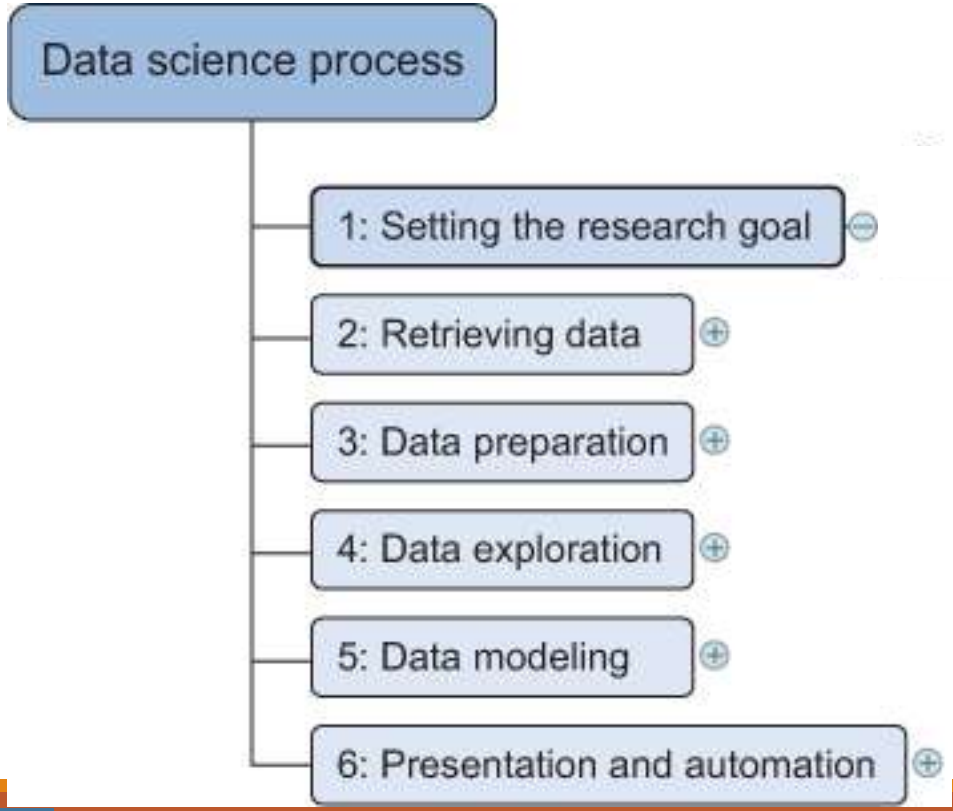
