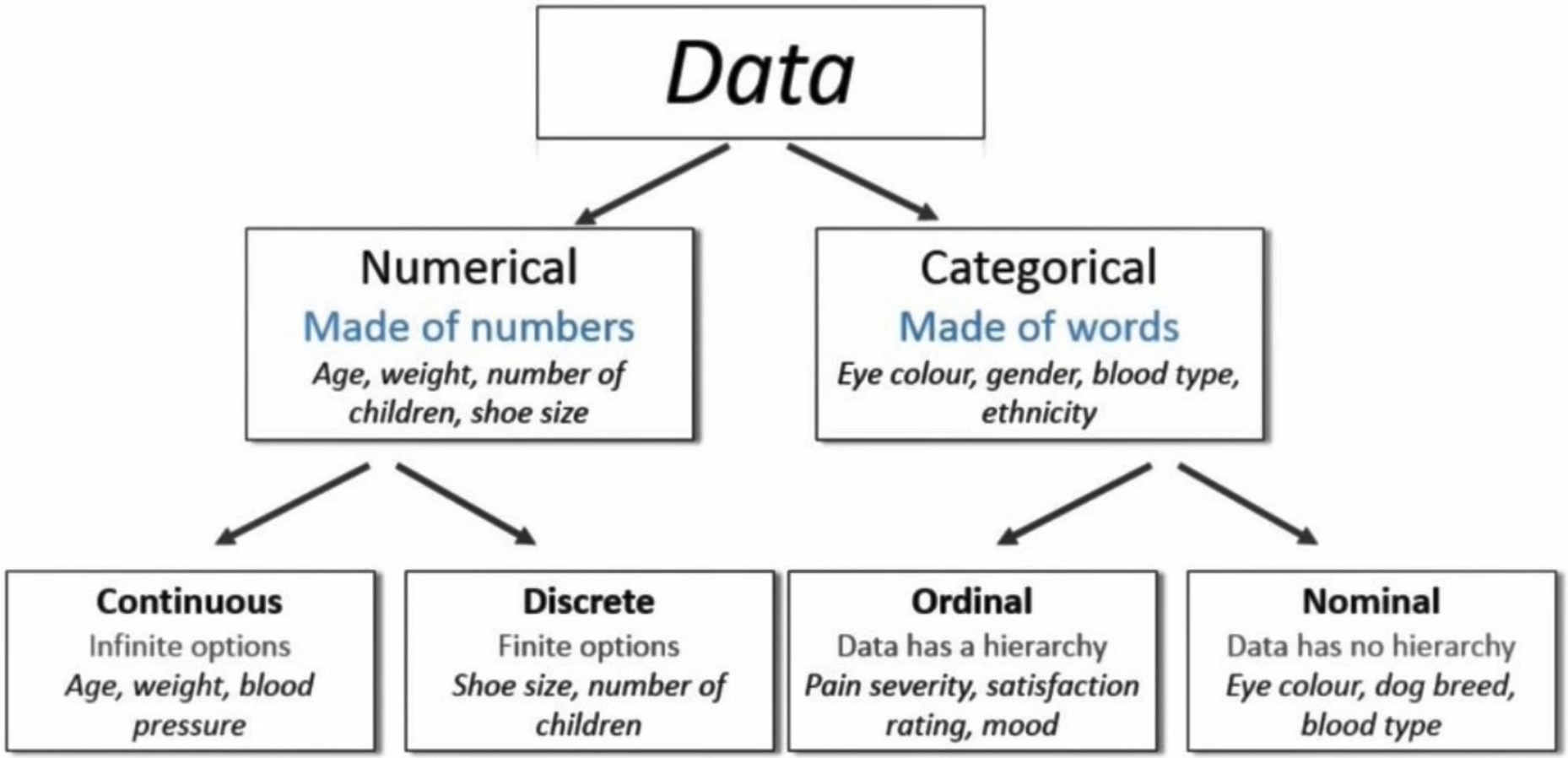
Data Analytics
and Visualization

19

Data Types

M. Vatandoost

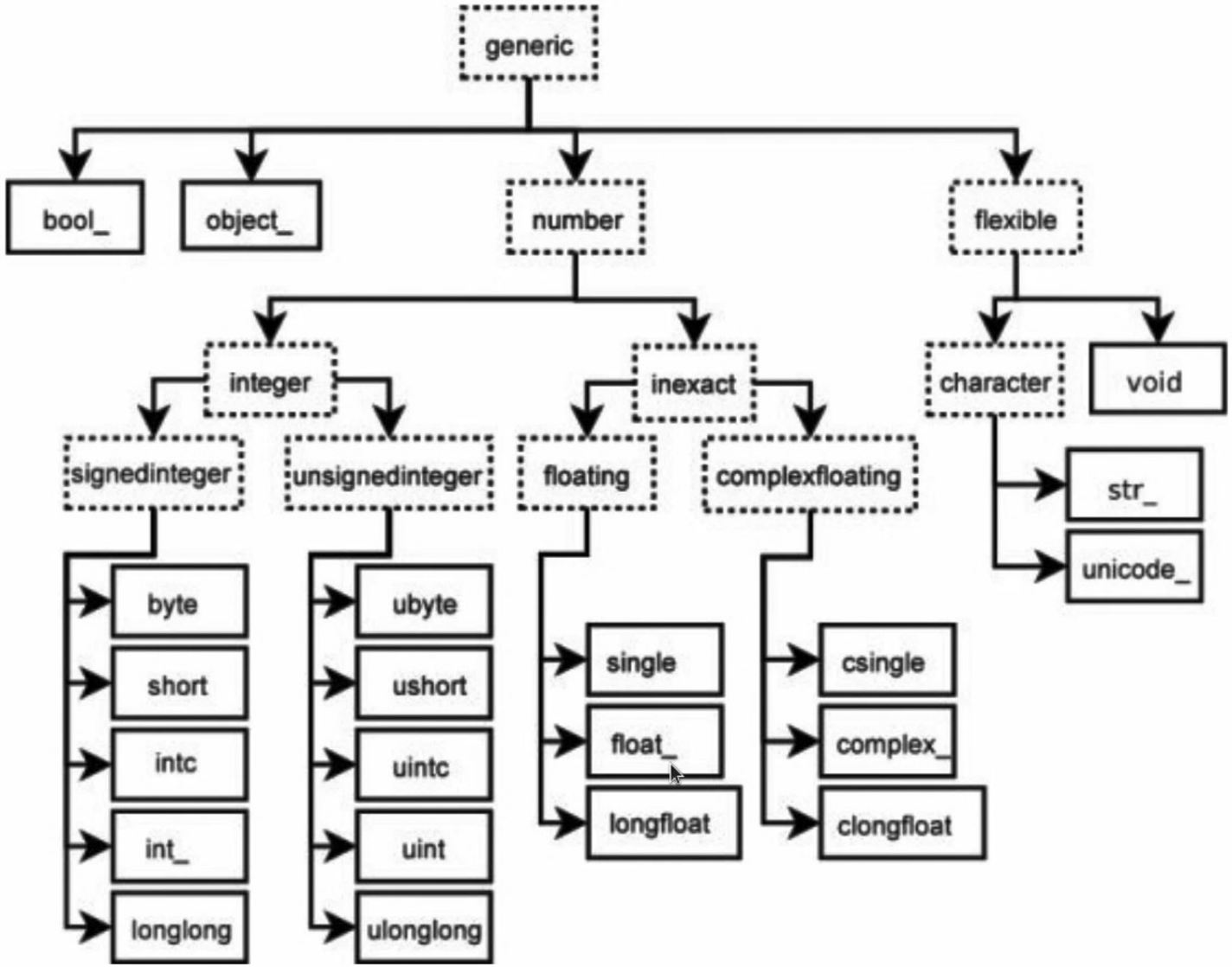

Data
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN
5	Tetris	GB	1989.0	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26	NaN	NaN
6	New Super Mario Bros.	DS	2006.0	Platform	Nintendo	11.28	9.14	6.50	2.88	29.80	89.0	65.0
7	Wii Play	Wii	2006.0	Misc	Nintendo	13.96	9.18	2.93	2.84	28.92	58.0	41.0
8	New Super Mario Bros. Wii	Wii	2009.0	Platform	Nintendo	14.44	6.94	4.70	2.24	28.32	87.0	80.0
9	Duck Hunt	NES	1984.0	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31	NaN	NaN

