CONTENTS

- Types of Data variable
  - Nominal
  - Binary
  - Ordinal
  - Numeric
    - Interval scaled
    - Ratio scaled
- Data pre-processing

Descriptive Data Summarization for Data Analysis

- Measuring the Central Tendency
  - Mean[Average]
  - Mode [Highest frequency Value]
  - Median [middle value of the ordered set]
  - Midrange [average of the largest and smallest values in the data set)]
- Measuring the Dispersion of Data

Measuring the Central Tendency: MEAN

- MEAN

Let x1,x2,...,xN be a set of N values or observations

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

- Mean of grouped data

- Weighted arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

The weights reflect the significance, importance, or occurrence frequency attached to their respective values.

Measuring the Central Tendency: MEDIAN

- Median
  - middle value of the ordered set

ODD N

65	55	89	56	35	14	56	55	87	45	92

N+1/2=11+1/2=6

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	**56**	56	65	87	89	92

Even N

65	55	89	56	35	14	56	55	87	45

N+1/2=10+1/2=**5.5**

We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	**55**	**56**	56	65	87	89

Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

Measuring the Central Tendency: MEDIAN

- Median of Grouped Data

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width,$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and *width* is the width of the median interval.

Example

Q. Suppose that the values for a given set of data are grouped into intervals.

The intervals and corresponding frequencies are as follows.

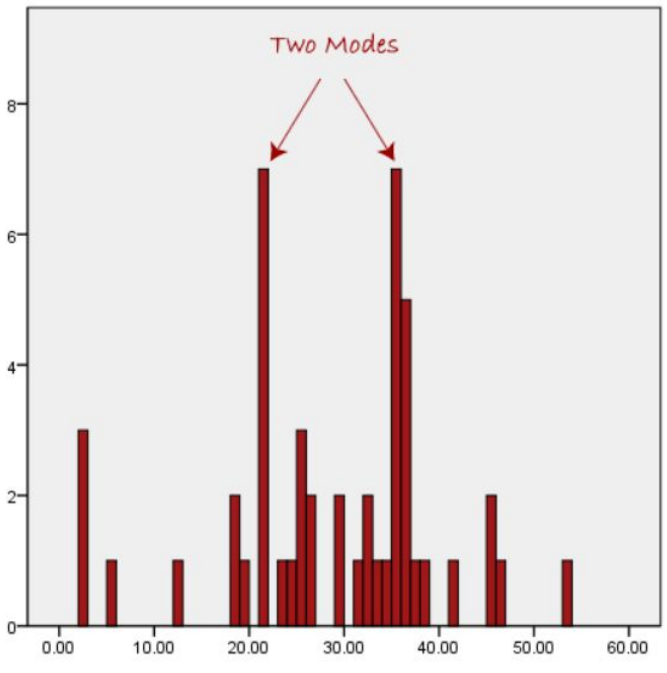
age	*frequency*
1-5	200
5-15	450
15-20	300
20-50	1500
50-80	700
80-110	44

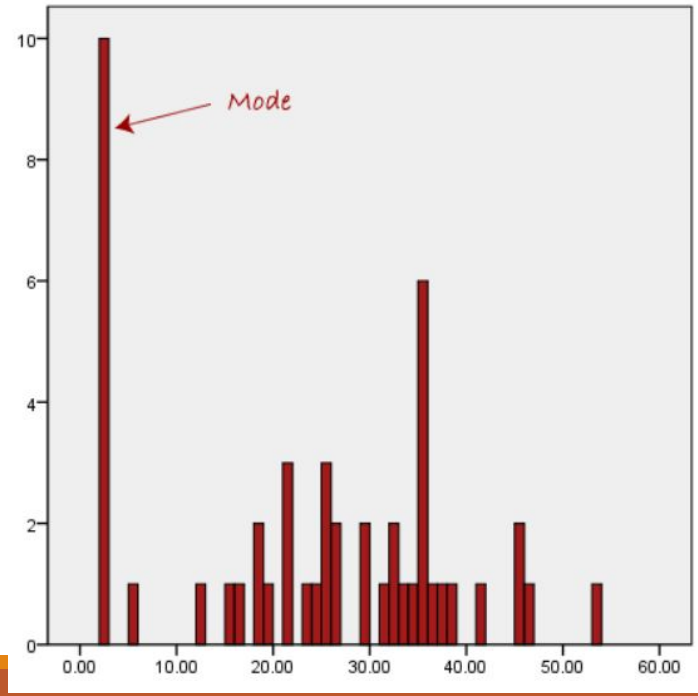
Compute an approximate median value for the data.

Answer: 32.94 Years

Measuring the Central Tendency: MODE

- Mode
  - The mode is the **most frequent score in our data set**.
  - On a histogram it represents the highest bar in a bar chart or histogram.
  - Problem:





Unimodal Distribution: If a data set has only **one value** that occurs most often, the set is called **unimodal.**

Bimodal Distribution: A data set that has **two values** that occur with the greatest frequency is referred to as **bimodal**.

Multimodal Distribution: When a set has **more than two values** that occur with the same greatest frequency, the set is called **multimodal**.

Measuring the Central Tendency: MODE

- Mode of grouped data

$$\text{Mode} = l + \left[\frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)}\right] h$$

where l = lower limit of the modal class

f_m = frequency of the modal class

f_1 = frequency of class preceding the modal class

f_2 = frequency of class succeeding the modal class

h = width of the modal class

Measuring the Central Tendency

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Measuring the Dispersion of Data

- Degree to which numerical data tend to spread is called the dispersion, or variance of the data
  - Range
  - Quartiles
  - Outliers
  - Boxplots

Range

Let x1,x2,...,xN be a set of observations for some attribute. The range of the set is the difference between the largest (max()) and smallest (min()) values.

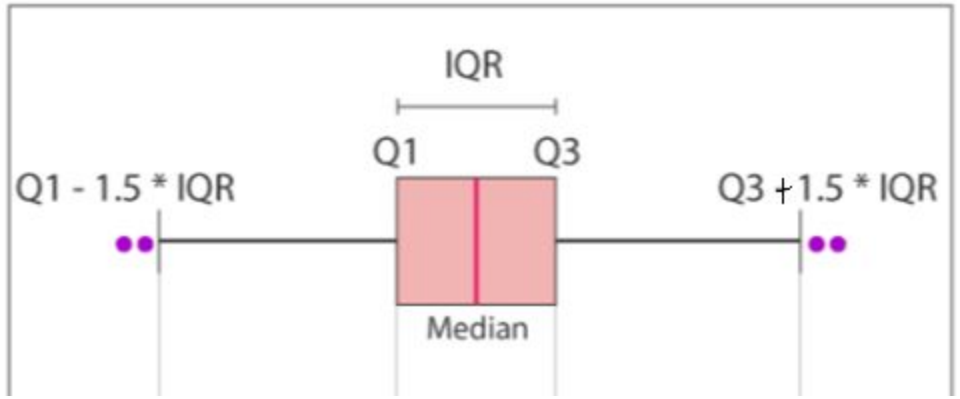
Quartiles

- The kth percentile of a set of data in numerical order is the value xi having the property that k percent of the data entries lie at or below xi . The median is the 50th percentile.
- The most commonly used percentiles other than the median are Quartiles.
  - The first quartile, denoted by Q1, is the 25th percentile
  - Second quartile (Q2) is 50^{th} percentile
  - Third quartile, denoted by Q3, is the 75th percentile.
- The quartiles, including the median, give some indication of the center, spread, and shape of a distribution.
- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (IQR) and is defined as IQR = Q3 −Q1

Quartiles

- Identifying suspected **outliers** is to single out **values falling at least 1.5×IQR above the third quartile or below the first quartile.**
- Q1, the median, and Q3 together contain no information about the endpoints (e.g., tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well. This is known as the **five-number summary.**
- The five-number summary of a distribution consists of the median, the quartiles Q1 and Q3, and the smallest and largest individual observations
  - Written in the order **Minimum, Q1, Median, Q3, Maximum.**

BoxPlots



- Boxplots are a popular way of visualizing a distribution.
- A boxplot incorporates the five-number summary as follows:
  - Typically, the ends of the box are at the quartiles, so that the box length is the interquartile range, IQR.
  - The median is marked by a line within the box.
  - Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.
  - When dealing with a moderate number of observations, it is worthwhile to plot potential outliers individually. To do this in a boxplot,
    - the whiskers are extended to the extreme low and high observations only if these values are less than 1.5 × IQR beyond the quartiles.
    - Otherwise, the whiskers terminate at the most extreme observations occurring within 1.5×IQR of the quartiles. The remaining cases are plotted individually. Boxplots can be used in the comparisons of several sets of compatible data

BoxPlots

- **Positively Skewed**: If the distance from the median to the maximum is greater than the distance from the median to the minimum, then the box plot is positively skewed.
- **Negatively Skewed**: If the distance from the median to minimum is greater than the distance from the median to the maximum, then the box plot is negatively skewed.
- **Symmetric**: The box plot is said to be symmetric if the median is equidistant from the maximum and minimum values.

