Machine Learning

Organizing Data: Frequency Distribution

Organizing Data

- **Frequency Distribution:**
  - A **frequency distribution** is the organization of raw data in table form, using classes and frequencies.
- As an example, consider the following data set:

49	**57**	**38**	**73**	**81**
74	**59**	**76**	**65**	**69**
54	**56**	**69**	**68**	**78**
65	**85**	**49**	**69**	**61**
48	**81**	**68**	**37**	**43**
78	**82**	**43**	**64**	**67**
51	**56**	**81**	**77**	**79**
85	**40**	**85**	**59**	**80**
60	**71**	**57**	**61**	**69**
61	**83**	**90**	**87**	**74**

Class limits	Tally	Frequency
35-41	///	3
42-48	///	3
49-55	////	4
56-62	𝍸 𝍸	10
63-69	𝍸 𝍸	10
70-76	𝍸	5
77-83	𝍸 𝍸	10
84-90	𝍸	5
	Total = 50	

Organizing Data

- **Categorical Frequency Distribution:**
  - The **categorical frequency distribution** is used for data that can be placed in specific categories, such as nominal- or ordinal-level data.
- Let us consider the following distribution of blood types:

Twenty-five army inductees were given a blood test				
A	**B**	**B**	**AB**	**O**
O	**O**	**B**	**AB**	**B**
B	**B**	**O**	**A**	**O**
A	**O**	**O**	**O**	**AB**
AB	**A**	**O**	**B**	**A**

A Class	B Tally	C Frequency	D Percent
A	𝍸	5	20
B	𝍸 //	7	28
O	𝍸 ////	9	36
AB	////	4	16
	Total	**25**	**100**

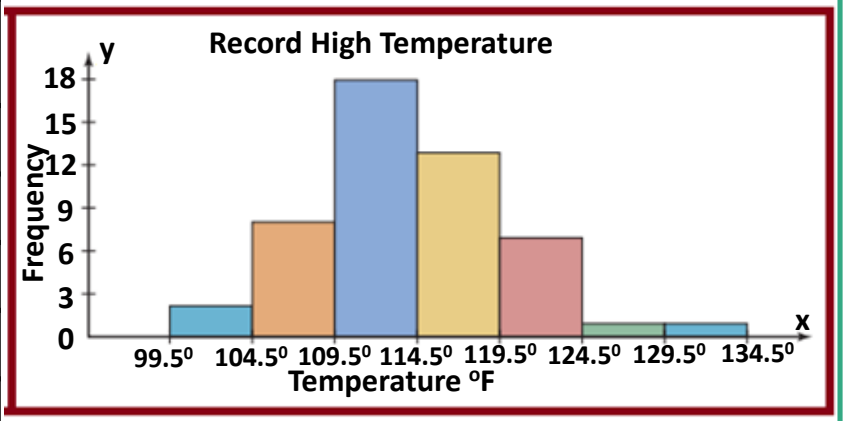
Machine Learning

Histograms, Frequency Polygons and Ogives

Histograms

The **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

Class Boundaries	Frequency
99.5-104.5	2
104.5-109.5	8
109.5-114.5	18
114.5-119.5	13
119.5-124.5	7
124.5-129.5	1
129.5-134.5	1



Histograms

- **Advantages and Disadvantages of Histograms:**

- **Advantages:**
  - Depicts the frequencies of observations occurring in certain ranges /intervals of values. The intervals must be adjacent
  - Accurate representation of the distribution of numerical data
  - Give a rough sense of the ***density*** of the underlying distribution of the data

