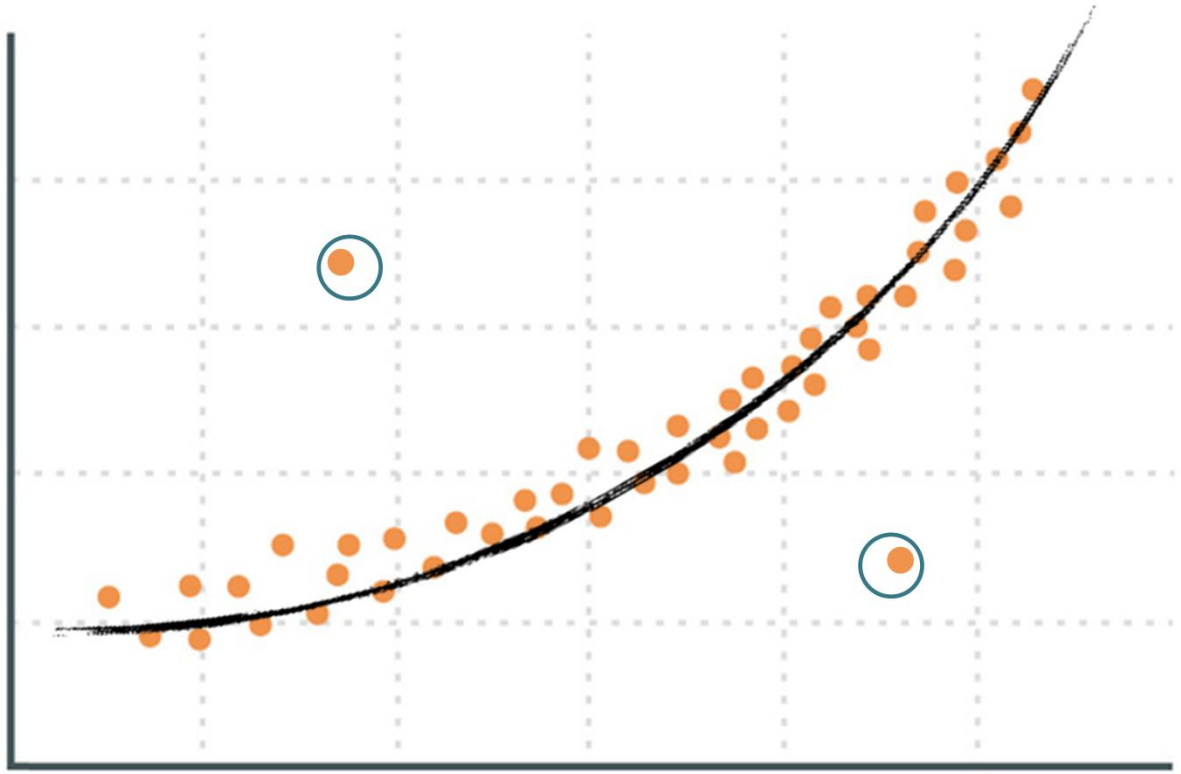
Data Visualisation

Visualising Numerical Data

Scatterplot



Income

Response Variable

Correlation, Not Causation

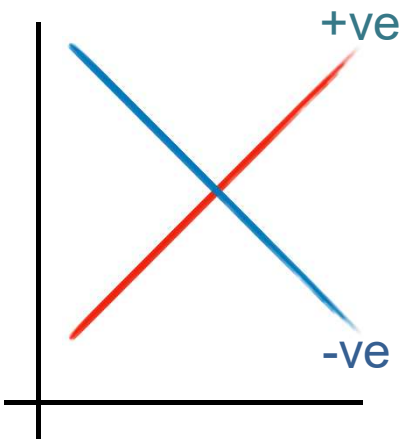
Age

Explanatory Variable

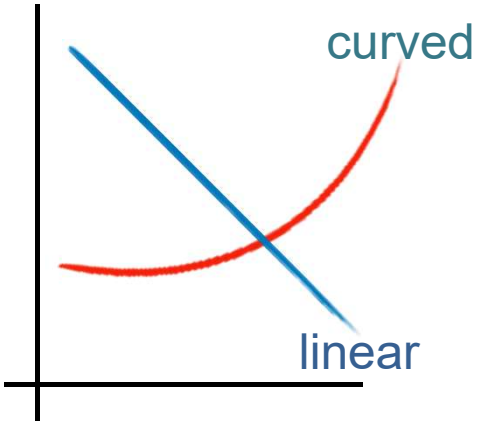
DICE ANALYTICS

Characteristics of Relationship

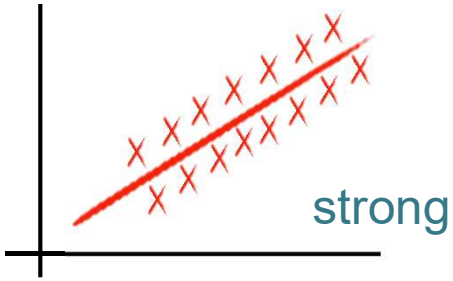
Direction

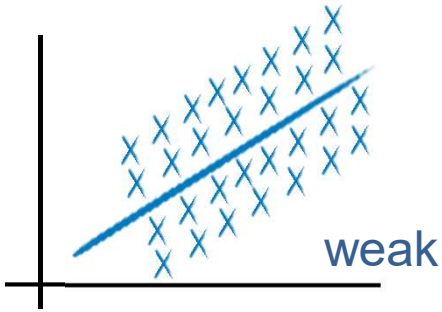


Shape

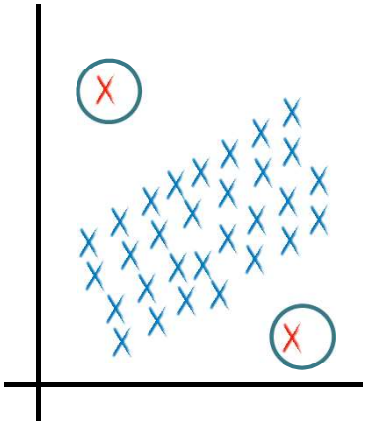


Strength

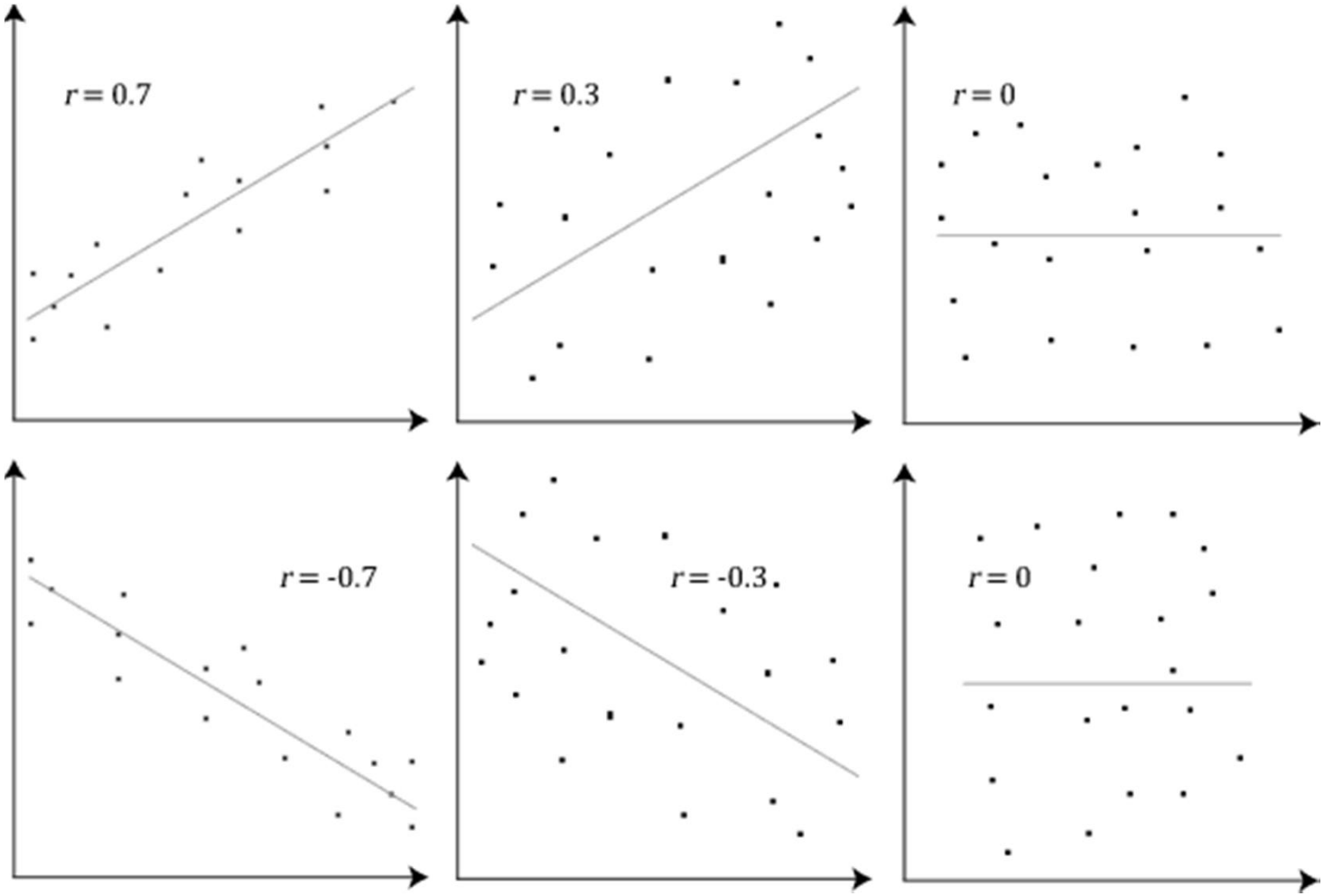




Outliers



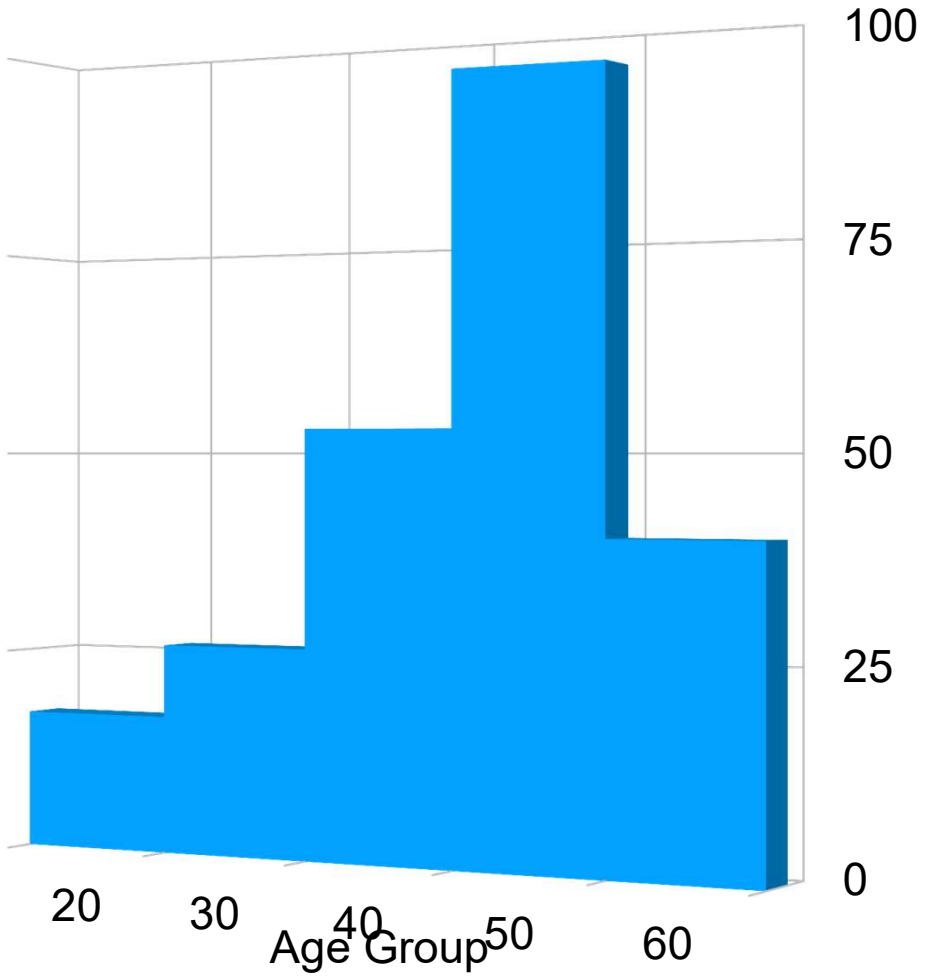
Correlation (example)



Histograms

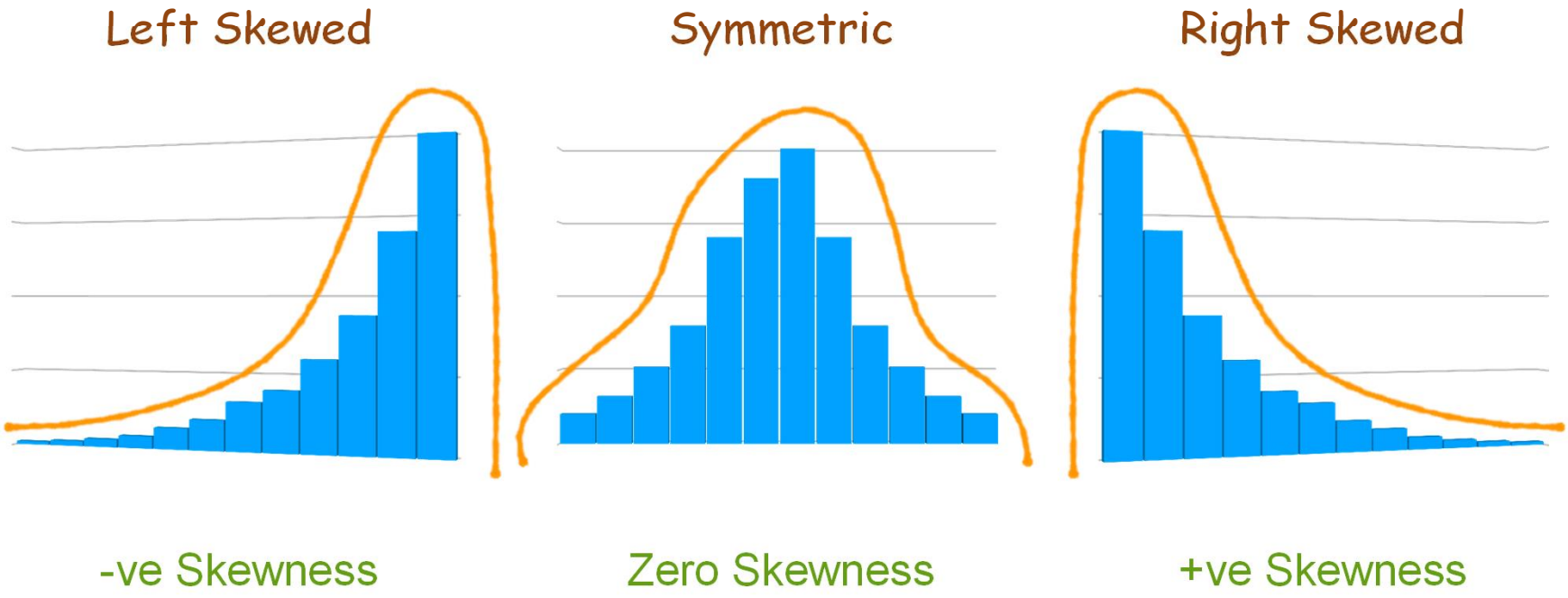
- Help to view *data density*
- Help to see *shape of distribution*

