National University
of computer and emerging sciences

INTRODUCTION TO DATA SCIENCE

COURSE INSTRUCTOR: MUHAMMAD SAIF UL ISLAM

Lecture Outline – Week#1

- Introductory words
- Introduction to the Course
- Discussion on Course outline
- Course plan, Assignments and Project
- Introduction to Data Science
- Applications of AI & Data Science
- Characteristics of Data Scientist
- Installing Python/Anaconda, Agent & Environment

About Myself

Muhammad Saif ul Islam

Education:

PhD Scholar (Computer Science)

- FAST-NUCES, LHR

Masters in Data Science - 2019

- FAST-NUCES, KHI

Bachelors in Computer Science -2017

- Bahria University, KHI

Work Experience:

IT Instructor – 5 Months

- IBA-BBSYDP

Python Developer – 7 Months

- Innovative Solutions

Sr. Operations Engineer – 1 Year

- Gfk Etilize

Lecturer – 2.5 years

- Mohammad Ali Jinnah University

Lecturer – 6 Months

- Beaconhouse National University

Lecturer – Since Spring 2023

- FAST- NUCES

About Myself

Certifications

Python:

•DAT210x: Programming with Python for Data Science
•Introduction to Python for Data Science
•Introduction to Data Science in Python
•Python for Everybody
•Python Data Structures

Database:

•Using Databases with Python
•Querying Data with Transact-SQL

Data Science:

•Data Science Essentials
•Python Project for Data Science
•Applied Plotting, Charting & Data Representation in Python
•Capstone: Retrieving, Processing, and Visualizing Data with Python
•Applied Machine Learning in Python
•Image Processing with Python

Web:

•Using Python to Access Web Data
•HTML5 Introduction

About Myself

Publications

Mustafa Khan, M., **Ul Islam, M. S.,** Siddiqui, A. A., & Qadri, M. T. (2023). Dual deterministic model based on deep neural network for the classification of pneumonia. *Intelligent Decision Technologies*, *17*(3), 641–654. https://doi.org/10.3233/idt-220192

Muhammad Saif ul Islam, Using deep learning based methods to classify salt bodies in seismic images, Journal of Applied Geophysics, Volume 178, 2020, 104054, ISSN 0926-9851, https://doi.org/10.1016/j.jappgeo.2020.104054.

M. Mehboob, M. S. Ali, **S. Ul Islam** and S. Sarmad Ali, "Evaluating Automatic CV Shortlisting Tool For Job Recruitment Based On Machine Learning Techniques," 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 2022, pp. 1-4, doi: 10.1109/MAJICC56935.2022.9994112.

M. S. ul Islam and H. Farooq, "Rating visual contents of website using brain computer interface," 2017 International Conference on Information and Communication Technologies (ICICT), Karachi, Pakistan, 2017, pp. 23-27, doi: 10.1109/ICICT.2017.8320159.

Students’ Introduction

Name?

Expectation:

- What do you expect from this course?
- What do you want to learn in this course?

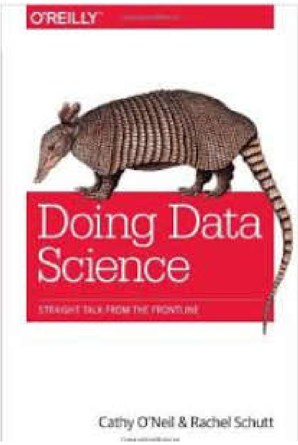
Course plan, Assignments and Quizzes

	Graded Assessment types	Weights (%)
1	Project	10%
2	Quiz	10%
3	Assignments	10%
4	Mid Exam	30%
5	Final Assessment	40%
	Total:	100%

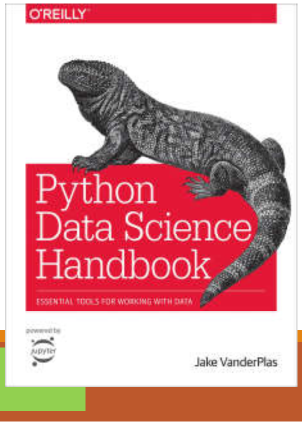
Resources

Books:

Doing Data Science by Oreilly


O'REILLY
Doing Data Science
STRAIGHT TALK FROM THE FRONTLINE
Cathy O'Neil & Rachel Schutt

Python Data Science Handbook: Essential Tools for Working with Data Book by Jake VanderPlas


O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by jupyter
Jake VanderPlas

Consulting Hours

Contact at:

Email: **saif.islam@lhr.nu.edu.pk**

Office Hours:

Room# MLO-007

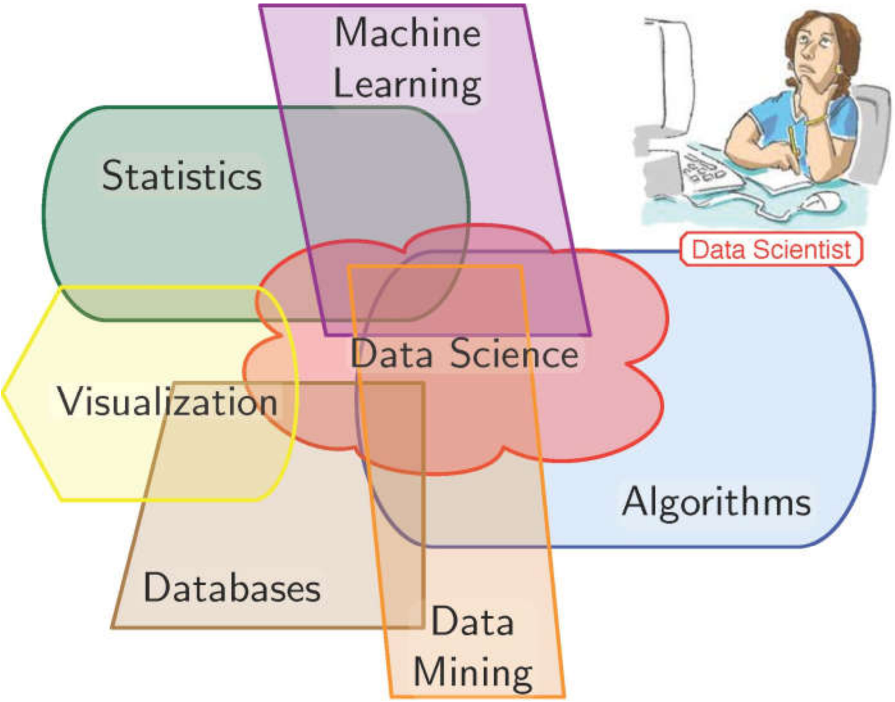
Mon, Wed 3:00PM-4:00PM

OFF Days: Sat, Sun

What is Data Science?

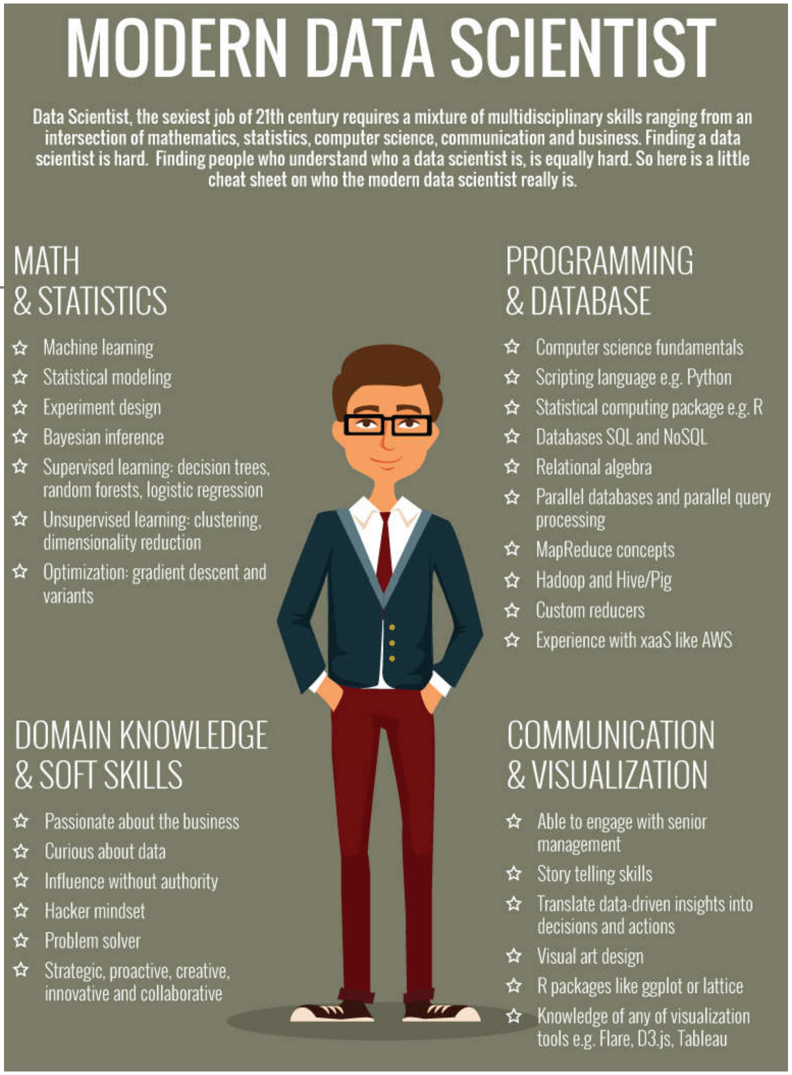
- Data Science is the process of slicing through massive chunks of data, processing and analyzing them for meaningful information that can help businesses get insights on concerns, customer experience, supply-chain and other prime aspects that would complement their business operations.
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

Data Science is Multidisciplinary



Who is Data Scientist?

Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and **decisions**



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Real Life Applications of AI & Data Science

- Marketing
- Finance
- Agriculture
- HealthCare
- Gaming
- Space Exploration
- Autonomous Vehicles
- Artificial Creativity

Real Life Applications of AI & Data Science
Marketing

AI generated content: An AI writing program called ‘WordSmith’ produced 1.5 billion pieces of content in 2016, and is expected to grow further in popularity in the coming years.

Smart Content Curation: Allows you to better engage visitors on your site by showing them content relevant to them. Cross selling, personalized messaging, recommendation etc.

Smart Search: Search engines read our minds and provide all possible results related to the item, Voice-search technology (Google, Amazon, Apple), Interpret consumer’s queries -Chatbots.

