



Data Preprocessing: Exploratory Data and Visualization

Lecture 1

Course Outline

- **Data Preprocessing: Data Exploratory and Visualization**
 - Data Exploration
 - Data Visualization

Outline

- What / Why is data preprocessing?
- What are data exploration and techniques involved?
- What are data analytic and visualization?

Data Preprocessing

- It is an arrangement of data content and format for facilitating the data analytic.
- This is an pivot process prior-recognized by the statistic, business intelligent, data mining and machine learning.

Why Data Preprocessing?

- Data collection always includes:
 1. Incomplete data / Missing value
 2. Inconsistent data or Duplicate data
 3. Noisy data / Outlier

Exploring Processes

1. Integration:

- From various data formats and sources will be collected into a single format and source.

2. Cleansing data from noise, incomplete and inconsistency

- Errata occurred while transmitting, entering or collecting should be carefully monitored before analyzing.

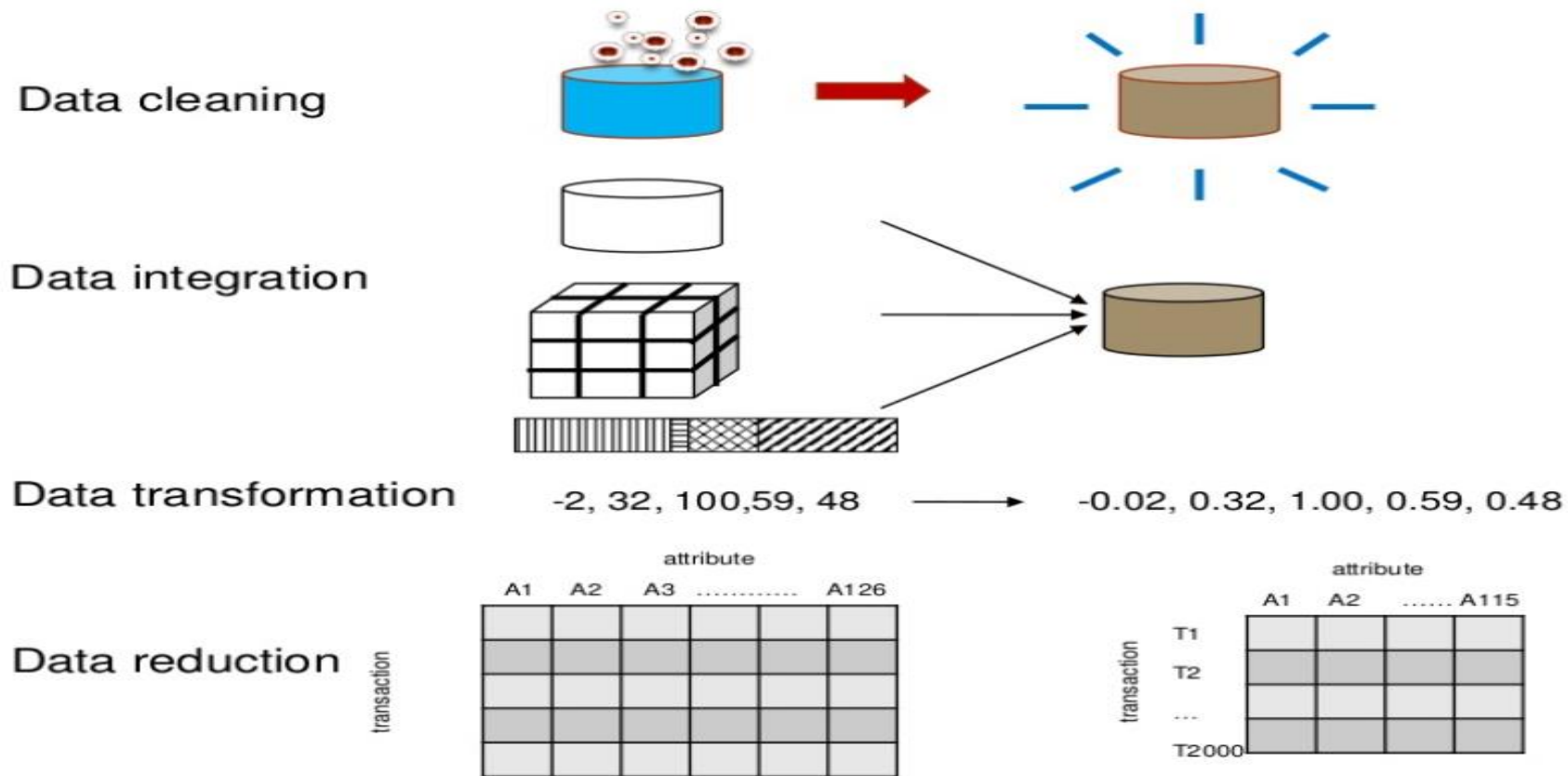
3. Transformation

- It is normalization process which can simplify the data into a reliable shape.

4. Reduction

- It is initiated in order to save the space to collect data and time to compute with meaningful reservation.

Exploring Processes



1. Integration



-  **From SQL Server**
Create a connection to a SQL Server table. Import data into Excel as a Table or PivotTable report.
-  **From Analysis Services**
Create a connection to a SQL Server Analysis Services cube. Import data into Excel as a Table or PivotTable report.
-  **From Windows Azure Marketplace**
Create a connection to a Microsoft Windows Azure DataMarket Feed. Import data into Excel as a Table or PivotTable report.
-  **From OData Data Feed**
Create a connection to an OData Data Feed. Import data into Excel as a Table or PivotTable report.