1) Skewness
2) Modality



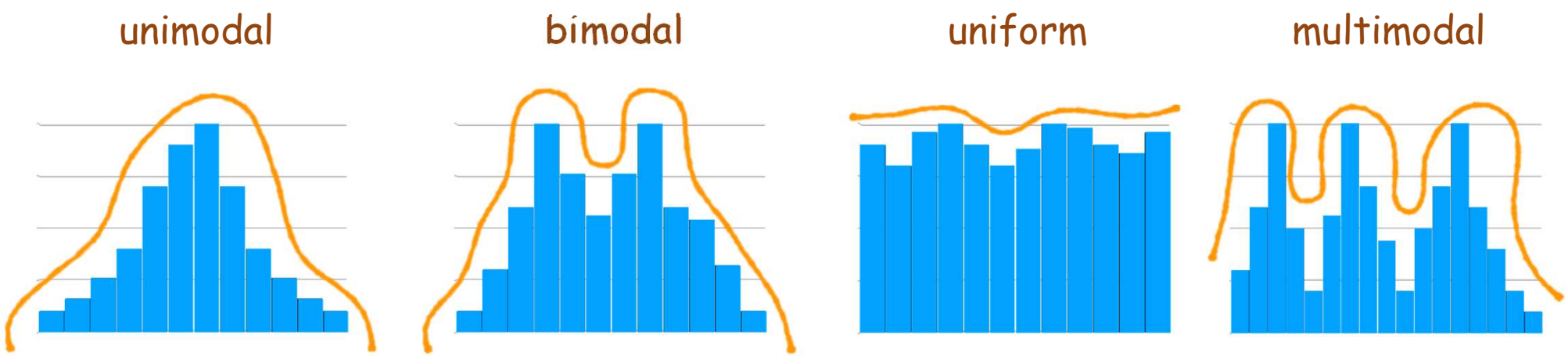
DICE ANALYTICS

Skewness

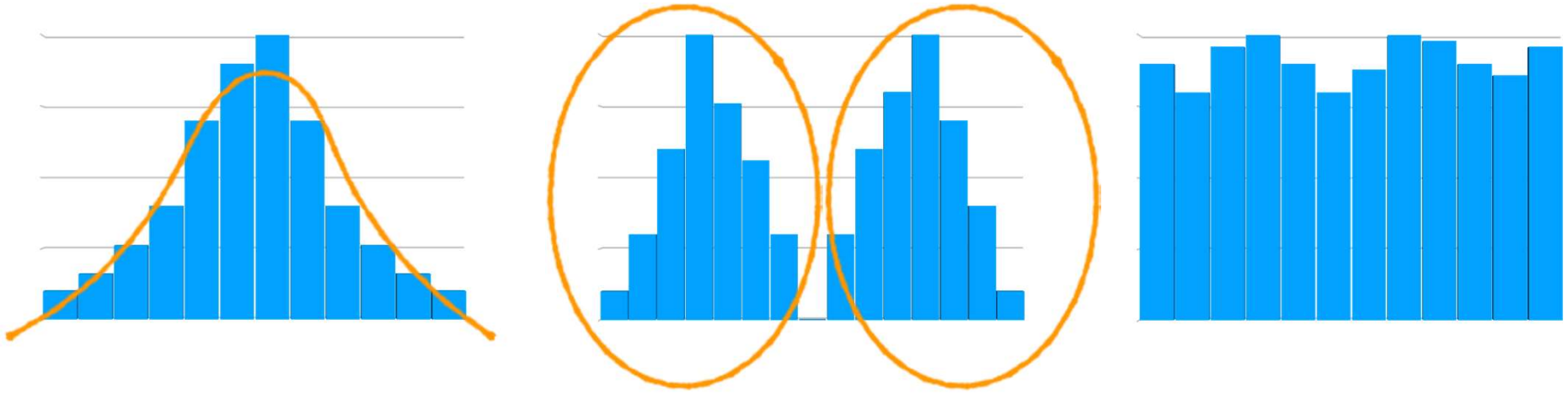


- Draw a smooth curve to see skewness
- Don’t rely on jagged edges

Modality



Modality (Example)

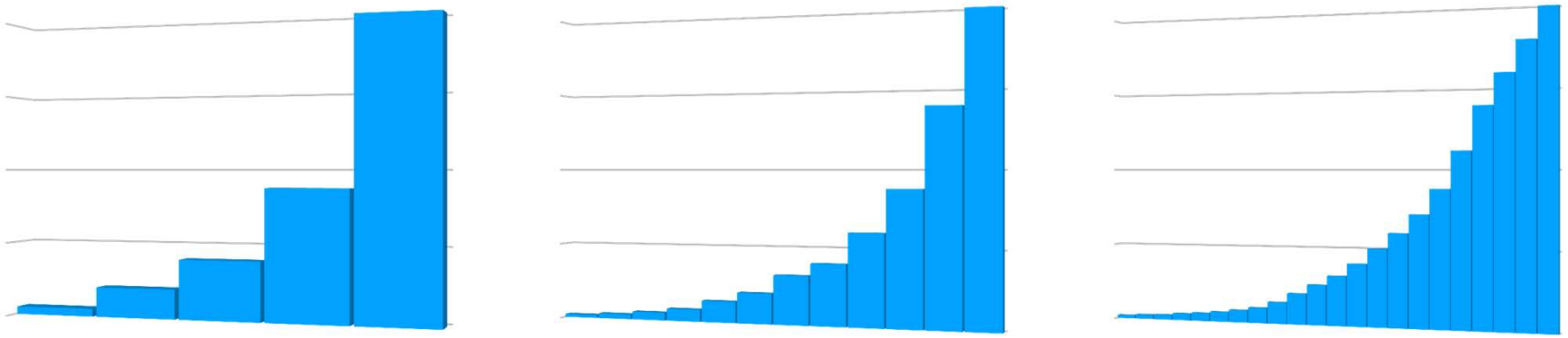


Normal Distribution

Two separate groups

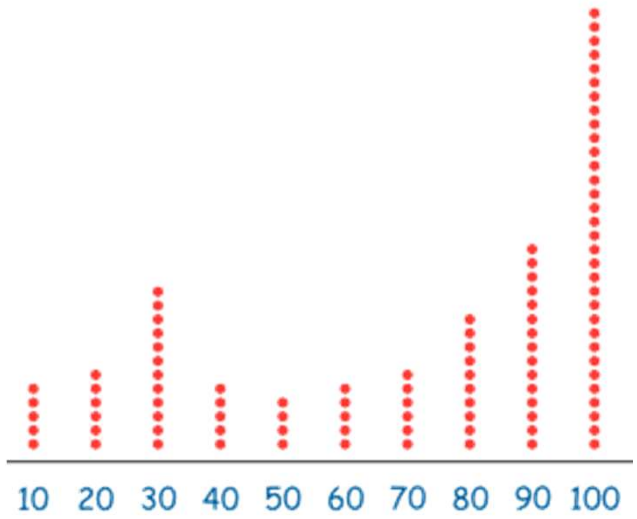
No trend

Binwidth

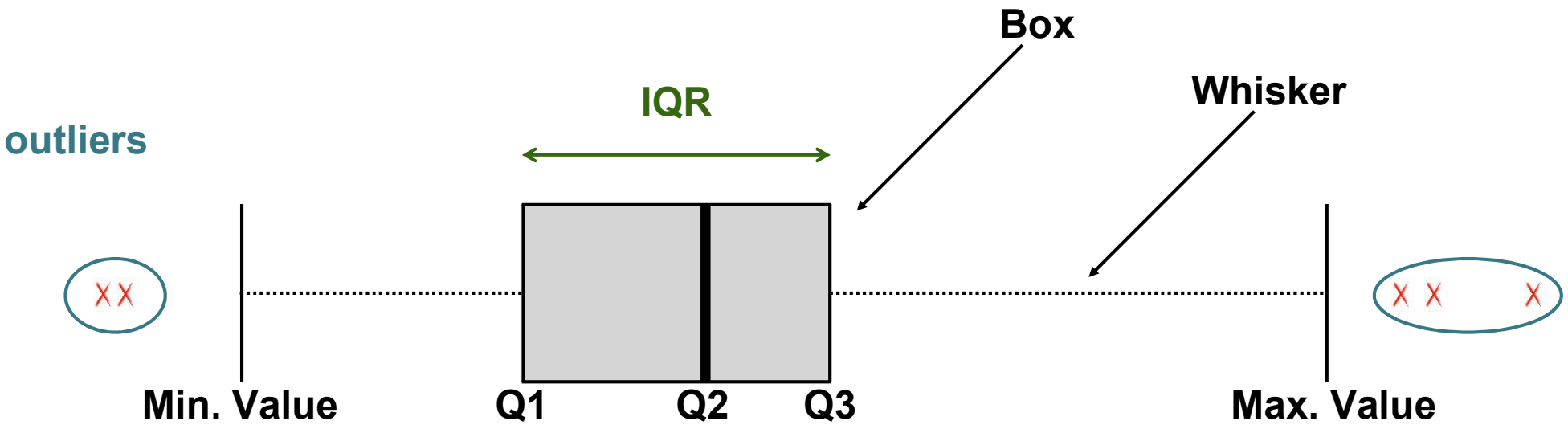


Dot Plots

- Useful when individual values are of interest
- Not suitable if sample size is big

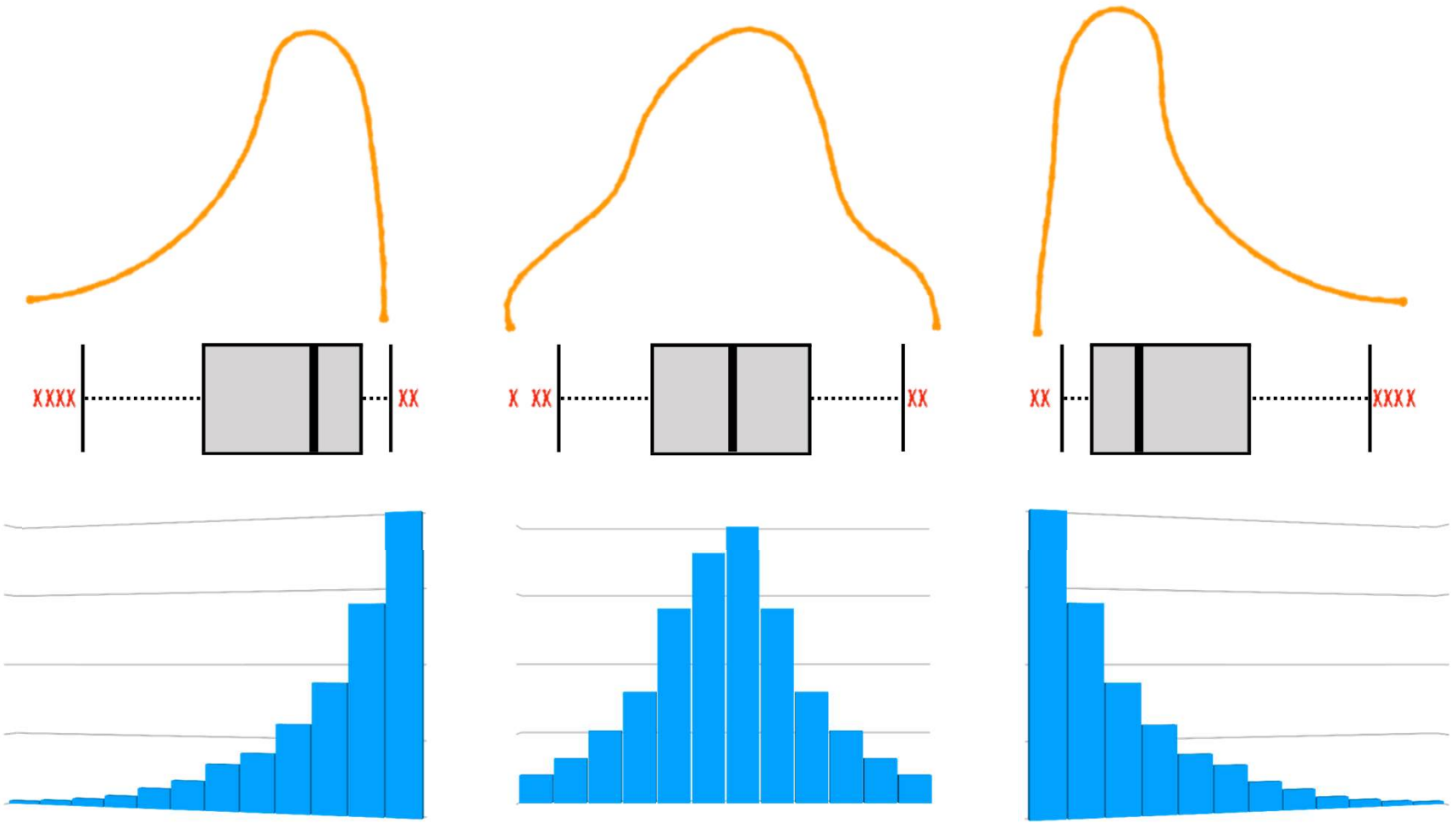


Box Plots

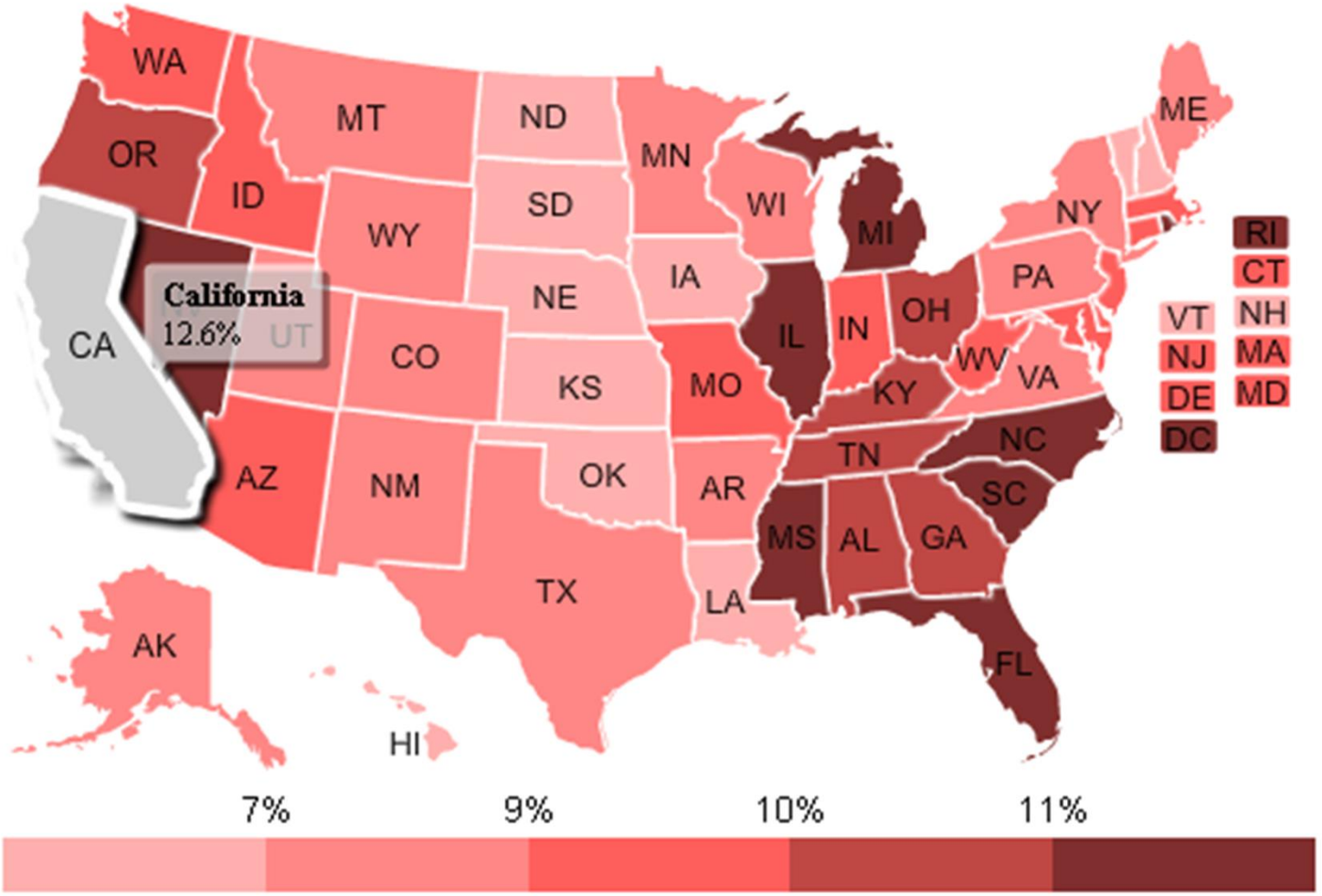


Min. Value :Lower Extreme (that's not an outlier)
Q1 :Lower Quartile (25% of observations)
Q2 :Median (50% of observations)
Q3 :Upper Quartile (75% of observations)
Max. Value :Upper Extreme (that's not an outlier)
IQR :Inter-Quartile Range = Q3 - Q1 (middle 50% of observations)