Histograms

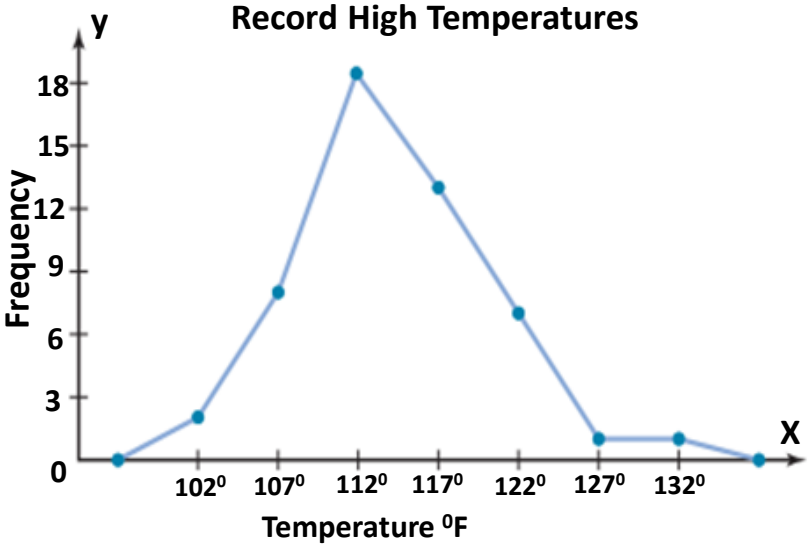
- **Advantages and Disadvantages of Histograms (Contd.):**

- **Disadvantages:**
  - Random Fluctuations in values
  - Alternative choices for ends of intervals give very different diagrams
  - Apparent multimodality can arise then vanish for different choices of intervals or for different small sample
  - Effects diminish with increasing size of data set

Frequency Polygon

The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.

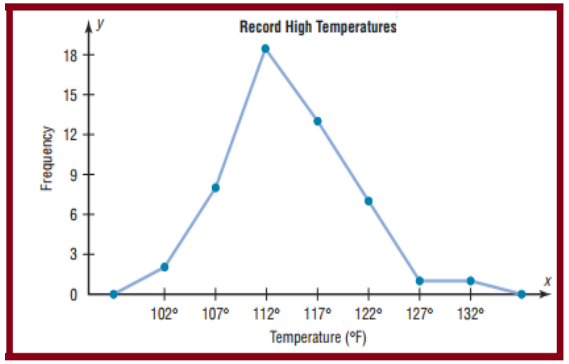
Class Boundaries	Midpoints	Frequency
99.5-104.5	102	2
104.5-109.5	107	8
109.5-114.5	112	18
114.5-119.5	117	13
119.5-124.5	122	7
124.5-129.5	127	1
129.5-134.5	132	1



Frequency Polygon

The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.

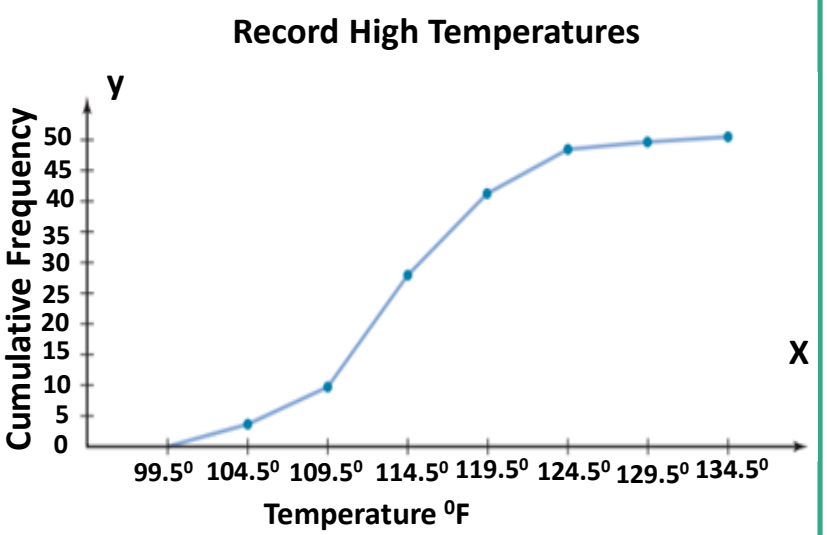
Class Boundaries	Midpoints	Frequency
99.5-104.5	102	2
104.5-109.5	107	8
109.5-114.5	112	18
114.5-119.5	117	13
119.5-124.5	122	7
124.5-129.5	127	1
129.5-134.5	132	1