Data Pre-processing

Why Pre-process the Data?

- For Quality
- Three elements of data quality: **accuracy, completeness, and consistency**
- Others:
  - **Timeliness**: Data must be available within a time frame that allows it to be useful for decision making.
  - **Believability**: Data values must be within the range of possible results in order to be useful for decision making.
  - **Value added**: Data must provide additional value in terms of information that offsets the cost of collecting and accessing it.
  - **Interpretability**: Data must not be so complex that the effort to understand the information it provides exceeds the benefit of its analysis.
  - **Accessability**: Data must be accessible so that the effort to collect it does not exceed the benefit from its use.

Major Tasks in Data Pre-processing

- Data cleaning
  - filling in missing values
  - smoothing noisy data
  - identifying or removing outliers
- Data Integration
- Data Reduction
  - dimensionality reduction
  - numerosity reduction.
- Data Transformation
  - Normalization
  - Discretization
  - Concept hierarchy generation

Data Cleaning

- Missing Values
  - Ignore the tuple
  - Fill in the missing value manually
  - Use a global constant to fill in the missing value
  - Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value
  - Use the attribute mean or median for all samples belonging to the same class as the given tuple
  - Use the most probable value to fill in the missing value

Data Cleaning

- Noisy Data
  - Binning: smooth a sorted data value by consulting its neighborhood
    - smoothing by bin means
    - smoothing by bin medians
    - smoothing by bin boundaries
  - Regression
    - Linear regression involves finding the “best” line to fit two attributes (or variables), so that one attribute can be used to predict the other.
    - Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface
  - Clustering

Binning methods for data smoothing

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

then partitioned into equal-frequency bins of size 3 (i.e., each bin contains three values)

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

minimum & maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Performs data discretization

- smoothing by bin means, each value in a bin is replaced by the mean value of the bin
- smoothing by bin medians can be employed, in which each bin value is replaced by the bin median

bins may be equal-width, where the interval range of values in each bin is constant.

Example of binning for data smoothing

Sorted data for Age: 3, 7, 8, 13, 22, 22, 22, 26, 26, 28, 30, 37

Data Integration

- Entity Identification Problem
- Redundancy and Correlation Analysis
- Tuple Duplication
  - For nominal data, we use the χ 2 (chi-square) test.
  - For numeric attributes, we can use the correlation coefficient and covariance
- Data Value Conflict Detection and Resolution

Data Reduction

- Dimensionality reduction
  - Numerosity reduction: parametric or non-parametric
    - Parametric: Regression and log-linear models
    - Non-Parametric: Histograms, Clustering, Sampling, Data cube aggregation
  - Data compression

Data Transformation

- Smoothing: remove noise from the data. Such techniques include binning, regression, and clustering
- Aggregation: where summary or aggregation operations are applied to the data.
- Generalization: Low-level or “primitive” (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
- Normalization: Attribute data are scaled so as to fall within a small specified range, such as −1.0 to 1.0, or 0.0 to 1.0.
- Attribute construction (or feature construction): New attributes are constructed and added from the given set of attributes to help the mining process.

Different normalization techniques

- **Min-Max normalization**: This technique scales the values of a feature to a range between 0 and 1. This is done by subtracting the minimum value of the feature from each value, and then dividing by the range of the feature.
- **Z-score normalization**: This technique scales the values of a feature to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the feature from each value, and then dividing by the standard deviation.
- **Decimal Scaling**: This technique scales the values of a feature by dividing the values of a feature by a power of 10.
- **Logarithmic transformation**: This technique applies a logarithmic transformation to the values of a feature. This can be useful for data with a wide range of values, as it can help to reduce the impact of outliers.
- **Root transformation**: This technique applies a square root transformation to the values of a feature. This can be useful for data with a wide range of values, as it can help to reduce the impact of outliers.

person_name	Salary	Year_of_ experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

The attributes salary and year_of_experience are on different scale and hence attribute salary can take high priority over attribute year_of_experience in the model.

Min-Max Normalization –

In this technique of data normalization, linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each value is replaced according to the following formula.

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)}\left(new\_max(A) - new\_min(A)\right) + new\_min(A)$$

Z-score normalization

In this technique, values are normalized based on mean and standard deviation of the data A. The formula used is:

v', v is the new and old of each entry in data respectively. σ_A, A is the standard deviation and mean of A respectively.

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

Data Normalization

Suppose a company wants to decide on a promotion based on the years of work experience of its employees. So, it needs to analyze a database that looks like this:

Employee Name	Years of Experience
ABC	8
XYZ	20
PQR	10
MNO	15

Suppose a company wants to compare the salaries of the new joiners. Here are the data values:

Employee Name	Salary
ABC	10,000
XYZ	25,000
PQR	8,000
MNO	15,000

Data Normalization

Data
3
5
5
8
9
12
12
13
15
16
17
19
22
24
25
134

Compute *Q*1 and *Q*3 for the data relating to the marks of 8 students in an examination given below

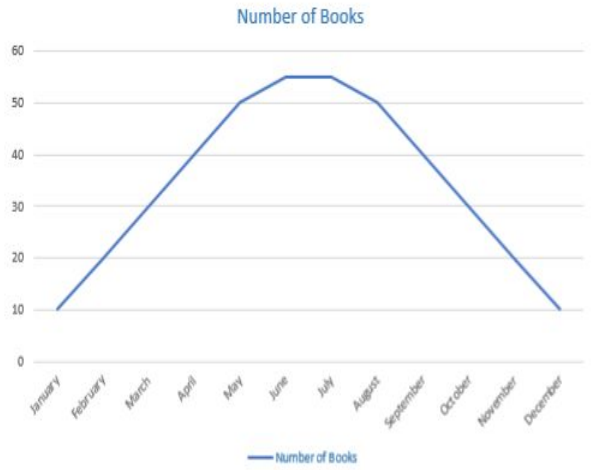
25, 48, 32, 52, 21, 64, 29, 57

Compute Q1 and Q3 for the data relating to age in years of 543 members in a village

Age in years	20	30	40	50	60	70	80
No. of members	3	61	132	153	140	51	3

Number of Books Read by Students

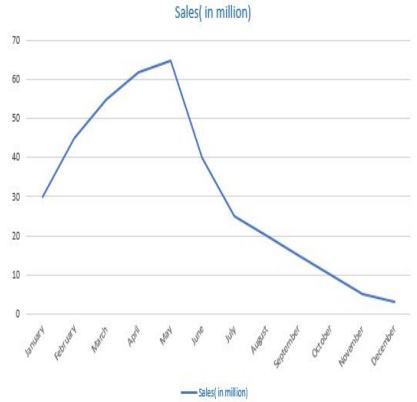
Months	Number of Books
January	10
February	20
March	30
April	40
May	50
June	55
July	55
August	50
Septembe	40
October	30
Novembe	20
Decembe	10



Normal Distribution

Amount of Sales in a Year

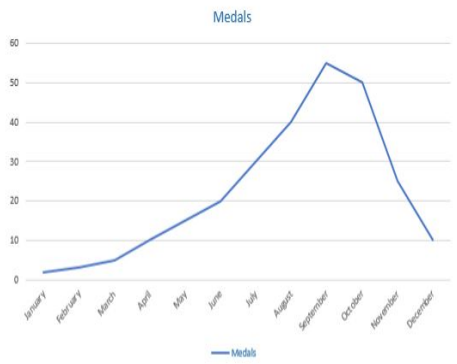
Months	Sales
January	30
February	45
March	55
April	60
May	65
June	40
July	25
August	20
Septembe	15
October	10
Novembe	5
Decembe	3



Right Skewed Distributior

Number of Medals won in a Year

Months	Medals
January	3
February	5
March	7
April	10
May	15
June	20
July	30
August	40
Septembe	55
October	50
Novembe	25
Decembe	10



Left Skewed Distribution

Decimal Scaling Normalization, Min –Max normalization,

Name	Salary	Salary after Decimal Scaling
ABC	10,000	0.1
XYZ	25, 000	0.25
PQR	8, 000	0.08
MNO	15,000	0.15

For 8 years of experience: **v'= 0**

For 10 years of experience: **v' = 0.16**

For 15 years of experience: **v' = 0.58**

For 20 years of experience: **v' = 1**

mean of this dataset is 21.2 also the standard deviation is 29.8.