Box Plots & Skewness

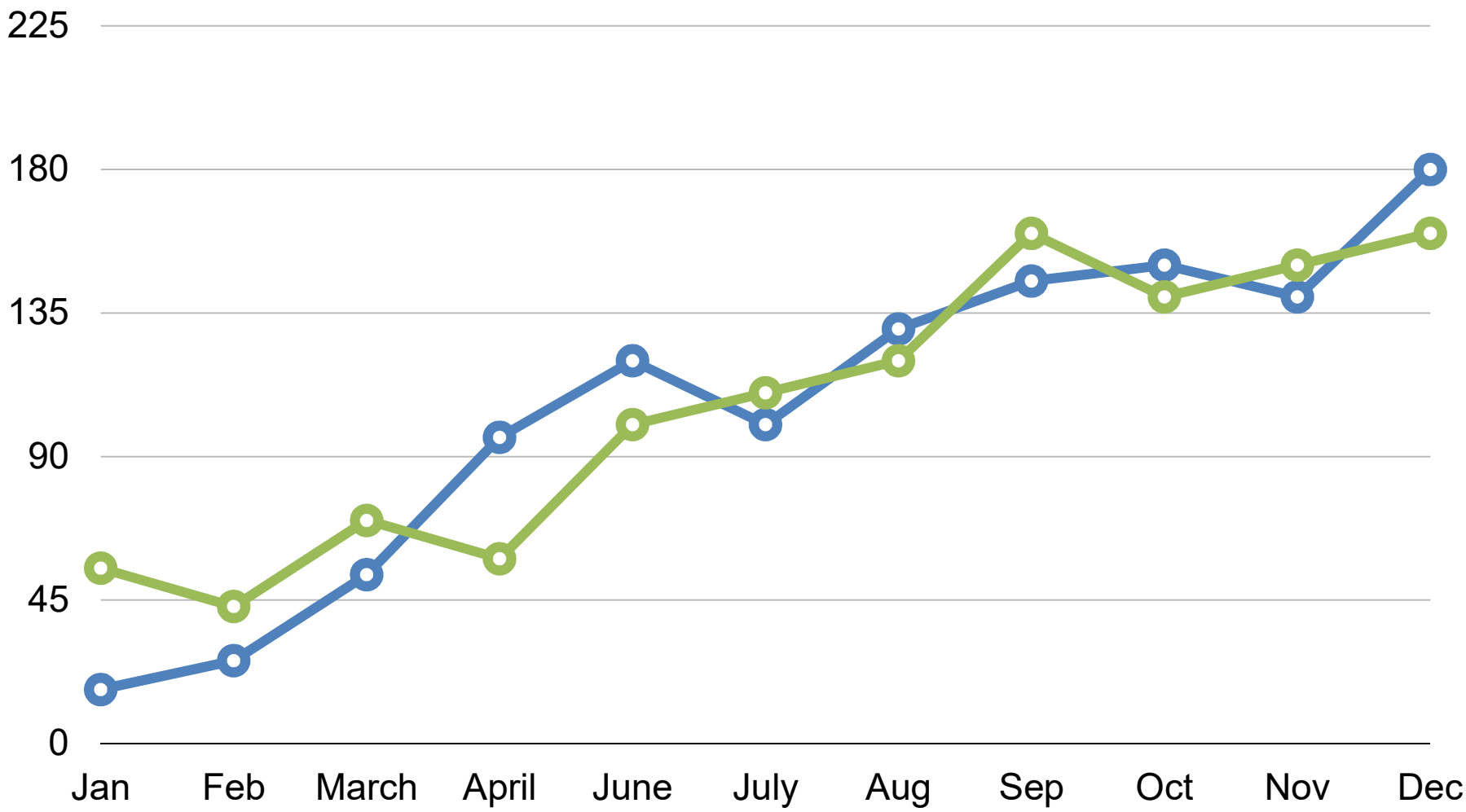


Intensity/Heat Maps



DICE ANALYTICS

Time Plots

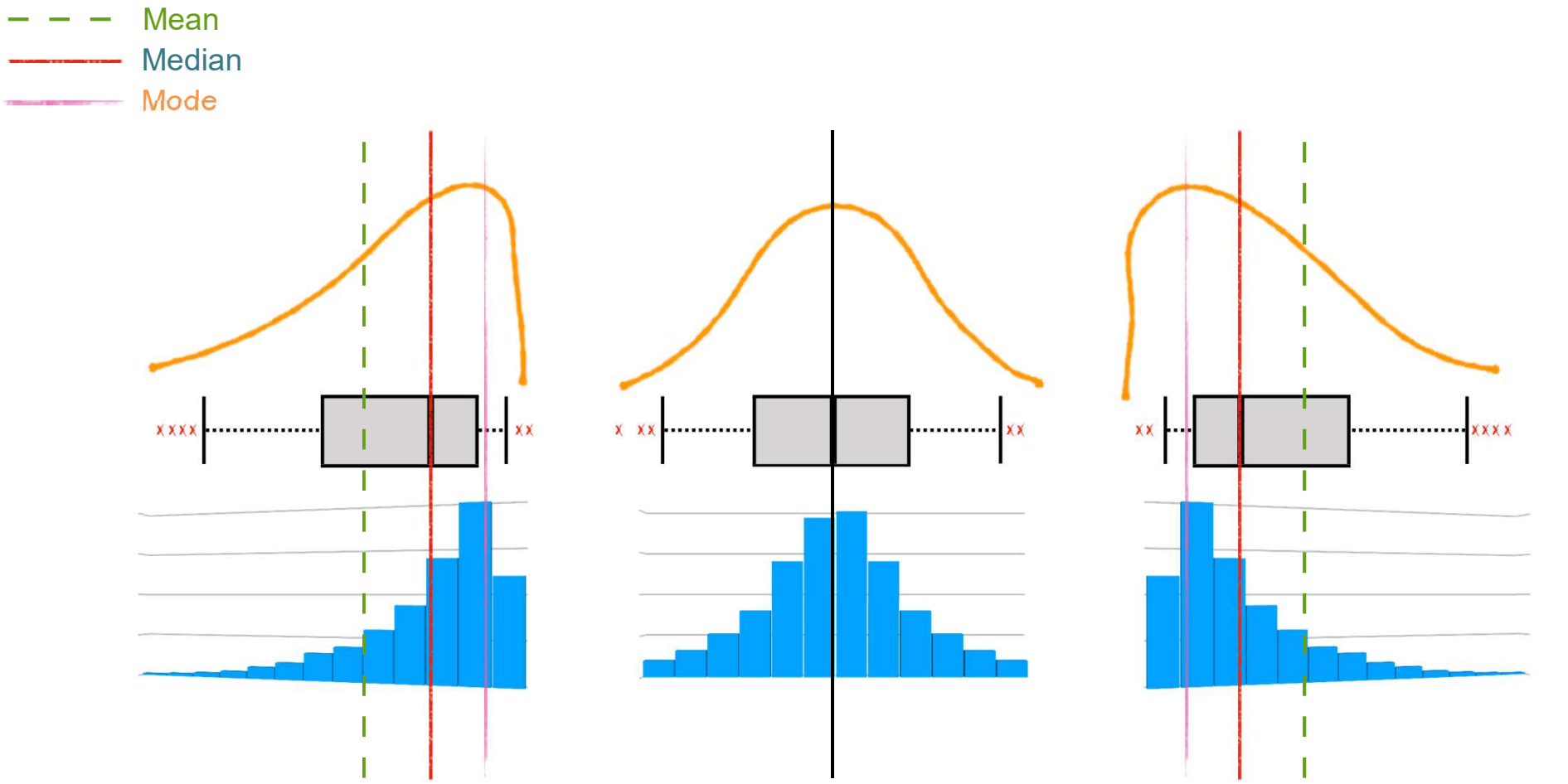


Measures of Center

Data : 56, 87, 34, 65, 77, 62, 90, 45, 77, 79

Mean	Arithmetic Average $\text{Mean} = \frac{56 + 87 + 34 + 65 + 77 + 62 + 90 + 45 + 77 + 79}{10}$ Mean = 67.2
Mode	Most frequent value/observation Mode = 77
Median	Midpoint of distribution (50^{th} percentile) $\text{Median} = \frac{77 + 62}{2} = 69.5$

Skewness vs Measures of Center



Mean < Median < Mode

Left Skewed

Mean = Median = Mode

Symmetric

Mean > Median > Mode

Right Skewed

DICE ANALYTICS

Parameter vs Statistic

- *Parameter*: Also called **Population Parameter**, is a numerical summary of the *population*.
- Denoted with letters by Greek alphabets. (Mean is μ)

- *Statistic*: Also called **Sample Statistic**, is a numerical summary of a *sample* taken from the population. It is a point estimate of population.
- Denoted with letters by Latin alphabets. (Mean is $\bar{X}$)

Measures of Spread

Range	**Variance**
Standard Deviation	**Inter-quartile Range**

Range

- Range = Max. Value - Min. Value
- **Data : 56, 87, 34, 65, 77, 62, 90, 45, 77, 79**
- Range = 90 - 34 = 56

Variance

- A measure of how much data (a variable) varies; how spread out a data set is about the mean.
- Average squared deviation from mean; has squared units of the variable
- Sample Variance $$s^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$
- Population Variance $$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Variance (Example)

- **Data : 56, 87, 34, 65, 77, 62, 90, 45, 77, 79**

$$s^2 = \frac{\sum (X - \bar{X})^2}{N-1} = \frac{(56 - 67.2)^2 + (87 - 67.2)^2 + \ldots\ldots + (79 - 67.2)^2}{10\text{-}1}$$

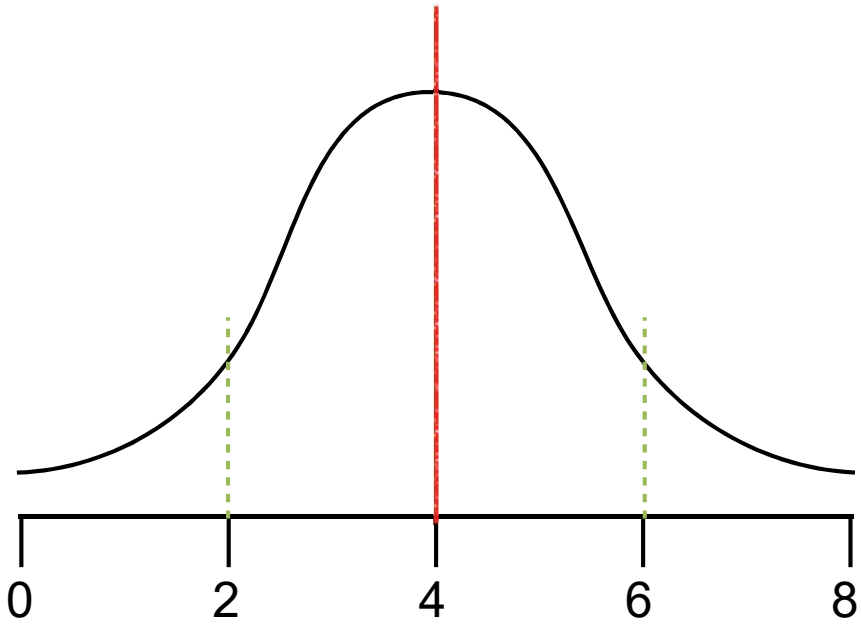
$$= \frac{2995.6}{9}$$

$$= 332.8$$

Sum of Squares

Why Square The Differences?

- Get rid of negatives, so that the negatives and positives do not cancel each other during addition.
- Increase larger deviations more than smaller ones so that they are weighed more heavily.



$(2-4) + (6-4) = -2 + 2 = 0$

Standard Deviation (SD)

- Square root of Variance
- It has the same units as the variable, which makes it useful in comparisons and calculations

- Sample SD

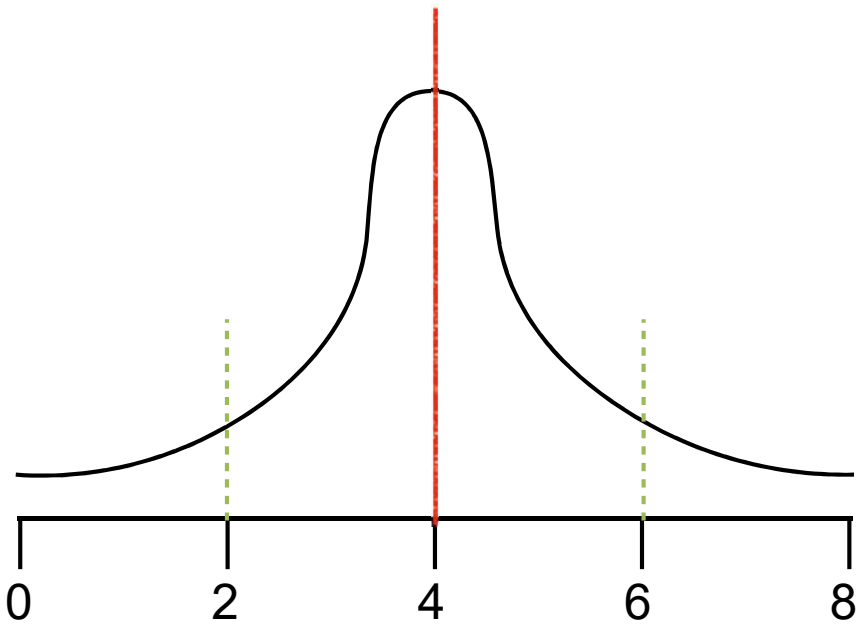
$$s = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