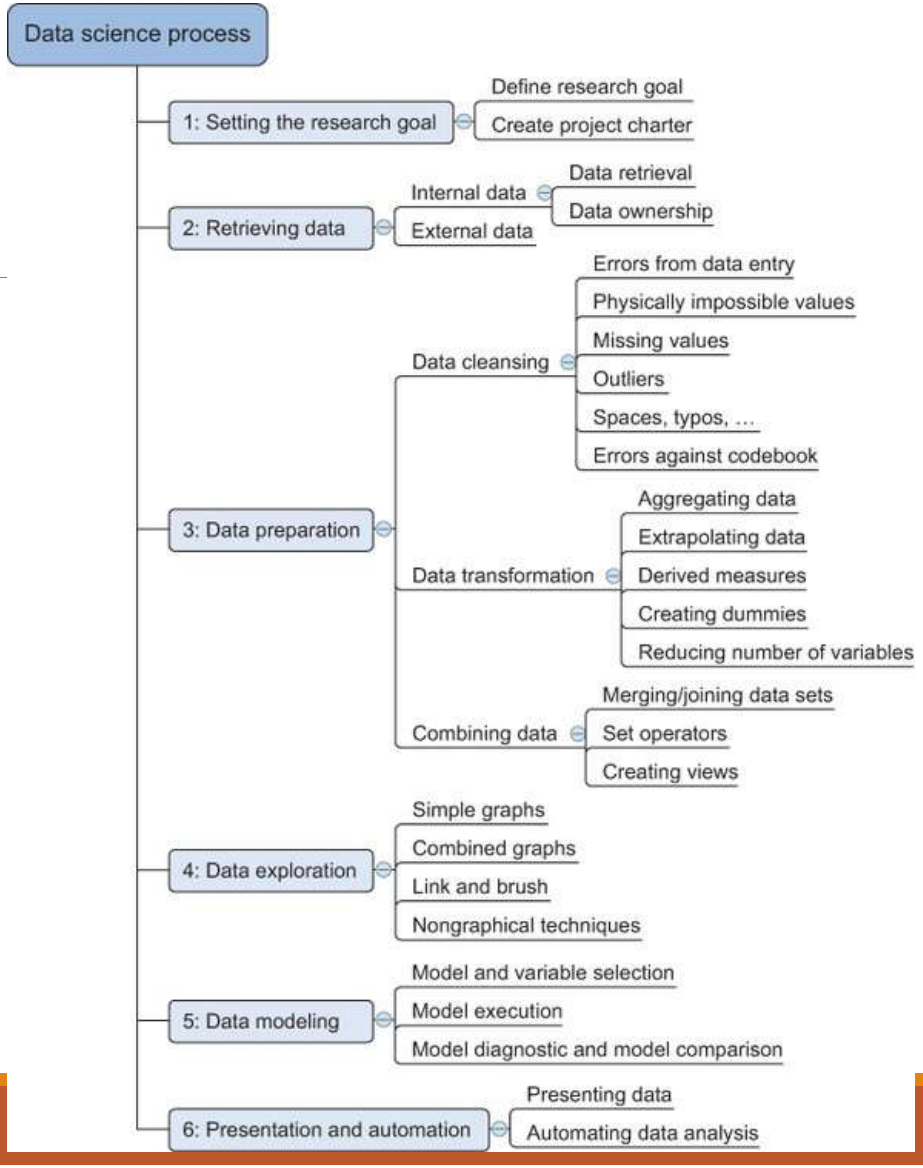
Data Science Workflow