Predictive analytics: Predicting the likelihood of a given customer to convert, predicting what price a customer is likely to convert at, or what customers are most likely to make repeat purchases. Propensity modeling.

Dynamic pricing: Dynamic pricing can nudge interested consumers into becoming customers by targeting only special offers only at those likely to need them in order to convert.

Real Life Applications of AI & Data Science
Banking & Finance

Recommendation Engines: In the banking sector, the system learns from the user’s behavior. Based on the previous actions, it can recommend appropriate investment strategies, credit card plans, and make other offers that would save the user a lot of time browsing through the website.

Fraud Detection and Prevention: Based on self-learning artificial technology and real-time behavioral profiling, the system can detect suspicious behavior and prevent frauds.

Trading: Investment companies have been relying on computers and data scientists to determine future patterns in the market. As a domain, trading and investments depend on the ability to predict the future accurately.

Predictive analytics: Uses real-time and historical data to deliver precise information that helps traders to quote a better price when selling and buying bonds for their clients.

Real Life Applications of AI & Data Science
Agriculture

Forecasted Weather data: The forecasted/ predicted data help farmers increase yields and profits without risking the crop. By implementing such practice helps to make a smart decision on time.

Monitoring Crop and Soil Health: Utilizing AI is an efficient way to conduct, or monitor identifies possible defects and nutrient deficiencies in the soil. With the image recognition approach, AI identifies possible defects through images captured by the camera.

Decrease pesticide usage: With the help of the AI, data are gathered to keep a check on the weed which helps the farmers to spray chemicals only where the weeds are. This directly reduced the usage of the chemical spraying an entire field.

AI Agriculture Bots: AI bots in the agriculture field can harvest crops at a higher volume and faster pace than human laborers. By leveraging computer vision helps to monitor the weed and spray them.

Real Life Applications of AI & Data Science
Health Care

Medical Imaging: With AI in medical imaging, treatments can be personalized, and results can be transmitted with ease. Doctors can also efficiently identify cardiovascular disorders along with other fractures and injuries. Cancer cells detection, brain tumor detection, pneumonia detection etc. are few example.

Robot Assisted Surgery: In orthopedic surgery, a form of AI-assisted robotics can analyze data from pre-op medical records to physically guide the surgeon's instrument in real-time during a procedure. It can also use data from actual surgical experiences to inform new surgical techniques.

Automated Diagnosis and Error Reduction: In 2017, a group at Stanford University tested an AI algorithm against 21 dermatologists on its ability to identify skin cancers. The clinical findings, as reported by Nature last year, "artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists."

Virtual Nurses: To interact with patients, ask them questions about their health, assess their symptoms, and direct them to the most effective care setting. Molly, etc.

Real Life Applications of AI & Data Science
Gaming

AlphaGo: DeepMind's AlphaGo is the first computer program to defeat a professional human Go player (GrandMaster)

AlphaZero: AI beats champion chess program 'StockFish' after teaching itself in four hours.

Intelligent behaviors in characters: In video **games**, **artificial intelligence** (**AI**) is used to generate responsive, adaptive or **intelligent** behaviors primarily in non-player characters (NPCs) similar to human-like **intelligence**

Adversarial searches: Examples are Chess, Checkers, Go, etc.

Real Life Applications of AI & Data Science
Space Exploration

Spacecraft Monitoring and Control: Machine learning algorithms have been used in monitoring the spacecraft, autonomous navigation of the spacecraft, controlling systems, and intelligently detecting objects in the route

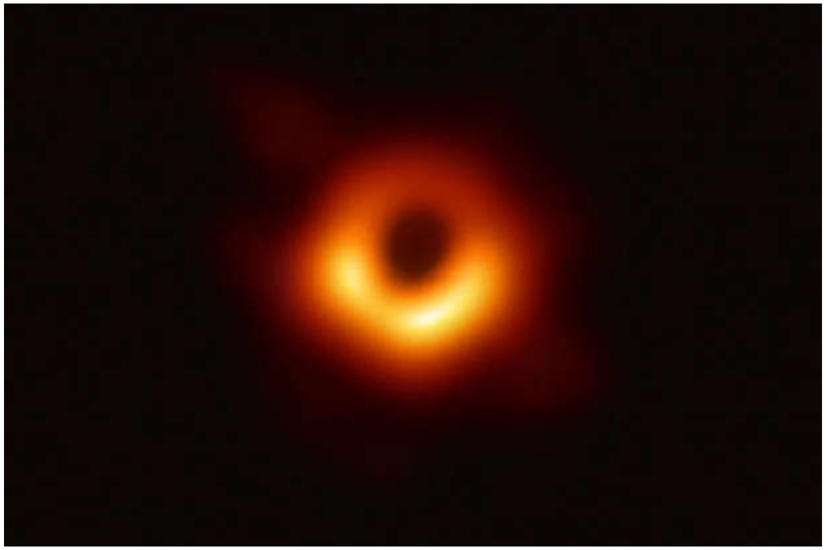
AI Based Assistants: AI-based assistants are being created to aid astronauts in their missions to Mars and beyond. These assistants are designed to understand and predicts the requirements of the crew and comprehend astronauts' emotions and their mental health.