- Population SD

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

DICE ANALYTICS

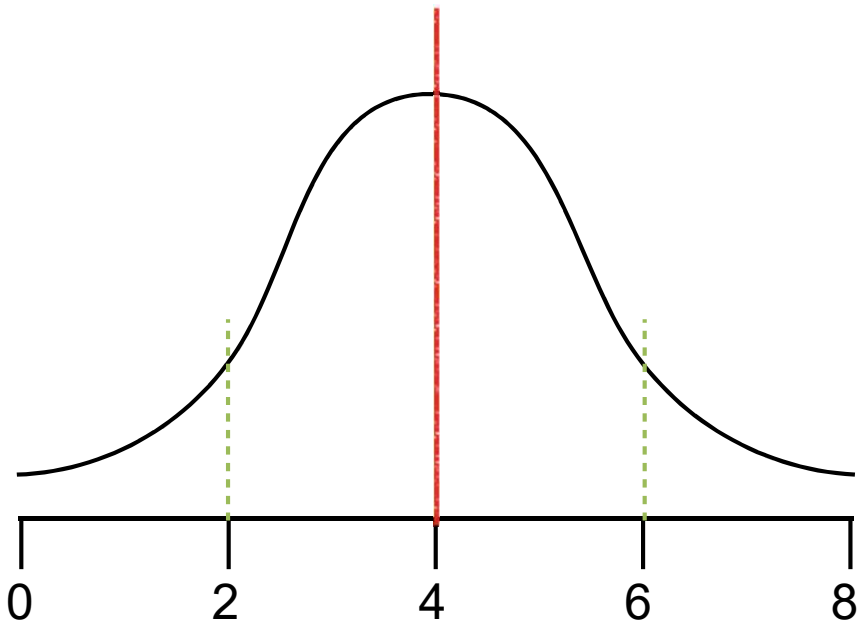
Spread



Less Spread

Low Variance

Low Deviation



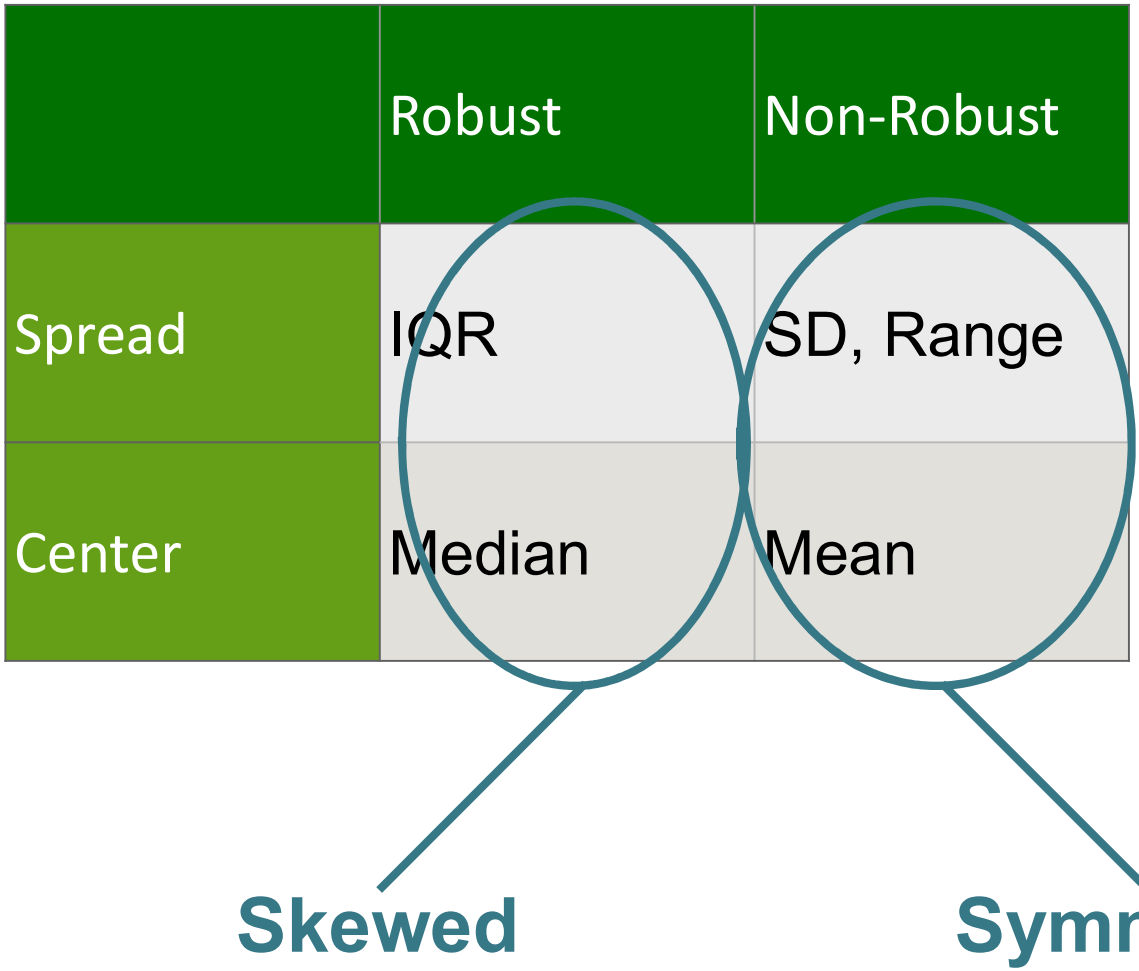
More Spread

High Variance

High Deviation

Robust Statistics

- Measures on which extreme observations or outliers have little effect



Normal Distribution

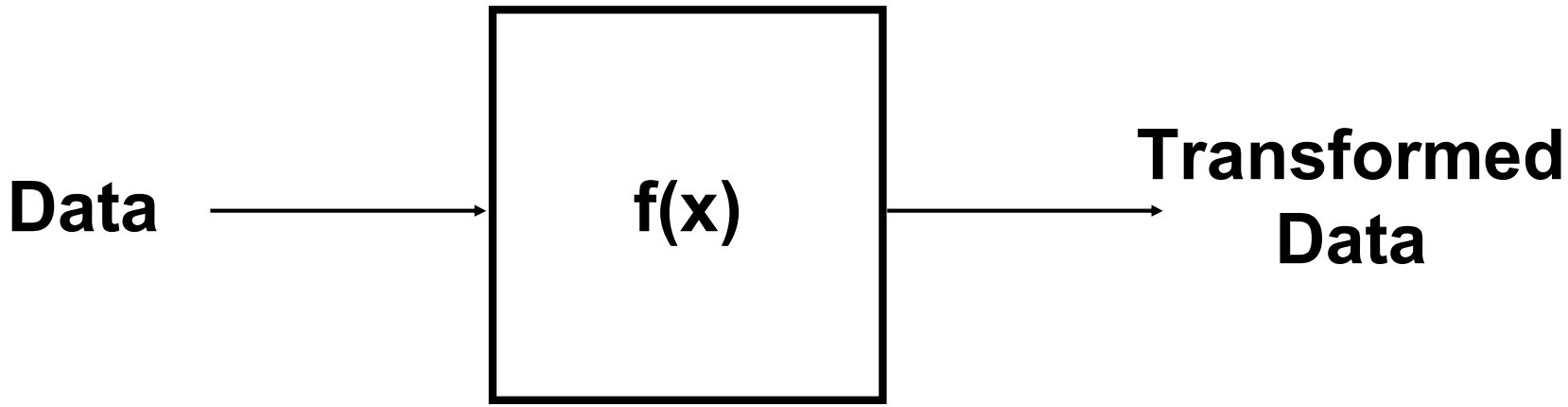
Mean Standard Deviation Variance

Non-Normal Distribution

Median IQR Range

Data Transformations

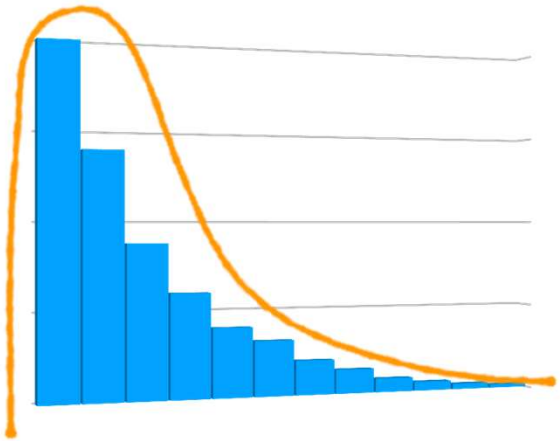
- Applying a Function f(x) to adjust scales of data.
- Done usually when data is skewed, so that it becomes easier to perform *modelling.*
- Done to convert non-linear relationship into a linear relationship.



(Natural) Log Transformation

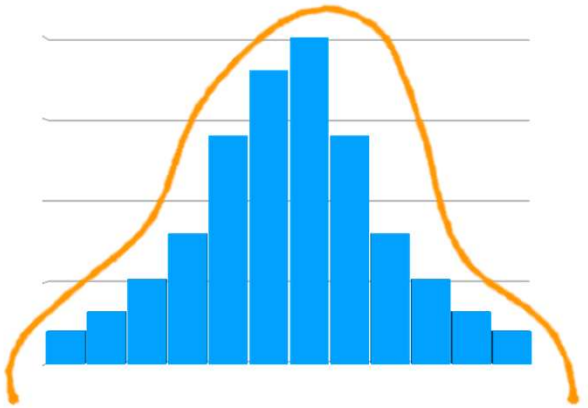
- To transform data that is positively skewed
- Usually done when data is concentrated near Zero (relative to the few large values in data)

Right Skewed



Natural Log

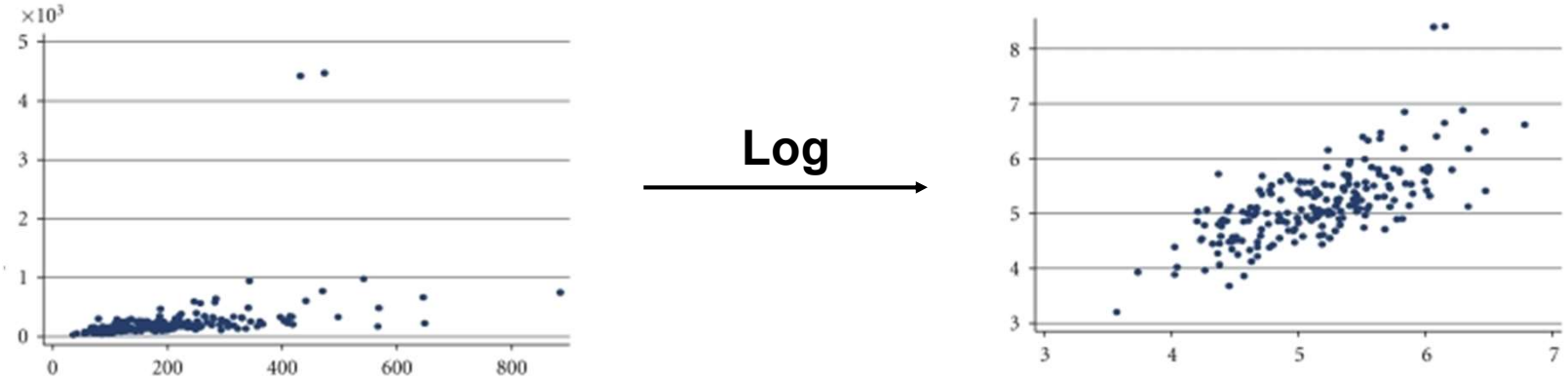
Symmetric



DICE ANALYTICS

Log Transformation

- To make the relationship between two variable more linear
- Most of the simple methods for modelling work only when relationship is linear



DICE ANALYTICS

Other Transformation

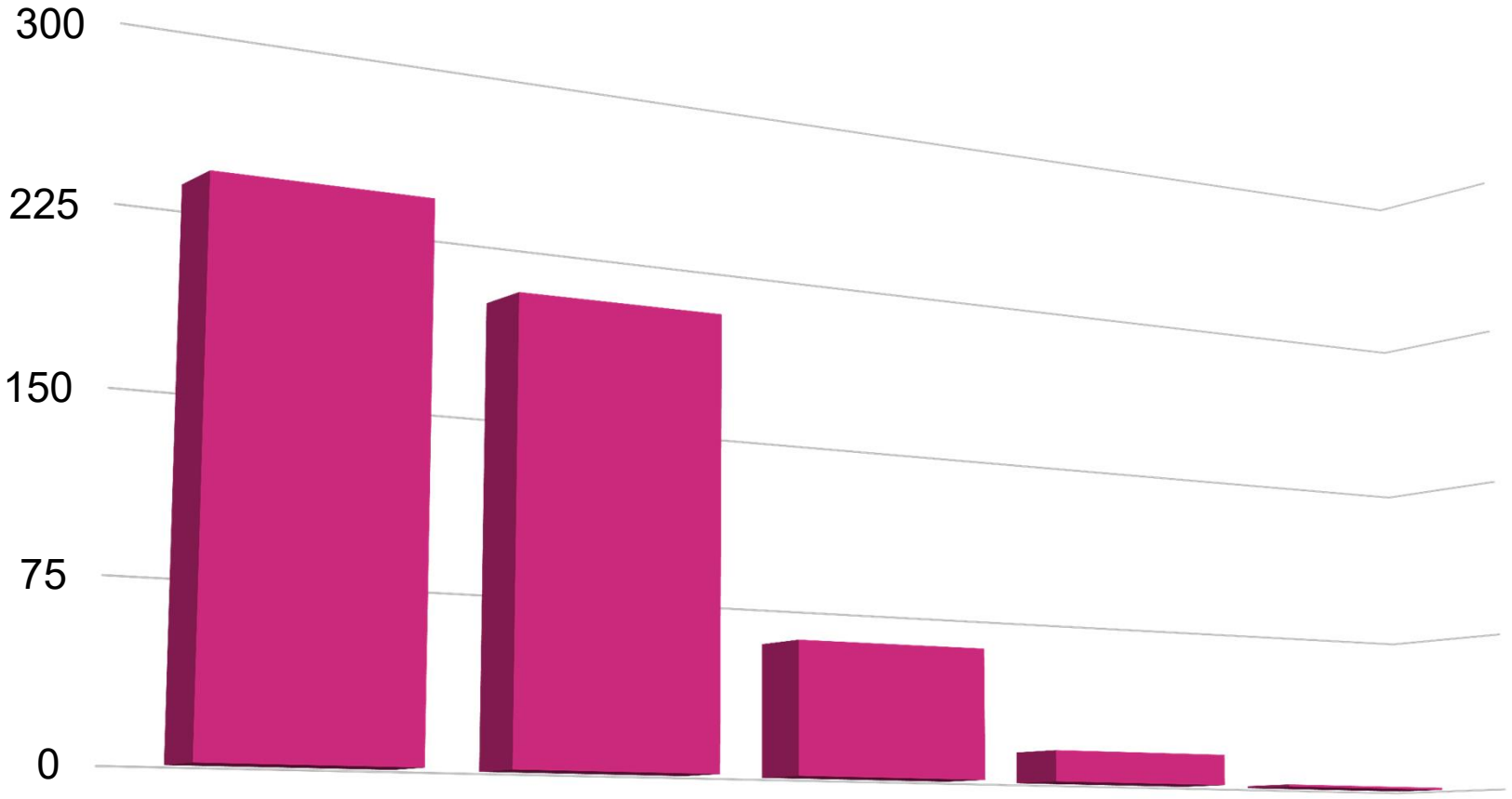
- You may use other transformations or create of your own
- For instance: Square Root, Square, Inverse