Book: https://livebook.manning.com/book/introducing-data-science/chapter-2/1

Data Science Process



Data Science Process A Big Picture



Step 1:
Defining research goals and creating a project charter

A project starts by understanding the *what*, the *why*, and the *how* of your project

Answering these three questions (what, why, how) is the goal of the first phase, so that everybody knows what to do and can agree on the best course of action.

Spend time understanding the goals and context of your research

An essential outcome is the research goal that states the purpose of your assignment in a clear and focused manner.

Step 1:
Defining research goals and creating a project charter

A project charter requires teamwork, and your input covers at least the following:

A clear research goal

The project mission and context

How you're going to perform your analysis

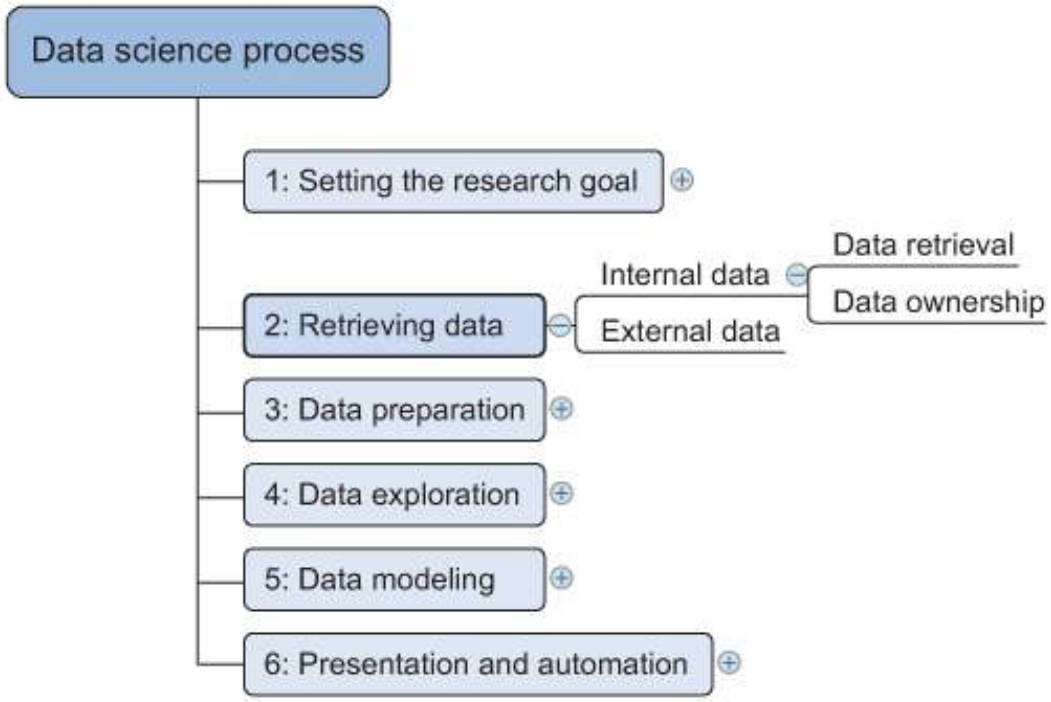
What resources you expect to use

Proof that it's an achievable project, or proof of concepts

Deliverables and a measure of success A timeline

Step 2:
Retrieving data

Figure 2.3. Step 2: Retrieving data

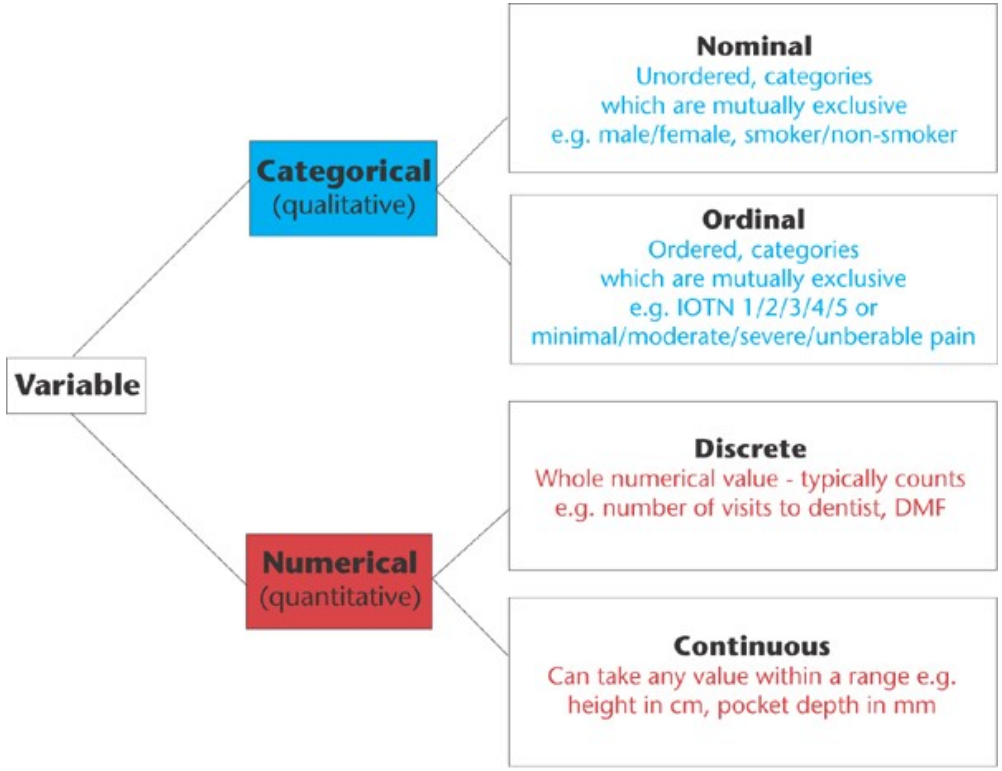


Data can be stored in many forms, ranging from simple text files to tables in a database.

