Data Transformations

Basic Statistics

DESCRIBING DATA

Given some large dataset, we'd like to compute a few quantities that intuitively summarizes the data. To begin with we'd like to know

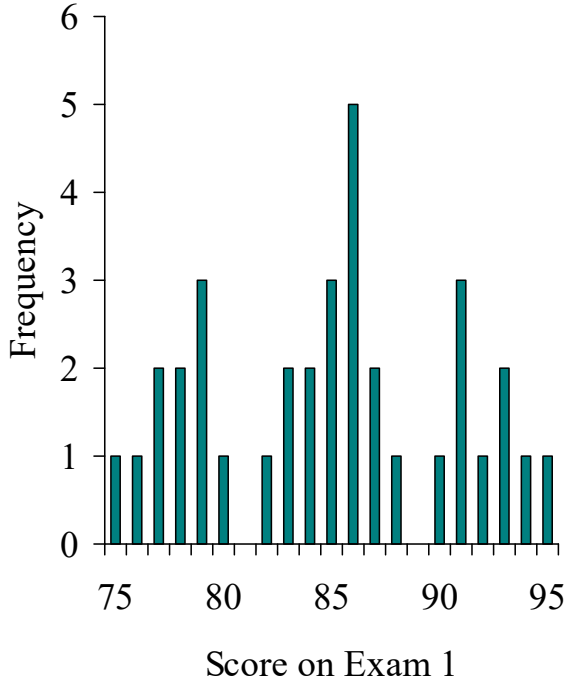
- what are typical values for our variables or attributes?
- how representative are these typical values?

Measures of Central Tendency

- A *measure of central tendency* is a descriptive statistic that describes the average, or typical value of a set of scores
- There are three common measures of central tendency:
  - the mode
  - the median
  - the mean

The Mode

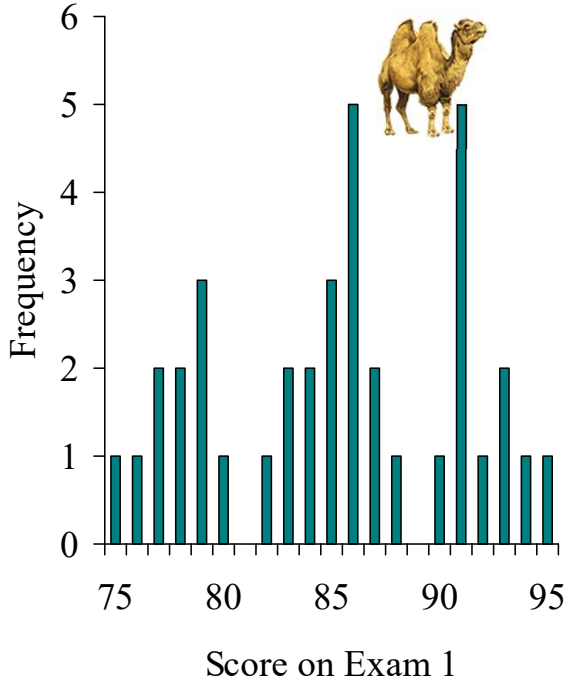
- The *mode* is the score that occurs most frequently in a set of data



Measures of Central Tendency

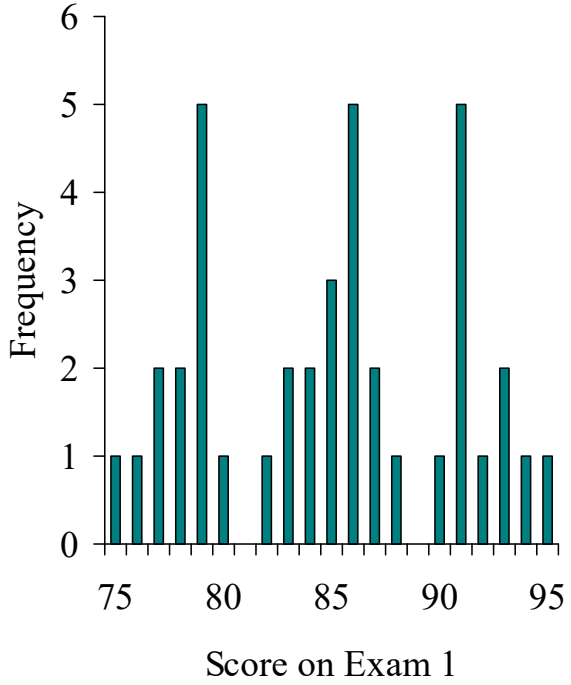
Bimodal Distributions

- When a distribution has two “modes,” it is called *bimodal*



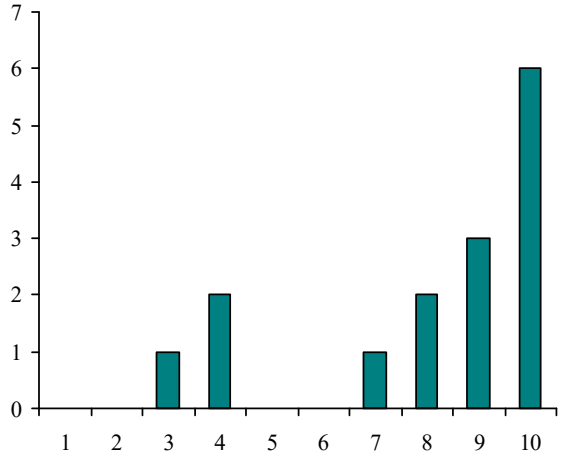
Multimodal Distributions

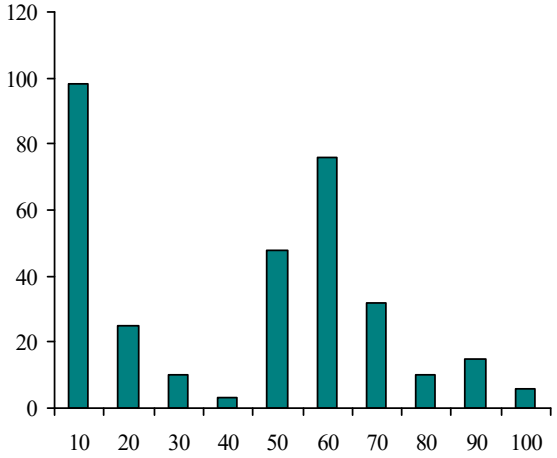
- If a distribution has more than 2 “modes,” it is called *multimodal*



When To Use the Mode

- The mode is not a very useful measure of central tendency
  - It is insensitive to large changes in the data set
    - That is, two data sets that are very different from each other can have the same mode





When To Use the Mode

- The mode is primarily used with nominally scaled data
  - It is the only measure of central tendency that is appropriate for nominally scaled data

The Median

- The *median* is simply another name for the 50th percentile
  - It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median

How To Calculate the Median

- Conceptually, it is easy to calculate the median
  - There are many minor problems that can occur; it is best to let a computer do it
- Sort the data from highest to lowest
- Find the score in the middle
  - middle = (N + 1) / 2
  - If N, the number of scores, is even the median is the average of the middle two scores

Median Example

- What is the median of the following scores:
  24 18 19 42 16 12
- Sort the scores:
  42 24 19 18 16 12
- Determine the middle score:
  middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5
- Median = average of 3rd and 4th scores:
  (19 + 18) / 2 = 18.5

Median Example

- What is the median of the following scores:
  10 8 14 15 7 3 3 8 12 10 9
- Sort the scores:
  15 14 12 10 10 9 8 8 7 3 3
- Determine the middle score:
  middle = (N + 1) / 2 = (11 + 1) / 2 = 6
- Middle score = median = 9