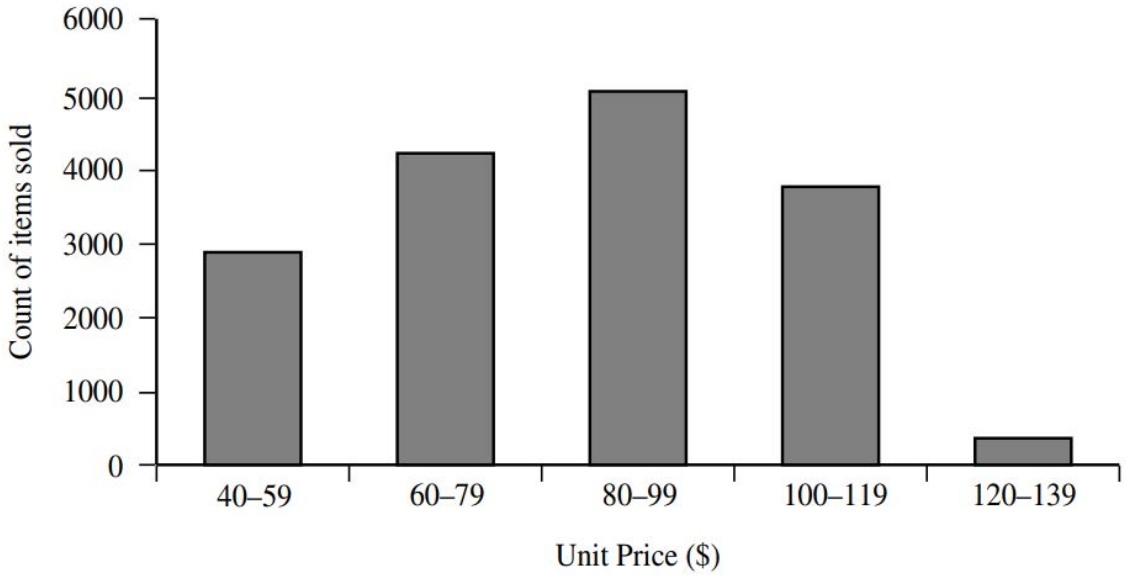
Graphic Displays of Basic Descriptive Data Summaries

- Histograms
- Quantile plots/q-q plots
- Scatter plots
- loess curves

Graphic Displays of Basic Descriptive Data Summaries

Histograms

- Graphical method for summarizing the distribution of a given attribute.
- A histogram for an attribute A partitions the data distribution of A into disjoint subsets, or buckets.
- Typically, the width of each bucket is uniform.
- Each bucket is represented by a rectangle whose height is equal to the count or relative frequency of the values at the bucket.

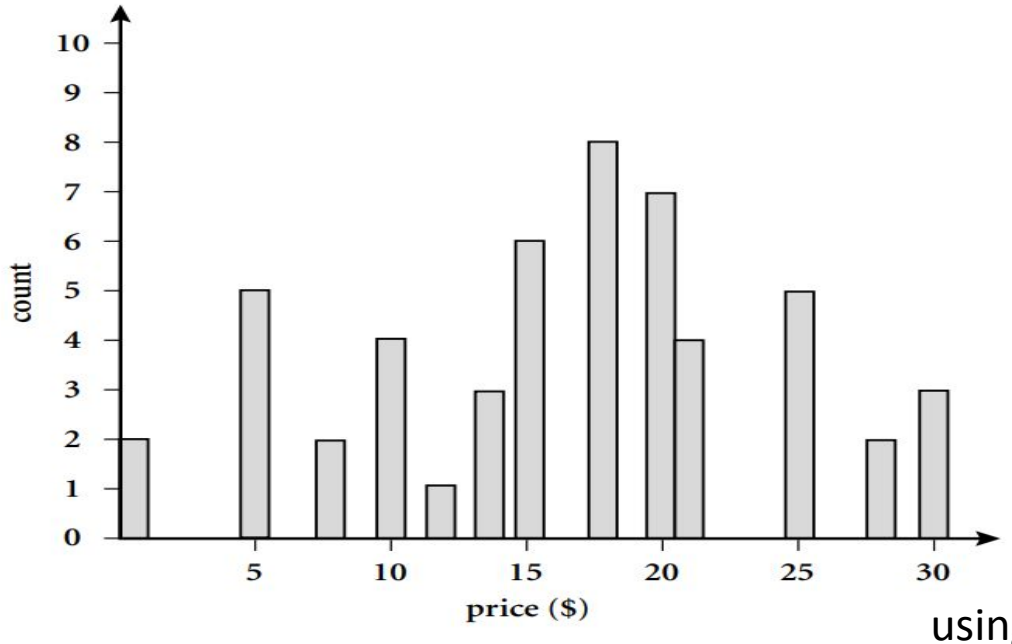


Example: Histograms

The following data are a list of prices of commonly sold items at *AllElectronics* (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Example: Histograms

The following data are a list of prices of commonly sold items at *AllElectronics* (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

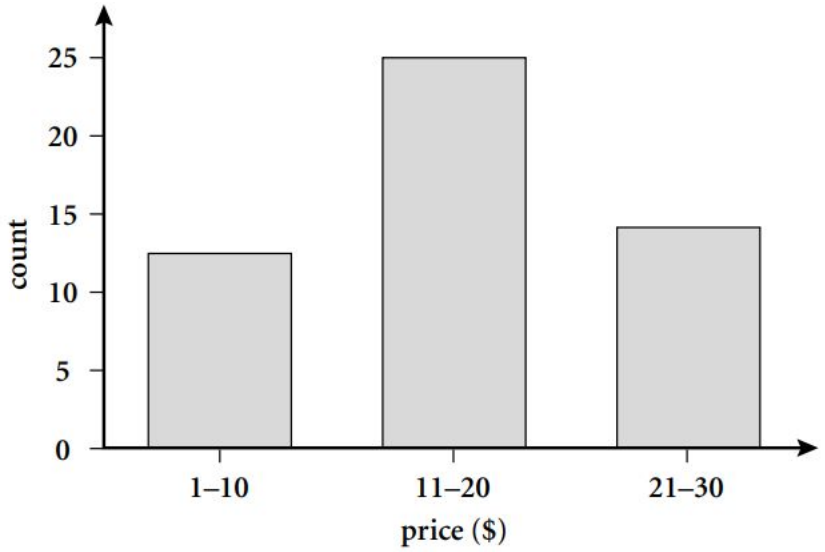


using singleton buckets

Histograms: Equal-Width

It is common to have each bucket denote a continuous range of values for the given attribute

- Equal-width: In an equal-width histogram, the width of each bucket range is uniform (such as below).



Example

The average amount of sleep, in hours, that 30 students in class 9 get on a weekday is given in the table below. [USE EQUAL WIDTH HISTOGRAM]

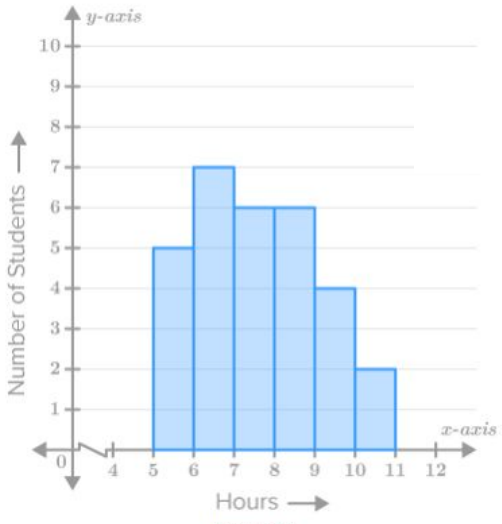
1. Construct a frequency table for the data.
2. Draw a histogram for the data.
3. Calculate the percentage of students who get an average of at least 8 hours of sleep on a weekday.

8	7	6	5	8	9	5	6	10	8
7	7	8	5	6	9	5	7	10	6
6	6	7	5	9	8	6	8	9	7

Example- Solution

i. The frequency table for the given data is:

Average amount of sleep (hours)	5-6	6-7	7-8	8-9	9-10	10-11
Number of students	5	7	6	6	4	2



iii. Percentage of students who get an average of at least 8 hours of sleep on a weekday $= \frac{12}{30} \times 100 = 40\%.$

Histograms: Equal-Width

- Advantages
  - natural order of the attribute-values is preserved.
  - Light on storage requirements
- Disadvantages:
  - High variance
  - Difficult to estimate errors

Histograms

It is common to have each bucket denote a continuous range of values for the given attribute

- Equal-frequency (or equi-depth): In an equal-frequency histogram, the buckets are created so that, roughly, the frequency of each bucket is constant (that is, each bucket contains roughly the same number of contiguous data samples). [Buckets have same height]
- V-Optimal: If we consider all of the possible histograms for a given number of buckets, the V-Optimal histogram is the one with the least variance. Histogram variance is a weighted sum of the original values that each bucket represents, where bucket weight is equal to the number of values in the bucket.
- MaxDiff: In a MaxDiff histogram, we consider the difference between each pair of adjacent values. A bucket boundary is established between each pair for pairs having the β−1 largest differences, where β is the user-specified number of buckets.