-  **From XML Data Import**
Open or map a XML file into Excel.
-  **From Data Connection Wizard**
Import data for an unlisted format by using the Data Connection Wizard and OLEDB.
-  **From Microsoft Query**
Import data for an unlisted format by using the Microsoft Query Wizard and ODBC. Functionality is limited for compatibility in previous versions.

2. Data Cleansing Techniques

1. Incomplete data / Missing value Handling

1.1 Ignore the tuple

1.2 Fill in the missing value with:

1.2.1 Default value or Constant value

1.2.2 Attribute mean

2. Inconsistent data or Duplicate data

3. Noisy data / Outlier

2. Data Cleansing Techniques

1. Incomplete data / Missing value Handling

1.1 Ignore the tuple

1.2 Fill in the missing value with:

1.2.1 Default value or Constant value

1.2.2 Attribute mean

Student	Math	Physics	Chemistry
Jane	69	71	68
Joe	50		82
Martha		46	56
Phil	69	88	96
Bruce	67	37	47
Mike	100		91
Tom	89	76	
Bill	78	53	86
Arjun		64	60
Amit	69	69	

Student	Math	Physics	Chemistry
Jane	69	71	68
Phil	69	88	96
Bruce	67	37	47
Bill	78	53	86

1.1

Student	Math	Physics	Chemistry
Jane	69	71	68
Joe	50	None	82
Martha	None		46
Phil	69	88	96
Bruce	67	37	47
Mike	100	None	91
Tom	89	76	None
Bill	78	53	86
Arjun	None	64	60
Amit	69	69	None

1.2.1

2. Data Cleansing Techniques

1. Incomplete data / Missing value Handling

1.1 Ignore the tuple

1.2 Fill in the missing value with:

1.2.1 Default value or Constant value

1.2.2 Attribute mean

Company Name	Revenue			Net Income			Net Income Margin			2018_1
	2016	2017	2018	2016	2017	2018	2016	2017	2018	
A	371	504	592	48	24	40	12.9%	4.8%	6.8%	6.8%
B	554	104		43	20	19	7.8%	19.2%	#DIV/0!	11.0%
C	388	823	582	12	21	14	3.1%	2.6%	2.4%	2.4%
D	805	510		46	26	12	5.7%	5.1%	#DIV/0!	11.0%
E	724	153	184	18	11	44	2.5%	7.2%	23.9%	23.9%
							Mean		11.0%	11.0%