Visualising Categorical Data

Frequency Table

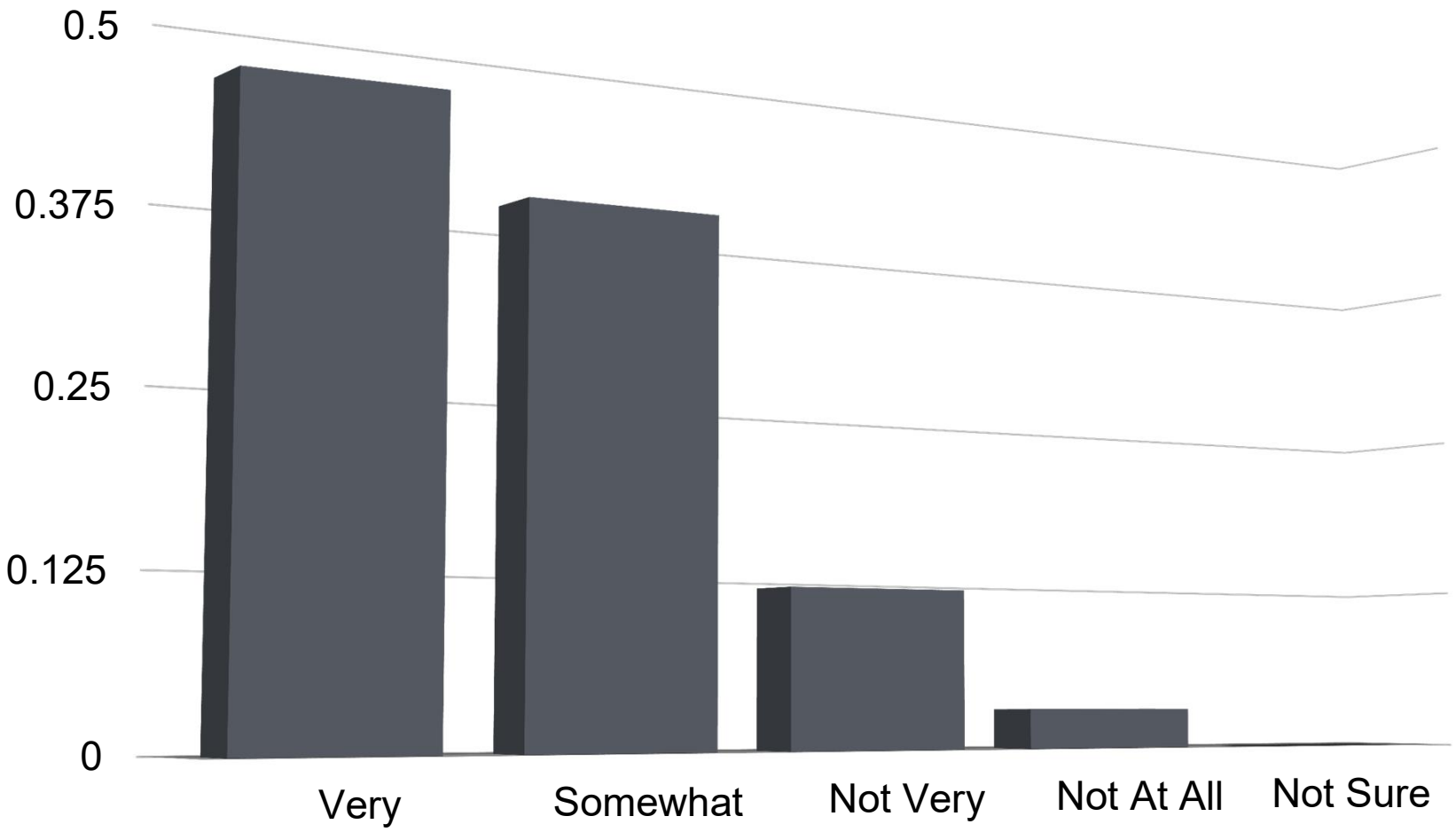
Difficulty Saving Energy	Frequency	Cumulative Frequency	Frequency (%)
Very	231	231	46%
Somewhat	196	427	39%
Not Very	58	485	12%
Not At All	14	499	3%
Not Sure	1	500	~0%
Total	**500**		

Bar Plot



Frequency

Bar Plot

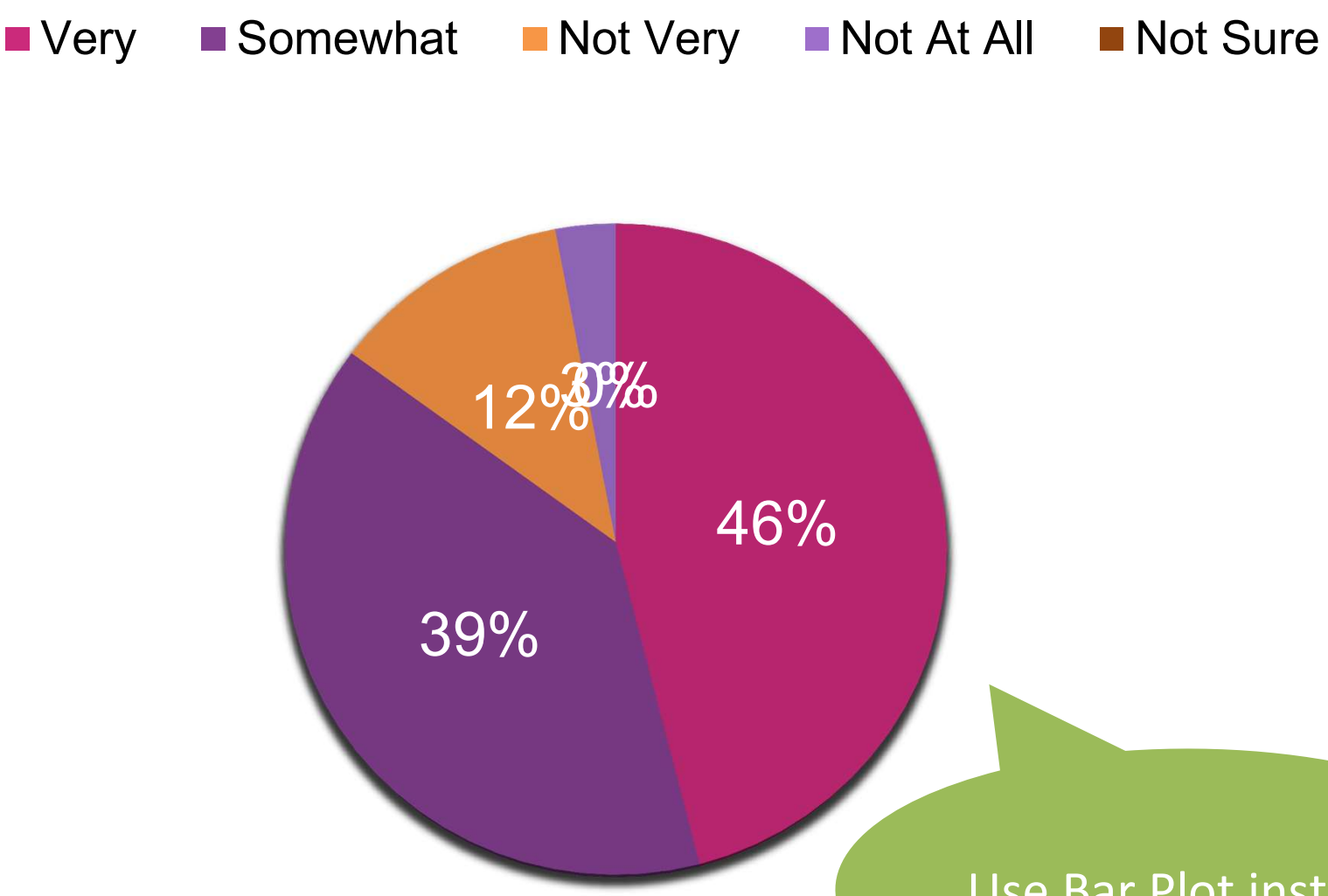


Frequency (%)

Bar Plot vs Histogram

- Bar Plot for Categorical Variables, Histogram for Numerical Variables
- X-axis in Histogram must be a Number Line
- Ordering of bars is not interchangeable in Histogram as compared to Bar Plot

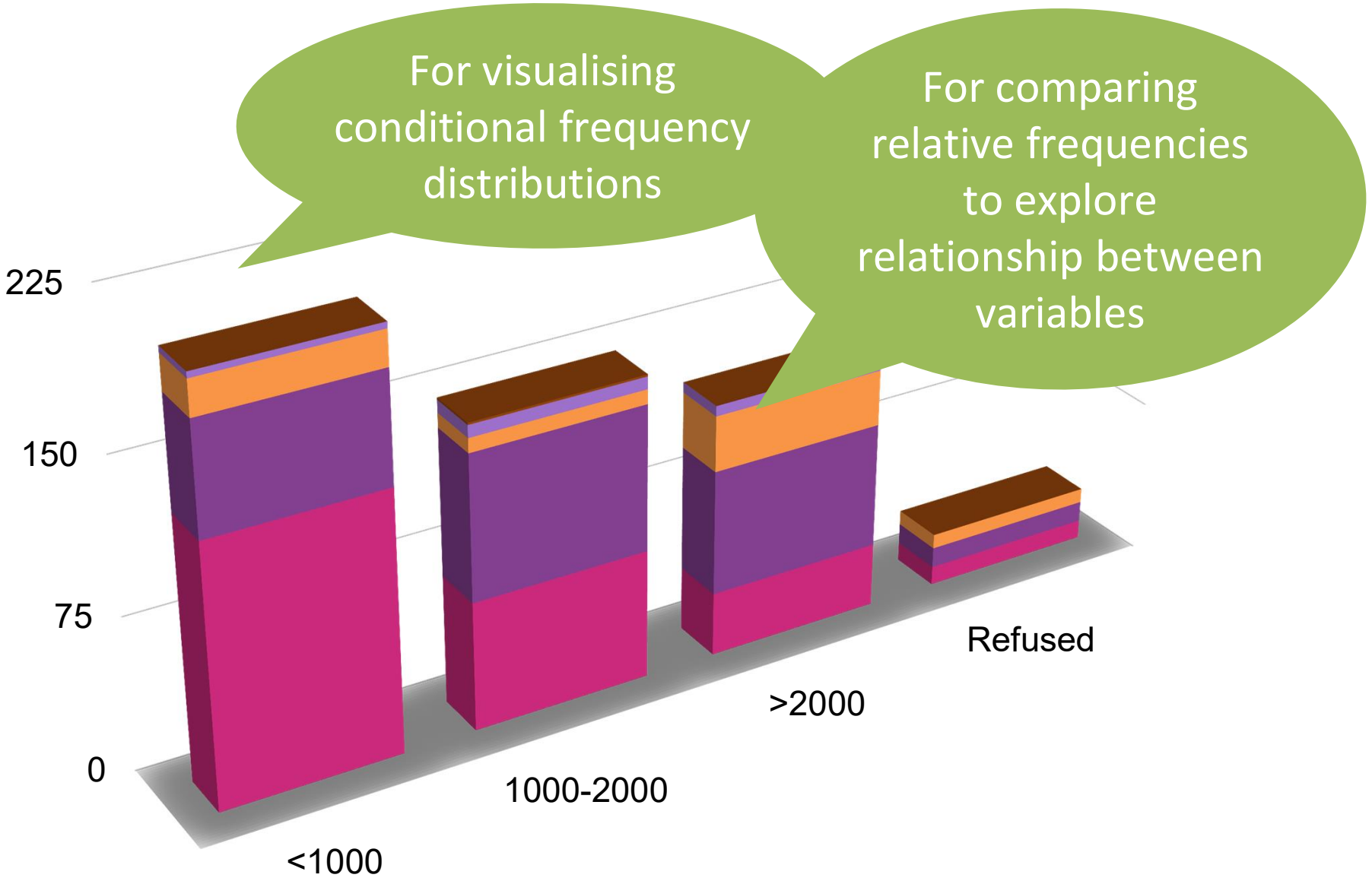
Pie Chart



Contingency Table

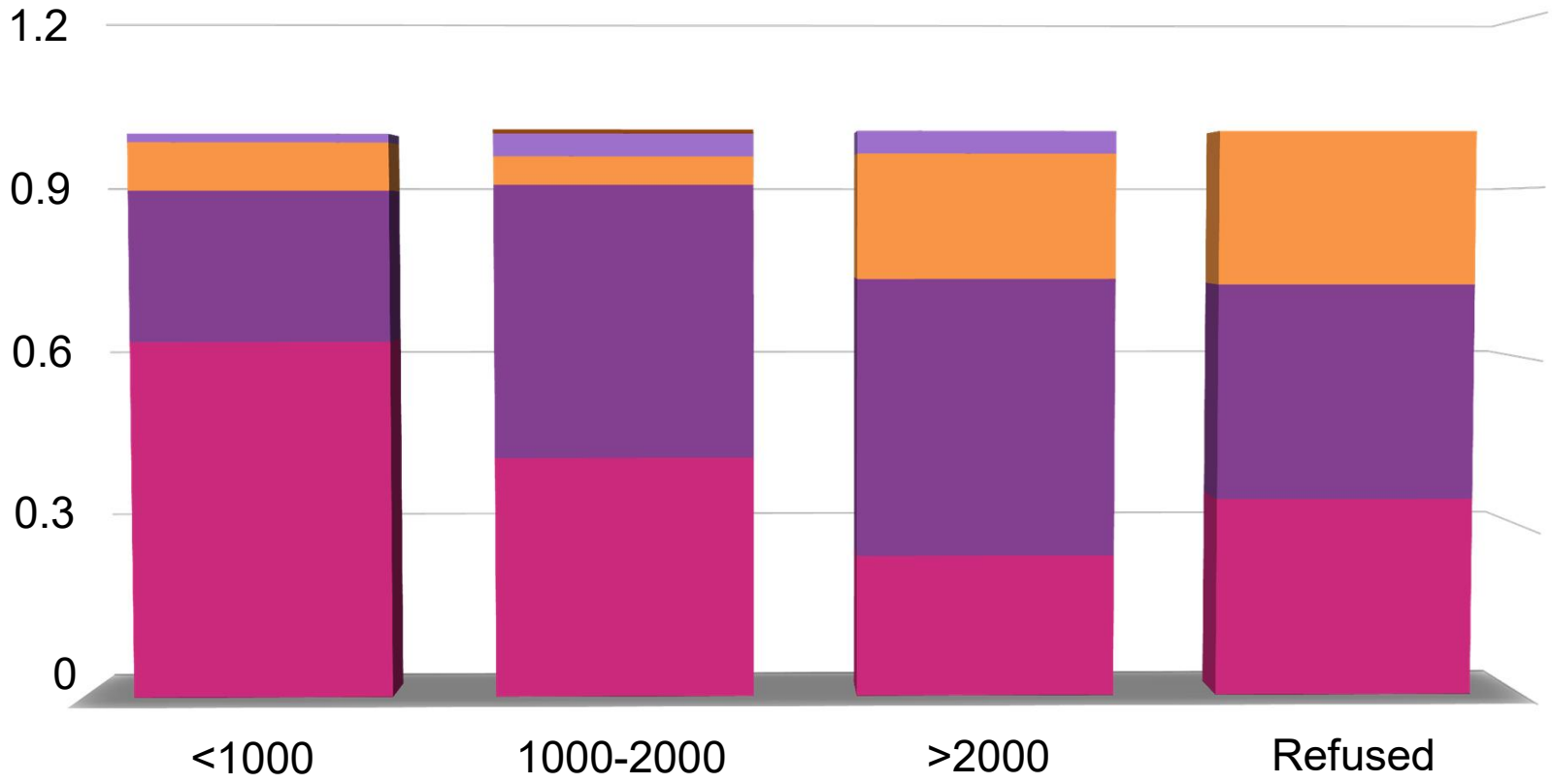
		Calories (cal)				
		<1000	1000-2000	>2000	Refused	Total
Difficulty Saving Energy	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not Very	17	7	27	7	58
	Not At All	3	6	5	0	14
	Not Sure	0	1	0	0	1
	Total	202	148	124	26	500