When To Use the Median

- The median is often used when the distribution of scores is either positively or negatively skewed
  - The few really large scores (positively skewed) or really small scores (negatively skewed) will not overly influence the median

Calculating the Mean

- Calculate the mean of the following data:
  1 5 4 3 2
- Sum the scores (ΣX):
  $1 + 5 + 4 + 3 + 2 = 15$
- Divide the sum ($\Sigma X = 15$) by the number of scores ($N = 5$):
  $15 / 5 = 3$
- Mean = $\bar{X} = 3$

The Mean

- The *mean* is:
  - the arithmetic average of all the scores
    $(\Sigma X)/N$
  - the number, m, that makes $\Sigma(X - m)$ equal to 0
  - the number, m, that makes $\Sigma(X - m)^2$ a minimum
- The mean of a population is represented by the Greek letter μ; the mean of a sample is represented by $\bar{X}$

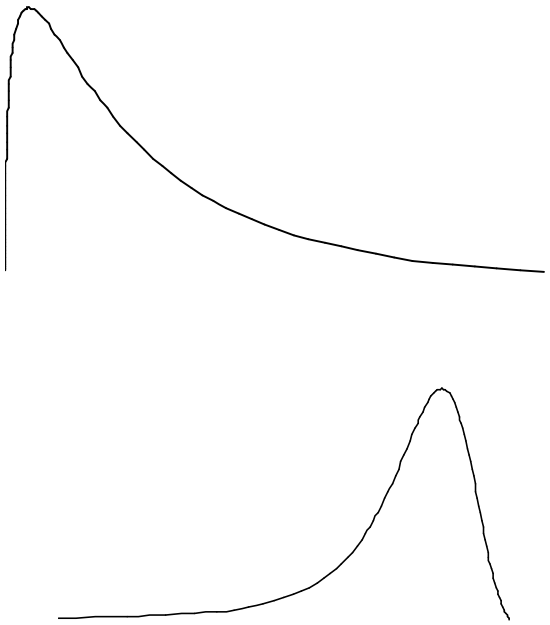
When To Use the Mean

- You should use the mean when
  - the data are interval or ratio scaled
    - Many people will use the mean with ordinally scaled data too
  - and the data are not skewed
- The mean is preferred because it is sensitive to every score
  - If you change one score in the data set, the mean will change

Relations Between the Measures of Central Tendency

- In symmetrical distributions, the median and mean are equal
  - For normal distributions, mean = median = mode
- In positively skewed distributions, the mean is greater than the median

- In negatively skewed distributions, the mean is smaller than the median

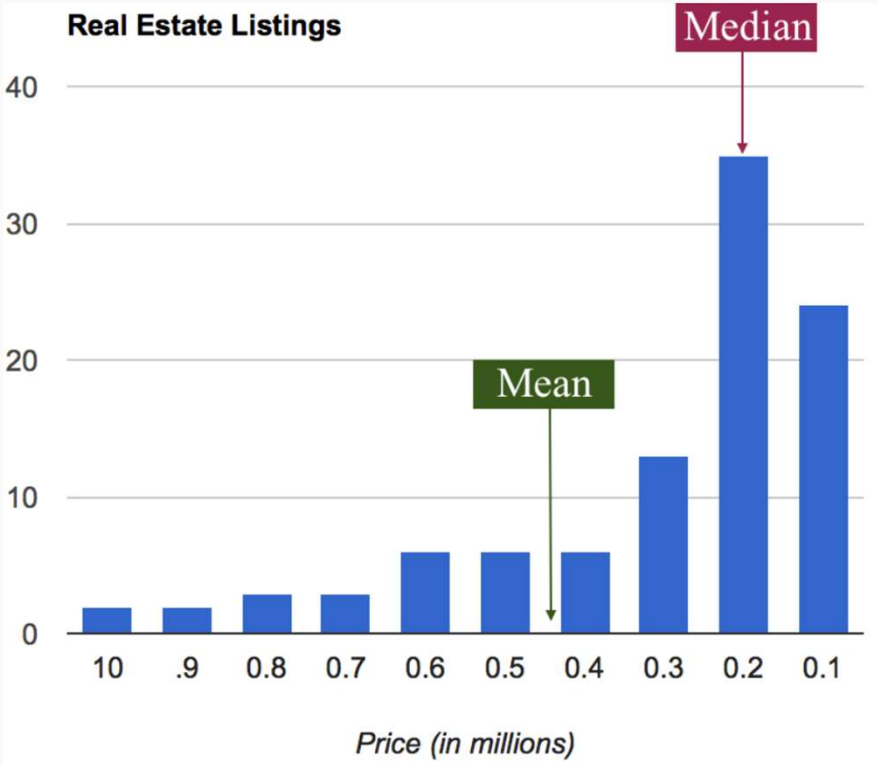


CENTRALITY

The mean is *sensitive to outliers*.

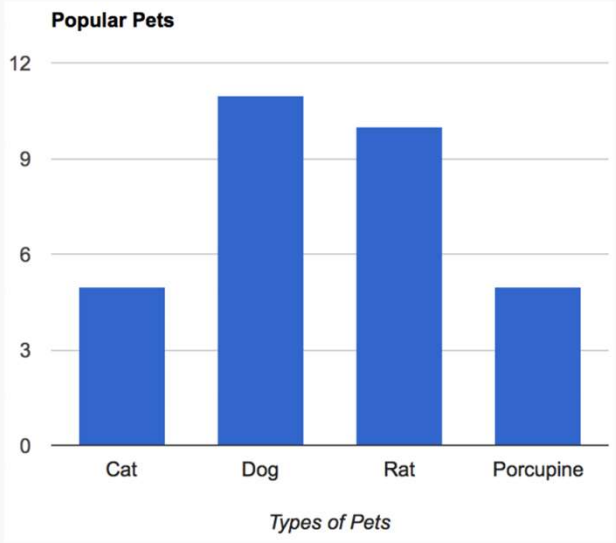
CENTRALITY

The mean is *sensitive to skewness (asymmetry) of distributions.*



CENTRALITY

For samples of categorical variables, neither mean or median make sense.



The **mode** might be a better way to find the most “representative” value.

Measures of Dispersion

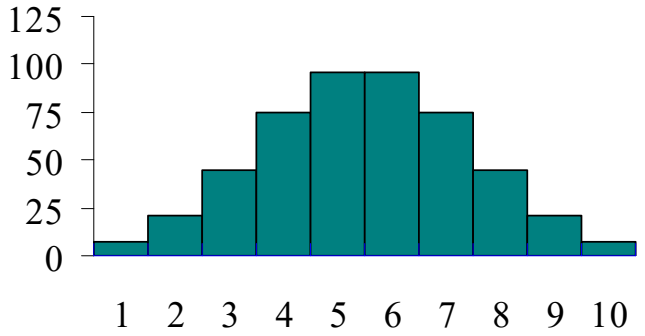
Definition

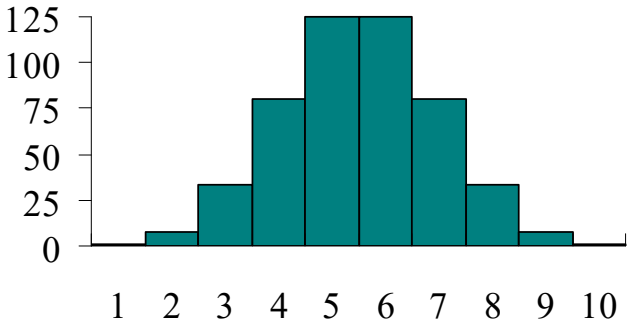
- *Measures of dispersion* are descriptive statistics that describe how similar a set of scores are to each other
  - The more similar the scores are to each other, the lower the measure of dispersion will be
  - The less similar the scores are to each other, the higher the measure of dispersion will be
  - In general, the more spread out a distribution is, the larger the measure of dispersion will be

Measures of Dispersion

- Which of the distributions of scores has the larger dispersion?

