Data Preprocessing: Concepts and Techniques

Lecture 02

Data Preprocessing

An Overview

- Data Quality
- Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Quality: Why Preprocess the Data?

Assessment of data quality: A comprehensive perspective

Precision: right or incorrect, precise or imprecise

Integrity: absent recordings, inaccessible, ...

Uniformity: certain alterations while others remain unchanged, inconsistencies, ...

Punctuality: timely updates?

Credibility: the reliability of data accuracy?

Comprehensibility: the ease with which data can be grasped?

Major Tasks in Data Preprocessing

Data cleaning
- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration
- Integration of multiple databases, data cubes, or files

Data reduction
- Dimensionality reduction
- Data compression

Data transformation and data discretization
- Normalization
- Concept hierarchy generation

Data Cleaning

- Real Data: flawed in various way, e.g., instrument faulty, human or computer error, transmission error
  - Incomplete Data: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., Occupation=“ ” (missing data)
  - Noisy Data: containing noise, errors, or outliers
    - e.g., Age=“−10” (an error)
  - Inconsistent Data: containing discrepancies in codes or names, e.g.,
    - Age=“42”, Birthday=“03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
    - discrepancy between duplicate records
  - Intentional Data:(e.g., disguised missing data)
    - Jan. 1 as everyone’s birthday?

Missing Data

- Unavailable Data (sometimes)
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - **equipment malfunction**

Weather Monitoring System:

Date	Temperature (°C)	Humidity (%)	Wind Speed (km/h)
2023-07-01	25.0	60	15
2023-07-02	26.5	58	10
2023-07-03	N/A	55	20
2023-07-04	N/A	62	12
2023-07-05	27.0	59	18

Explanation: On July 3rd and 4th, the temperature sensor malfunctioned, resulting in no temperature data being recorded for those days.

Missing Data

Missing data may be due to

- **inconsistent with other recorded data and thus deleted: Sales Database**

Transaction ID	Customer ID	Product ID	Quantity	Price	Total
001	1001	2001	2	$50.00	$100.00
002	1002	2002	-50	$30.00	N/A
003	1003	2003	3	$20.00	$60.00
004	1004	2004	1	$15.00	$15.00

Explanation: Transaction 002 was flagged and the total value was deleted because it showed an impossible purchase quantity of -50 units, which is inconsistent with other recorded data.

- **data not entered due to misunderstanding**

Hospital Patient Records:

Patient ID	Age	Gender	Smoking Status	Diagnosis
001	50	Female	N/A	Diabetes
002	37	Male	Non-smoker	Asthma

Explanation: The "smoking status" field is missing for several patients because a nurse misunderstood the form and left this field blank for those entries.

Missing Data

Missing data may be due to

- **certain data may not be considered important at the time of entry**

Customer Database:

Customer ID	Name	Email	Occupation
1	Alice	alice@example.com	N/A
2	Bob	bob@example.com	N/A
3	Charlie	charlie@example.com	Engineer
4	Daisy	daisy@example.com	Teacher

Explanation: The company initially did not record customers' occupation, considering it unimportant. Later, when analyzing customer demographics, this data was found missing for initial entries.

- **not register history or changes of the data**

Product Inventory System:

Date	Product ID	Product Name	Price
2023-07-15	3001	Widget A	$12
2023-08-01	3001	Widget A	$15
2023-08-15	3002	Widget B	$20

Explanation: The price of Widget A is updated regularly, but past prices are not stored. If analysis requires historical pricing data, only the most recent price is available, and previous prices are missing.

- In such cases, it may be necessary to infer missing data.

How to Handle Missing Data?

Student Grades Data:

Student ID	Age	Exam Score	Grade
1	20	85	A
2	21	78	B
3	21	N/A	N/A
4	22	88	A
5	23	90	A

Ignore the tuple:
usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
Student ID 2 does not have grade in the data given. So, we can ignore that row.

Student ID	Age	Exam Score	Grade
1	20	85	A
2	21	78	B
4	22	88	A
5	23	90	A

How to Handle Missing Data?

Student Grades Data:

Student ID	Age	Exam Score	Grade
1	20	85	A
2	21	N/A	B
3	N/A	78	C
4	22	88	A
5	23	90	A

Fill in it automatically with

a global constant : e.g., "unknown", a new class?!