Python basic data types

- Numbers
  - Integers
  - Floating points
  - Complex
- Boolean
- String
- List
- Tuple
- Dictionary

Numpy scalar data types



Pandas data types

Pandas `dtype` mapping

Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

Statistical Inference

Introduction to Data

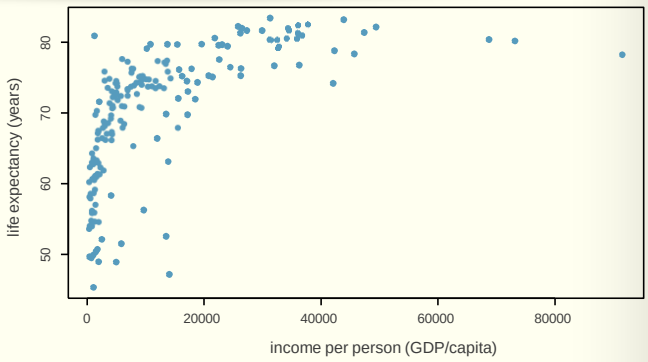
Behnam Bahrak

Visualizing Numerical Data

- Scatterplot
- Histogram
- Dot plot
- Box plot
- Intensity map

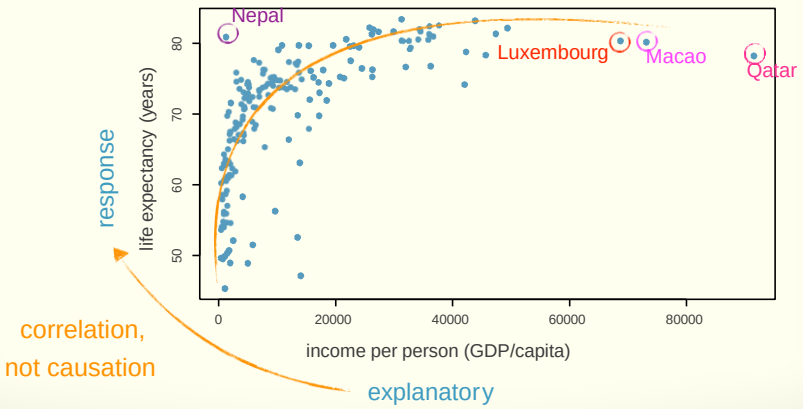
Scatterplot

data	income /person	life expectancy
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
⋮	⋮	⋮
Zimbabwe	545.3	58.142

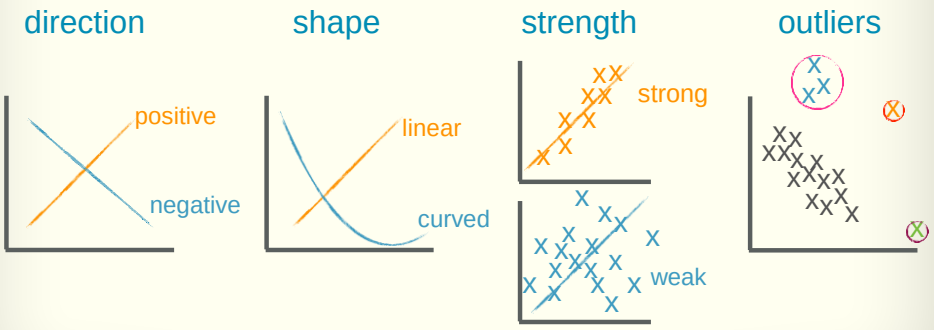


- *Scatterplots* are useful for visualizing the relationship between two numerical variables.