Segmented Bar Plot



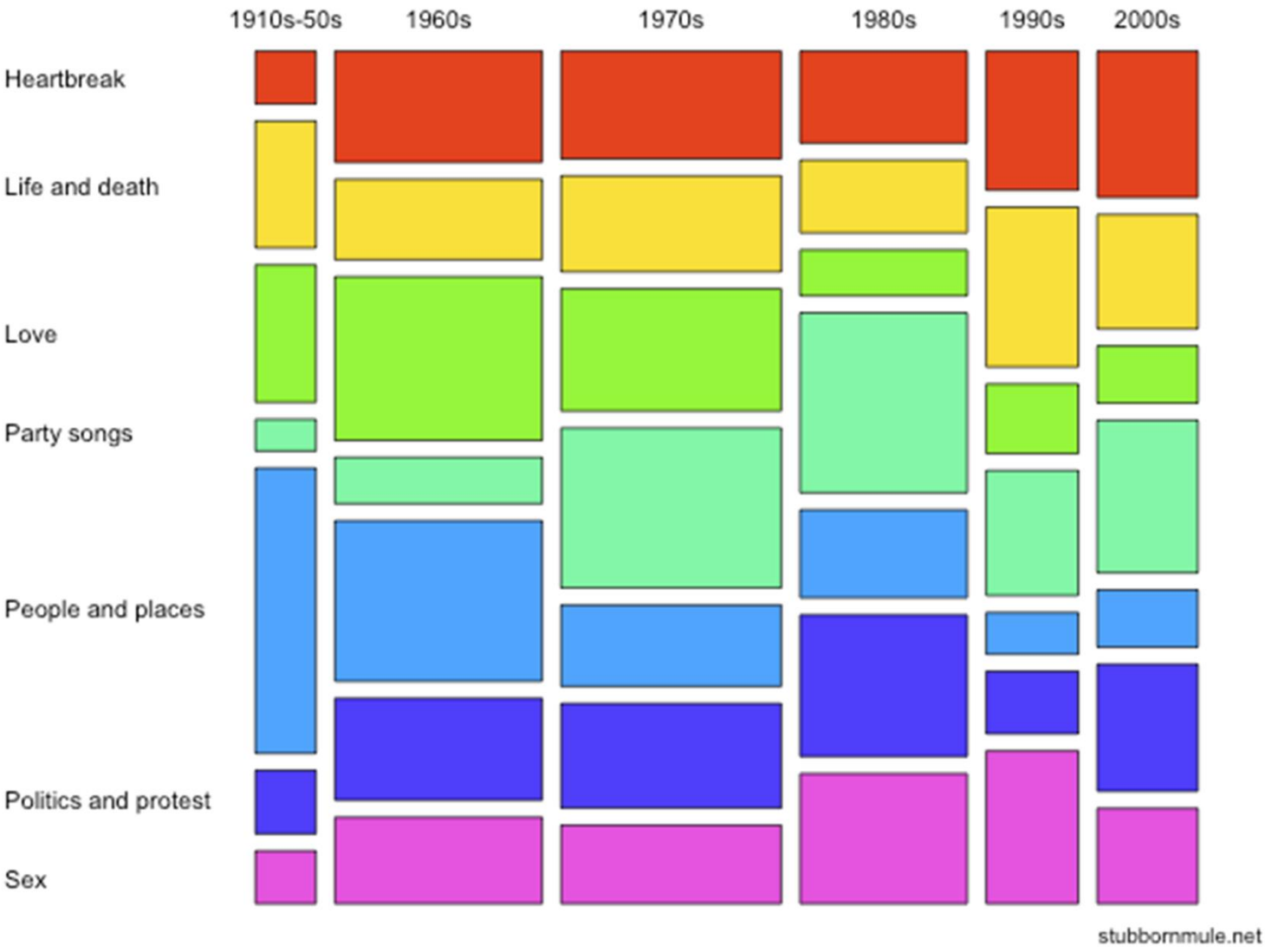
DICE ANALYTICS

Relative Frequency Segmented Bar Plot

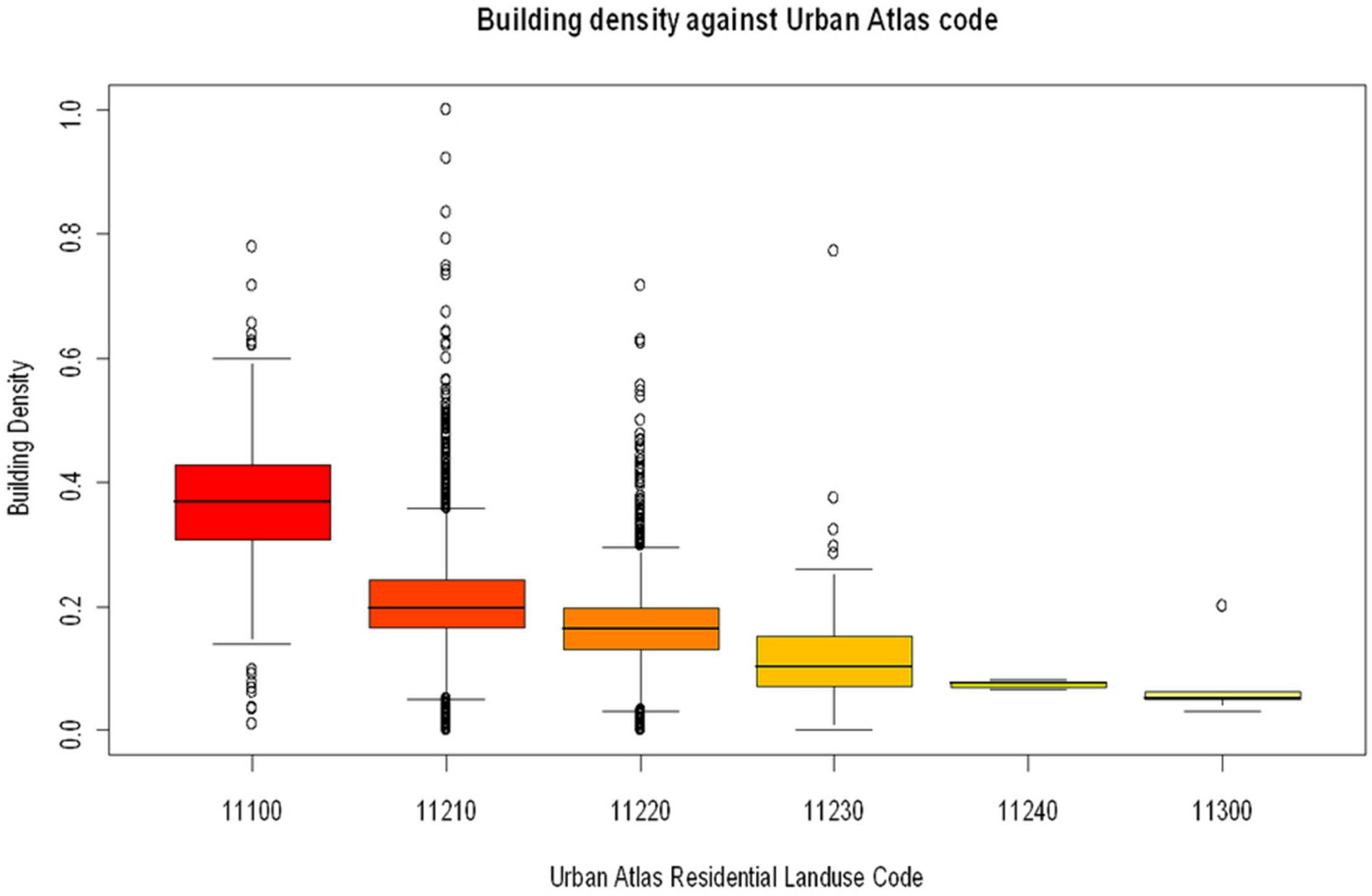


DICE ANALYTICS

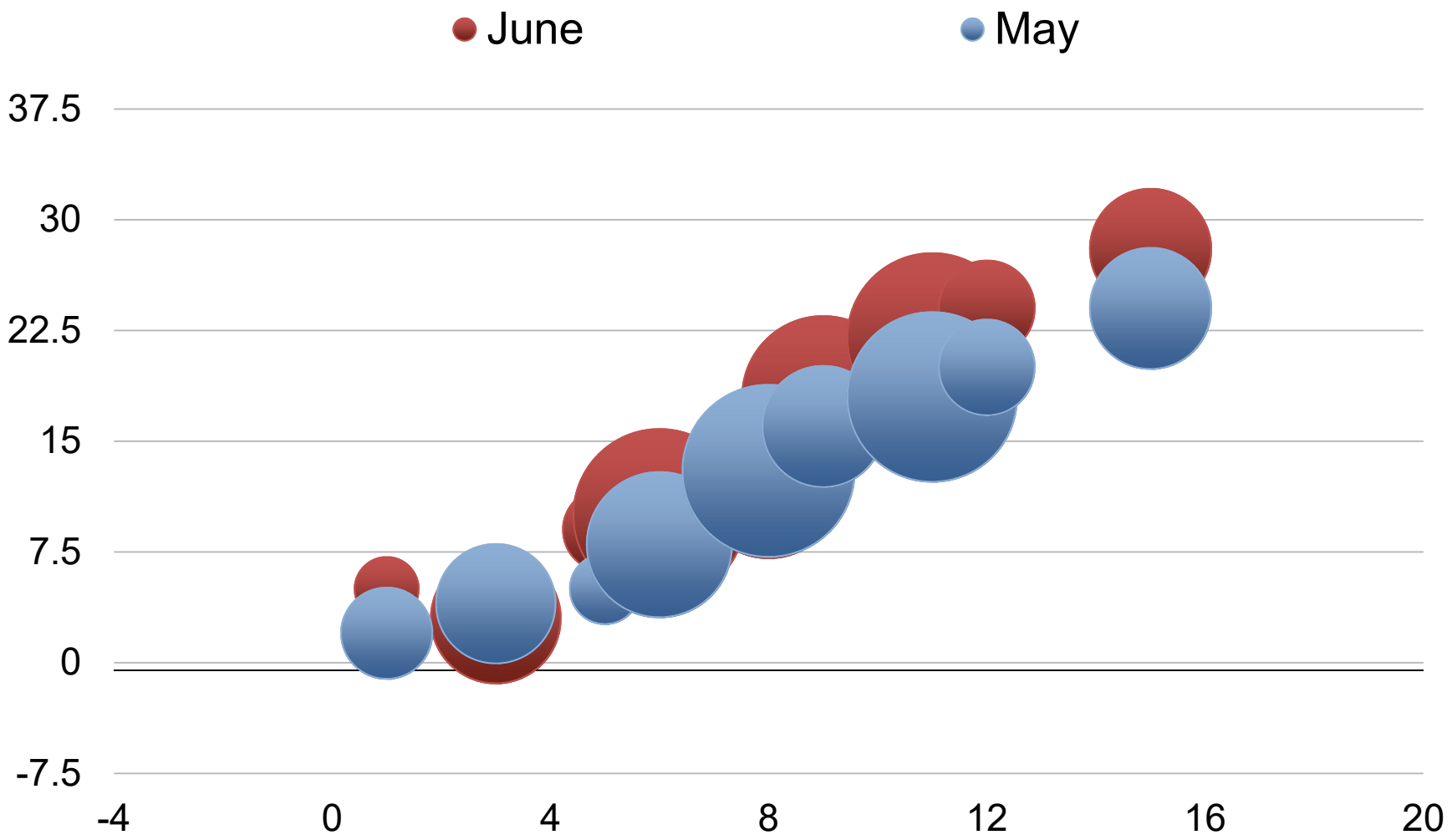
Mosaic Plot



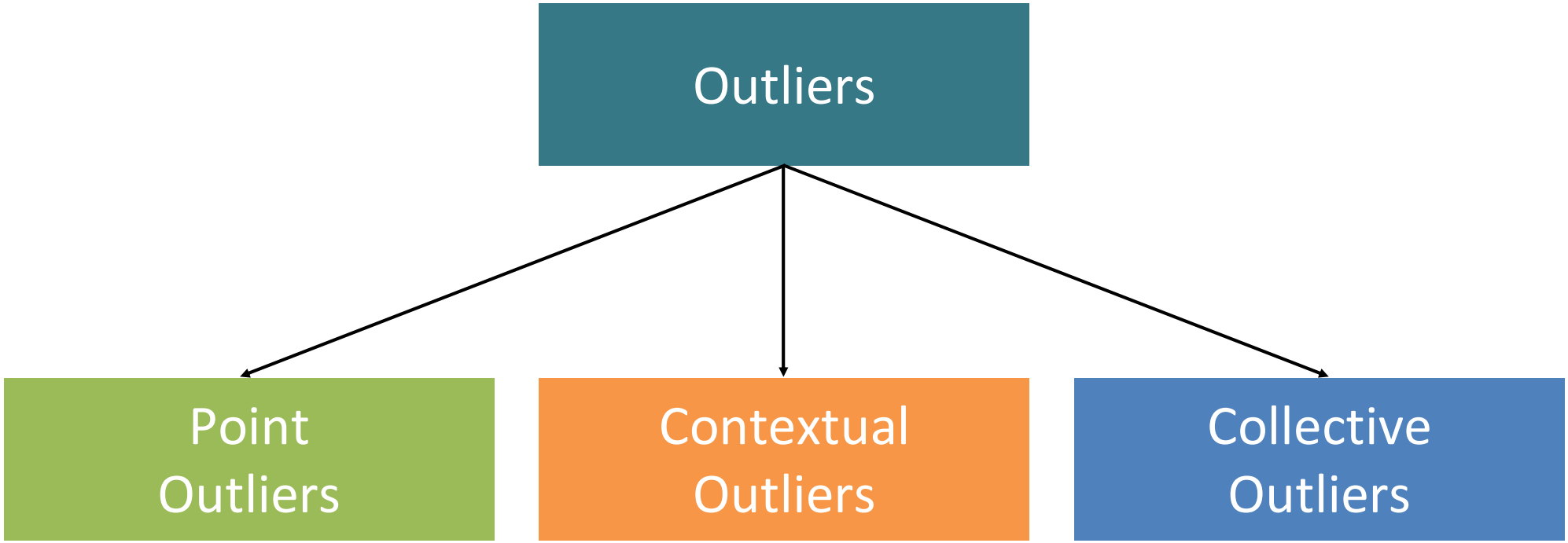
Side-by-Side Box Plots



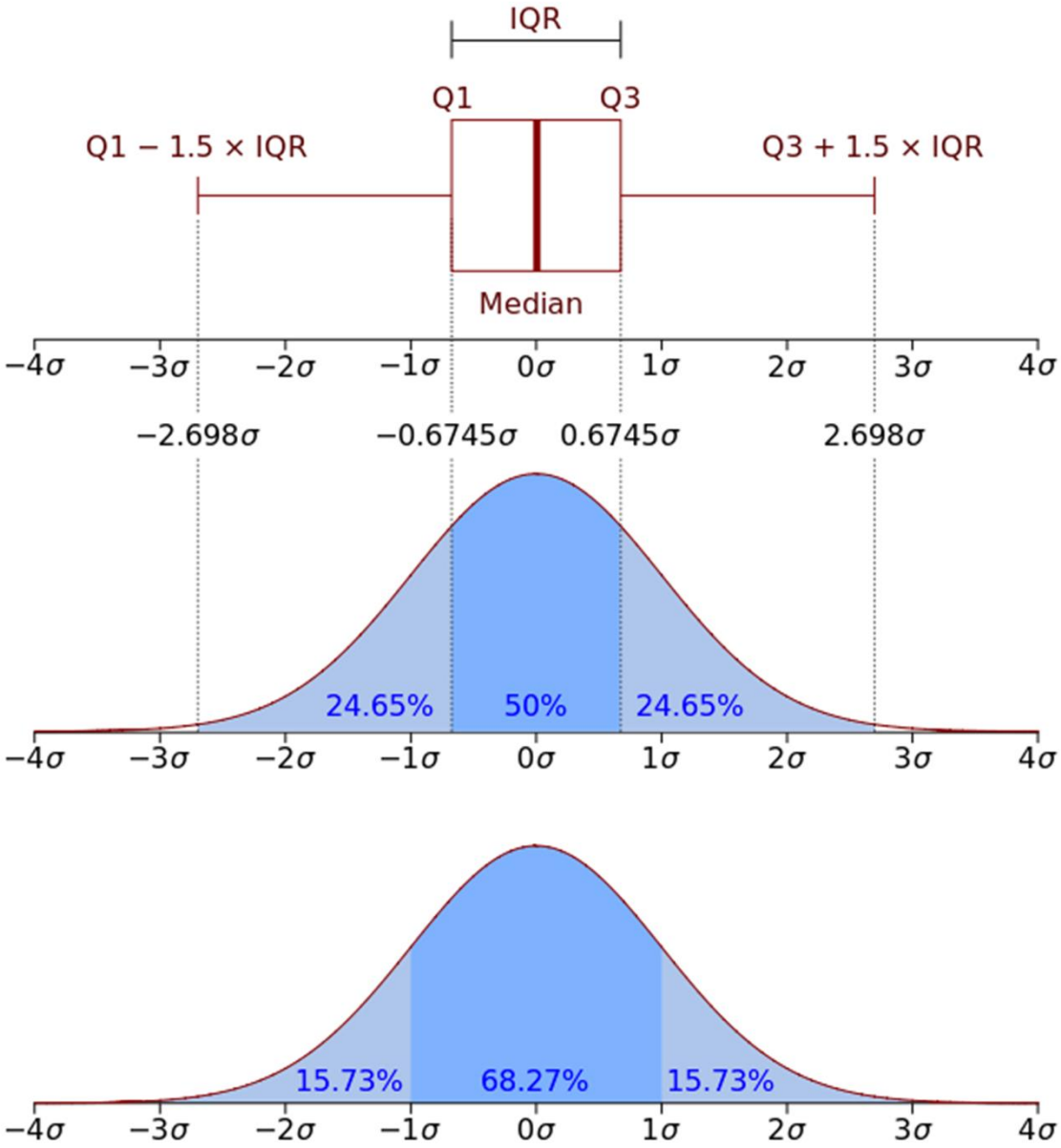
Bubble Plot



Outliers



Outliers



Principles of Analytic Graphs

- Show comparisons
- Show causality, mechanism, explanation, systematic structure
- Show multivariate data
- Integrate multiple modes of evidence
- Describe and document the evidence
- Content is king

(From JHU)

DICE ANALYTICS

Why do EDA

- To understand data properties
- To find patterns in data
- To suggest modelling strategies
- To "debug" analyses
- To communicate results

(From JHU)

Probability Basics

Random Process

- It is a function of time
- We know all the possible outcomes
- But we don’t know which outcome will occur exactly
- Also known as ***Stochastic Process***

Probability

- Quantifiable likelihood (chance) of the occurrence of an event expressed as odds, or a fraction of 1.
- Notation : P(A) = 0.3, read as Probability of Event A is 0.3 or 30%

- Example : What is the probability that we get a Head after a coin toss?