- The upper distribution has more dispersion because the scores are more spread out
  - That is, they are less similar to each other





Measures of Dispersion

- There are three main measures of dispersion:
  - The range
  - The semi-interquartile range (SIR)
  - Variance / standard deviation

The Range

- The *range* is defined as the difference between the largest score in the set of data and the smallest score in the set of data, $X_L - X_S$
- What is the range of the following data:
  4 8 1 6 6 2 9 3 6 9
- The largest score (X_L) is 9; the smallest score (X_S) is 1; the range is $X_L - X_S = 9 - 1 = 8$

When To Use the Range

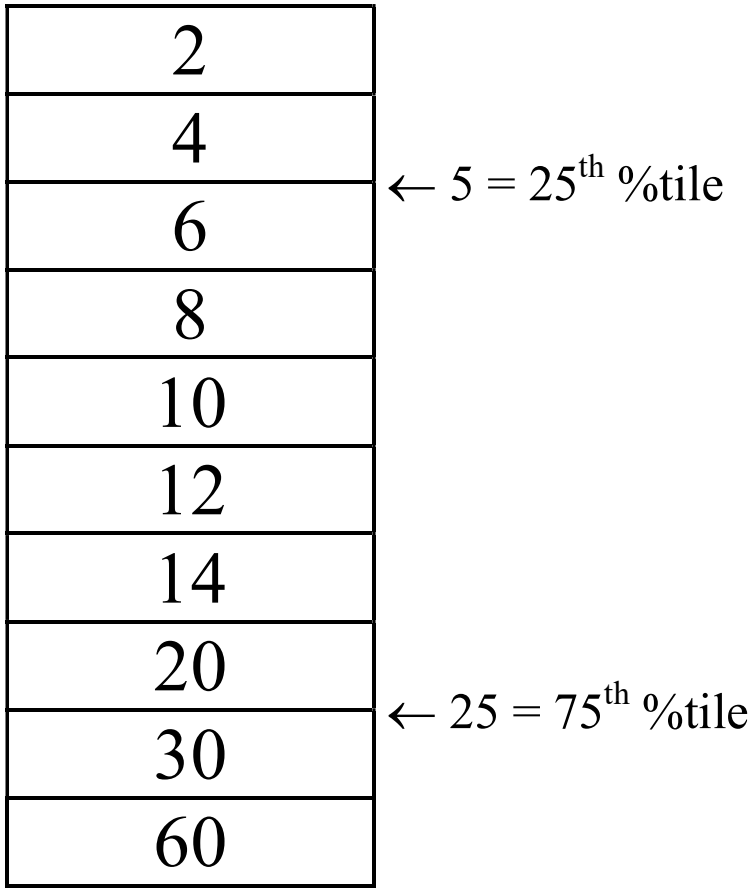
- The range is used when
  - you have ordinal data or
  - you are presenting your results to people with little or no knowledge of statistics
- The range is rarely used in scientific work as it is fairly insensitive
  - It depends on only two scores in the set of data, X_L and X_S
  - Two very different sets of data can have the same range:
    1 1 1 1 9 vs 1 3 5 7 9

The Semi-Interquartile Range

- The *semi-interquartile range* (or *SIR*) is defined as the difference of the first and third quartiles divided by two
  - The first quartile is the 25^{th} percentile
  - The third quartile is the 75^{th} percentile
- $SIR = (Q_3 - Q_1) / 2$

SIR Example

- What is the SIR for the data to the right?
- 25 % of the scores are below 5
  - 5 is the first quartile
- 25 % of the scores are above 25
  - 25 is the third quartile
- SIR = $(Q_3 - Q_1) / 2 = (25 - 5) / 2 = 10$



When To Use the SIR

- The SIR is often used with skewed data as it is insensitive to the extreme scores

Variance

- *Variance* is defined as the average of the square deviations:

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

What Does the Variance Formula Mean?

- First, it says to subtract the mean from each of the scores
  - This difference is called a *deviate* or a *deviation score*
  - The deviate tells us how far a given score is from the typical, or average, score
  - Thus, the deviate is a measure of dispersion for a given score

What Does the Variance Formula Mean?

- Why can't we simply take the average of the deviates? That is, why isn't variance defined as:

$$\sigma^2 \neq \frac{\sum (X - \mu)}{N}$$

This is not the formula for variance!

What Does the Variance Formula Mean?

- One of the definitions of the *mean* was that it always made the sum of the scores minus the mean equal to 0
- Thus, the average of the deviates must be 0 since the sum of the deviates must equal 0
- To avoid this problem, statisticians square the deviate score prior to averaging them
  - Squaring the deviate score makes all the squared scores positive

What Does the Variance Formula Mean?

- Variance is the mean of the squared deviation scores
- The larger the variance is, the more the scores deviate, on average, away from the mean
- The smaller the variance is, the less the scores deviate, on average, from the mean

Standard Deviation

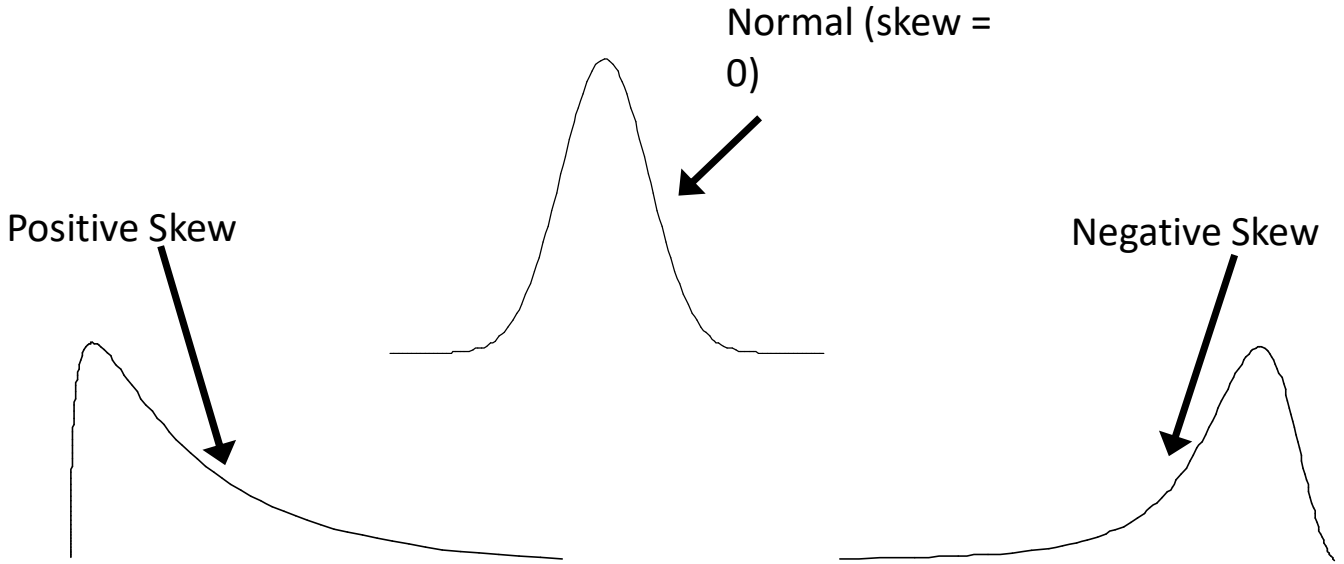
- When the deviate scores are squared in variance, their unit of measure is squared as well
  - E.g. If people's weights are measured in pounds, then the variance of the weights would be expressed in pounds^2 (or squared pounds)
- Since squared units of measure are often awkward to deal with, the square root of variance is often used instead
  - The standard deviation is the square root of variance