1.2.2

2. Data Cleansing Techniques

1. Incomplete data / Missing value Handling

1.1 Ignore the tuple

1.2 Fill in the missing value with:

1.2.1 Default value or Constant value

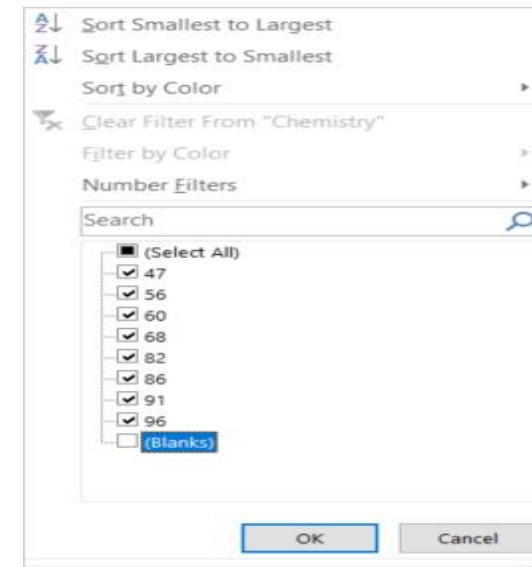
1.2.2 Attribute mean

2. Inconsistent data or Duplicate data

3. Noisy data / Outlier

TODO Task#1

1. Select the attribute(s) concerned
2. Click menu item as follows:
 - Sort & Filter
 - Filter
3. Untick "Blanks"



2. Data Cleansing Techniques

1. Incomplete data / Missing value Handling

1.1 Ignore the tuple

1.2 Fill in the missing value with:

1.2.1 Default value or Constant value

1.2.2 Attribute mean

2. Inconsistent data or Duplicate data

3. Noisy data / Outlier

TODO Task#2

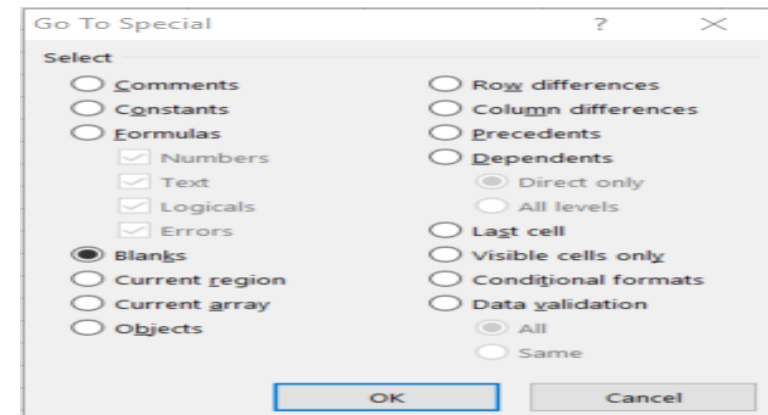
1. Select entire data set

2. Click menu item as follows:

📁 Find & Select

📁 Go to Special

3. Choose "Blanks"



4. Assign a default value (E.g. None)

5. Press buttons: Ctrl + Enter

2. Data Cleansing Techniques

1. Incomplete data / Missing value Handling

1.1 Ignore the tuple

1.2 Fill in the incomplete data with:

1.2.1 Default value or Constant value

1.2.2 **Attribute mean**

2. Inconsistent data or Duplicate data

3. Noisy data / Outlier

TODO Task#3

1. Compute the attribute mean with the formula:

Syntax

AVERAGEIF(range, criteria, [average_range])

The AVERAGEIF function syntax has the following arguments:

- **Range** Required. One or more cells to average, including numbers or names, arrays, or references that contain numbers.
- **Criteria** Required. The criteria in the form of a number, expression, cell reference, or text that defines which cells are averaged. For example, criteria can be expressed as 32, "32", ">32", "apples", or B4.
- **Average_range** Optional. The actual set of cells to average. If omitted, range is used.

2. Create a new column

3. Detect error and Replace mean with the formula:

Syntax

IFERROR(value, value_if_error)

The IFERROR function syntax has the following arguments:

- **Value** Required. The argument that is checked for an error.
- **Value_if_error** Required. The value to return if the formula evaluates to an error. The following error types are evaluated: #N/A, #VALUE!, #REF!, #DIV/0!, #NUM!, #NAME?, or #NULL!.