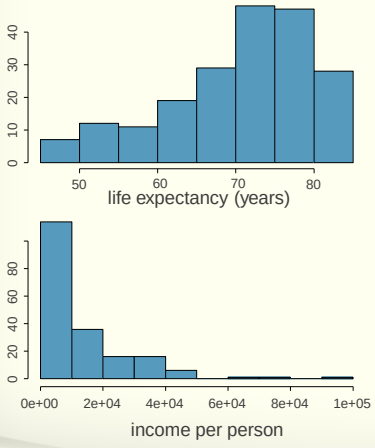
Scatterplot



Evaluating the relationship



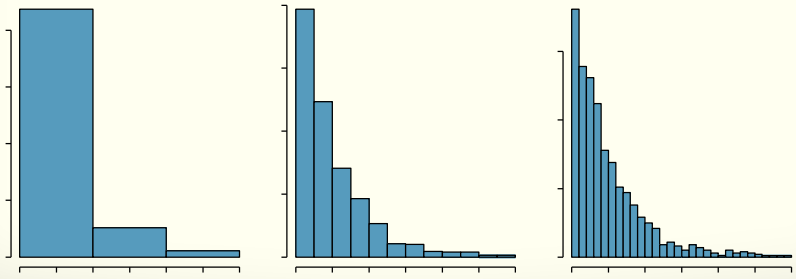
Histogram



- Histograms provide a view of the data density.
- Histograms are especially convenient for describing the shape of the data distribution.
- The chosen bin width can alter the story the histogram is telling

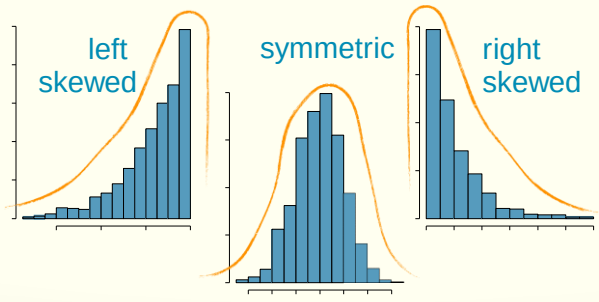
Bin Width

- When the bin width is too wide, we might lose interesting details.
- When the bin width is too narrow, it might be difficult to get an overall picture of the distribution.

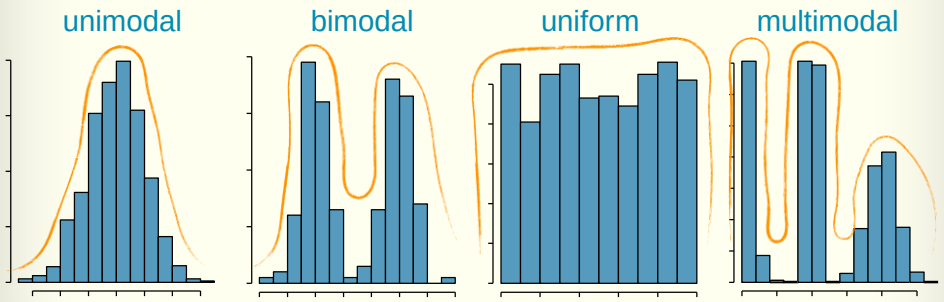


Skewness

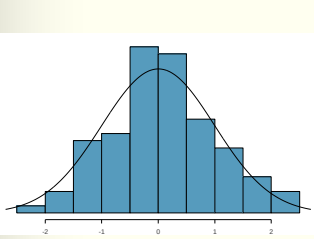
- Distributions are skewed to the side of the long tail