Use a global constant like "unknown" for missing values.

Student ID	Age	Exam Score	Grade
1	20	85	A
2	21	unknown	B
3	unknown	78	C
4	22	88	A
5	23	90	A

How to Handle Missing Data?

- **Student Grades Data:**

Student ID	Age	Exam Score	Grade
1	20	85	A
2	21	N/A	B
3	N/A	78	C
4	22	88	A
5	23	90	A

- **Fill in it automatically with**
  - **the attribute mean**

Student ID	Age	Exam Score	Grade
1	20	85	A
2	21	85.25	B
3	21.5	78	C
4	22	88	A
5	23	90	A

Age mean = 21.5,
Exam Score mean = 85.25

How to Handle Missing Data?

Student Grades Data:

Student ID	Age	Exam Score	Grade
1	20	85	A
2	21	N/A	B
3	N/A	78	C
4	22	88	A
5	23	90	A

Fill in it automatically with

the attribute mean for all samples belonging to the same class: smarter

Student ID	Age	Exam Score	Grade
1	20	85	A
2	21	**85**	B
3	**22**	78	C
4	22	88	A
5	23	90	A

How to Handle Missing Data?

Fill in it automatically with
the most probable value: inference-based such as Bayesian formula or decision tree
(Will be explained with examples and hands-on in classification topic)

Fill in the missing value manually —------ but this process will be —---- tedious + infeasible

Numerical

Fill in Missing Values: Given the dataset:

Age	Salary
25	5000
30	?
35	8000
?	10000
27	5500

Noisy Data

- Noise: random error or variance in a measured variable
  - A temperature sensor records 100°C in one reading while nearby sensors show around 22°C. The 100°C reading is likely random noise.
- Incorrect attribute values may be due to
  - **faulty data collection instruments**
    - A humidity sensor reads 120% (which is impossible) due to a malfunction.
  - **data entry problems**
    - An income field incorrectly shows -$5,000 due to a typing error.
  - **data transmission problems**
    - A transaction record shows missing data (e.g., Amount is N/A) because of a transmission error.
  - **inconsistency in naming convention**
    - "john doe" vs. "John Doe" in user records, leading to inconsistent naming.
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Numerical

Smooth Noisy Data: Given the dataset:

Temperature
20
22
21
100
23
20
25
19

Apply a moving average filter with a window size of 3 to smooth the data.

Data Cleaning - IQR (Interquartile Range) method

The IQR method is used to identify and remove outliers from a dataset.
It is based on the range within which the middle 50% of data values lie.
Outliers are values that fall outside this range.

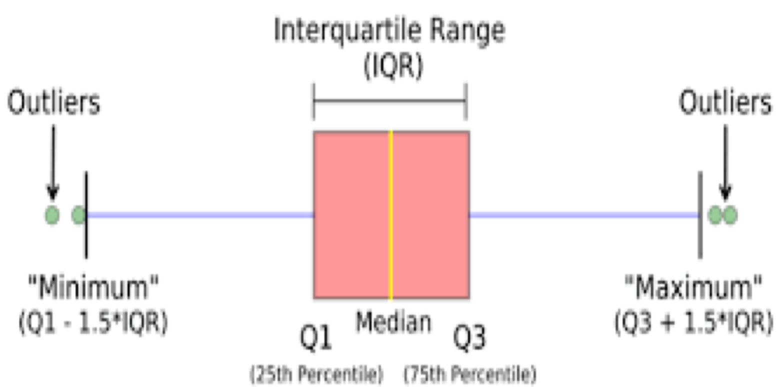
Quartiles divide a dataset into four equal parts. Q1 (First Quartile) is the 25th percentile, and Q3 (Third Quartile) is the 75th percentile. Here's how to calculate them:

Steps to Calculate Q1 and Q3:

Sort the Data: Arrange your data in ascending order.

Determine the Position of Q1 and Q3:

- **Q1 Position:** Position of Q1=(N+1)/4
- **Q3 Position:** Position of Q3=3×(N+1)/4
- Where N is the number of data points.