2. Data Cleansing Techniques

1. Incomplete data / Missing value Handling

1.1 Ignore the tuple

1.2 Fill in the missing value with:

1.2.1 Default value

1.2.2 Attribute mean

2. Inconsistent data or Duplicate data

3. Noisy data / Outlier:

Student	Math	Physics	Chemistry
Jane	69	71	68
Joe	50	35	82
Maritha	91	46	56
Phil	69	88	96
Bruce	67	37	47
Mike	100	87	91
Tom	89	76	76
Bill	78	53	86
Arjun	65	64	60
Amit	69	69	54
Bill	78	53	86
Phil	68	88	96

Student	Math	Physics	Chemistry
Jane	69	71	68
Joe	50	35	82
Maritha	91	46	56
Phil	69	88	96
Bruce	67	37	47
Mike	100	87	91
Tom	89	76	76
Bill	78	53	86
Arjun	65	64	60
Amit	69	69	54
Phil	68	88	96

2. Data Cleansing Techniques

1. Incomplete data / Missing value Handling

1.1 Ignore the tuple

1.2 Fill in the missing value with:

1.2.1 Default value

1.2.2 Attribute mean

2. Inconsistent data or Duplicate data

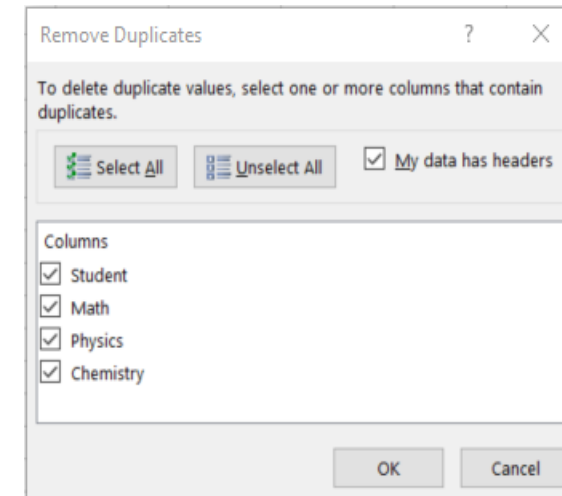
3. Noisy data / Outlier

TODO Task#4

1. Select entire data set

2. Click tab: DATA

3. Click icon menu: Remove Duplicates



4. Choose entire column(s) concerned duplicated.

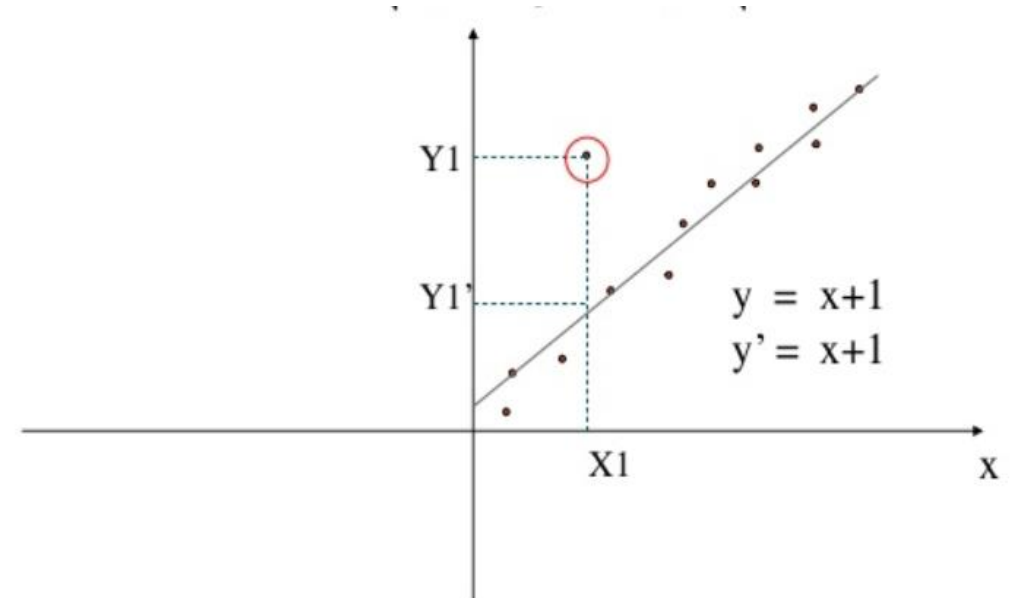
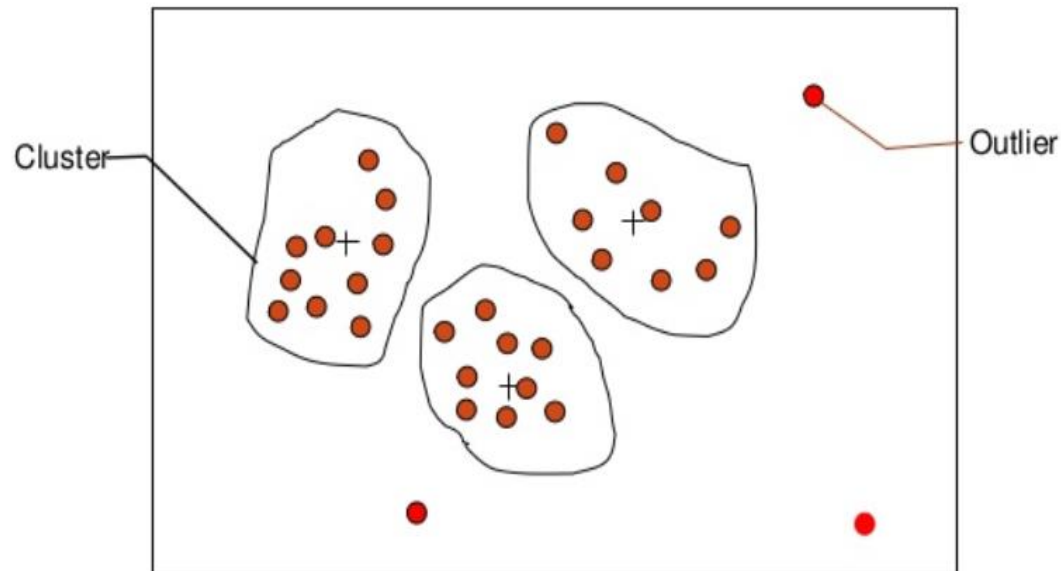
2. Data Cleansing Techniques

3. Noisy data / Outlier

3.1 Clustering

3.2 Regression

3.3 Box Plot (Consider Inter Quartile Range Upper / Lower Limit)



2. Data Cleansing Techniques

3. Noisy data / Outlier

3.1 Clustering

3.2 Regression