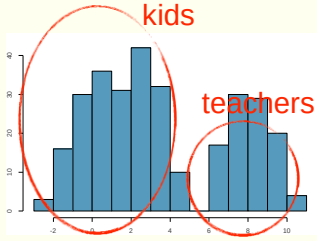
Modality



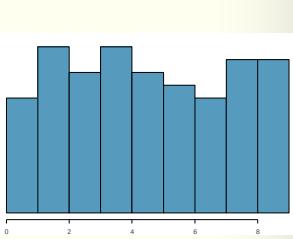
Modality



normal

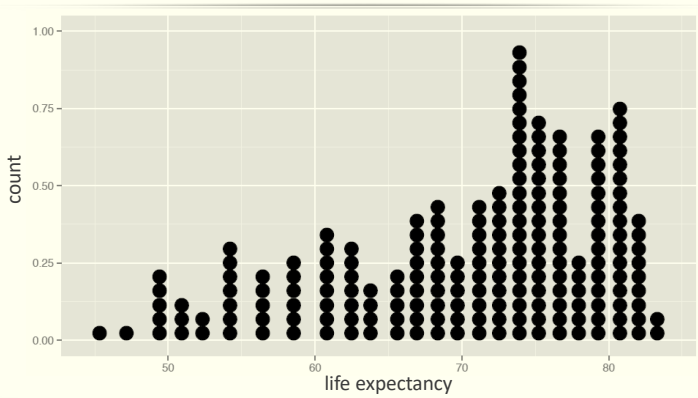


heights at elementary school



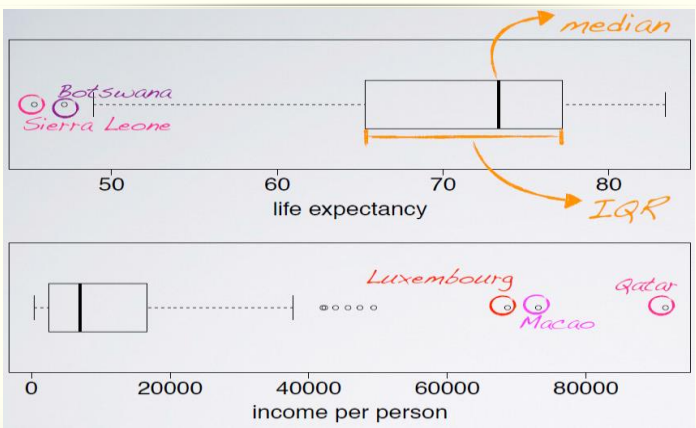
last digit of national ID

Dotplot



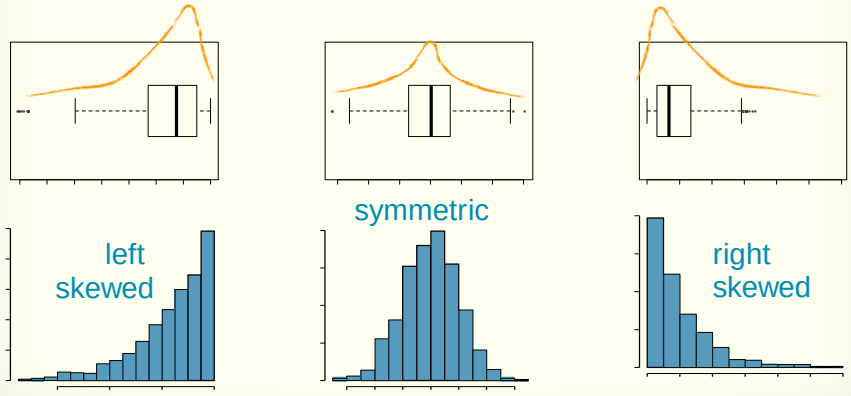
- Useful when individual values are of interest
- Can get busy as the sample size increases

Box plot

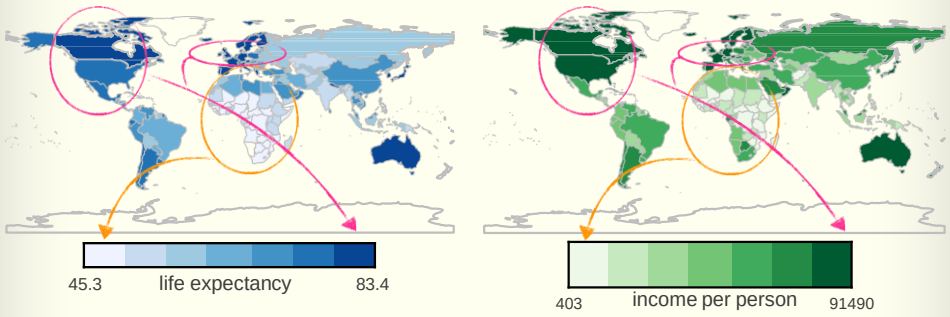


- Useful for highlighting outliers, median, IQR.