Find Q1 and Q3:

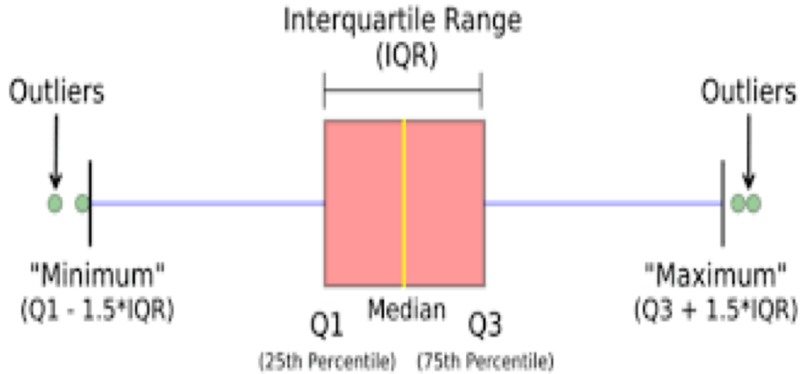
- **If the position is an integer:** The quartile value is the data point at that position.
- **If the position is not an integer:** Interpolate between the two closest data points.

Data Cleaning - IQR (Interquartile Range) method

Calculate the IQR:

Find the first quartile (Q1) and the third quartile (Q3).

Compute the IQR as IQR=Q3−Q1



Determine the bounds:

Calculate the lower bound as Lower Bound=Q1−1.5×IQR

Calculate the upper bound as Upper Bound=Q3+1.5×IQR

Identify outliers:

Any data point outside the lower and upper bounds is considered an outlier.

Remove the outlier from the data set.

Numerical

Identify or Remove Outliers: Given the dataset:

Scores
80
86
79
150
88
90
84
92
25

Identify the outlier using the IQR (Interquartile Range) method and remove it.

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Example- Integration of Multiple Databases

Database 1:

ID	Age
1	34
2	45
3	27
4	40
5	19

Database 2:

ID	Name
1	John
2	Alice
3	Bob
4	Rahul
5	Harry

Integrating the databases to form a single dataset

ID	**Name**	**Age**
1	John	34
2	Alice	45
3	Bob	27
4	Rahul	40
5	Harry	19

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - Object identification: The same attribute or object may have different names in different databases
  - Derivable data: One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by correlation analysis and covariance analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Reduction Strategies

Data reduction:

Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Why data reduction?

A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data Reduction Strategies

Data reduction strategies

- Dimensionality reduction
  - Wavelet transforms

    Example: Compressing an image by converting it into wavelet coefficients and keeping only the important ones.
  - Principal Components Analysis (PCA)

    Example: Reducing a dataset of 100 features to just 10 principal components.
  - Feature subset selection, feature creation

    Example: Using only the height and weight from a health dataset, ignoring other less important features.

- Numerosity reduction (Data Reduction)
  - Regression and Log-Linear Models

    Example: Using a simple line to predict house prices based on size.
  - Histograms, clustering, sampling

    Example: Creating a bar chart to show the frequency of different age groups in a survey.
  - Data cube aggregation

    Example: Summarizing sales data by week instead of daily.

- Data compression

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: Scaled to fall within a smaller, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Discretization: Concept hierarchy climbing

Min-max normalization

Min-Max Normalization rescales feature values to a specified range, usually [0, 1].

Sometimes it rescales feature values to [-1, 1].

It is used to ensure that all features contribute equally to the analysis by standardizing their range.

$$\text{Normalized Value} = \frac{X - min(X)}{max(X) - min(X)}$$

Where:

X is the original value.

Min(X) is the minimum value of the feature.

Max(X) is the maximum value of the feature.

Person	Age
A	23
B	45
C	30
D	50
E	40

Example

Person	Age
A	23
B	45
C	30
D	50
E	40

Minimum Age (Min) = 23

Maximum Age (Max) = 50

Apply Min-Max Normalization:

For each value:

Normalized Age = (Age−23)/(50−23)

Person A: Normalized Age=(23−23)/(50−23)=0/27=0

Person B: Normalized Age=(45−23)/(50−23)=22/27≈0.81

Person C: Normalized Age=(30−23)/(50−23)=7/27≈0.26

Person D: Normalized Age=50−23)/(50−23)=27/27=1

Person E: Normalized Age=(40−23)/(50−23)=17/27≈0.63

Person	Age	Normalized Age
A	23	0.00
B	45	0.81
C	30	0.26
D	50	1.00
E	40	0.63

Numerical

Normalization: Given the dataset:

Value
100
400
200
500

Normalize the values to a range of [0, 1].