Space Imaging and Exploration: According to the European Space Agency (ESA), satellites can produce over 150 terabytes of data per day. With the use of AI technologies, one can reduce the mission costs, extend battery life, and can analyze a vast amount of imaging data produced by the satellites. Example: Earth Observer 1 (EO-1) satellite, SKICAT, ENVISAT etc.

With the help of Google's trained model, NASA also managed to discover two obscure planets — **Kepler-90i and Kepler-80g**.

The creation of the algorithm that made the **first black hole image** possible was led by MIT grad student **Katie Bouman**

Real Life Applications of AI & Data Science
Space Exploration (Continue..)



Real Life Applications of AI & Data Science
Autonomous Vehicles

Waymo: n April 2017, Waymo started a limited trial of a self-driving taxi service in Phoenix, Arizona. On December 5, 2018, the service launched a commercial self-driving car service called "Waymo One"; users in the Phoenix metropolitan area use an app to request a pick-up

Advanced Driver Assistance Systems (ADAS): Camera-based machine vision systems, radar-based detection units, driver condition evaluation and sensor fusion engine control units (ECUs).

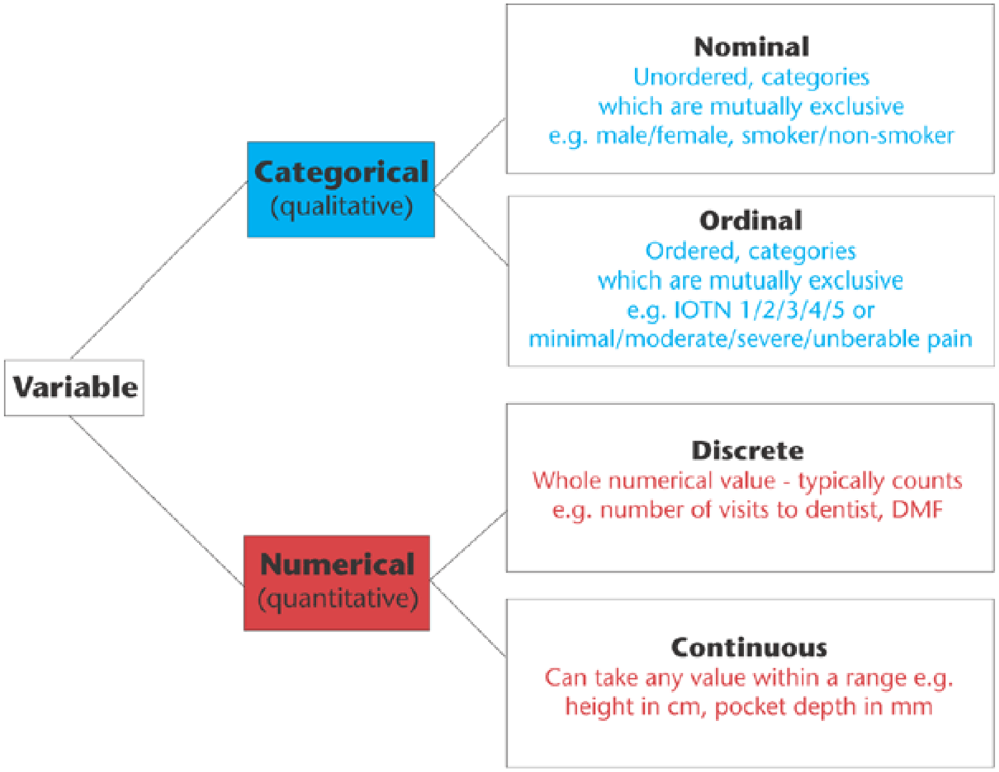
Infotainment human-machine interface: Speech recognition and gesture recognition, eye tracking and driver monitoring, virtual assistance and natural language interfaces.

Real Life Applications of AI & Data Science
Artificial Creativity

ChatGPT: ChatGPT (Chat Generative Pre-trained Transformer) is a chatbot launched by OpenAI in November 2022. It is built on top of OpenAI's GPT-3 family of large language models, and is fine-tuned (an approach to transfer learning) with both supervised and reinforcement learning techniques.

- Question answer
- Solving math equations
- Writing texts (basic academic articles, literary texts, movie script, etc.)
- Interlingual translation
- Summarizing text and detecting keywords in text
- Classification
- Making recommendations
- Explaining what anything does (for example, explaining what a code block does)

Types of Data (Arial View)



Variable

Categorical
(qualitative)

Nominal
Unordered, categories
which are mutually exclusive
e.g. male/female, smoker/non-smoker

Ordinal
Ordered, categories
which are mutually exclusive
e.g. IOTN 1/2/3/4/5 or
minimal/moderate/severe/unberable pain

Numerical
(quantitative)