Standard Deviation

- Standard deviation = $\sqrt{}$variance
- Variance = standard deviation^2

Measure of Skew

- *Skew* is a measure of symmetry in the distribution of scores



Measure of Skew

- The following formula can be used to determine skew:

$$s^3 = \frac{\frac{\sum\left(X-\overline{X}\right)^3}{N}}{\sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{N}}}$$

Measure of Skew

- If $s^3 < 0$, then the distribution has a negative skew
- If $s^3 > 0$ then the distribution has a positive skew
- If $s^3 = 0$ then the distribution is symmetrical
- The more different s^3 is from 0, the greater the skew in the distribution

Part II Data Transformation

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values so that each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization/ Standardizaion: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Feature Scaling

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

- to transform the values of features or variables in a dataset to a similar scale.
- to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.
- dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.
- Standardization
- Normalization

Feature Scaling: Distance based alg

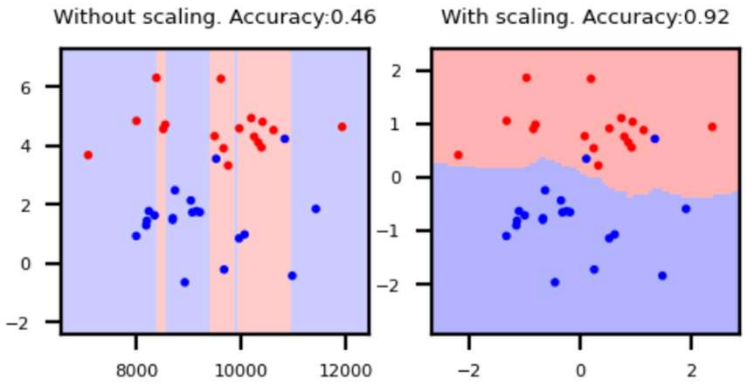
	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

- Distance AB before scaling => $\sqrt{(40-60)^2+(3-3)^2}=20$
- Distance BC before scaling => $\sqrt{(40-40)^2+(4-3)^2}=1$
- Distance AB after scaling => $\sqrt{(1.1+1.5)^2+(1.18-1.18)^2}=2.6$
- Distance BC after scaling => $\sqrt{(1.1-1.1)^2+(0.41+1.18)^2}=1.59$

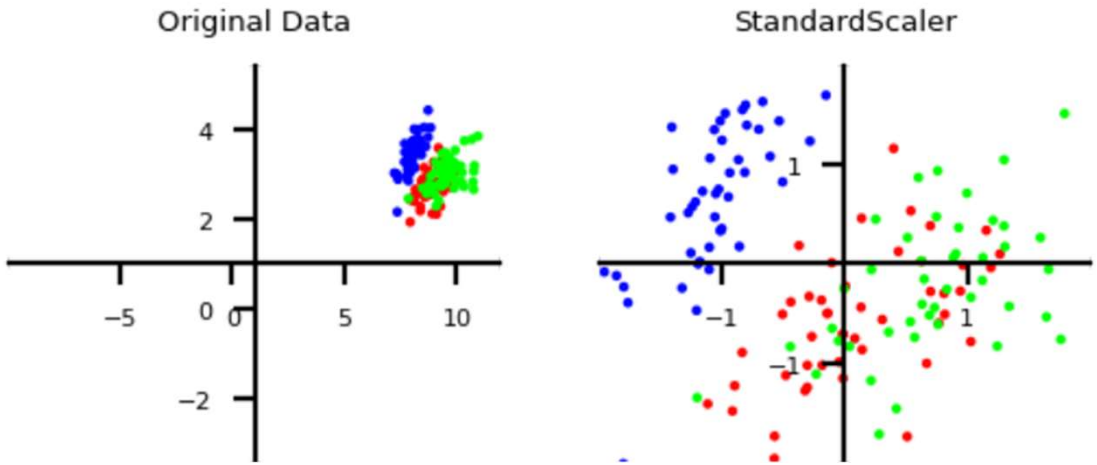
Why do we need scaling?

- KNN: Distances depend mainly on feature with larger values
- SVMs: (kernelized) dot products are also based on distances
- Linear model: Feature scale affects regularization
    - Weights have similar scales, more interpretable



Scaling

- Use when different numeric features have different scales (different range of values)
    - Features with much higher values may overpower the others
- Goal: bring them all within the same range
- Different methods exist



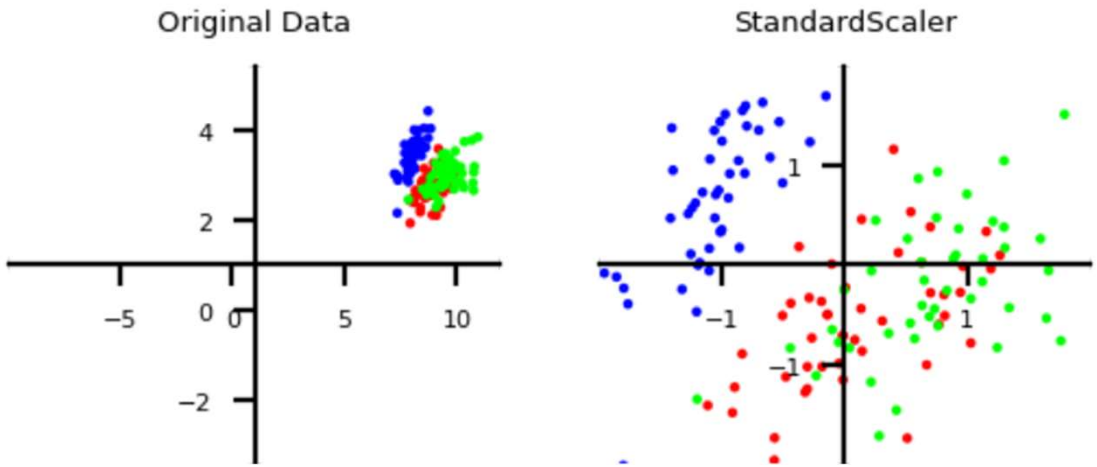
Standardization

- values are centered around the mean with a unit standard deviation.
- mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

Standard scaling (standardization)

- Generally most useful, assumes data is more or less normally distributed
- Per feature, subtract the mean value μ, scale by standard deviation σ
- New feature has $\mu = 0$ and $\sigma = 1$, values can still be arbitrarily large

$$\mathbf{x}_{new} = \frac{\mathbf{x} - \mu}{\sigma}$$



Data Values
13
16
19
22
23
38
47
56
58
63
65
70
71

The mean value in the dataset is 43.15 and the standard deviation is 22.13.