Data Discretization Methods

Typical methods: All the methods can be applied recursively

Customer ID	Age
1	23
2	45
3	31
4	52
5	37

Binning (Top-down split, unsupervised)

Binning is a data preprocessing technique that transforms numerical variables into categorical ones by dividing the range of the variable into bins.

Typically top-down, where the range of the variable is divided into intervals (bins) in an unsupervised manner.

Customer ID	Age	Age Group
1	23	20-29
2	45	40-49
3	31	30-39
4	52	50-59
5	37	30-39

Data Discretization Methods

Histogram analysis (Top-down split, unsupervised)

It involves creating histograms to understand the distribution of data points within bins.

Approach: This method can be seen as a top-down split where data is divided into bins based on the value range in an unsupervised manner.

Customer ID	Age
1	23
2	45
3	31
4	52
5	37

Customer ID	Age	Age Group
1	23	20-29
2	45	40-49
3	31	30-39
4	52	50-59
5	37	30-39

Age	Frequency
20-29	1
30-39	2
40-49	1
50-59	1

Data Discretization Methods

Clustering analysis (unsupervised, top-down split or bottom-up merge)

Clustering is the task of dividing a set of objects into groups (clusters) so that objects in the same cluster are more similar to each other than to those in other clusters.

Approach:

Top-down split: Methods like divisive clustering start with all data points in one cluster and split them recursively.

Bottom-up merge: Methods like agglomerative clustering start with each data point as its own cluster and merge them recursively.

Customer ID	Age	Purchase Amount
1	23	200
2	45	500
3	31	150
4	52	700
5	37	300

Clusters: Assuming 2 clusters based on age and purchase amount:

Cluster 1: Customers 1, 3, 5 (younger, lower purchases)

Cluster 2: Customers 2, 4 (older, higher purchases)

Data Discretization Methods

Decision-tree analysis (supervised, top-down split)

Decision-tree analysis involves using a tree-like model to make decisions based on the values of input features. It is a supervised learning method.

Approach: The method starts at the root and splits the data recursively into subsets based on feature values (top-down split).

Customer ID	Age	Purchase Amount	Repeat Purchase
1	23	200	No
2	45	500	No
3	31	150	No
4	52	700	Yes
5	37	300	No

Decision Tree:

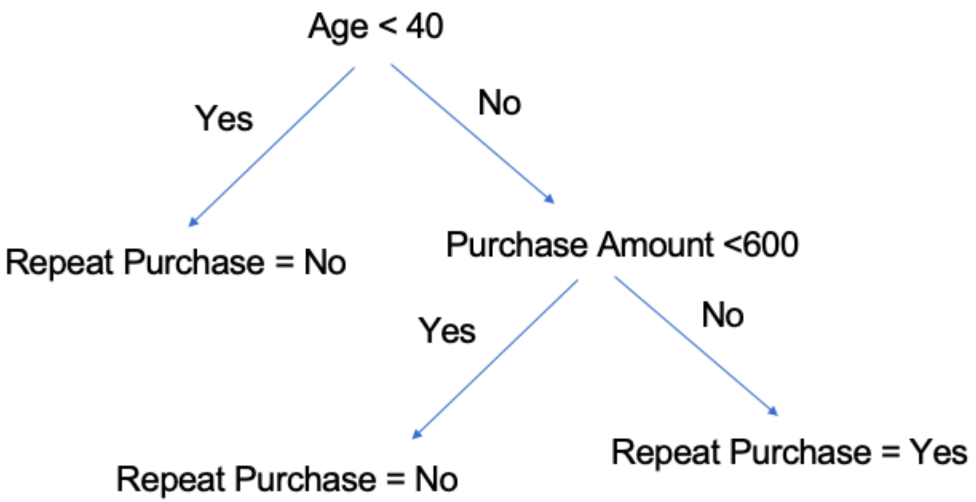
Node 1 (Root): Age

Age < 40: Repeat Purchase = No

Age ≥ 40: Purchase Amount

Amount < 600: Repeat Purchase = No

Amount ≥ 600: Repeat Purchase = Yes



Data Discretization Methods

Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

Correlation analysis examines the relationship between two or more variables. The chi-squared ($\chi 2$) test is often used to test the independence of two categorical variables.