The objective now is acquiring all the data you need.

This may be difficult, and even if you succeed, data is often like a diamond in the rough: it needs polishing to be of any use to you.

Types of Data (Arial View)



Acquiring Data

Open data site	Description
Data.gov	The home of the US Government's open data
https://open-data.europa.eu/	The home of the European Commission's open data
Freebase.org	An open database that retrieves its information from sites like Wikipedia, MusicBrains, and the SEC archive
Data.worldbank.org	Open data initiative from the World Bank
Aiddata.org	Open data for international development
Open.fda.gov	Open data from the US Food and Drug Administration

Table 2.1. A list of open-data providers that should get you started

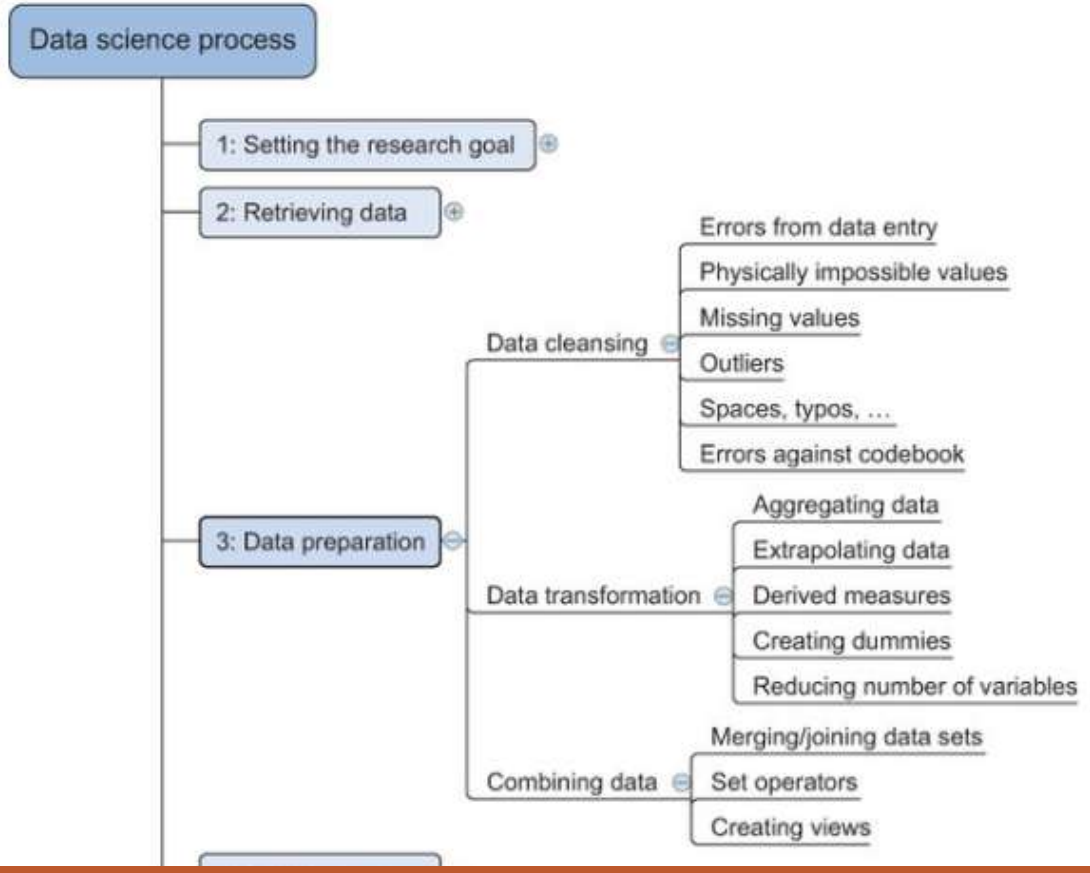
Acquiring Data

Open data site	Description
Kaggle	The platform supports open and accessible data formats.
UCI Machine Learning Repository	University of California Irvine hosts 440 data set as a service to the machine learning community.
Academic Torrents	Academic Torrents is a site that is geared around sharing the data sets from scientific papers.
Quandl	Quandl is a repository of economic and financial data. Some of the datasets are free, while others are up for purchase

Table 2.1. A list of open-data providers that should get you started

Step 3:
Cleansing, integrating, and transforming data

Figure 2.4. Step 3: Data preparation



Data Cleansing

General solution: Try to fix the problem early in the data acquisition chain or else fix it in the program	
Error description	**Possible solution**
Errors pointing to false values within one data set	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
Errors pointing to inconsistencies between data sets	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation

Table 2.2. An overview of common errors

Outliers

Value	Count
Good	1598647
Bad	1354468
Godo	15
Bade	1

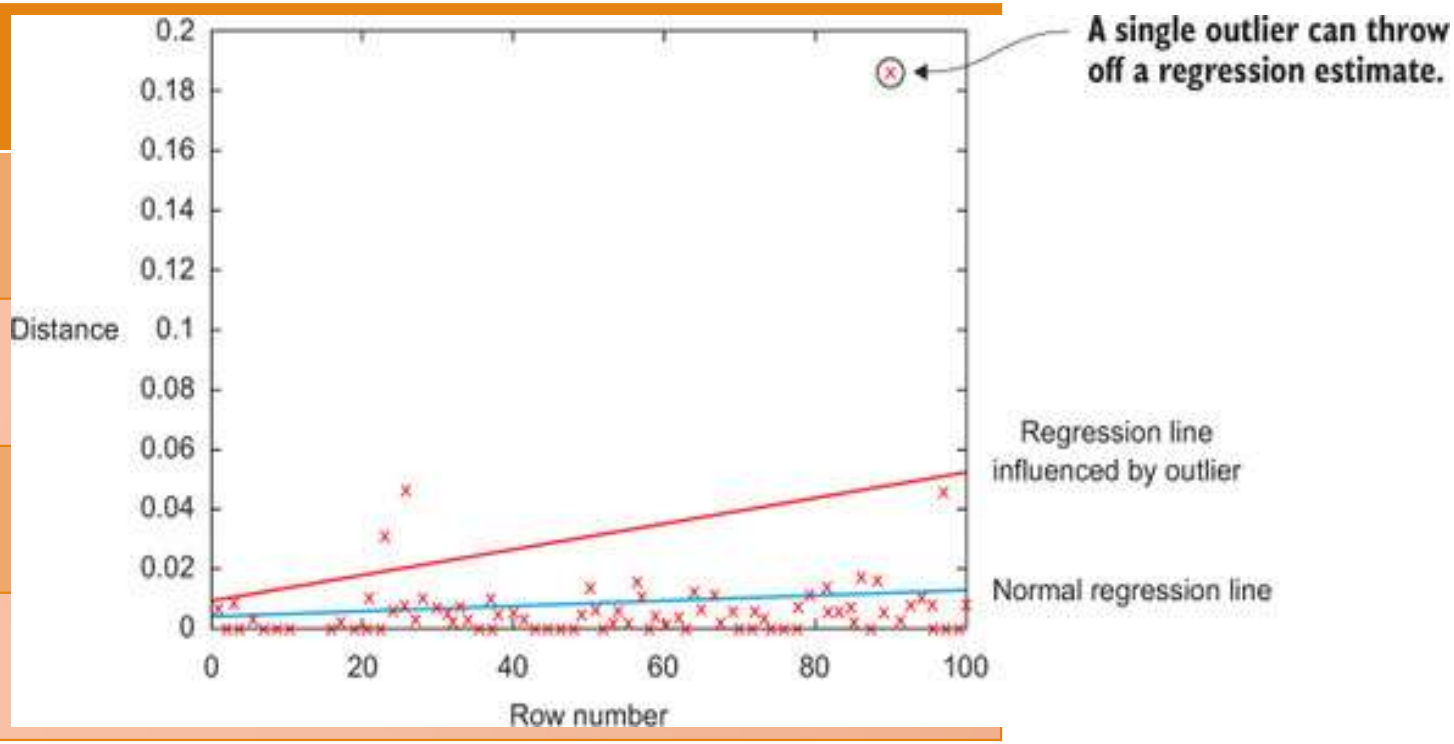


Table 2.3. Detecting outliers on simple variables with a frequency table

Handling Missing Values

Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to null	Easy to perform	Not every modeling technique and/or implementation can handle null values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions

Table 2.4. An overview of techniques to handle missing data

Step 3:
Cleansing, integrating, and transforming data

Data should be cleansed when acquired for many reasons:

Not everyone spots the data anomalies. Decision-makers may make costly mistakes on information based on incorrect data from applications that fail to correct for the faulty data.

If errors are not corrected early on in the process, the cleansing will have to be done for every project that uses that data.

Data errors may point to a business process that isn't working as designed.

Data errors may point to defective equipment, such as broken transmission lines and defective sensors.

Data errors can point to bugs in software or in the integration of software that may be critical to the company.

Step 3:
Cleansing, integrating, and transforming data

Combining data from different data sources

You can perform two operations to combine information from different data sets.

The first operation is ***joining***: enriching an observation from one table with information from another table.

The second operation is ***appending*** or ***stacking***: adding the observations of one table to those of another table.