Ogive

The **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

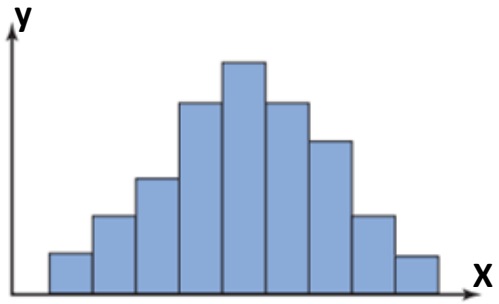


Machine Learning

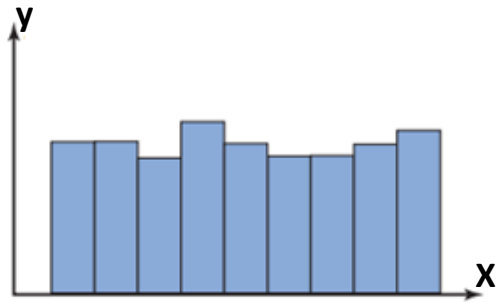
Distribution Shapes

Distribution Shapes

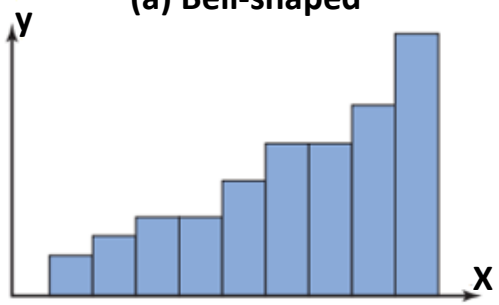
A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution.