To normalize the first value of **13**, we would apply the formula shared earlier

- $\mathbf{x_{new} = (x_i - \bar{x}) / s} = (13 - 43.15) / 22.13 = \mathbf{-1.36}$

Data Values	Standardized
13	-1.36
16	-1.23
19	-1.09
22	-0.96
23	-0.91
38	-0.23
47	0.17
56	0.58
58	0.67
63	0.90
65	0.99
70	1.21
71	1.26

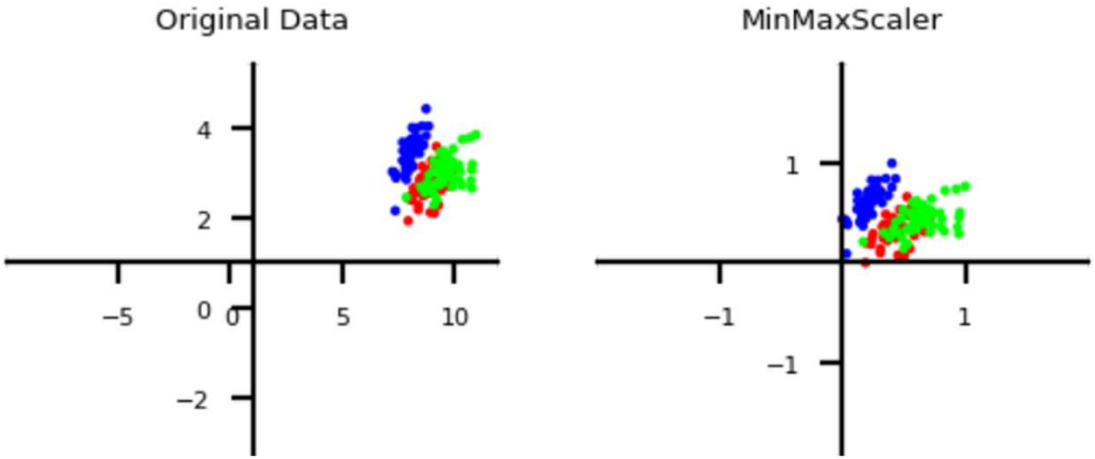
Normalization

- to adjust the values of features in a dataset to a common scale.
- to facilitate data analysis and modeling, and to reduce the impact of different scales on the accuracy of machine learning models.
- **Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1.**

Min-max scaling

- Scales all features between a given min and max value (e.g. 0 and 1)
- Makes sense if min/max values have meaning in your data
- Sensitive to outliers

$$\mathbf{x}_{new} = \frac{\mathbf{x} - x_{min}}{x_{max} - x_{min}} \cdot (max - min) + min$$



Example

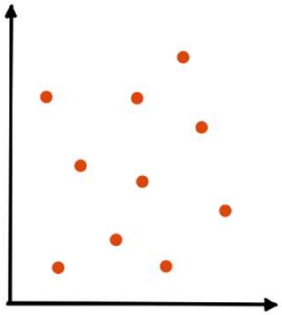
$$\mathbf{x}_{new} = \frac{\mathbf{x} - x_{min}}{x_{max} - x_{min}} \cdot (max - min) + min$$

- normalize the following data set, **200**, **300**, **400**, **600**, **1000** to a new range [0, 1], then using min-max normalization
- xmin = 200, xmax = 1000,
- min = 0, max = 1

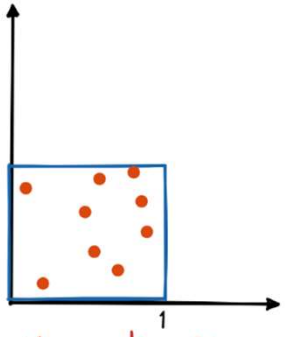
$$\frac{200 - 200}{1000 - 200}(1 - 0) + 0 = 0$$

$$\frac{300 - 200}{1000 - 200}(1 - 0) + 0 = \frac{100}{800} = 0.125$$

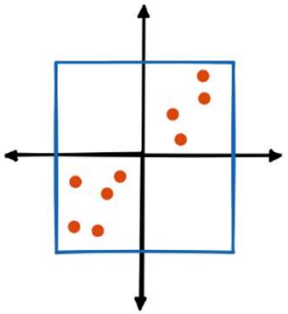
$$\frac{400 - 200}{1000 - 200}(1 - 0) + 0 = \frac{200}{800} = 0.25$$



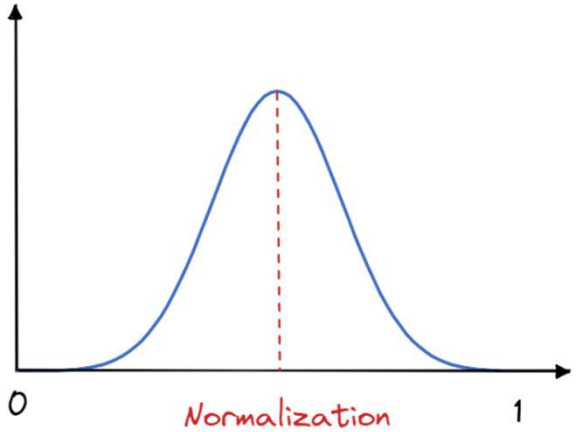
Actual Data



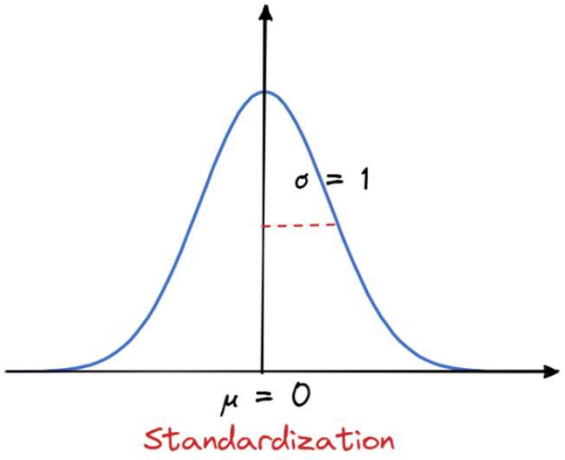
Normalization



Standardization



Normalization



Standardization

Standardisation

	Age	Salary
0	0.758874	7.494733e-01
1	-1.711504	-1.438178e+00
2	-1.275555	-8.912655e-01
3	-0.113024	-2.532004e-01
4	0.177609	6.632192e-16
5	-0.548973	-5.266569e-01
6	0.000000	-1.073570e+00
7	1.340140	1.387538e+00
8	1.630773	1.752147e+00
9	-0.258340	2.937125e-01

Max-Min Normalization

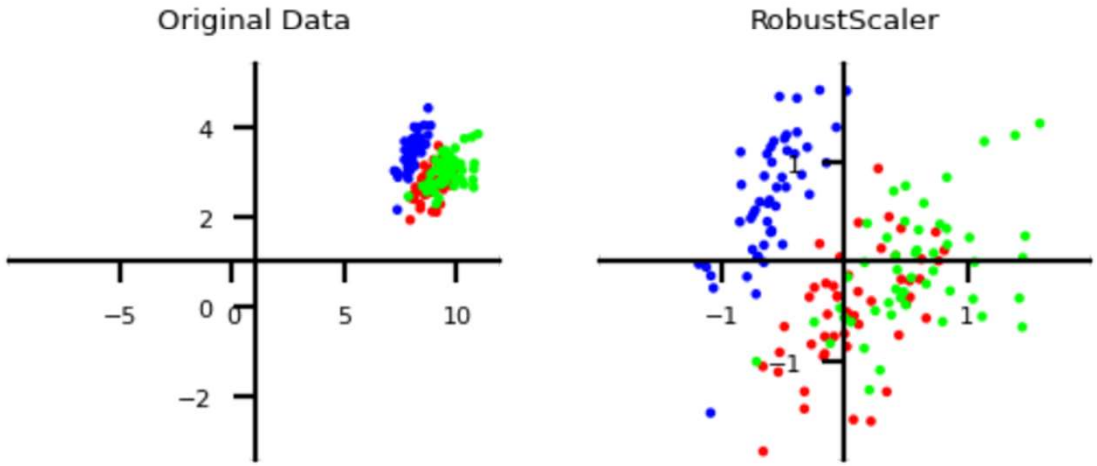
	Age	Salary
0	0.739130	0.685714
1	0.000000	0.000000
2	0.130435	0.171429
3	0.478261	0.371429
4	0.565217	0.450794
5	0.347826	0.285714
6	0.512077	0.114286
7	0.913043	0.885714
8	1.000000	1.000000
9	0.434783	0.542857