  P(H) = 0.5

Frequentist Interpretation

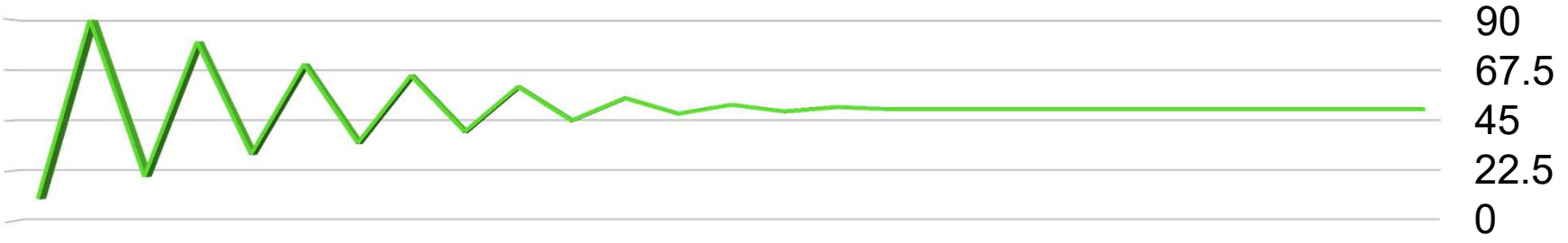
- It is an **interpretation** of **probability**; it defines an event's **probability** as the limit of its relative frequency in a large number of trials.

Bayesian Interpretation

- It is an **interpretation** of the concept of **probability**, in which, instead of frequency or propensity of some phenomenon, **probability** is interpreted as reasonable expectation representing a state of knowledge or as quantification of a personal belief.

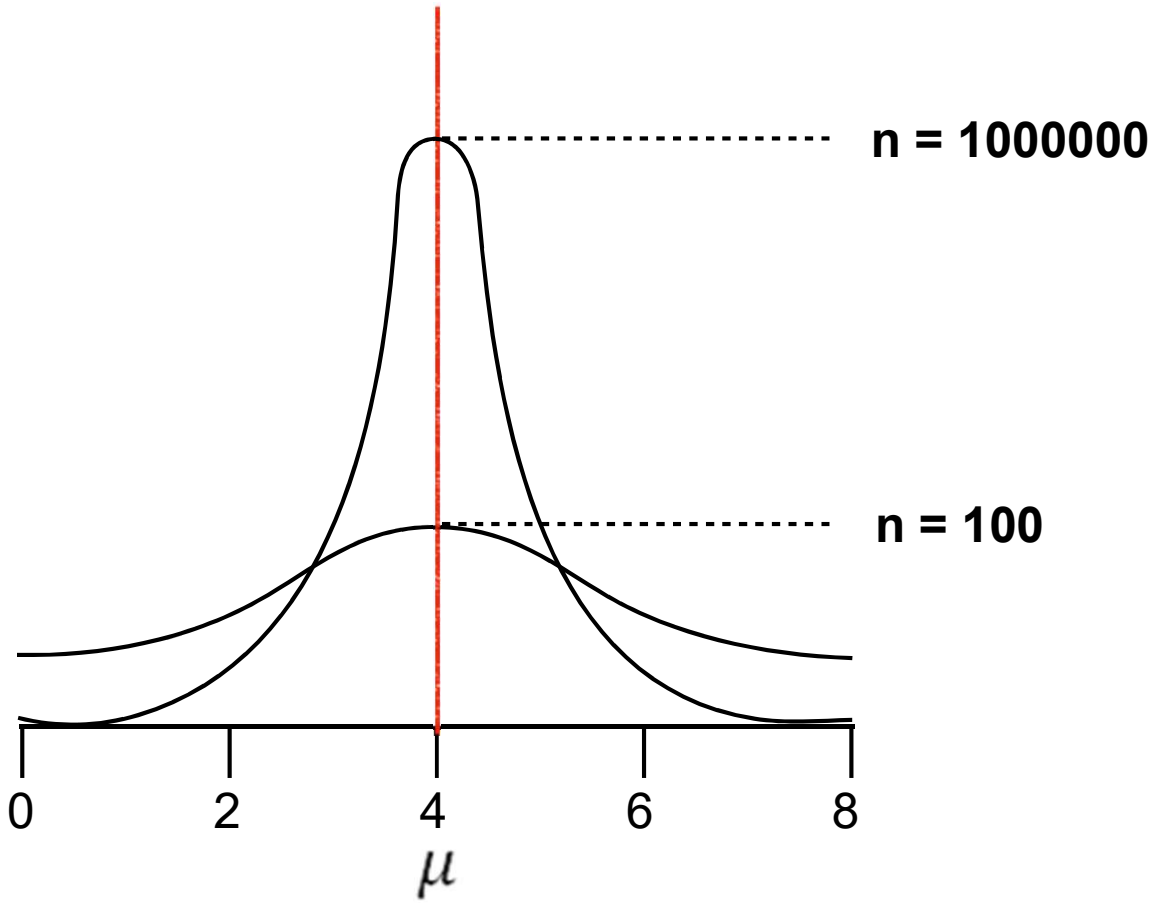
Law of Large Numbers

- The **law of large numbers** is a principle of probability according to which the frequencies of events with the same likelihood of occurrence even out, given enough trials or instances.
- As the **number** of experiments increases, the actual ratio of outcomes will converge on the theoretical, or expected, ratio of outcomes.



DICE ANALYTICS

Law of Large Numbers



PMF

- **PMF** stands for probability mass function.
- As it name suggest, it gives the probability of each number in the data set or you can say that it basically gives the count or frequency of each element.
- It is calculated for Discrete Random Variables.

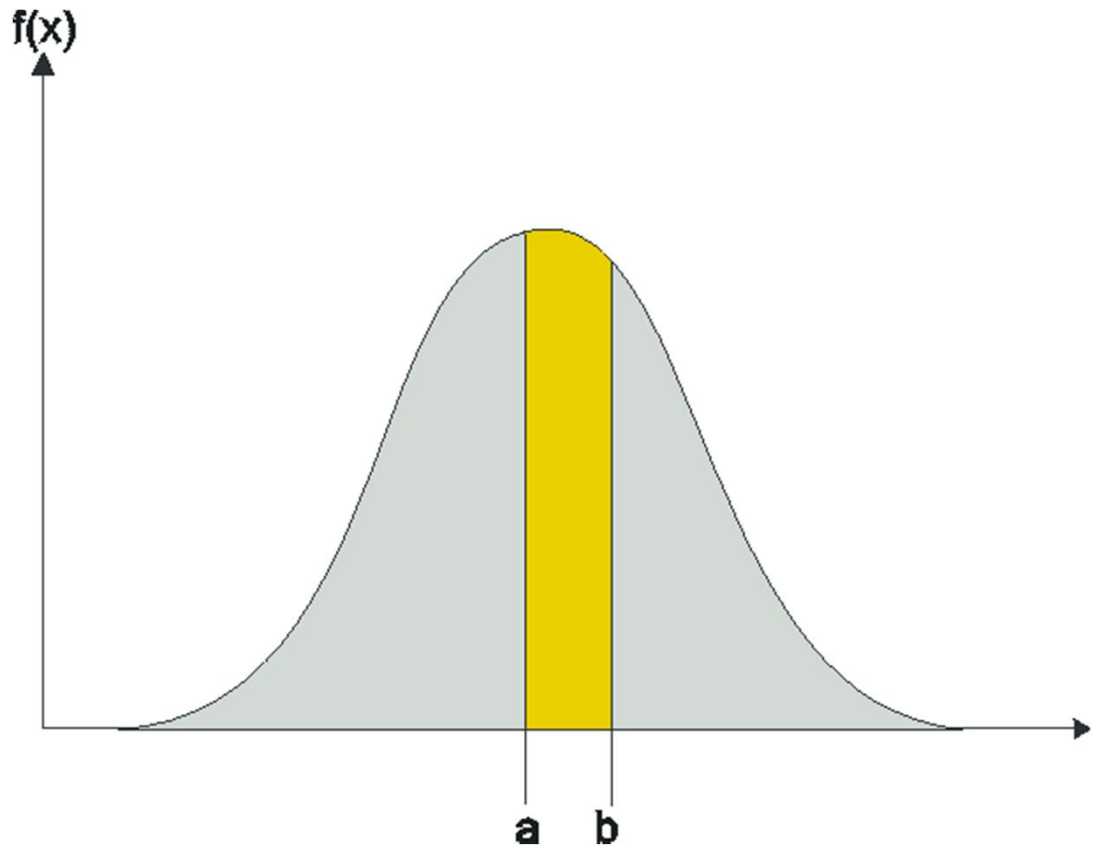
1	2	7	5	6
7	2	3	4	5
0	1	5	7	3
1	2	5	6	7
6	1	0	3	4

PMF

Value	Count	PMF
0	2	2/25
1	4	4/25
2	3	3/25
3	3	3/25
4	2	2/25
5	4	4/25
6	3	3/25
7	4	4/25

PDF

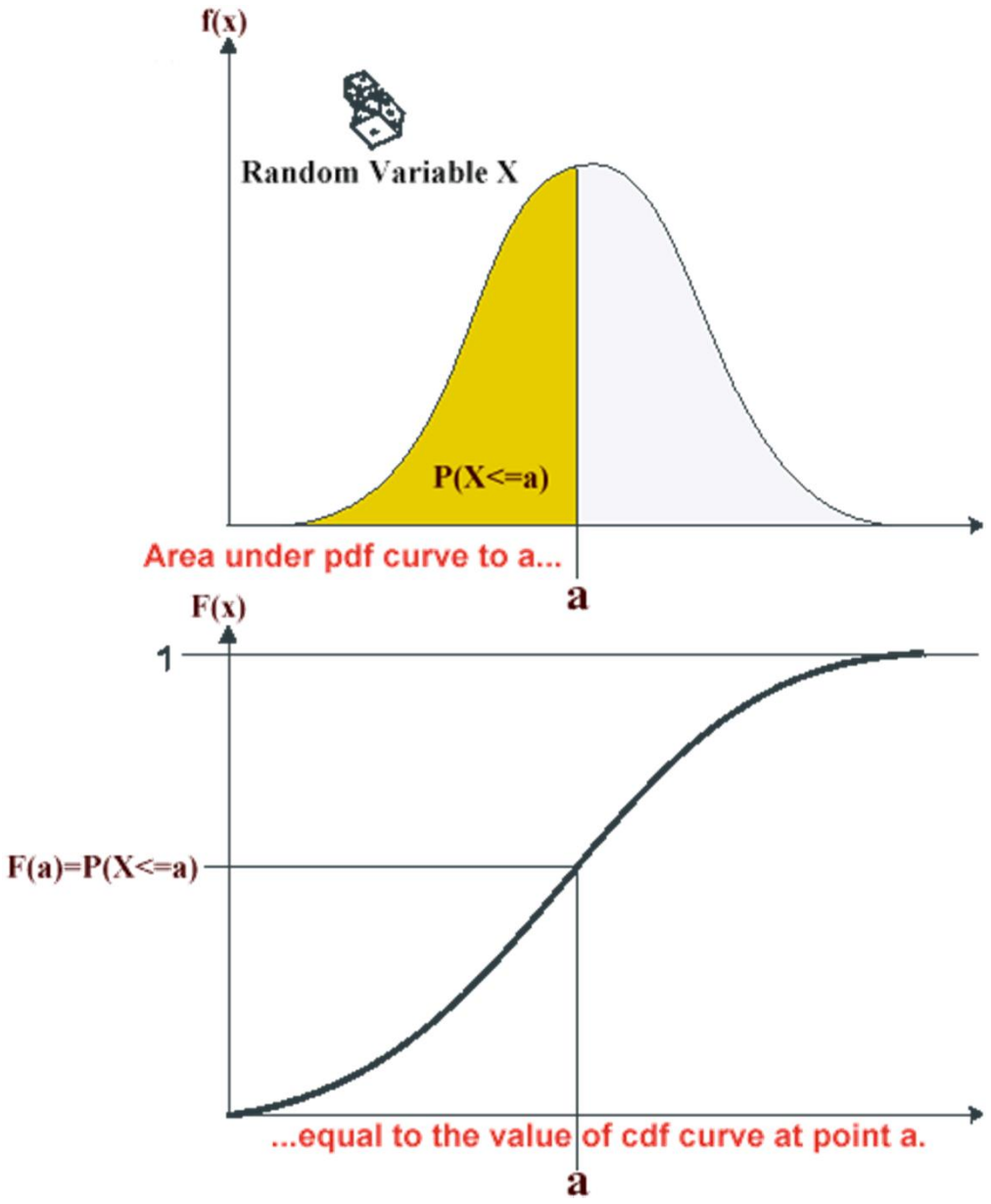
- A function of a continuous random variable, whose integral across an interval gives the probability that the value of the variable lies within the same interval.



CDF

- **CDF** stands for cumulative distributive function.
- It is a function that calculates the cumulative sum of all the values that are calculated by PMF.
- It basically sums the previous one.
- It can be calculated for both discrete random variable and continuous random variable.