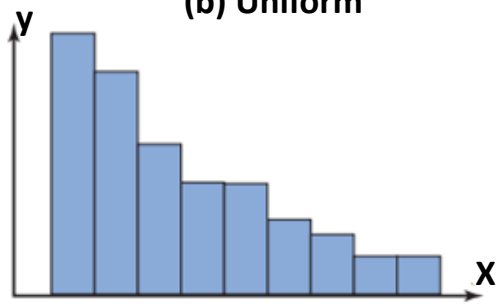
(a) Bell-shaped



(b) Uniform

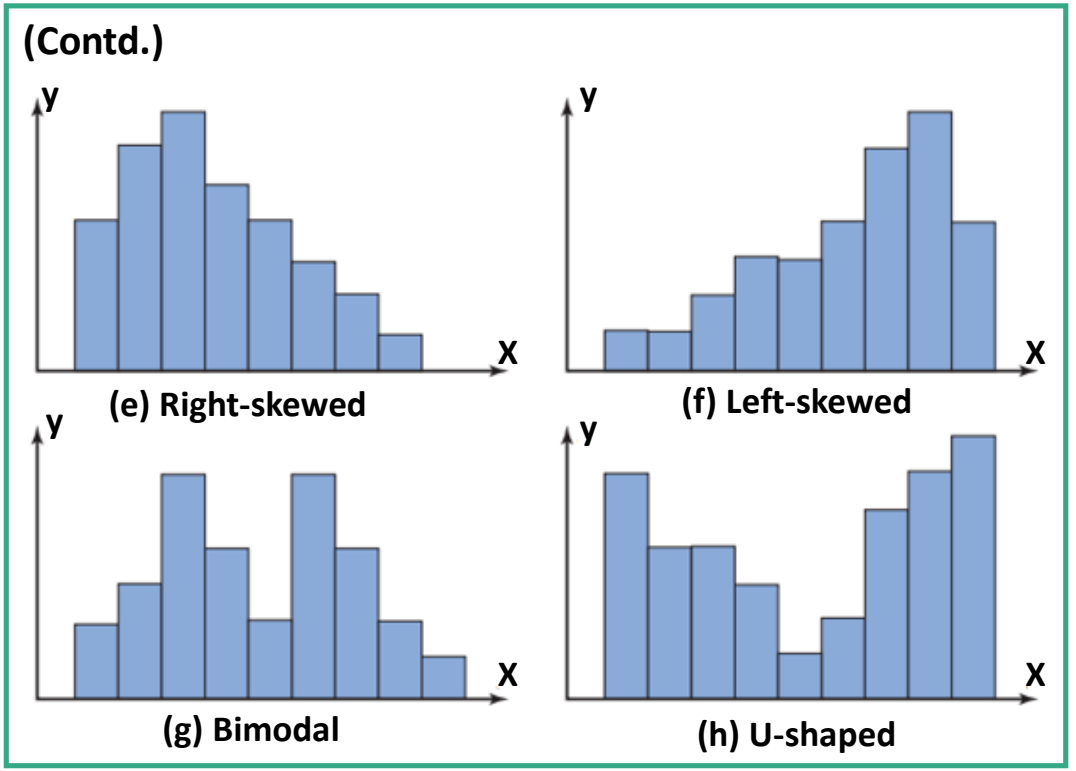


(c) J-shaped



(d) Reverse J-shaped

Distribution Shapes



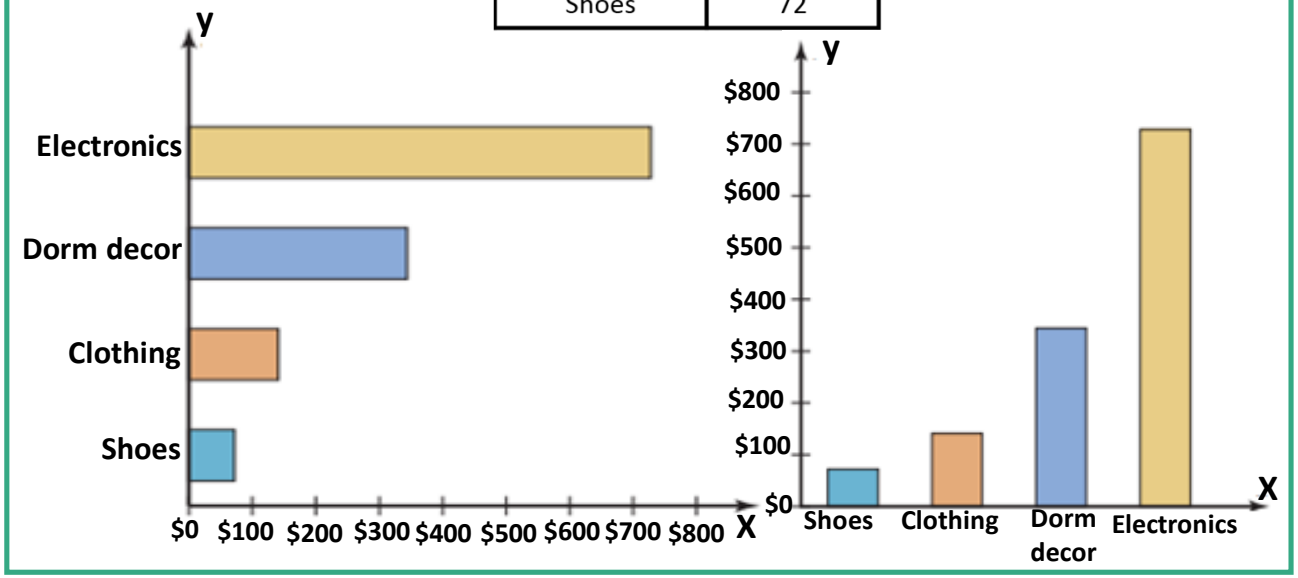
Machine Learning

Other Types of Graphs: Bar, Pareto, Time Series and Pie

Bar Graphs

A **bar graph** represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

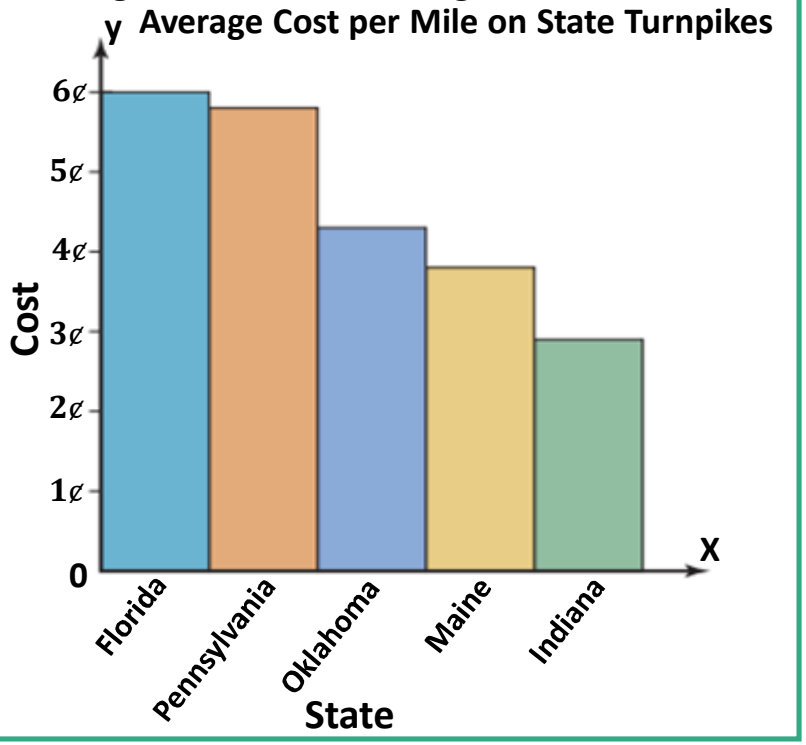
College Spending for First-Year Students	
Electronics	$728
Dorm décor	344
Clothing	141
Shoes	72



Pareto Charts

A **Pareto chart** is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

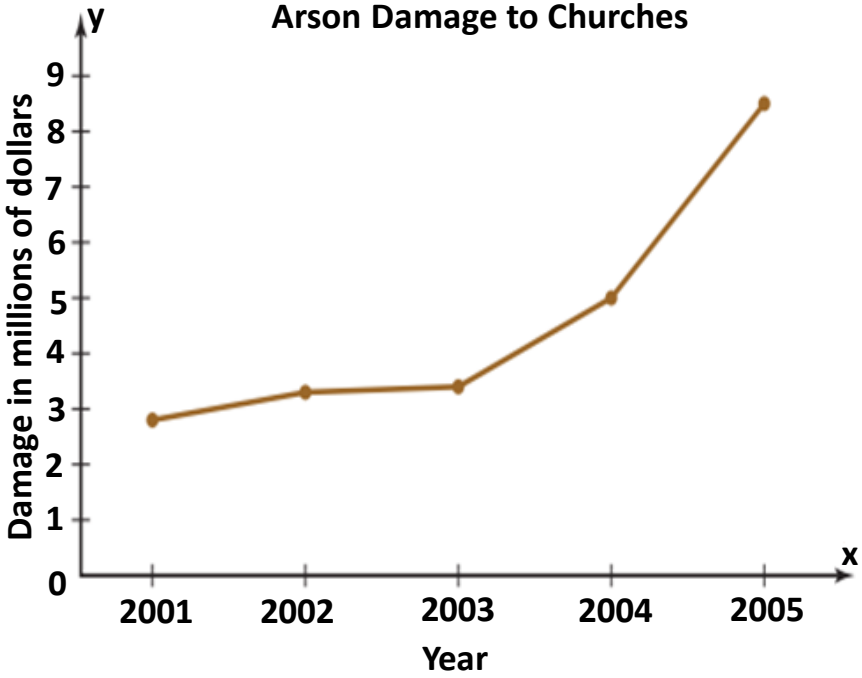
State wise average cost per mile-	
State	**Number**
Florida	6.0¢
Pennsylvania	5.8
Oklahoma	4.3
Maine	3.8
Indiana	2.9



Time Series Graph

A **time series graph** represents data that occur over a specific period of time.

Year	Damage (in millions)
2001	$2.8
2002	3.3
2003	3.4
2004	5.0
2005	8.5



Pie Graph

A **pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

Class	Frequency	Percent
A	5	20
B	7	28
O	9	36
AB	4	16
	25	100

For each class,

A $\frac{5}{25} \cdot 360° = 72°$

B $\frac{7}{25} \cdot 360° = 100.8°$

O $\frac{9}{25} \cdot 360° = 129.6°$

AB $\frac{4}{25} \cdot 360° = 57.6°$

Blood Types for Army Inductees