Step 3:
Cleansing, integrating, and transforming data

Combining data from different data sources

Figure 2.7. Joining two tables on the Item and Region keys

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

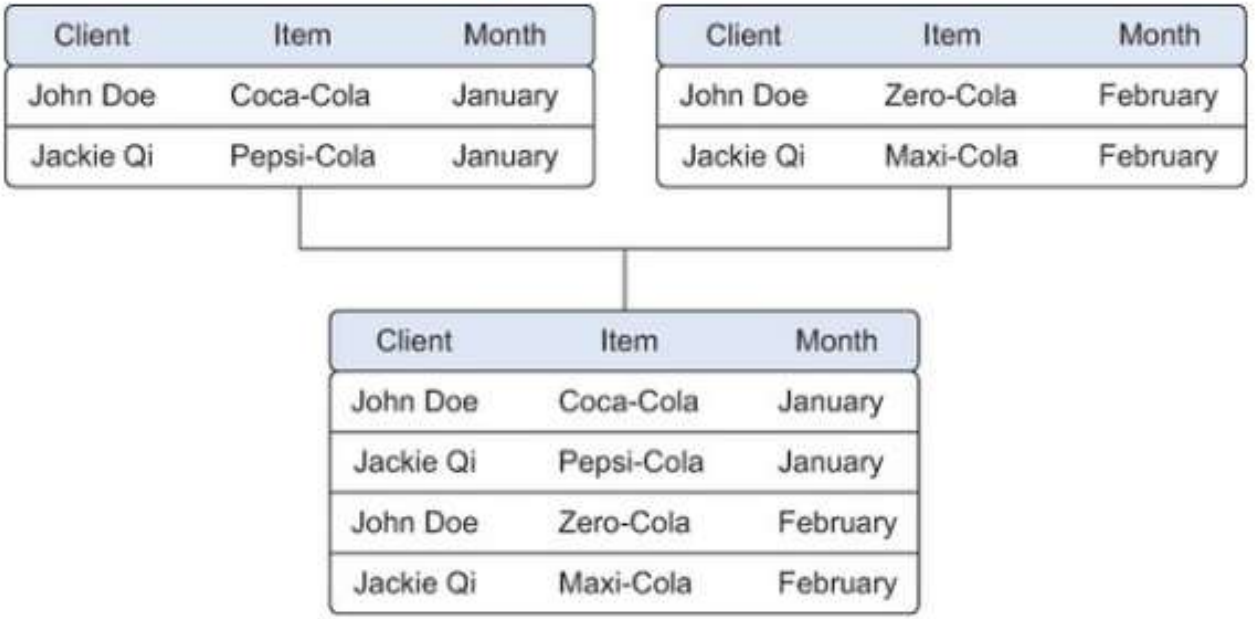
Client	Region
John Doe	NY
Jackie Qi	NC

Client	Item	Month	Region
John Doe	Coca-Cola	January	NY
Jackie Qi	Pepsi-Cola	January	NC

Step 3:
Cleansing, integrating, and transforming data

Combining data from different data sources

Figure 2.8. Appending data from tables is a common operation but requires an equal structure in the tables being appended.



Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

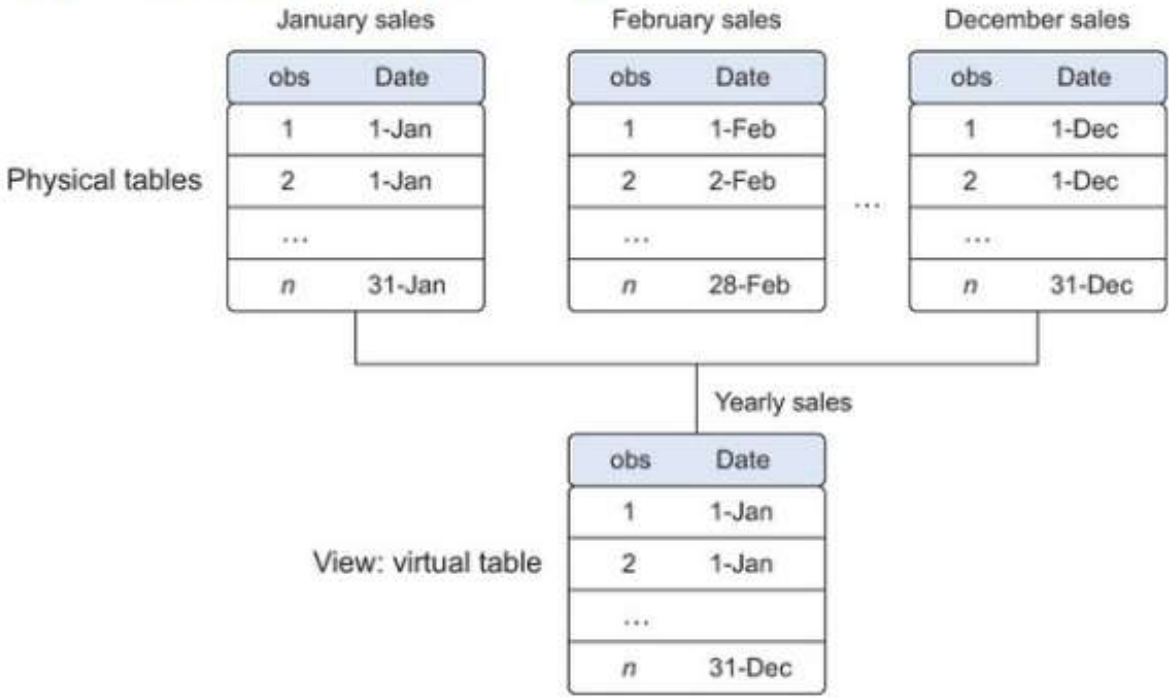
Client	Item	Month
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

Step 3:
Cleansing, integrating, and transforming data

Combining data from different data sources

Figure 2.9. A view helps you combine data without replication.