V-Optimal Histogram

- a V-optimal histogram have a Sort Value of Value, a Source Value of Frequency, and a Partition Class of Serial.
- In practice, almost all histograms used in research or commercial products are of the Serial class, meaning that sequential sort values are placed in either the same bucket, or sequential buckets.
- For example, values 1, 2, 3 and 4 will be in buckets 1 and 2, or buckets 1, 2 and 3, but never in buckets 1 and 3

Consider set of data ofa list of integers:

1, 3, 4, 7, 2, 8, 3, 6, 3, 6, 8, 2, 1, 6, 3, 5, 3, 4, 7, 2, 6, 7, 2

Computing the value and frequency pairs **(1, 2), (2, 4), (3, 5), (4, 2), (5, 1), (6, 4), (7, 3), (8, 2)**

V-Optimal Histogram

- The V-optimality rule states that the cumulative weighted variance of the buckets must be minimized.

The weighted variance *[of the set p={0.38, 0.42} with weights W={0.50,0.50}]* equals 0.5(0.38-.0.40)^2+0.5(0.42-0.40)^2 =0.0004.

$$s^2 = \frac{V_1}{V_1^2 - V_2} \sum_{i=1}^{N} w_i \left(x_i - \mu^*\right)^2,$$

V-Optimal Histogram

Option 1: Bucket 1 contains values 1 through 4. Bucket 2 contains values 5 through 8.

Bucket 1:
Average frequency 3.25
Weighted variance 2.28

Bucket 2:
Average frequency 2.5
Weighted variance 2.19

Sum of Weighted Variance 4.47

V-Optimal Histogram

Option 2: Bucket 1 contains values 1 through 2. Bucket 2 contains values 3 through 8.

Bucket 1:
Average frequency 3
Weighted variance 1.41

Bucket 2:
Average frequency 2.83
Weighted variance 3.29

Sum of Weighted Variance 4.70

The first choice is better,
Bucket 1: Range (1–4), Average Frequency 3.25
Bucket 2: Range (5–8), Average Frequency 2.5

Advantages of V-optimality vs. equi-width or equi-depth

- V-optimal histograms do a better job of estimating the bucket contents.
- The partition rule used in VOptimal histograms attempts to have the smallest variance possible among the buckets, which provides for a smaller error.
- Research demonstrated that the most accurate estimation of data is done with a VOptimal histogram using value as a sort parameter and frequency as a source parameter.

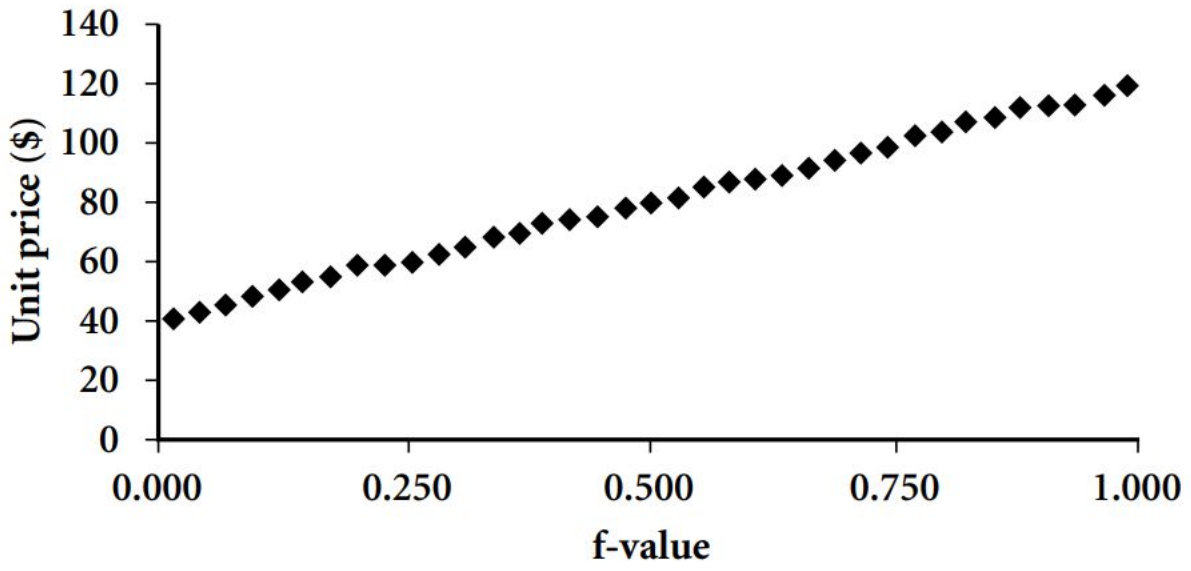
Disavantages of V-optimality vs. equi-width or equi-depth

- V-optimal histogram is a difficult structure to update. Any changes to the source parameter could potentially result in having to re-build the histogram entirely, rather than updating the existing histogram. An equi-width histogram does not have this problem.

- Computing the V-optimal histogram is computationally expensive to compute compared to other types of histograms.

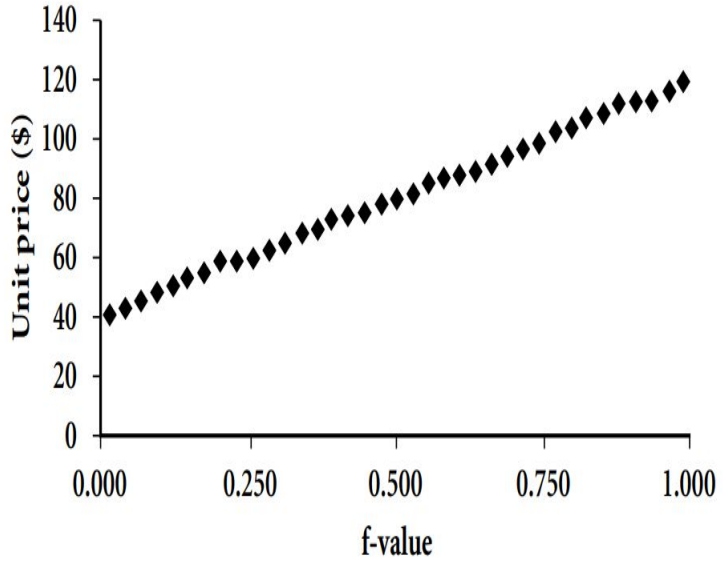
Graphic Displays of Basic Descriptive Data Summaries

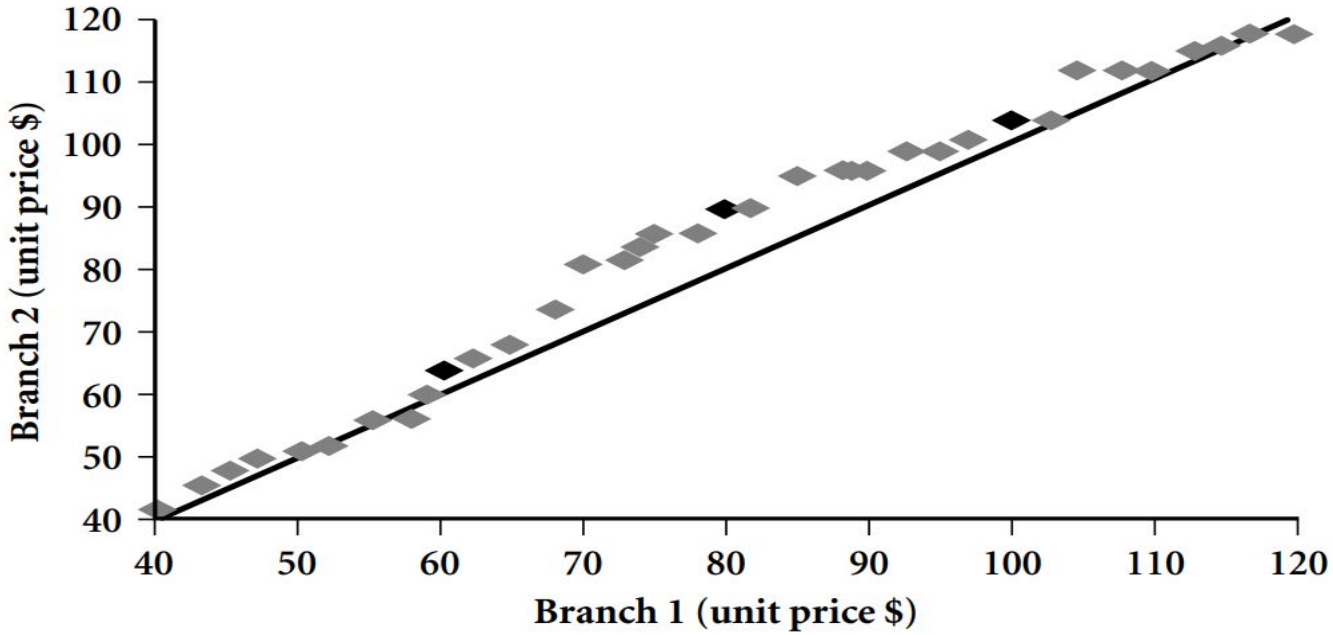
- Quantile plots
  - simple and effective way to have a first look at a univariate data distribution.
  - displays all of the data for the given attribute
  - it plots quantile information



Graphic Displays of Basic Descriptive Data Summaries

- q-q plots [quantile-quantile plot]
  - graphs the quantiles of one univariate distribution against the corresponding quantiles of another
  - view whether there is a shift in going from one distribution to another.