Difference

Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points

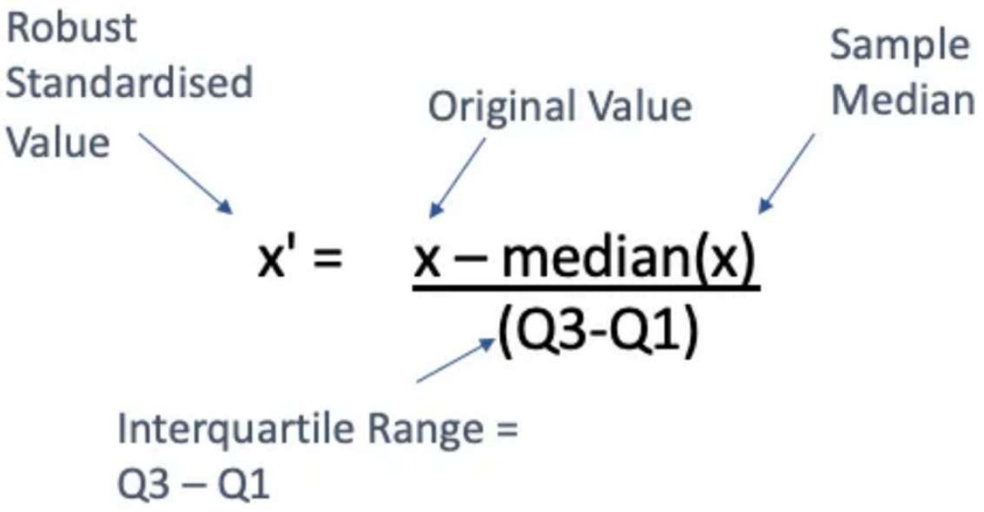
Robust scaling

- Subtracts the median, scales between quantiles q_{25} and q_{75}
- New feature has median 0, $q_{25} = -1$ and $q_{75} = 1$
- Similar to standard scaler, but ignores outliers



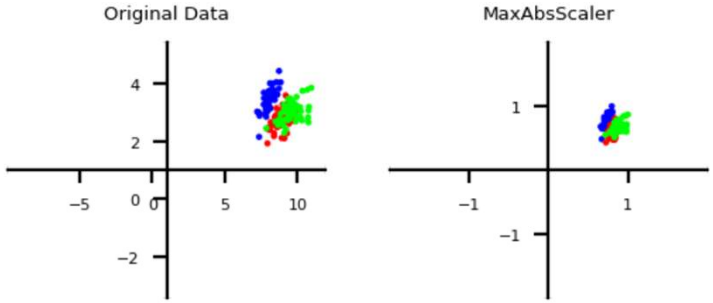
Robust Scalar transforms x to x' by subtracting each value of features by the **median** and dividing it by the **interquartile range** between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile).

The formula for Robust Scalar is:



Maximum Absolute scaler

- For sparse data (many features, but few are non-zero)
    - Maintain sparseness (efficient storage)
- Scales all values so that maximum absolute value is 1
- Similar to Min-Max scaling without changing 0 values



$$xscaled = \frac{x}{max(x)}$$

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling**

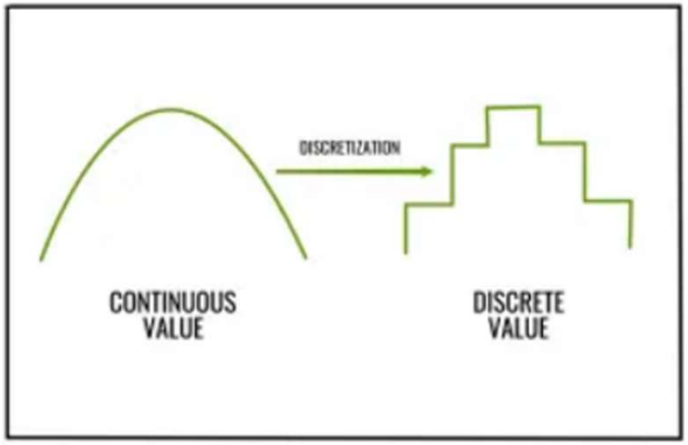
$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that Max}(|v'|) < 1$$

-10, 201, 301, -401, 501, 601, 701 Maximum absolute value: 701 Divide the given data by 1000 (i.e j=3) ***Result:*** *The normalized data is: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701*

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization/ Standardization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Discretization



- converting attributes values of continuous data into a finite set of intervals with minimum data loss.
- supervised discretization
- unsupervised discretization.

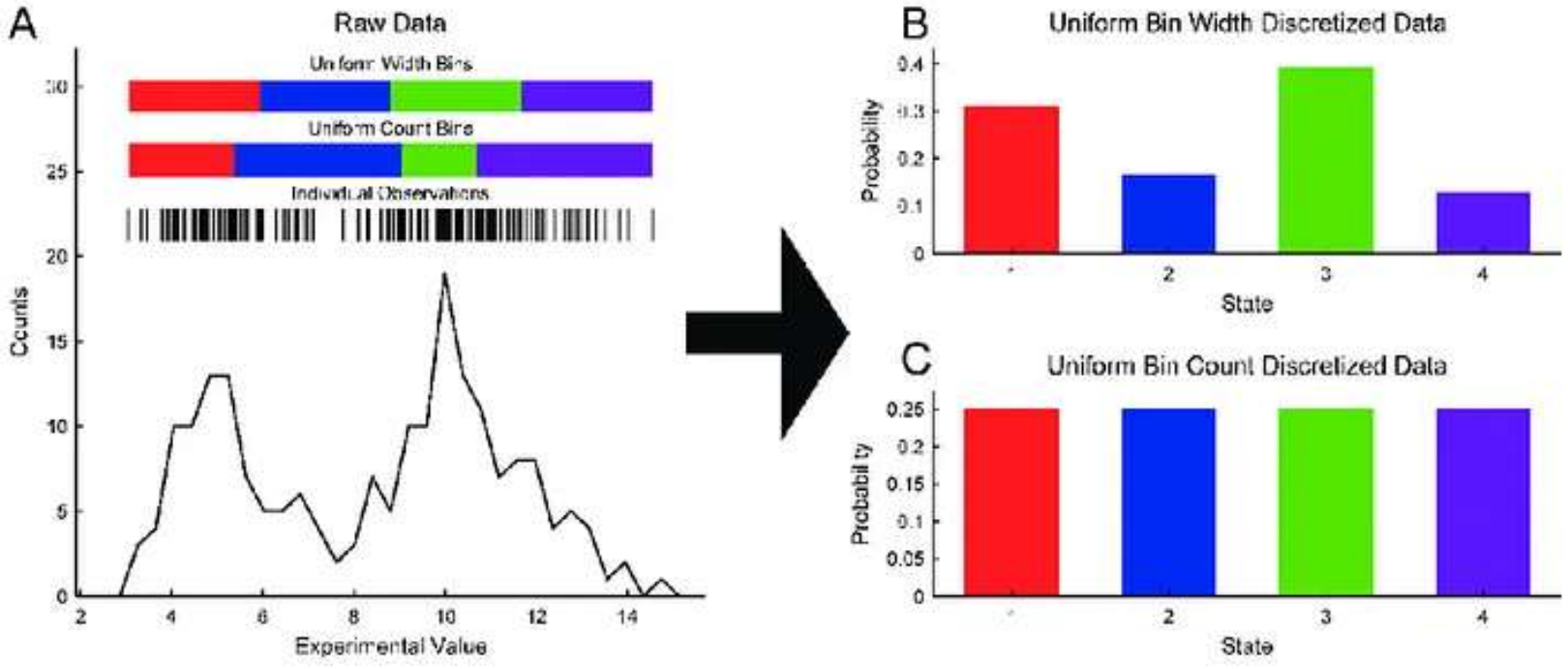
Age	1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