3.3 **Box Plot (Consider**

Inter Quartile Range Upper / Lower Limit)

Outlier			
Minimum	QUARTILE.INC(ARRAY, 0)	0.8333	
Q1	QUARTILE.INC(ARRAY, 1)	22.25	
Median	QUARTILE.INC(ARRAY, 2)	28	
Q3	QUARTILE.INC(ARRAY, 3)	36	
Maximum	QUARTILE.INC(ARRAY, 4)	62	
Mean	AVERAGE(ARRAY)	29.23413	
range	Maximum - Minimum	61.1667	
IQR	Q3 - Q1	13.75	
IQR x 1.5	IQR x 1.5	20.625	
Lower Limit	Q1 - (IQR x 1.5)	1.625	Outlier < Lower Limit
Upper Limit	Q3 + (IQR x 1.5)	56.625	Outlier > Upper Limit

TODO Task#5

1. Select entire data set
2. Click menu item as follows:
 - Conditional Formatting
 - New Rule

New Formatting Rule

Select a Rule Type:

- Format all cells based on their values
- Format only cells that contain
- Format only top or bottom ranked values
- Format only values that are above or below average
- Format only unique or duplicate values
- Use a formula to determine which cells to format

Edit the Rule Description:

Format only cells with:

Cell Value not between =\$F\$21 and =\$F\$22

Preview: AaBbCcYyZz

3. Transformation

1. Min-Max Normalization

Age in Min-Max Norm =	Age - Minimum	$\times ((\text{new_max} - \text{new_min}) + \text{new_min})$
	Maximum - Minimum	
new_max		1
new_min		0
Minimum	MIN (ARRAY)	
Maximum	MAX (ARRAY)	

2. Z-Score Normalization

Age in Z Score (0, 1) Norm =	Age - Mean
	Standar Deviation
Mean	AVERAGE (ARRAY)
Standar Deviation	STDEV (ARRAY)

3.