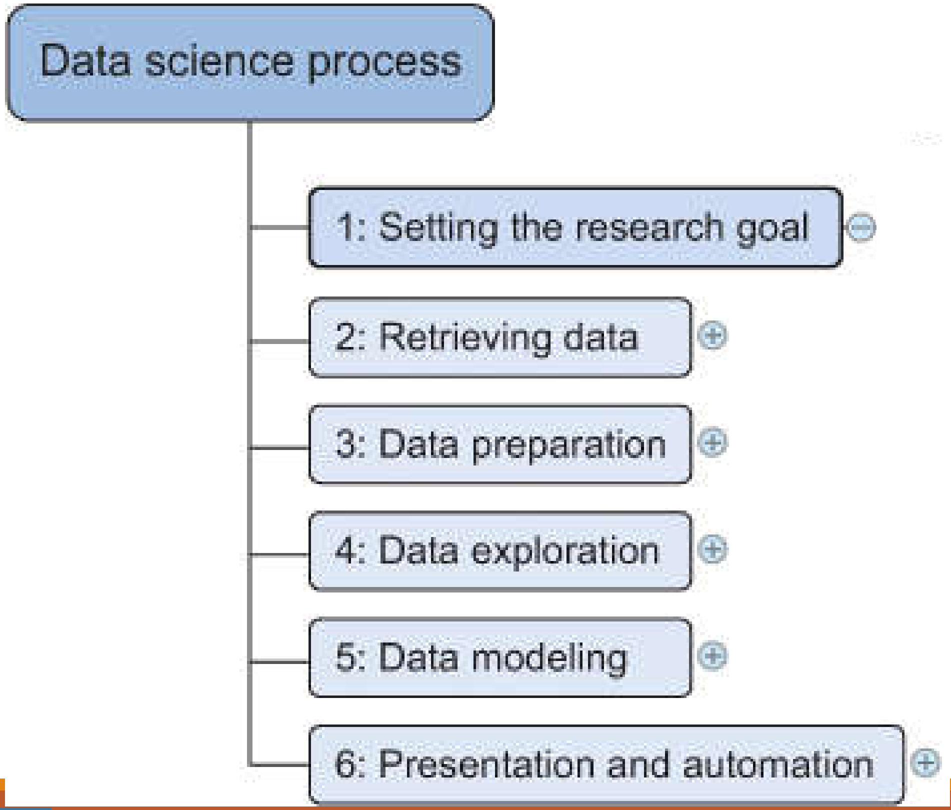
Discrete
Whole numerical value - typically counts
e.g. number of visits to dentist, DMF

Continuous
Can take any value within a range e.g.
height in cm, pocket depth in mm