Data Discretization Methods

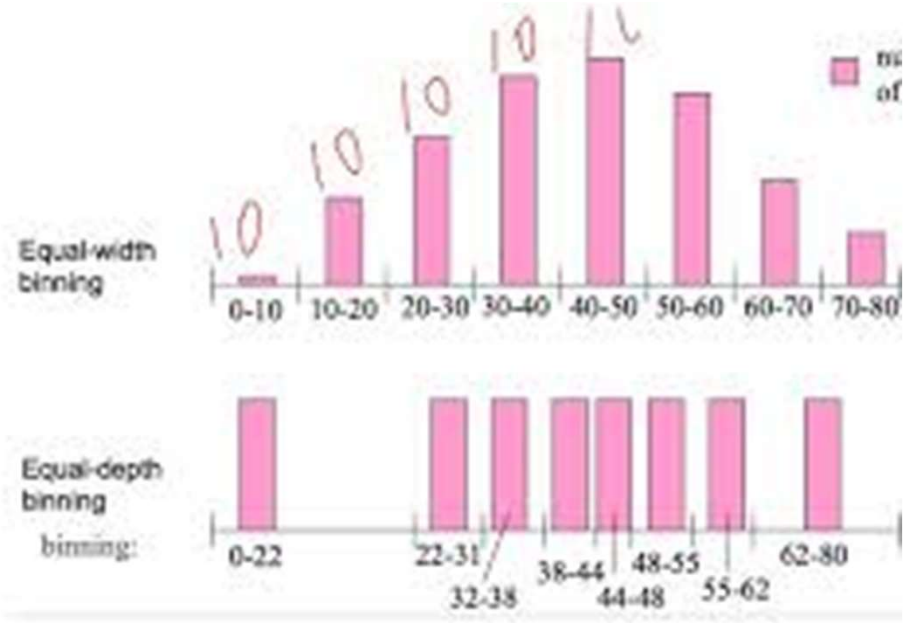
- Typical methods: All the methods can be applied recursively
  - Binning
  - Histogram analysis
  - Clustering analysis
  - Correlation (e.g., χ^2) analysis

Data Discretization Methods



Discretization: Binning

- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky



Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Discretization: Histogram analysis

- **1,1,4,4,4,4,7,7,9,9,9,9,9,11, 13,13,13,17,17,17,17,17,17, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30,30, 30.**

A histogram is a graph that shows the frequency of numerical data using rectangles. The height of a rectangle (the vertical axis) represents the distribution frequency of a variable

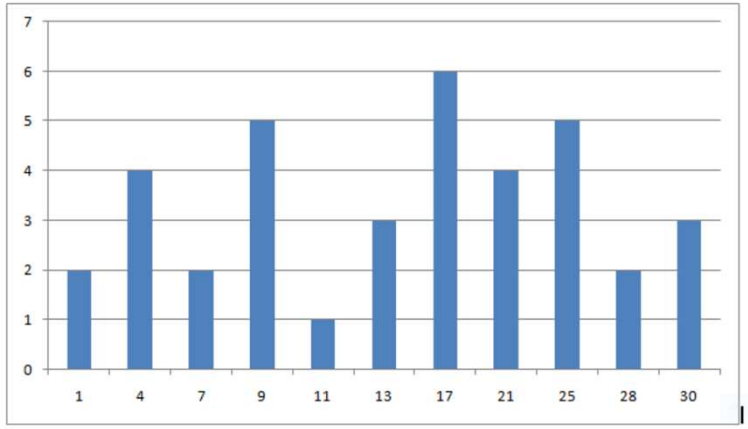


Figure 1 Histogram using price where one bucket represents one value

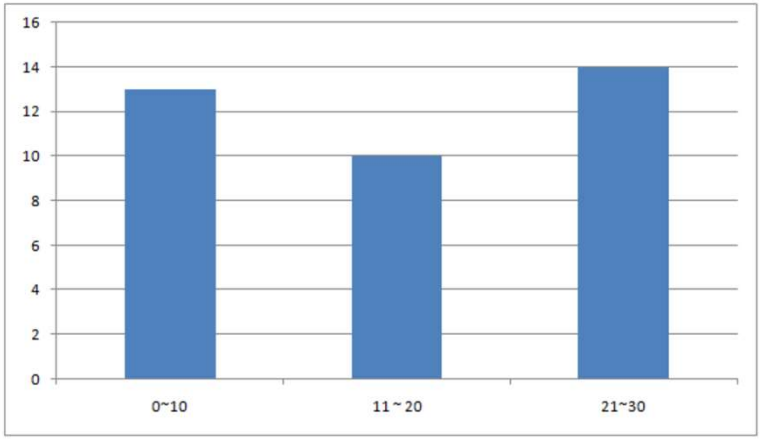
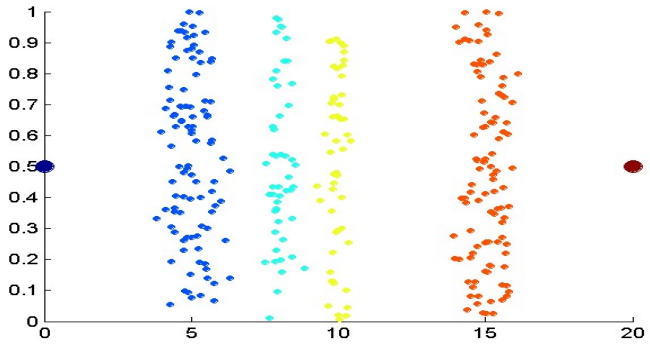


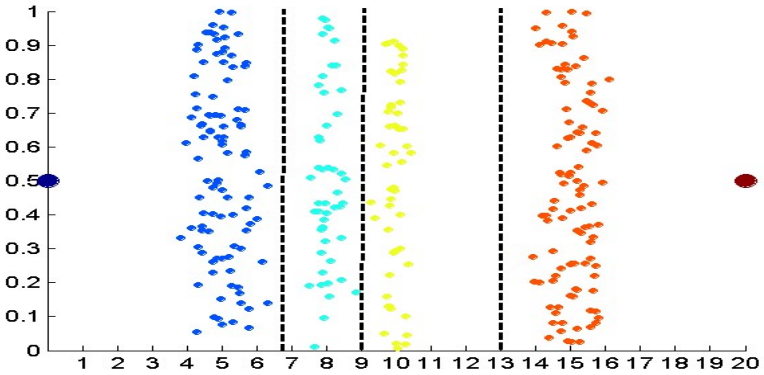
Figure 2 : Equal width Histogram

Discretization : clustering



Data

A clustering algorithm can be applied to discretize a numeric attribute, A, by partitioning the values of A into clusters or groups based on similarity, and store cluster representation (e.g., centroid and diameter) only



K-means clustering leads to better results

Discretization by Correlation Analysis

- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
  - Merge performed recursively, until a predefined stopping condition

https://medium.com/@nithin_rajan/data-discretization-using-chimerge-55c8ade3cfda

ChiMerge Discretization

Sample	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

- Statistical approach to Data Discretization

- Applies the Chi-square method to determine the similarity of data between two intervals.