7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

Step 1: **Order the items from smallest to largest**.

3.77

4.25

4.50

5.19

5.89

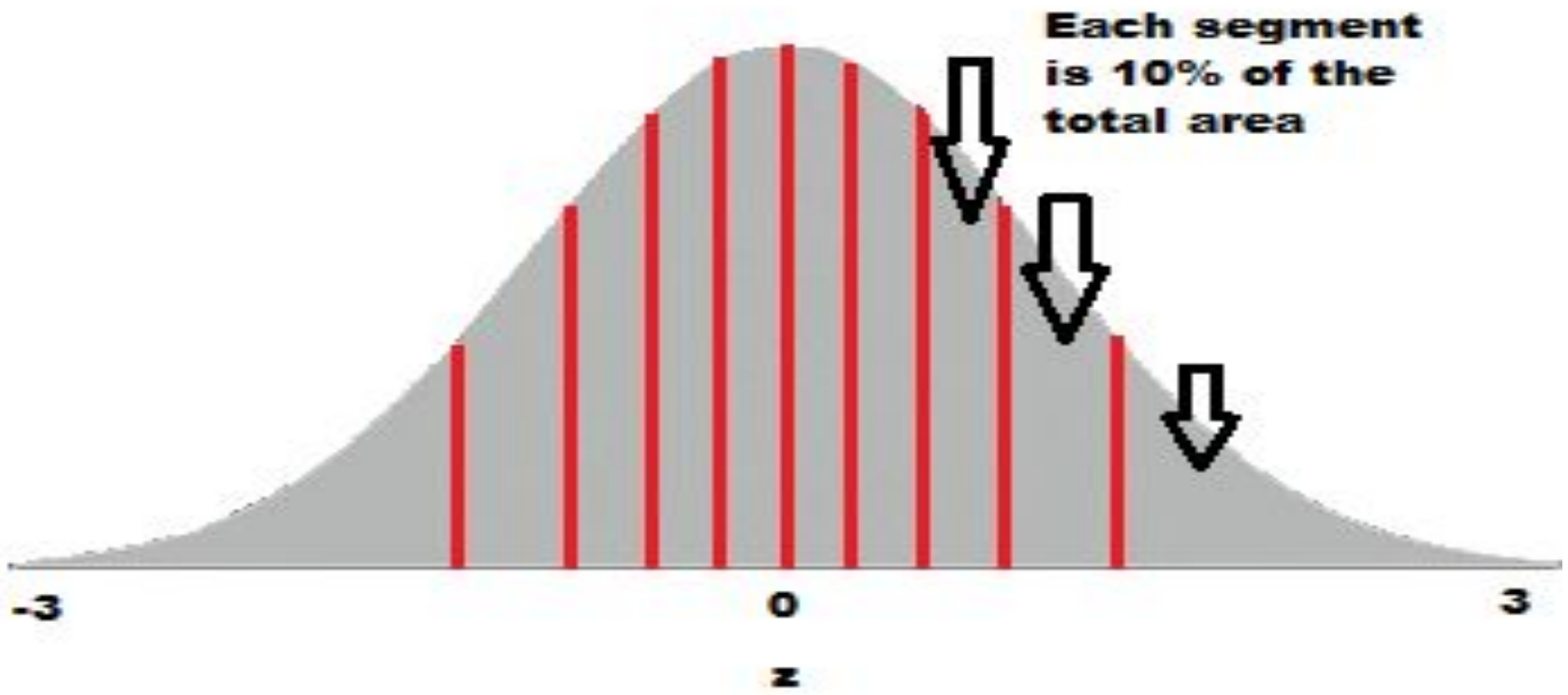
5.79

6.31

6.79

7.19

Step 2: **Draw a normal distribution curve.** Divide the curve into n+1 segments.



Z-Score Formula

The statistical formula for a value's z-score is calculated using the following formula:

$z = (x - \mu) / \sigma$

Where:

z = Z-score

x = the value being evaluated

μ = the mean

σ = the standard deviation

Step 3: **Find the z-value (cut-off point) for each segment**

10% = -1.28

20% = -0.84

30% = -0.52

40% = -0.25

50% = 0

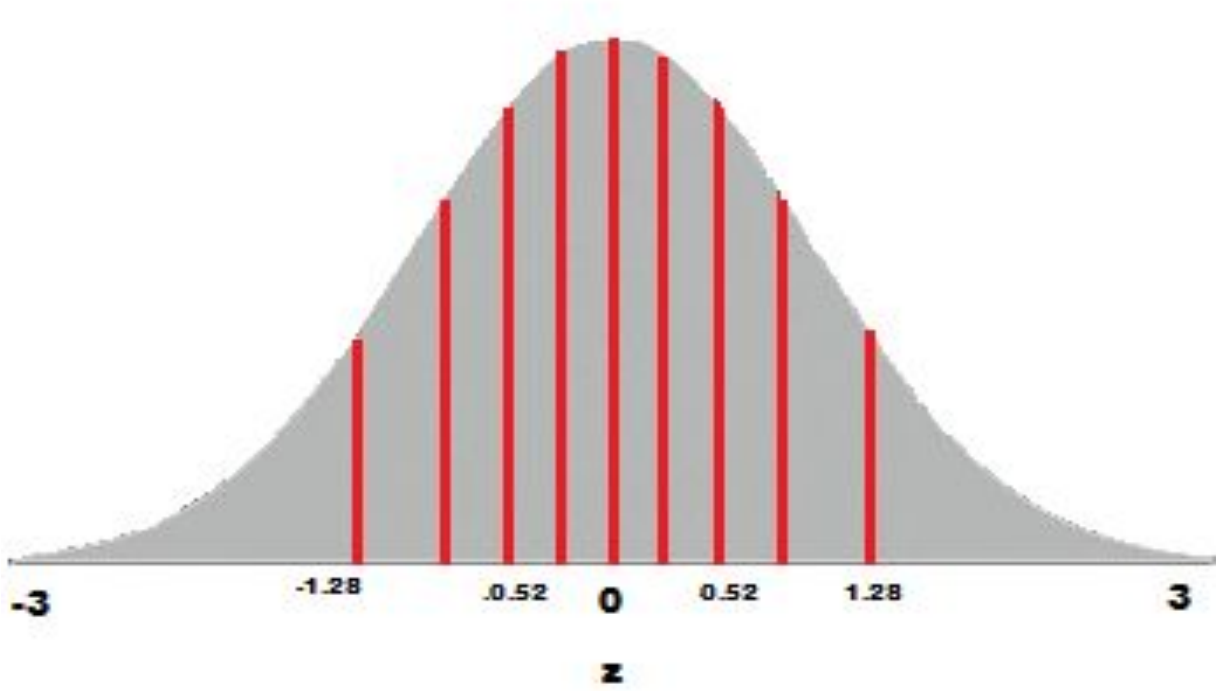
60% = 0.25

70% = 0.52

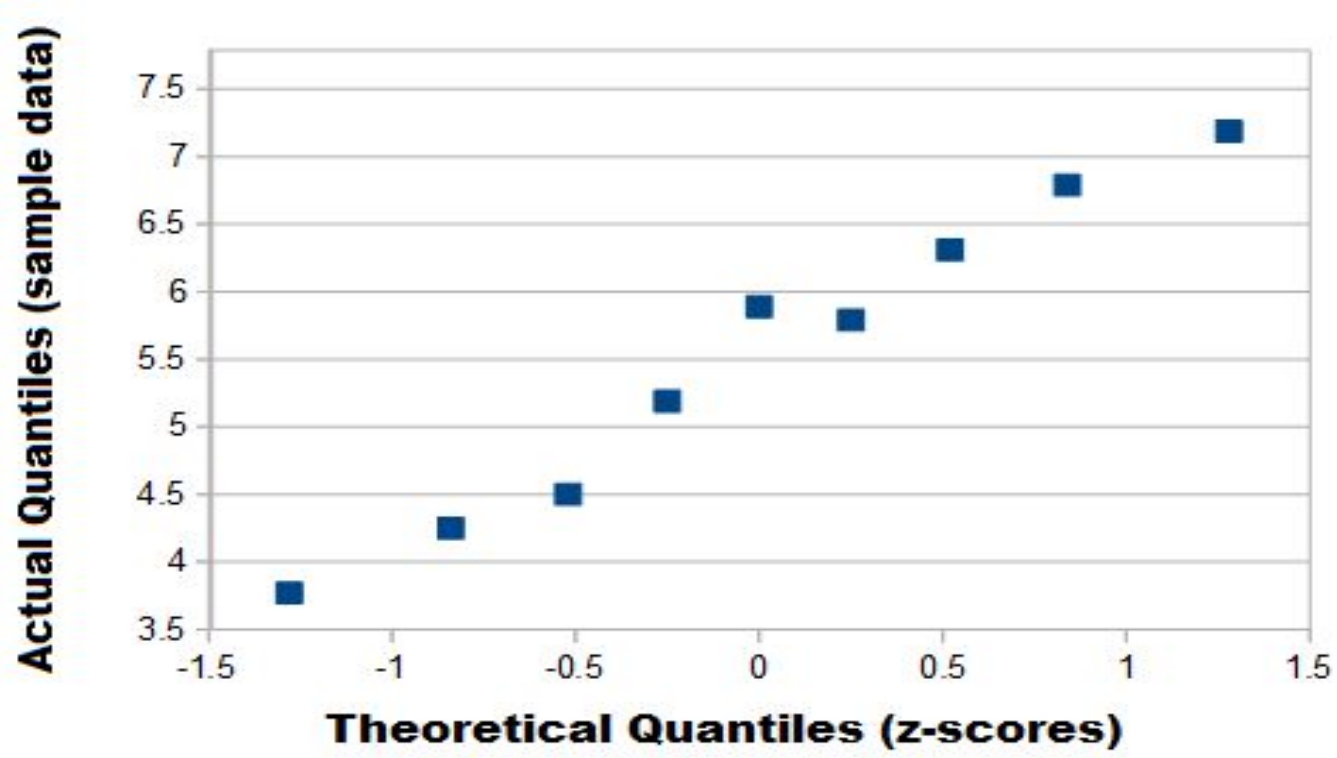
80% = 0.84

90% = 1.28

100% = 3.0

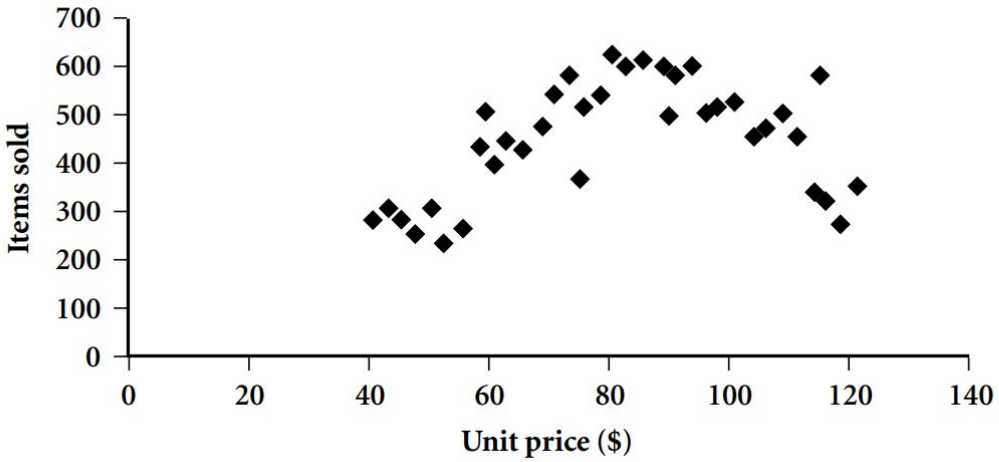


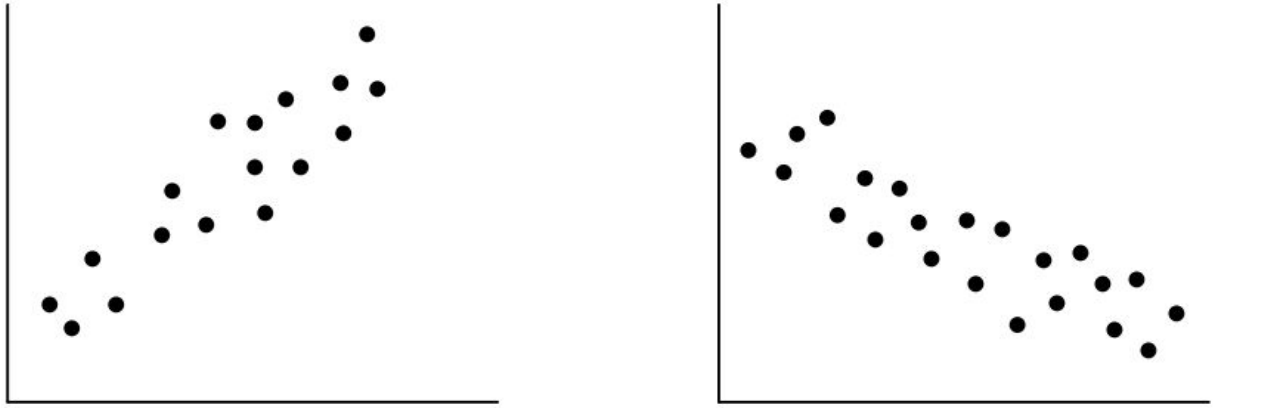
Step 4: Plot data set values (Step 1) against normal distribution cut-off points (Step 3)



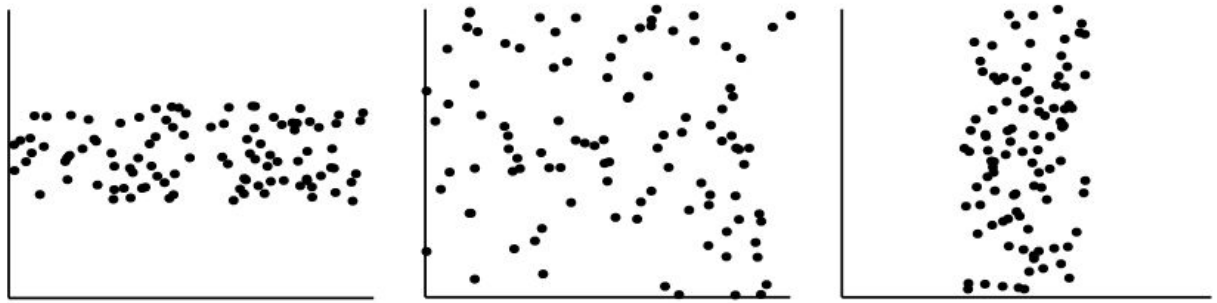
Graphic Displays of Basic Descriptive Data Summaries

- Scatter plots
  - if there appears to be a relationship, pattern, or trend between two numerical attributes.





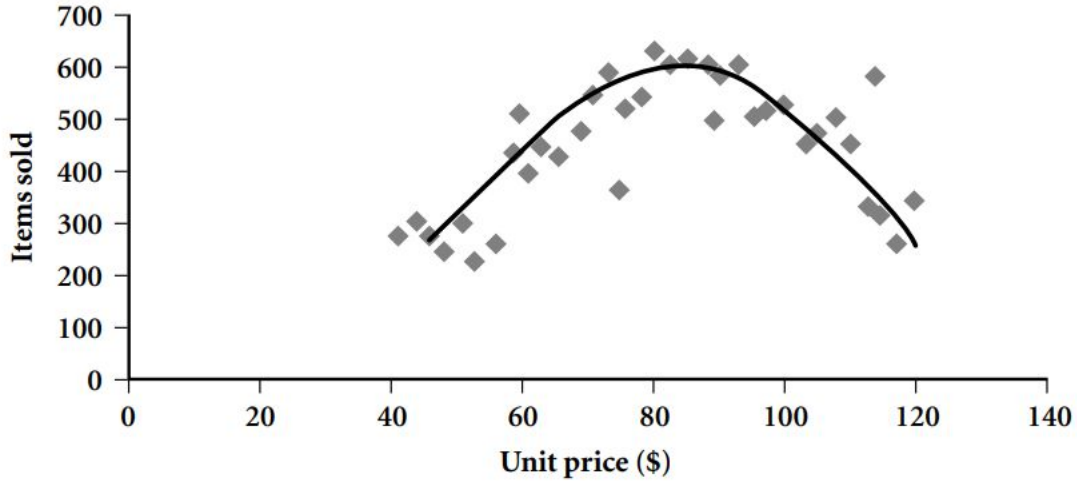
Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.



Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

Graphic Displays of Basic Descriptive Data Summaries

- loess curves
  - adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
  - loess is short for “local regression.”



Basic Data Analytics Techniques

There are several different analytical methods and techniques data analysts can use to process data and extract information. Some of the most popular methods are listed below.

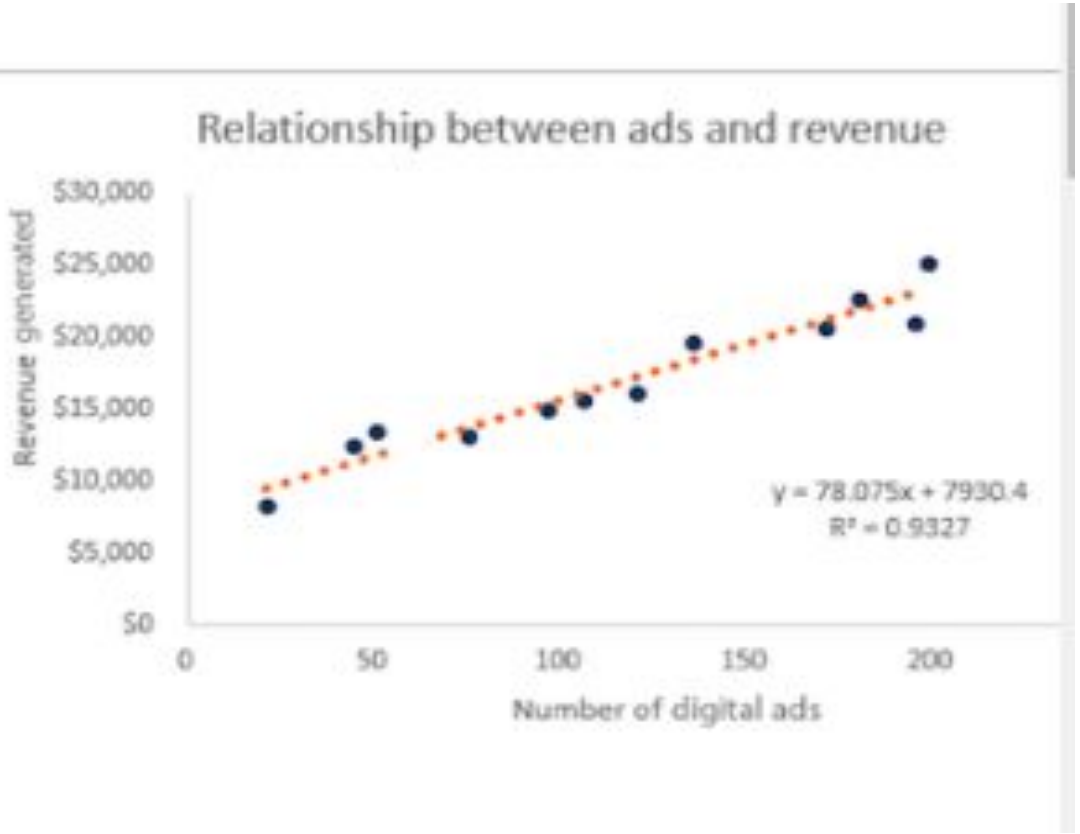
- Regression analysis
- Monte Carlo simulation
- Factor analysis
- Cohort analysis
- Cluster analysis
- Time series analysis

Basic Data Analytics Technique: Regression analysis

- Regression analysis is used to estimate the **relationship between a set of variables**
- Find **correlation** between **dependent** variable (outcome you want to measure or predict) and any number of **independent** variables (factors which may have an impact on the dependent variable).
- The aim of regression analysis is to estimate how one or more variables might impact the dependent variable, in order to identify trends and patterns. This is especially useful for making predictions and forecasting future trends.
- Example: Expenditure vs sales of a company

Regression analysis

Data	Digital ads	Revenue
Jan	21	$8,350.0
Feb	180	$22,755.0
Mar	50	$13,455.0
Apr	195	$21,100.0
May	96	$15,000.0
Jun	44	$12,500.0
Jul	171	$20,700.0
Aug	135	$19,722.0
Sep	120	$16,115.0
Oct	75	$13,100.0
Nov	106	$15,670.0
Dec	198	$25,300.0
Totals	**1,391**	**$203,767.0**
Average	**116**	**$16,980.6**



Basic Data Analytics Technique: Monte Carlo simulation

- When making decisions or taking certain actions, there are a range of different possible outcomes.
- Monte Carlo simulations model the probability of different outcomes happening.
- It essentially considers a range of possible outcomes and then calculates how likely it is that each particular outcome will be realized.
- The Monte Carlo method is used by data analysts to conduct advanced risk analysis, allowing them to better forecast what might happen in the future and make decisions accordingly.
- It is used for risk mitigation and loss prevention, these simulations incorporate multiple values and variables and often have greater forecasting capabilities than other data analytics approaches.

Monte Carlo Simulation Example

A work schedule for a research and development project. One can use the Monte Carlo Simulation to analyze the impact of risks that will affect project.

Activities	Optimistic	Pessimistic	Most Likely
Choose a Topic	4	7	5
Develop Research Plan	5	7	6
Complete the Research	7	9	8
Report	2	4	3

Monte Carlo Simulation Example

calculate the duration of each activity by using PERT Formula

PERT Estimate = (Optimistic Estimate+ 4 x Most likely Estimate + Pessimistic Estimate) / 6

Activities	Optimistic	Pessimistic	Most Likely	PERT Estimate
Choose a Topic	4	7	5	5,2
Develop Research Plan	5	7	6	6
Complete the Research	7	9	8	8
Report	2	4	3	3

Monte Carlo Simulation Example

Total Completion Time of the project is = 5,2 +6 + 8 + 3 = 22,2 Months.

For the best case, completion time of the project is ;

Total Completion Time = 4 +5 + 7 + 2 = 18 Months.

For the worst case, completion time of the project is ;

Total Completion Time = 7 +7 + 9 + 4 = 27 Months.

Now run the Monte Carlo Simulation by using Excel or software and get the chances of completion of the project.

Duration of Project (Months)	Possibility of Completion (%)
18	10%
19	17%
20	24%
21	28%
22	35%
23	46%
24	57%
25	69%
26	88%
27	100%

- It enables to make realistic forecasts or manage activities that involve uncertainty.
- It enables to get accurate results by exploring thousands of combinations with "what-if" analysis.
- In this simulation, it's possible to model interdependent relationships between input variables.
- It helps to improve the quality of decisions.
- It helps to make forecasts for budget, schedule, and other important project work.
- This tool provides graphical and visual data. This helps to improve communication among project team and stakeholders.
- This tool shows the inputs which have the biggest effect on the result.

Basic Data Analytics Technique: Factor Analysis

- Factor analysis is a technique used to reduce a large number of variables to a smaller number of factors.
- Factor analysis entails taking a large data set and shrinking it to a smaller data set. The goal of this maneuver is to attempt to discover hidden trends that would otherwise have been more difficult to see.
- Use factor analysis to group them into factors that belong together or that are strongly correlated. This is known as covariance.
- Example: Survey Forms

Factor Analysis

Factor analysis uses the correlation structure amongst observed variables to model a smaller number of unobserved, latent variables known as factors.

Analysts often refer to the observed variables as indicators because they literally indicate information about the factor.

Factor analysis treats these indicators as linear combinations of the factors in the analysis plus an error.

The procedure assesses how much of the variance each factor explains within the indicators.

socioeconomic status (SES) is a factor that can't measure directly. However, one can assess **occupation**, **income**, and **education** levels. These variables all relate to socioeconomic status. People with a particular socioeconomic status tend to have similar values for the observable variables. If the factor (SES) has a strong relationship with these indicators, then it accounts for a large portion of the variance in the indicators.