Data Science Workflow

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

Data Science Process



Data science process

1: Setting the research goal

2: Retrieving data

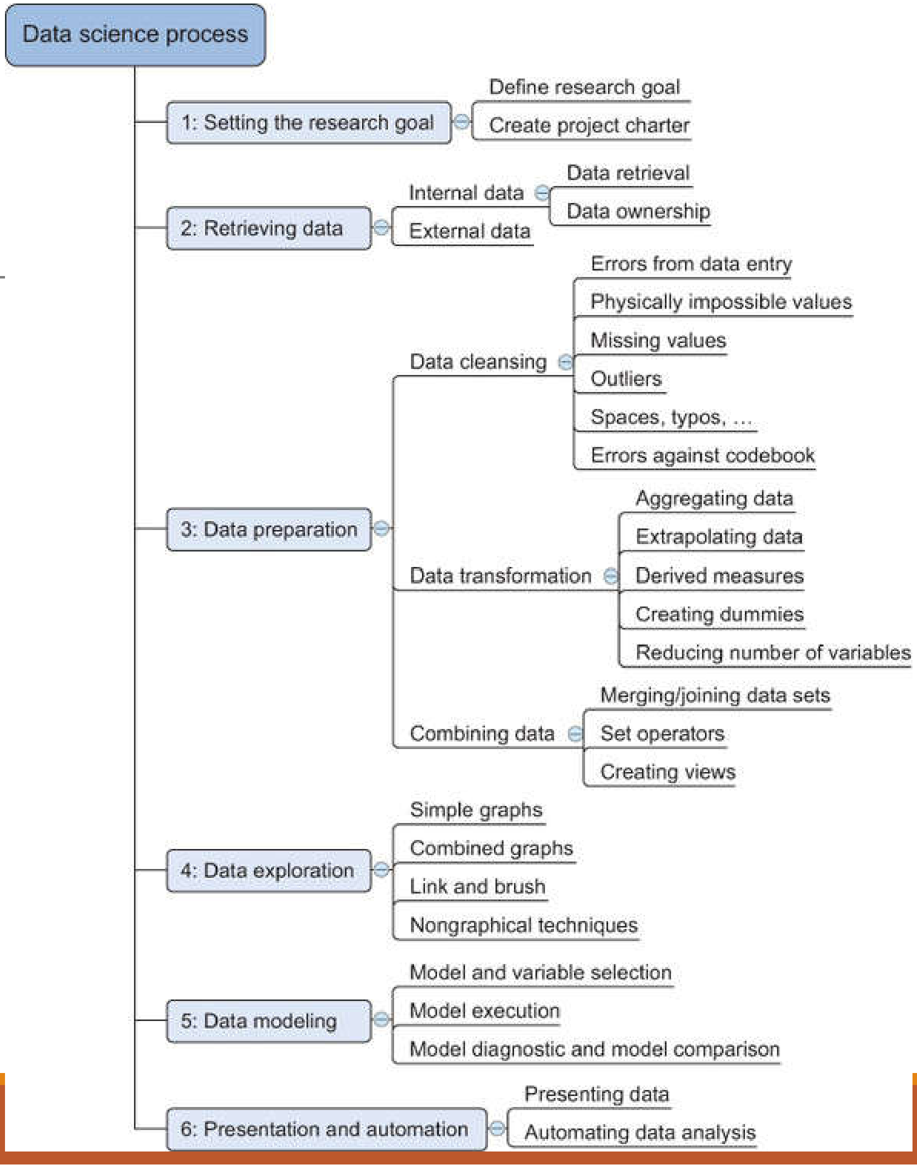
3: Data preparation

4: Data exploration

5: Data modeling

6: Presentation and automation

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

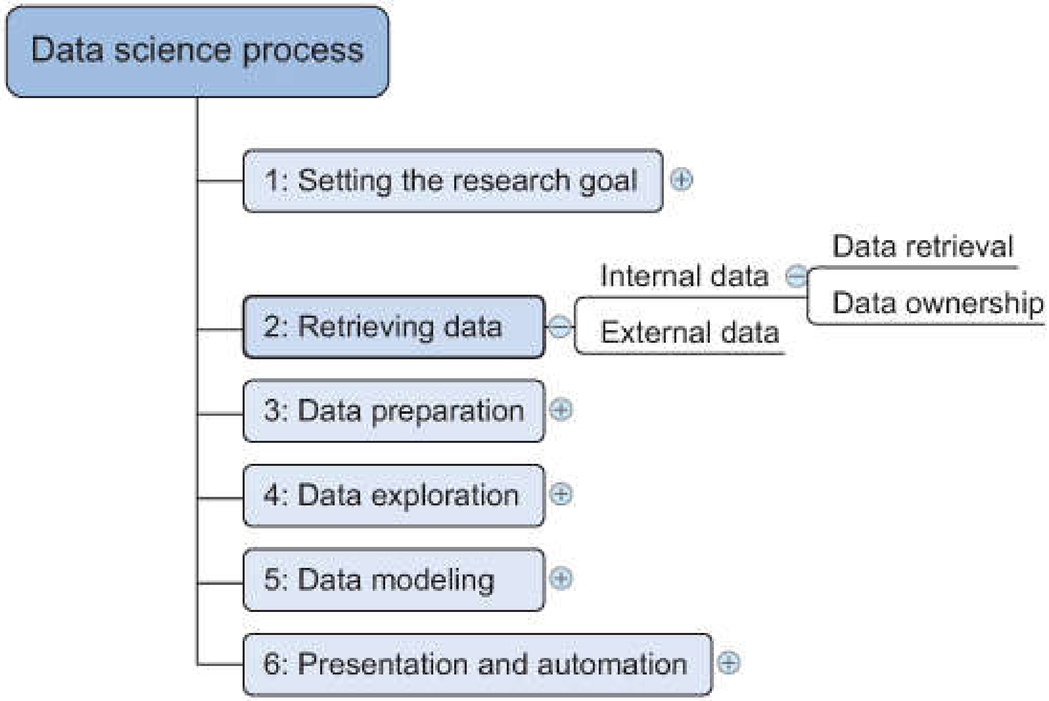
What resources you expect to use

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.