Age in Decimal Scaling =	Age
	10^j
j	2

TODO Task#6

1. Select an attribute to be normalized
2. Compute three transformations according to the given formulas

4. Reduction

1. Dimensionality reduction

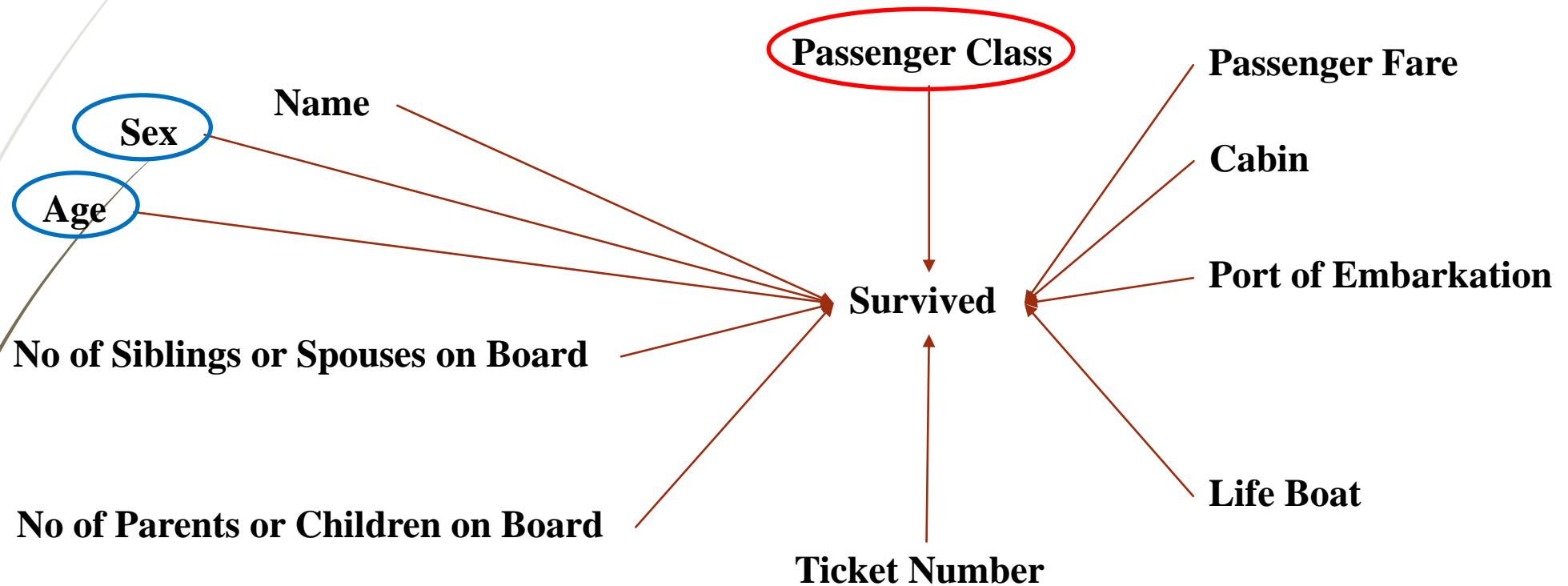
- The data attributes can be selected, reduced maintained based on at least one rational criteria in order to optimized cost.

2. Data aggregation

- The data attribute can be aggregated and lead to the data summarize
- It can be analysis if data at multiple level of abstraction.