Basic Data Analytics Technique: Cohort Analysis

- Cohort analysis is the process of breaking a data set into groups of similar data, often broken into a customer demographic. This allows data analysts and other users of data analytics to further dive into the numbers relating to a specific subset of data.
- Cohort analysis is a subset of behavioral analytics that takes the data from a given dataset and rather than looking at all users as one unit, it breaks them into related groups for analysis. These related groups, or cohorts, usually share common characteristics or experiences within a defined time-span.
- This is useful because it allows companies to tailor their service to specific customer segments (or cohorts)

App Launched ↓ % Active users after App Launches →

Cohort	Users	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Jan 25	1,098	100%	33.9%	23.5%	18.7%	15.9%	16.3%	14.2%	14.5%	Retention over user lifetime		12.1%
Jan 26	1,358	100%	31.1%	18.6%	14.3%	16.0%	14.9%	13.2%	12.9%			
Jan 27	1,257	100%	27.2%	19.6%	14.5%	12.9%	13.4%	13.0%	10.8%	11.4%		
Jan 28	1,587	100%	26.6%	17.9%	14.6%	14.8%	14.9%	13.7%	11.9%			
Jan 29	1,758	100%	26.2%	20.4%	16.9%	14.3%	12.7%	12.5%				
Jan 30	1,624	100%	26.4%	18.1%	13.7%	15.4%	11.8%					
Jan 31	1,541	100%	23.9%	19.6%	15.0%	14.8%						
Feb 01	868	100%	24.7%	16.9%	15.8%							
Feb 02	1,143	Retention over product lifetime		18.5%								
Feb 03	1,253											
All Users	**13,487**	**100%**	**27.0%**	**19.2%**	**15.4%**	**14.9%**	**14.0%**	**13.3%**	**12.5%**	**13.1%**	**12.2%**	**12.1%**

From the retention table – Triangular chart, one can infer the following

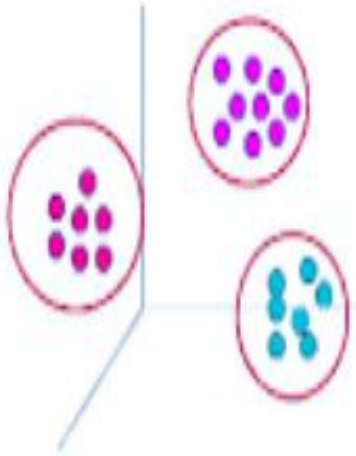
- 1358 users launched an app on Jan 26. Day 1 retention was 31.1%, day 7 retention was 12.9%, and day 9 retention was 11.3%. So on the 7^{th} day after using the app, 1 in 8 users who launched an app on Jan 26 were still active users on the app.
- Out of all of the new users during this time range (13,487 users), 27% users are retained on day 1, 12.5% on day 7, and 12.1% on day 10.


100%
80%
60%
40%
20%
0%
% of users retained
27.0%
19.2%
15.4%
14.9%
14.0%
13.3%
12.5%
13.1%
12.2%
12.1%
Day 0
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Day 7
Day 8
Day 9
Day 10
Day

Basic Data Analytics Technique: Cluster Analysis

- Cluster analysis is an exploratory technique that seeks to identify structures within a dataset.
- The goal of cluster analysis is to sort different data points into groups (or clusters) that are internally homogeneous and externally heterogeneous.
  - data points within a cluster are similar to each other, and dissimilar to data points in another cluster
- Clustering is used to gain insight into how data is distributed in a given dataset, or as a preprocessing step for other algorithms.
- Example: In marketing, cluster analysis is commonly used to group a large customer base into distinct segments, allowing for a more targeted approach to advertising and communication.

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Retail companies often use clustering to identify groups of households that are similar to each other.
For example, a retail company may collect the following information on households:

- Household income
- Household size
- Head of household Occupation
- Distance from nearest urban area

- **Cluster 1:** Small family, high spenders
- **Cluster 2**: Larger family, high spenders
- **Cluster 3**: Small family, low spenders
- **Cluster 4**: Large family, low spenders

Basic Data Analytics Technique: Time Series Analysis

- Time series analysis is a statistical technique used to identify trends and cycles over time.
- Time series data is a sequence of data points which measure the same variable at different points in time (for example, weekly sales figures or monthly email sign-ups).
- Time series analysis tracks data over time and solidifies the relationship between the value of a data point and the occurrence of the data point.
- By looking at time-related trends, analysts are able to forecast how the variable of interest may fluctuate in the future.
- Mainly used for financial forecasts.

When conducting time series analysis, the main patterns you'll be looking out for in your data are:

Trends: Stable, linear increases or decreases over an extended time period.

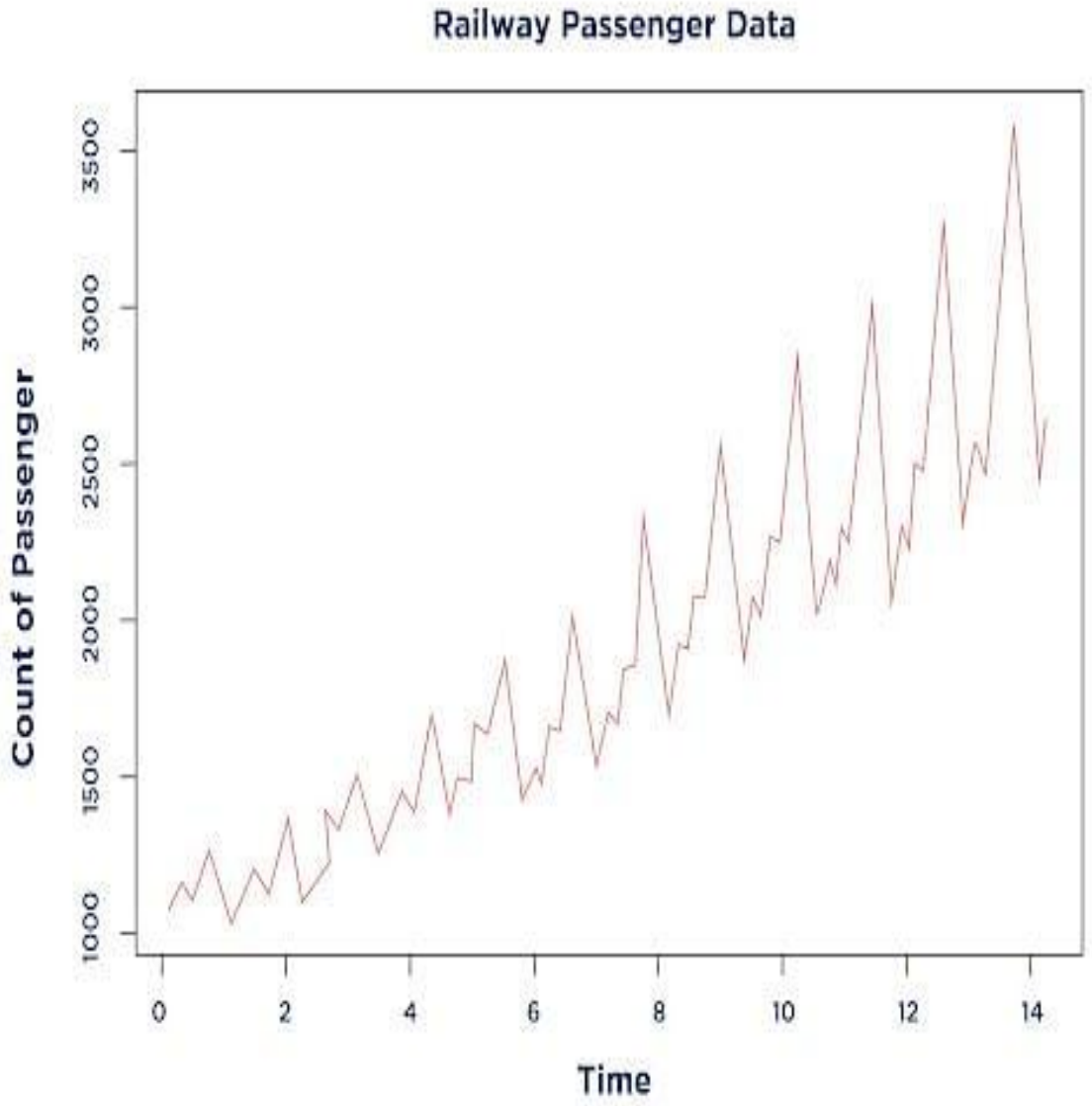
Seasonality: Predictable fluctuations in the data due to seasonal factors over a short period of time. For example, you might see a peak in swimwear sales in summer around the same time every year.

Cyclic patterns: Unpredictable cycles where the data fluctuates. Cyclical trends are not due to seasonality, but rather, may occur as a result of economic or industry-related conditions.

Time Series Analysis

Time series analysis can be used in -

- Rainfall measurements
- Automated stock trading
- Industry forecast
- Temperature readings
- Sales forecasting



The following observations can be derived from the given data.

1. Trend: Over time, an increasing or decreasing pattern has been observed. The total number of passengers has risen over time.
2. Seasonality: Cyclic patterns are the ones that repeat after a certain interval of time. In the case of the railway passenger, you can see a cyclic pattern with a high and low point that is visible throughout the interval.