Determining the skewness from a box plot



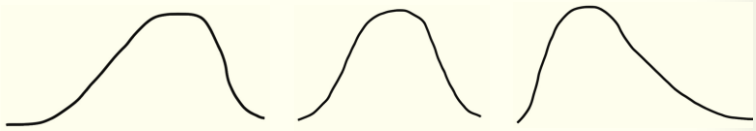
Intensity Map



- Useful for highlighting the spatial distribution.

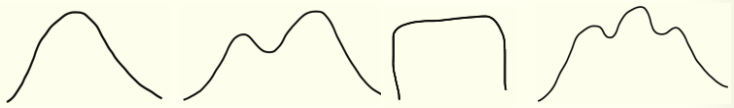
Shapes of Numerical Distributions

Skewness



left skewed | symmetric | right skewed

Modality



unimodal | bimodal | uniform | multimodal

Measures of Center

- Mean: arithmetic average
  - Sample mean: $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n}$
  - Population mean: μ
- Median: midpoint of the distribution
  - 50th percentile
- Mode: most frequent observation

sample statistic

point estimate

population parameter

Statistical Inference

Introduction to Data

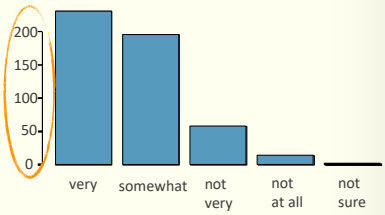
Behnam Bahrak

Describing Categorical Variables

- Contingency tables
- Bar plots
- Segmented bar
- Mosaic plots
- Pie charts

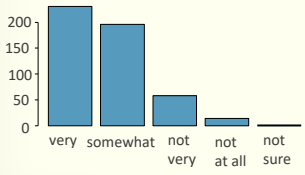
Frequency Table & Bar Plot

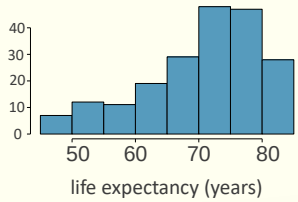
Difficulty saving money	Counts	Frequencies
Very	231	46%
Somewhat	196	39%
Not very	58	12%
Not at all	14	3%
Not sure	1	~0%
Total	500	100%



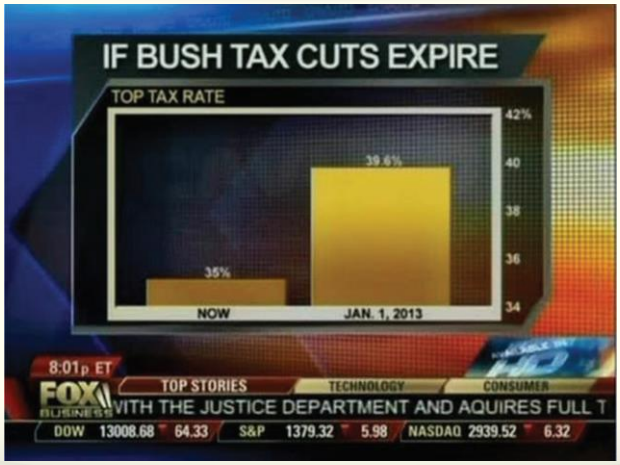
Bar Plots vs. Histograms

- Bar plots for categorical variables, but histograms for numerical variables
- x-axis on a histogram is a number line, and the ordering of the bars are not interchangeable