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

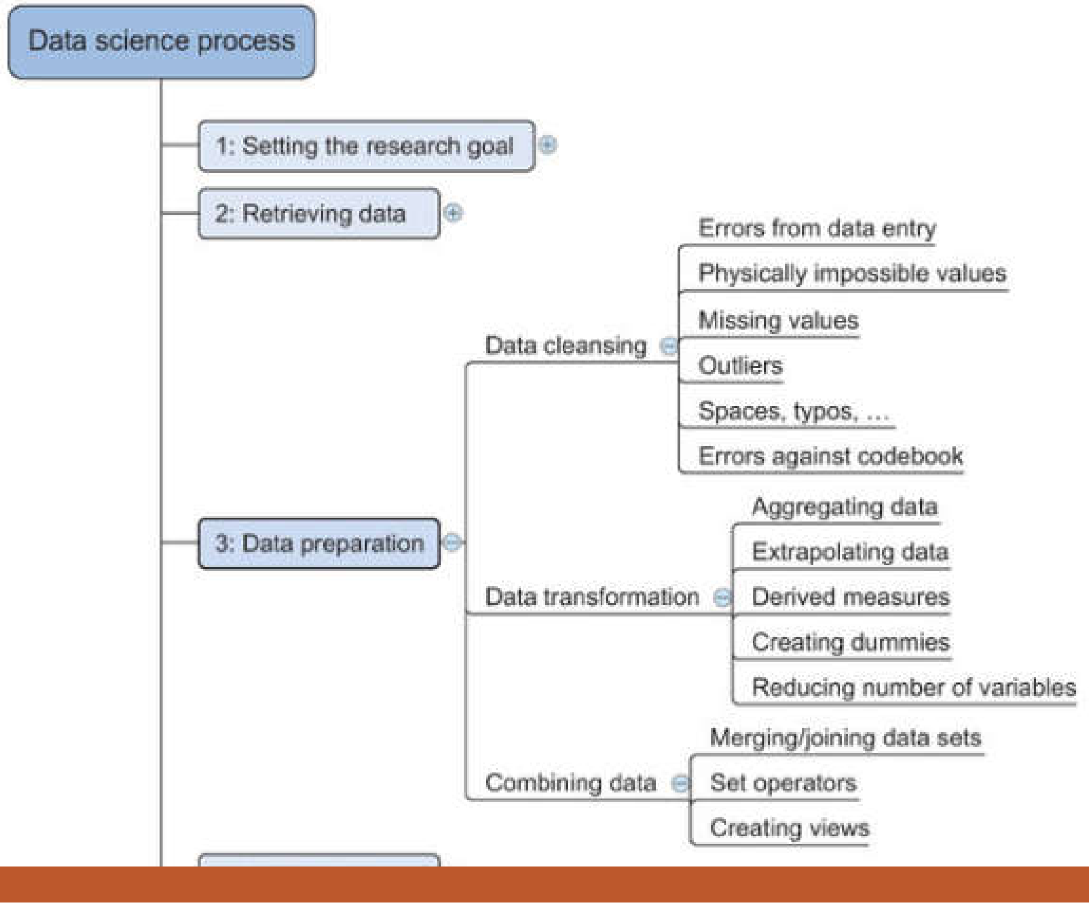
Acquiring Data

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

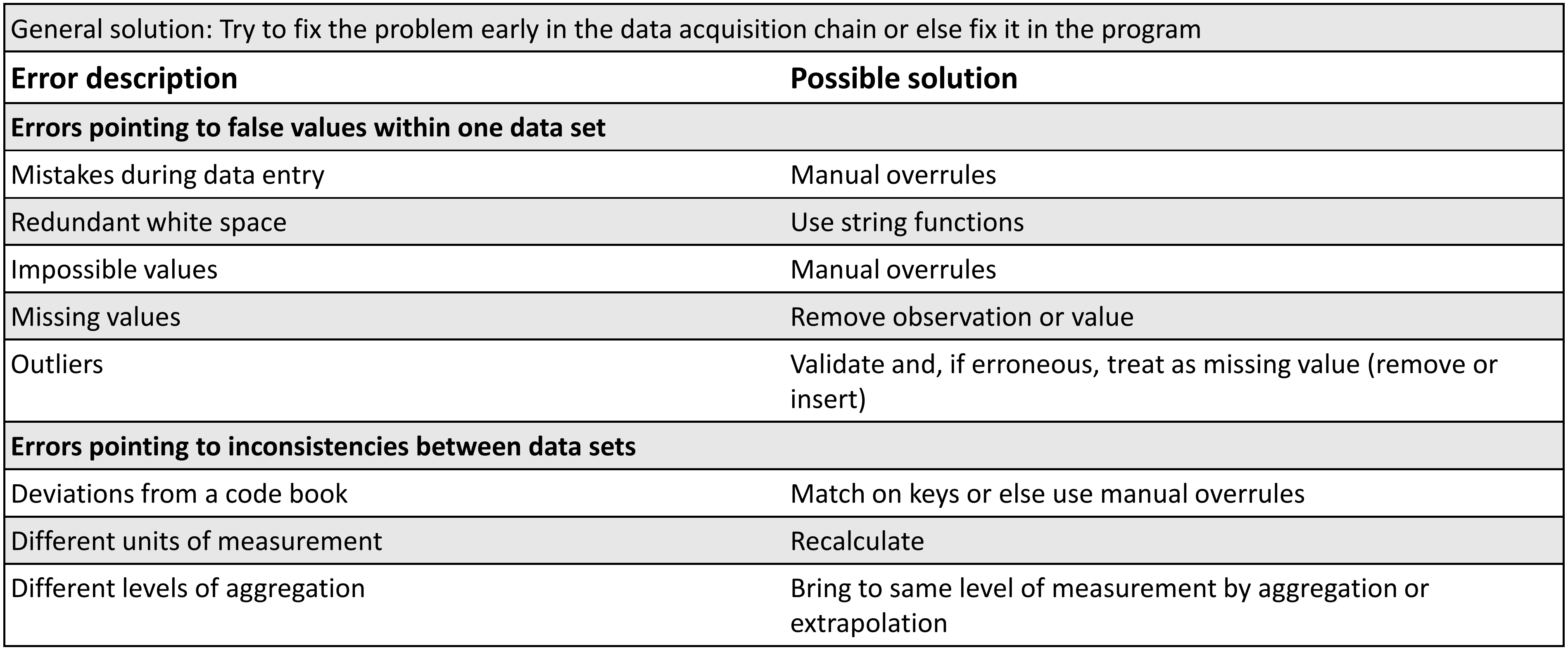
Figure 2.4. Step 3: Data preparation



Data science process

- 1: Setting the research goal
- 2: Retrieving data
- 3: Data preparation
  - Data cleansing
    - Errors from data entry
    - Physically impossible values
    - Missing values
    - Outliers
    - Spaces, typos, …
    - Errors against codebook
  - Data transformation
    - Aggregating data
    - Extrapolating data
    - Derived measures
    - Creating dummies
    - Reducing number of variables
  - Combining data
    - Merging/joining data sets
    - Set operators
    - Creating views