Approach: In a bottom-up merge, variables that have a significant correlation are grouped together, which can be seen in methods like hierarchical clustering that use correlation measures to merge clusters.

Correlation (e.g., χ^2) analysis

Step 1: Create a Contingency Table

A contingency table is used to display the frequency distribution of the variables. It shows how often certain combinations of categories occur.

1. Identify the variables: In this case, the variables are "Age Group" and "Repeat Purchase".

2. Organize the data: Arrange the data into a table format, showing the counts of each combination of the variables.

Customer ID	Age	Repeat Purchase
1	23	No
2	45	Yes
3	31	No
4	52	Yes
5	37	No
6	28	No
7	55	Yes
8	49	Yes

Age Group	**Repeat Purchase: Yes**	**Repeat Purchase: No**	**Row Total**
20-29	0	2	2
30-39	0	2	2
40-49	2	0	2
50-59	2	0	2
Column Total	4	4	8

Correlation (e.g., χ^2) analysis

Step 2: Calculate Expected Frequencies

Expected frequencies are calculated to determine what the frequencies would be if there was no association between the variables.

Formula for expected frequency: Expected frequency=(Row total×Column total)/Grand total

Apply the formula: For each cell in the table, calculate the expected frequency.

Age Group	Repeat Purchase: Yes	Repeat Purchase: No	Row Total
20-29	0	2	2
30-39	0	2	2
40-49	2	0	2
50-59	2	0	2
Column Total	4	4	8

Age Group	Expected Yes	Expected No
20-29	1	1
30-39	1	1
40-49	1	1
50-59	1	1

Correlation (e.g., χ^2) analysis

Step 3: Calculate the Chi-Squared Statistic

The chi-squared statistics measures how the observed frequencies deviate from the expected frequencies

Formula for chi-squared: $$\chi^2 = \frac{(Oi - Ei)^2}{Ei}$$

Oi is the observed frequency and Ei is the expected frequency

Calculate the statistic: Compute the chi-squared value for each cell and sum them up.

Age Group	Observed (Yes)	Expected (Yes)	(O-E)²/E (Yes)	Observed (No)	Expected (No)	(O-E)²/E (No)
20-29	0	1	1	2	1	1
30-39	0	1	1	2	1	1
40-49	2	1	1	0	1	1
50-59	2	1	1	0	1	1