PDF vs CDF

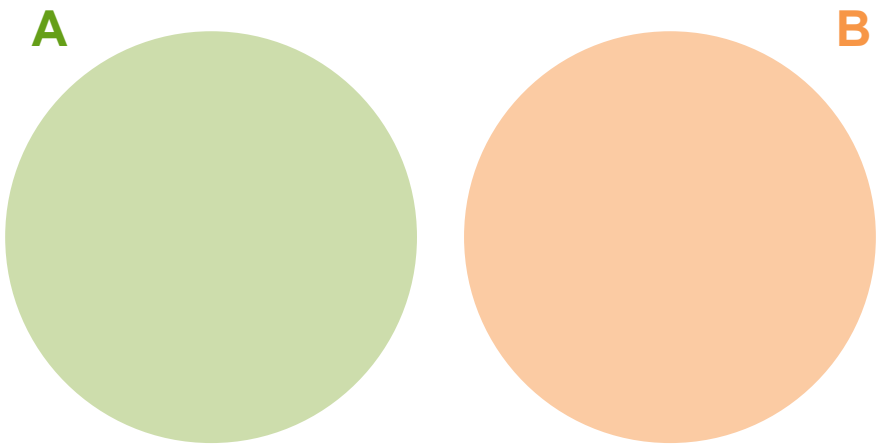


DICE ANALYTICS

Disjoint Events

- Events that are *mutually exclusive* and cannot happen at the same time.

- Example : Result of a coin toss can either be Head or Tail, not both.

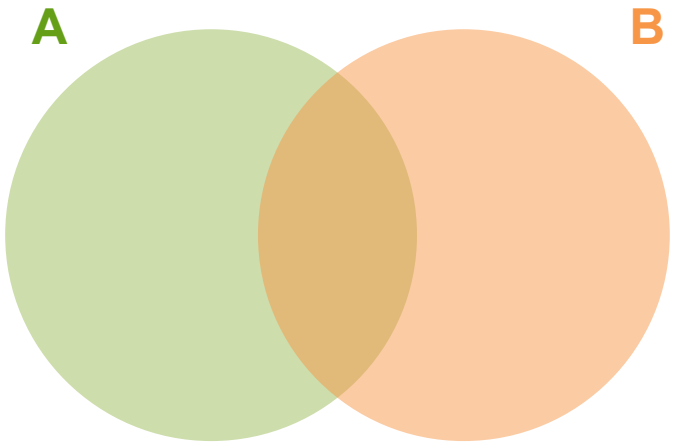


P (A and B) = 0

Non-Disjoint Events

- Events that are not *mutually exclusive* and can happen at the same time.

- Example : A person can like both Biryani and Pizza



P (A and B) ≠ 0

Union of Events

- Disjoint Events : P(A or B) = P(A) + P(B)

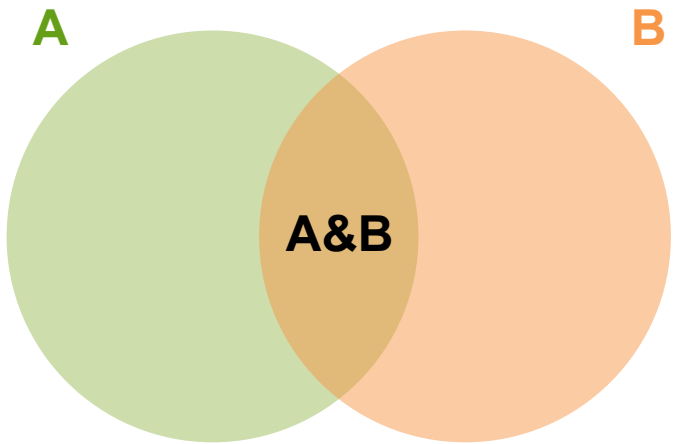
Because P(A and B) = 0

- Non-Disjoint Events : P(A or B) = P(A) + P(B) - P(A and B)

Because $P(A \text{ and } B) \neq 0$

General Addition Rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



Sample Space

- Collection of all possible outcomes of a trial
- Example : What is the sample space of an experiment where two coins are tossed?

S = {HH, HT, TH, TT}

Probability Distributions

- It lists all possible outcomes in the sample space, and the probabilities with which they occur.

One Toss	H	T
Probability	0.5	0.5

Two Tosses	HH	HT	TH	TT
Probability	0.25	0.25	0.25	0.25

Rules :

1. All events must be disjoint
2. Each probability must be between 0 and 1
3. All probabilities must add up to 1

Complimentary Events

- Two Events that are *mutually exclusive* and their probabilities add up to 1.

One Toss	H	T
Probability	0.5	0.5

Complimentary

Two Tosses	HH	HT	TH	TT
Probability	0.25	0.25	0.25	0.25

Complimentary

DICE ANALYTICS

Disjoint vs Complimentary Events

- Will sum of probabilities of two disjoint events be always equal to 1? NO

- Will sum of probabilities of two complimentary events be always equal to 1? YES

Independent Process

- Two processes are said to be independent if knowing the outcome of one provides no useful information about the other.

DICE ANALYTICS

Determining Dependence

- If observed differences between conditional probabilities is high, we can say that there is dependence.

- If sample size is large enough, even small difference in probabilities can be interpreted as dependence.

Product Rule

- If A and B are independent, then

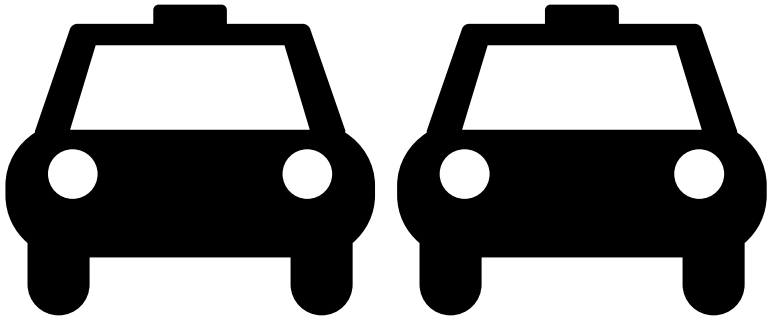
$$P(A \text{ and } B) = P(A) \times P(B)$$

- If $A_1, A_2, \ldots\ldots, A_k$ are independent, then

$$P(A_1 \text{ and } A_2 \text{ and } \ldots A_k) = P(A_1) \times P(A_2) \times \ldots. \times P(A_k)$$

Disjoint vs Independence

Disjoint & Dependent



Dependent & Independent

DICE ANALYTICS

Study

RESULTS:		Objective Social Class Position		
		Working Class	Upper Middle Class	Total
Subjective Social Class Identity	Poor	0	0	0
	Working Class	8	0	8
	Middle Class	32	13	45
	Upper Middle Class	8	37	45
	Upper Class	0	0	0
	Total	48	50	98

Marginal Probability

What is the probability that a student's objective social position is upper middle class?

P (obj UMC) = 50 / 98 = 0.51

Note that the counts used to calculate Marginal Probability came from the *margins* of Contingency Table.

Joint Probability

What is the probability that a student's objective position *and* subjective identity are both upper middle class?

P (obj UMC & sub UMC) = 37 / 98 = 0.38

Note that the counts used to calculate Joint Probability came from the *intersection* of Contingency Table.

Conditional Probability

What is the probability that a student who is objectively in the working class associates himself with upper middle class?

P (sub UMC | obj WC) = 8 / 48 = 0.17

Note that we first *conditioned* on the working class and then calculated probability using counts only in this column.

Bayes' Theorem

$$P(A \mid B) = \frac{P(A \,\&\, B)}{P(B)}$$

$$\textbf{P (sub UMC | obj WC)} = \frac{\textbf{P (sub UMC \& obj WC)}}{\textbf{P (obj WC)}} = \frac{\textbf{8 / 98}}{\textbf{48 / 98}} = \textbf{0.17}$$

DICE ANALYTICS

General Product Rule

$$P(A \& B) = P(A \mid B) \times P(B)$$

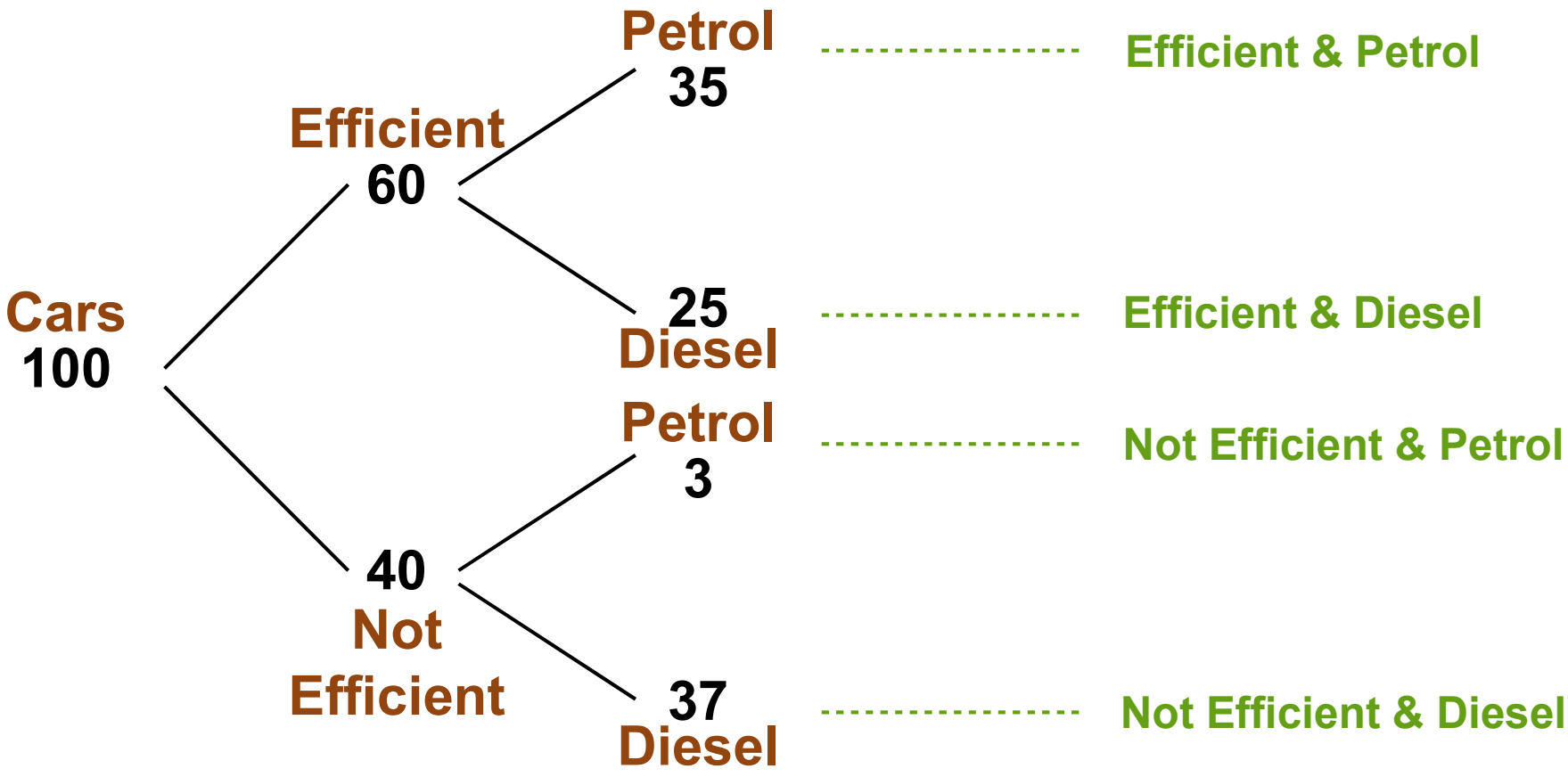
Useful when we don't know whether events are independent or dependent

Probability Trees

Study

There are 100 cars. 60 are fuel efficient while 40 are not fuel efficient. Of the 60 fuel efficient cars, 35 are Petrol while 25 are Diesel. Of the rest, 3 are Petrol while 37 are Diesel.

Probability Tree



Probability Tree (Examples)

If the car is Petrol, what is the probability that it is Fuel Efficient?

$$P(\text{Efficient} \mid \text{Petrol}) = \frac{35}{35 + 3} = 0.92$$

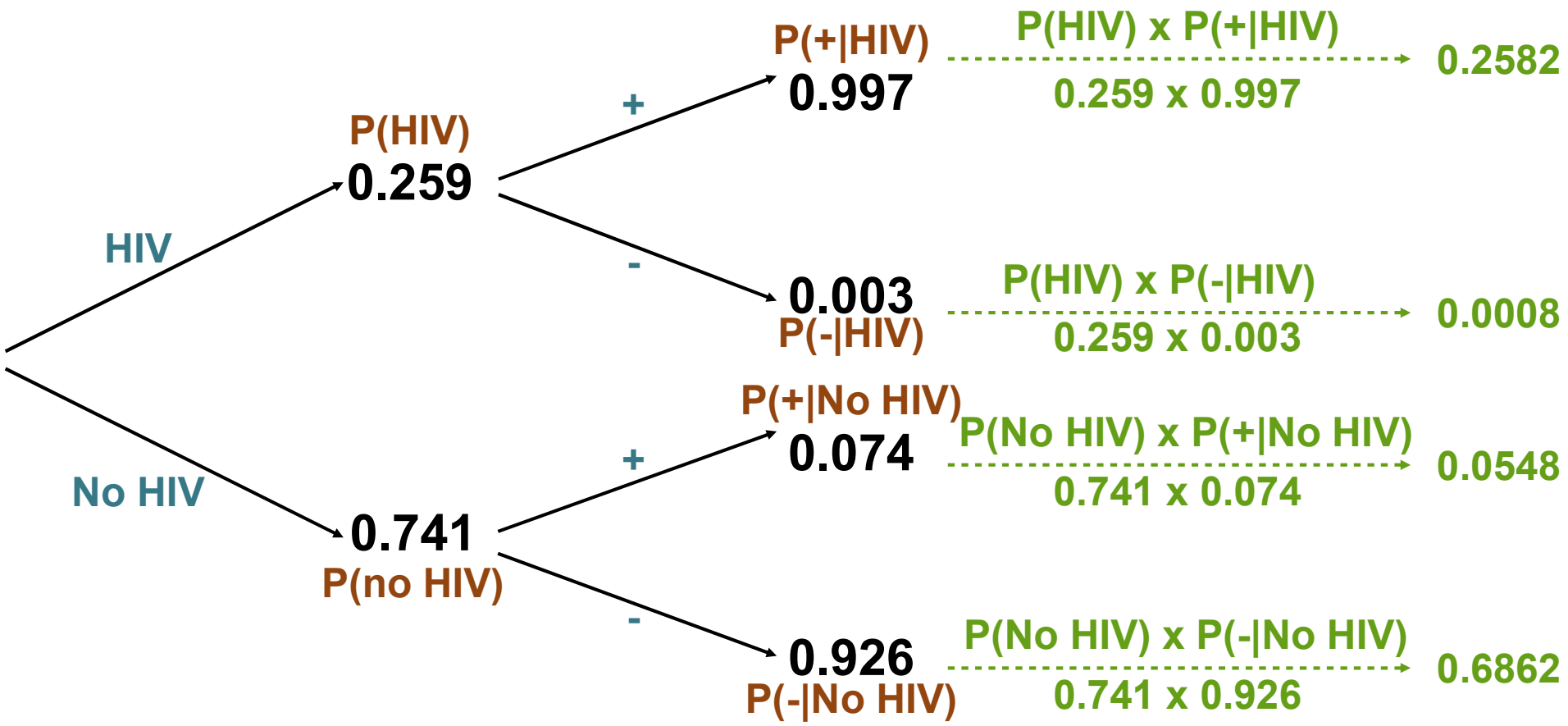
Not that this is a Conditional Probability

Have we made use of Bayes' Theorem here?

Study

As of 2009, Swaziland, has the highest HIV prevalence in the world. 25.9% of this country's population is infected with HIV. The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate.

Probability Tree



Note that P(HIV) x P(+|HIV) means P(HIV & +), recall General Product Rule!

DICE ANALYTICS

Probability Tree (Examples)

If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

$$P(HIV \mid +) = \frac{P(HIV \,\&\, +)}{P(+)} = \frac{0.2582}{0.2582 + 0.0548} = 0.82$$

Not that this is a Conditional Probability

Have we made use of Bayes' Theorem here?

DICE ANALYTICS