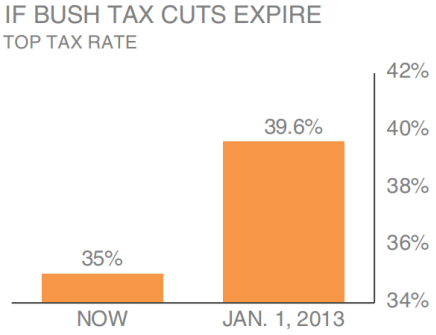


Bar Plot Abuse

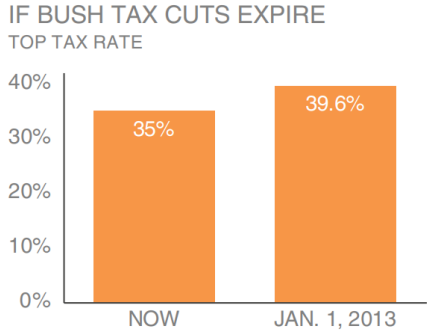


Bar Plot Abuse

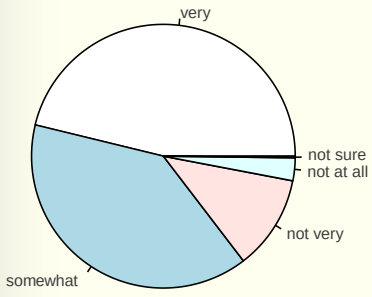
Non-zero baseline: as originally graphed

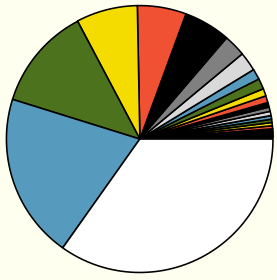


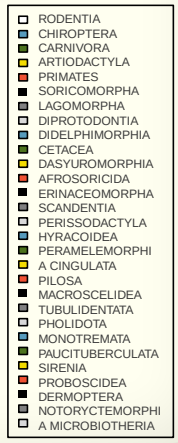
Zero baseline: as it should be graphed



~~Pie Chart?~~ NO!

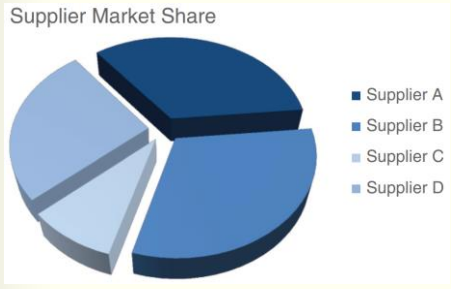


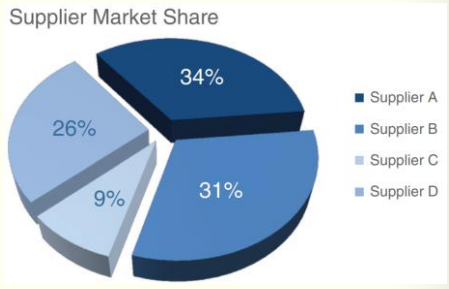




- A pie chart is actually much less informative than a bar plot.

To be avoided: Pie Charts





Contingency Table

		Income				
		< $40K	$40-80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

- A table that summarizes data for two categorical variables is called a contingency table.

Relative Frequency

		Income				
		< $40K	$40K - $80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

< $40K: 128/202 = 63% find it very difficult to save

$40K-$80K: 63/148 = 43%

$80K: 31/124 = 25%

Refused: 9/26 = 35%

⇒ feelings about difficulty of saving money and income are associated (dependent)

Contingency Table

- In January 1971, a Gallup Poll asked, "A proposal has been made in Congress to require the US government to bring home all US troops before the end of the year. Would you like to have your congressman vote for or against this proposal?" Guess the results, for respondents in each education category.

	Elementary Education	High School Education	College Education	Total
For Withdrawal				73%
Against Withdrawal				27%
Total	100%	100%	100%	100%

Contingency Table

- In January 1971, a Gallup Poll asked, "A proposal has been made in Congress to require the US government to bring home all US troops before the end of the year. Would you like to have your congressman vote for or against this proposal?" Guess the results, for respondents in each education category.