Sum the chi-squared values for each cell: $\chi^2 = 1+1+1+1+1+1+1+1 = 8$

The Chi Square Distribution

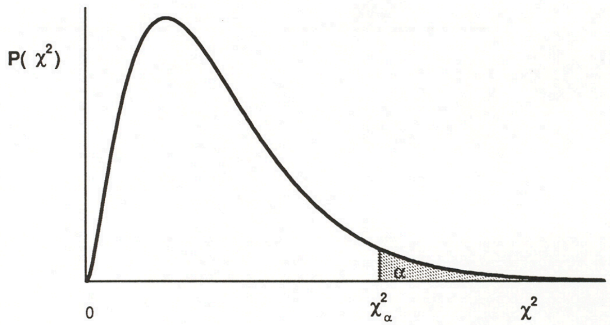


Figure J.1: The χ^2 distribution

	Level of Significance α								
df	0.200	0.100	0.075	0.050	0.025	0.010	0.005	0.001	0.0005
1	1.642	2.706	3.170	3.841	5.024	6.635	7.879	10.828	12.116
2	3.219	4.605	5.181	5.991	7.378	9.210	10.597	13.816	15.202
3	4.642	6.251	6.905	7.815	9.348	11.345	12.838	16.266	17.731
4	5.989	7.779	8.496	9.488	11.143	13.277	14.860	18.467	19.998
5	7.289	9.236	10.008	11.070	12.833	15.086	16.750	20.516	22.106
6	8.558	10.645	11.466	12.592	14.449	16.812	18.548	22.458	24.104
7	9.803	12.017	12.883	14.067	16.013	18.475	20.278	24.322	26.019
8	11.030	13.362	14.270	15.507	17.535	20.090	21.955	26.125	27.869
9	12.242	14.684	15.631	16.919	19.023	21.666	23.589	27.878	29.667
10	13.442	15.987	16.971	18.307	20.483	23.209	25.188	29.589	31.421
11	14.631	17.275	18.294	19.675	21.920	24.725	26.757	31.265	33.138
12	15.812	18.549	19.602	21.026	23.337	26.217	28.300	32.910	34.822
13	16.985	19.812	20.897	22.362	24.736	27.688	29.820	34.529	36.479
14	18.151	21.064	22.180	23.685	26.119	29.141	31.319	36.124	38.111
15	19.311	22.307	23.452	24.996	27.488	30.578	32.801	37.698	39.720
16	20.465	23.542	24.716	26.296	28.845	32.000	34.267	39.253	41.309
17	21.615	24.769	25.970	27.587	30.191	33.409	35.719	40.791	42.881
18	22.760	25.989	27.218	28.869	31.526	34.805	37.157	42.314	44.435
19	23.900	27.204	28.458	30.144	32.852	36.191	38.582	43.821	45.974
20	25.038	28.412	29.692	31.410	34.170	37.566	39.997	45.315	47.501
21	26.171	29.615	30.920	32.671	35.479	38.932	41.401	46.798	49.013
22	27.301	30.813	32.142	33.924	36.781	40.289	42.796	48.269	50.512
23	28.429	32.007	33.360	35.172	38.076	41.639	44.182	49.729	52.002
24	29.553	33.196	34.572	36.415	39.364	42.980	45.559	51.180	53.480
25	30.675	34.382	35.780	37.653	40.646	44.314	46.928	52.620	54.950
26	31.795	35.563	36.984	38.885	41.923	45.642	48.290	54.053	56.409
27	32.912	36.741	38.184	40.113	43.195	46.963	49.645	55.477	57.860
28	34.027	37.916	39.380	41.337	44.461	48.278	50.994	56.894	59.302
29	35.139	39.087	40.573	42.557	45.722	49.588	52.336	58.302	60.738
30	36.250	40.256	41.762	43.773	46.979	50.892	53.672	59.704	62.164
40	47.269	51.805	53.501	55.759	59.342	63.691	66.766	73.403	76.097
50	58.164	63.167	65.030	67.505	71.420	76.154	79.490	86.662	89.564
60	68.972	74.397	76.411	79.082	83.298	88.380	91.952	99.609	102.698
70	79.715	85.527	87.680	90.531	95.023	100.425	104.215	112.319	115.582
80	90.405	96.578	98.861	101.880	106.629	112.329	116.321	124.842	128.267
90	101.054	107.565	109.969	113.145	118.136	124.117	128.300	137.211	140.789
100	111.667	118.498	121.017	124.342	129.561	135.807	140.170	149.452	153.174

Correlation (e.g., χ^2) analysis

Step 4: Calculate Degrees of Freedom

Degrees of freedom are used to determine the critical value from the chi-squared distribution.

Formula for degrees of freedom: Degrees of freedom=(No of rows−1)×(No of columns−1)

Calculate: Degrees of freedom=(4−1)×(2−1)=3

Step 5: Perform Chi-Squared Test

Compare the calculated chi-squared statistic to the critical value from the chi-squared distribution table at the desired significance level (e.g., 0.05) with the calculated degrees of freedom.

Find the critical value: For 3 degrees of freedom at the 0.05 significance level, the critical value for $\chi 2$ is approximately 7.815.

Compare the values: Since the calculated $\chi 2=8$ is greater than 7.815, reject the null hypothesis that the variables are independent.

Analyse the correlation (e.g., χ^2) for the given data

Outcome	Treatment A	Treatment B	Control
Recovered	30	50	20
Not Recovered	20	40	40

Summary

- Data quality: accuracy, completeness, consistency, timeliness, believability, interpretability
- Data cleaning: e.g. missing/noisy values, outliers
- Data integration from multiple sources:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Data transformation and data discretization
  - Normalization