- F:attribute
- K:class label

Discretization by Correlation Analysis

- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Exploit the correlation between intervals and class labels.
  - Interval – Class contingency tables
  - If two adjacent intervals have low χ^2 values (less correlated to the class labels), merge them to form a larger interval. (separating them does not offer more information on how to classify objects)
  - Merge performed recursively, until a predefined stopping condition is met

ChiMerge Discretization Example

Sample	F	K	Intervals
1	1	1	{0,2}
2	3	2	{2,5}
3	7	1	{5,7.5}
4	8	1	{7.5,8.5}
5	9	1	{8.5,10}
6	11	2	{10,17}
7	23	2	{17,30}
8	37	1	{30,38}
9	39	2	{38,42}
10	45	1	{42,45.5}
11	46	1	{45.5,52}
12	59	1	{52,60}

- Sort and order the attribute that you want to group (in this example attribute F).
- Start: having every unique value in the attribute be in its own interval.

ChiMerge Discretization Example

Sample	K=1	K=2	
2	0	1	1
3	1	0	1
total	1	1	2

$E_{11} = (1/2)*1 = .05$
$E_{12} = (1/2)*1 = .05$
$E_{21} = (1/2)*1 = .05$
$E_{22} = (1/2)*1 = .05$

$$X^2 = (0-.5)^2/.5 + (0-.5)^2/.5 + (0-.5)^2/.5 + (0-.5)^2/.5 = \mathbf{2}$$

Sample	K=1	K=2	
3	1	0	1
4	1	0	1
total	2	0	2

$E_{11} = (1/2)*2 = 1$
$E_{12} = (0/2)*2 = 0$
$E_{21} = (1/2)*2 = 1$
$E_{22} = (0/2)*2 = 0$

$$X^2 = (1-1)^2/1+(0-0)^2/0+ (1-1)^2/1+(0-0)^2/0 = \mathbf{0}$$

Sig Level 0.1 with df=1 from Chi square distribution X^2 critical value = 2.7024. Keep merging until all $X^2 > 2.7024$

ChiMerge Discretization Example

Sample	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

- Begin calculating the Chi square test on every pair of adjacent intervals

- Interval/class contingency tables:

Sample	K=1	K=2	
2	0	1	1
3	1	0	1
total	1	1	2

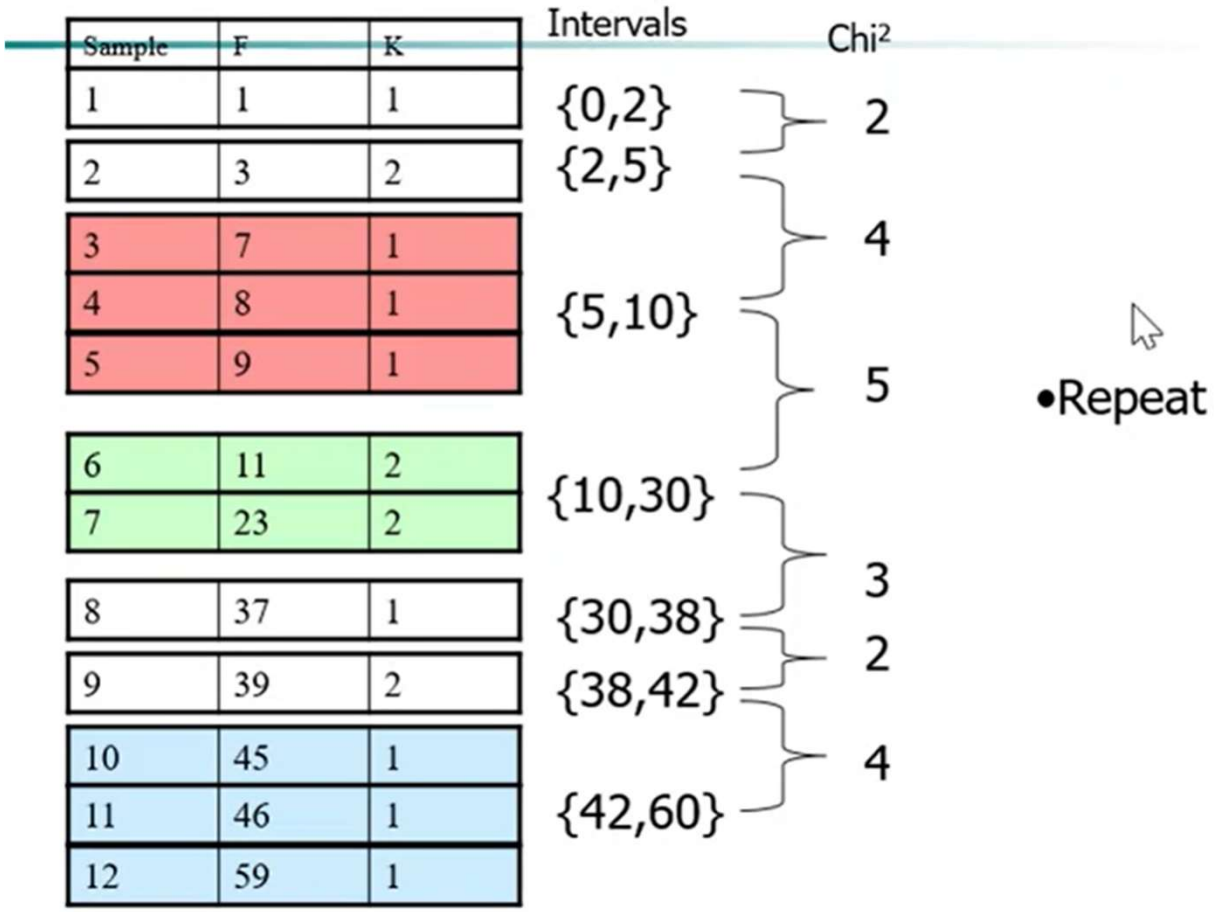
Sample	K=1	K=2	
3	1	0	1
4	1	0	1
total	2	0	2

ChiMerge Discretization Example

Sample	F	K	Intervals	Chi^2
1	1	1	{0,2}	2
2	3	2	{2,5}	2
3	7	1	{5,7.5}	0
4	8	1	{7.5,8.5}	0
5	9	1	{8.5,10}	2
6	11	2	{10,17}	0
7	23	2	{17,30}	2
8	37	1	{30,38}	2
9	39	2	{38,42}	2
10	45	1	{42,45.5}	0
11	46	1	{45.5,52}	0
12	59	1	{52,60}	

- Calculate all the Chi-square value for all intervals
- Merge the intervals with the smallest Chi values

ChiMerge Discretization Example



ChiMerge Discretization Example