	Elementary Education	High School Education	College Education	Total
For Withdrawal	80%	75%	60%	73%
Against Withdrawal	20%	25%	40%	27%
Total	100%	100%	100%	100%

Simpson's Paradox

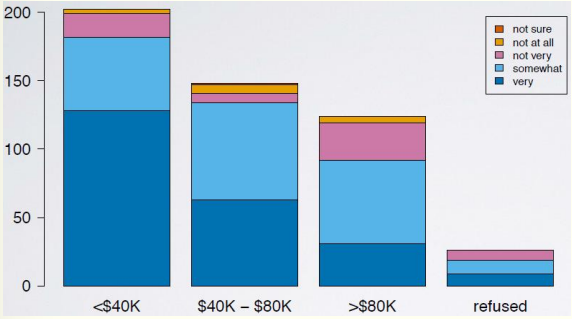
- A phenomenon in which a trend appears in different groups of data but disappears or reverses when the groups are combined.

Major	Women acceptance rate	Men acceptance rate
Computer science	27%	25%
Economics	26%	22%
Engineering	32%	26%
Medicine	24%	24%
Veterinary Medicine	16%	12%
Total	**23%**	**24%**

Simpson's Paradox

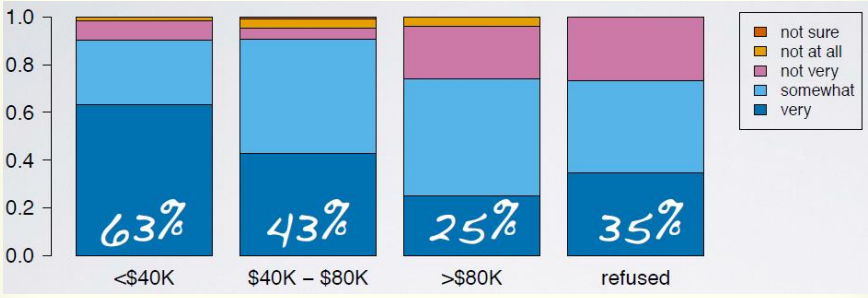
	Male	Female
Major A	$\frac{50}{150} = 30\%$	$\frac{180}{360} = 50\%$
Major B	$\frac{200}{250} = 80\%$	$\frac{36}{40} = 90\%$
Combined	$\frac{250}{400} = 62.5\%$	$\frac{216}{400} = 54\%$

Segmented Bar Plot

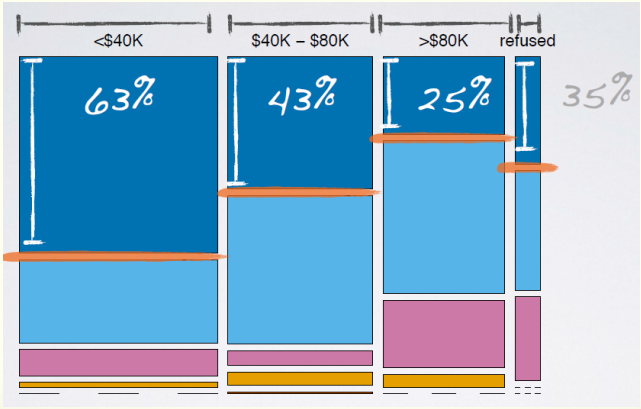


- Useful for visualizing conditional frequency distributions
- Compare relative frequencies to explore the relationship between the variables

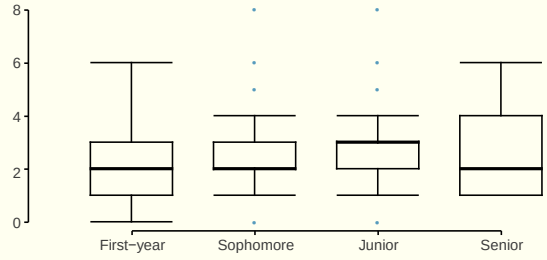
Relative Frequency Segmented Bar Plot



Mosaic Plot



side-by-side box plots



- Does there appear to be a relationship between class year and number of societies students are in?