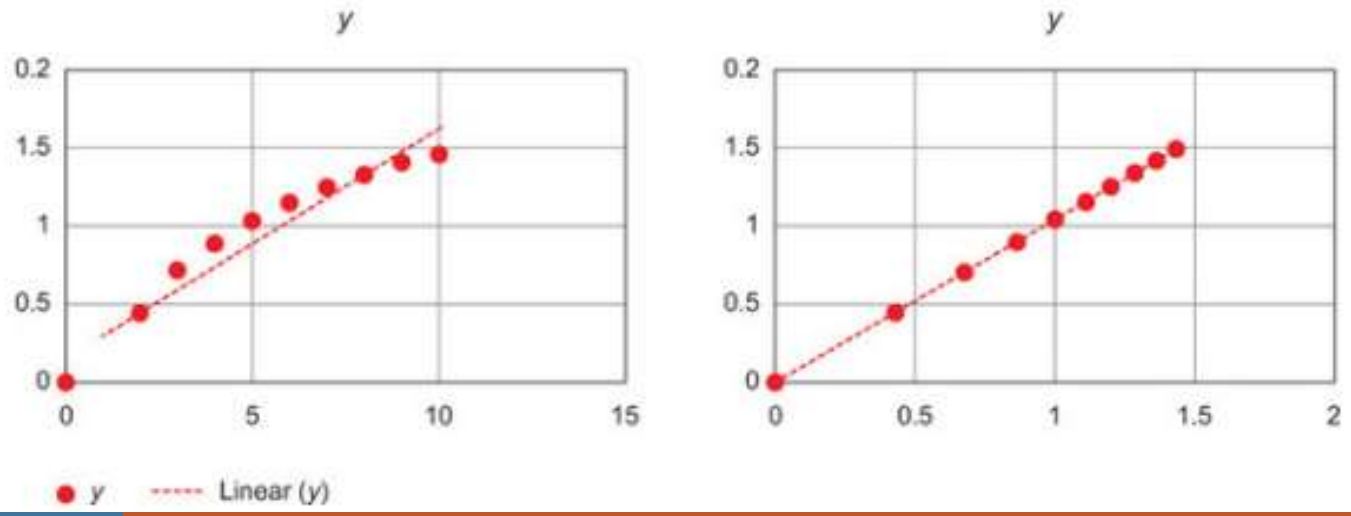


Step 3:
Cleansing, integrating, and transforming data

Transforming the Data

Certain models require their data to be in a certain shape.

Transforming your data so it takes a suitable form for data modeling.



Step 3:
Cleansing, integrating, and transforming data

Reducing the number of variables

Sometimes you have too many variables and need to reduce the number because they don't add new information to the model.

Having too many variables in your model makes the model difficult to handle, and certain techniques don't perform well when you overload them with too many input variables.

For instance, all the techniques based on a Euclidean distance perform well only up to 10 variables.

Step 3:
Cleansing, integrating, and transforming data

Turning variables into dummies

Variables can be turned into dummy variables.

Dummy variables can only take two values: true(1) or false(0).

They're used to indicate the absence of a categorical effect that may explain the observation.

In this case you'll make separate columns for the classes stored in one variable and indicate it with 1 if the class is present and 0 otherwise. not exclusive to, economists.

Step 3:
Cleansing, integrating, and transforming data

Turning variables into dummies

Customer	Year	Gender	Sales
1	2015	F	10
2	2015	M	8
1	2016	F	11
3	2016	M	12
4	2017	F	14
3	2017	M	13

Customer	Year	Sales	Male	Female
1	2015	10	0	1
1	2016	11	0	1
2	2015	8	1	0
3	2016	12	1	0
3	2017	13	1	0
4	2017	14	0	1