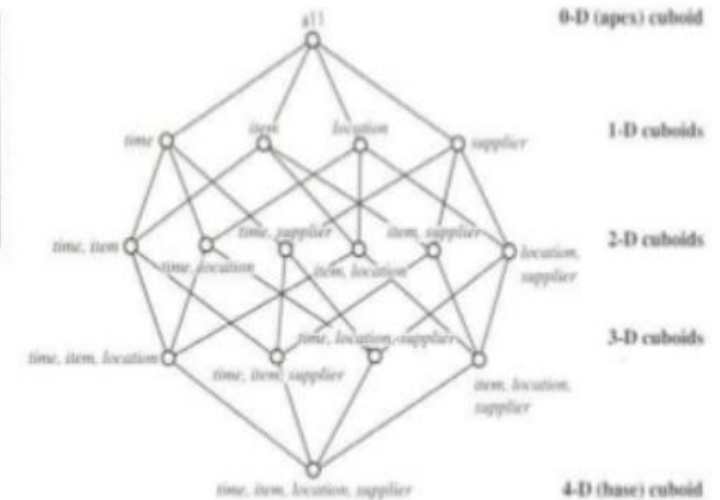
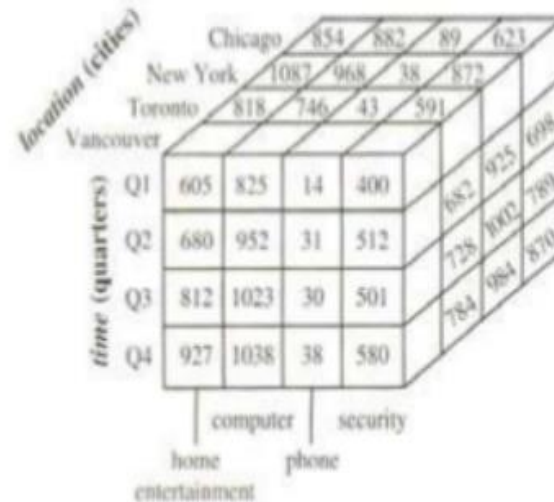
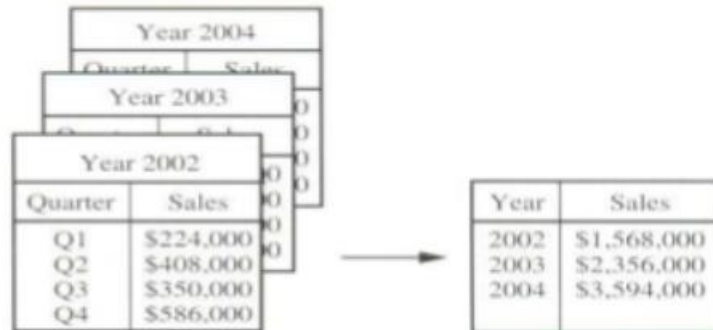
4. Reduction

1. Dimensionality reduction



4. Reduction

2. Data aggregation




4. Reduction

2. Data aggregation

Count of Survived		
Survived	Sex	Total
No	Female	127
	Male	682
No Total		809
Yes	Female	339
	Male	161
Yes Total		500
Grand Total		1309

TODO Task#7

1. Select all concerned attributes
2. At tab "Insert" , click menu icon:
 Pivot Table

?

×

Create PivotTable

Choose the data that you want to analyze

☒ Select a table or range

Table/Range: 'titanic adj'!\$A\$1:\$G\$1310

☐ Use an external data source

Choose Connection...

Connection name:

Choose where you want the PivotTable report to be placed

☐ New Worksheet

☒ Existing Worksheet

Location: 'titanic adj'!\$U\$2

Choose whether you want to analyze multiple tables

☐ Add this data to the Data Model

OK

Cancel

References

- <https://medium.freecodecamp.org/i-ranked-all-the-best-data-science-intro-courses-based-on-thousands-of-data-points-db5dc7e3eb8e>
- <https://www.slideshare.net/NontawatB/03-data-preprocessing-34542881>
- http://rstudio-pubs-static.s3.amazonaws.com/339331_347a7471810d4f4c80b1ead13db9a2ad.html