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

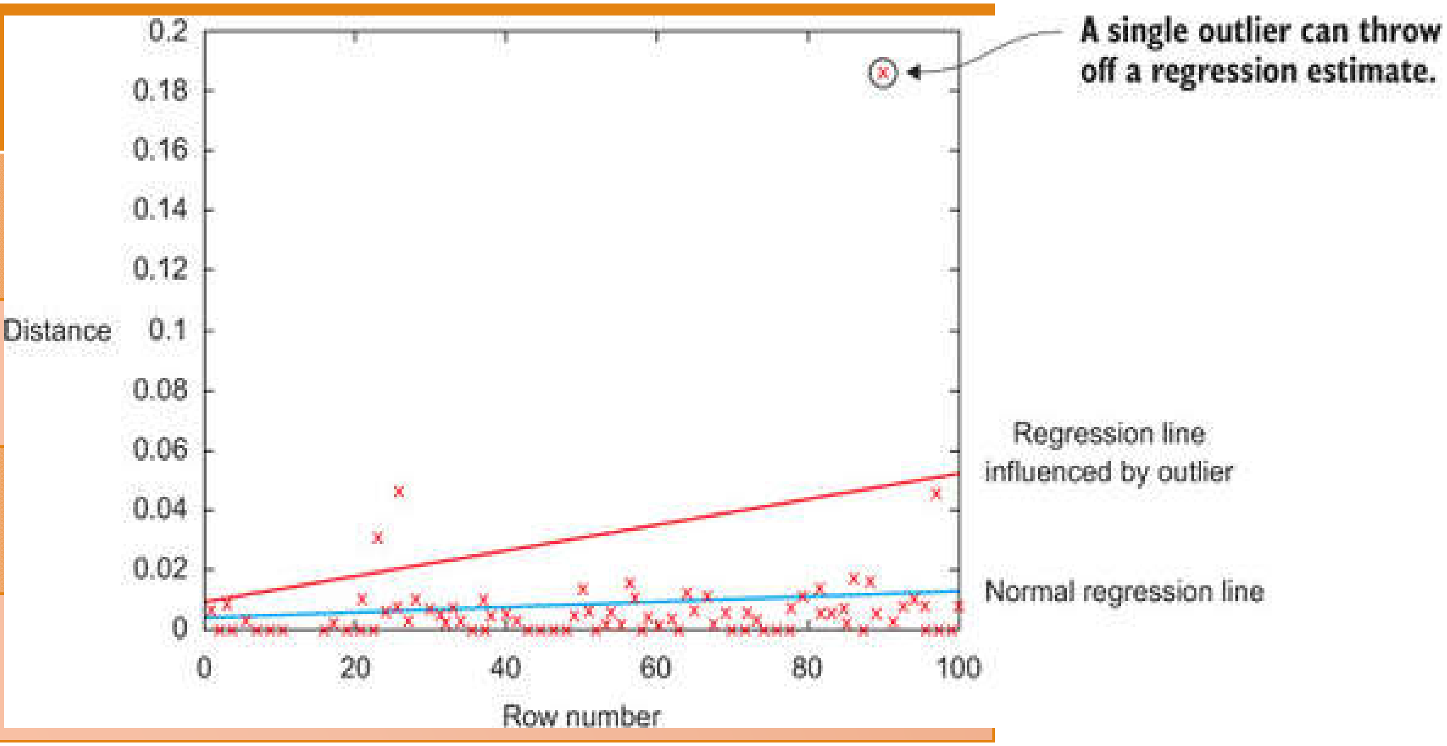


Table 2.3. Detecting outliers on simple variables with a frequency table

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.

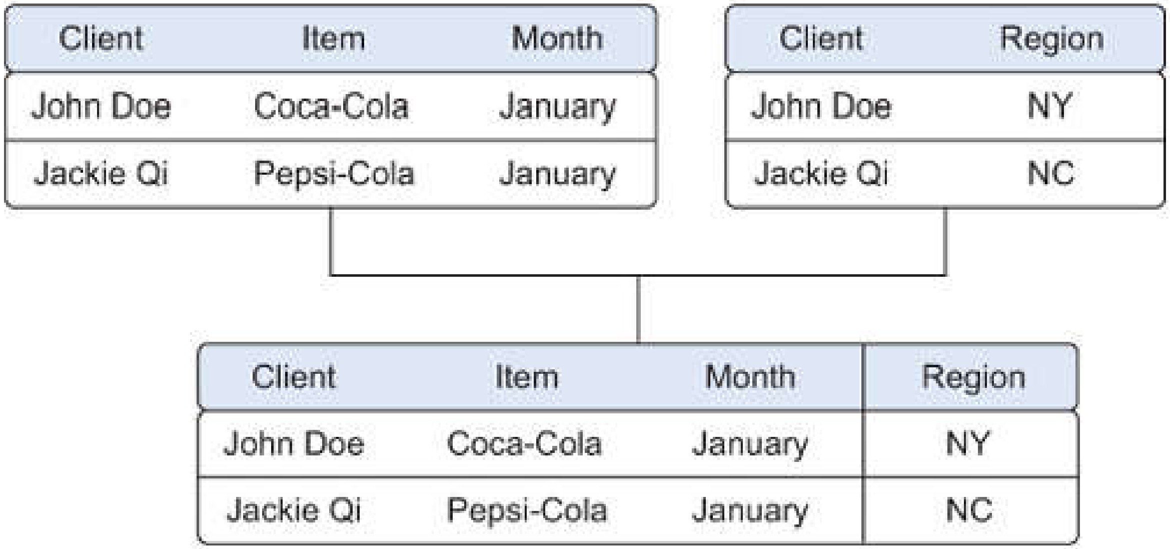
The first operation is ***joining***: enriching an observation from one table with information from another table.

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

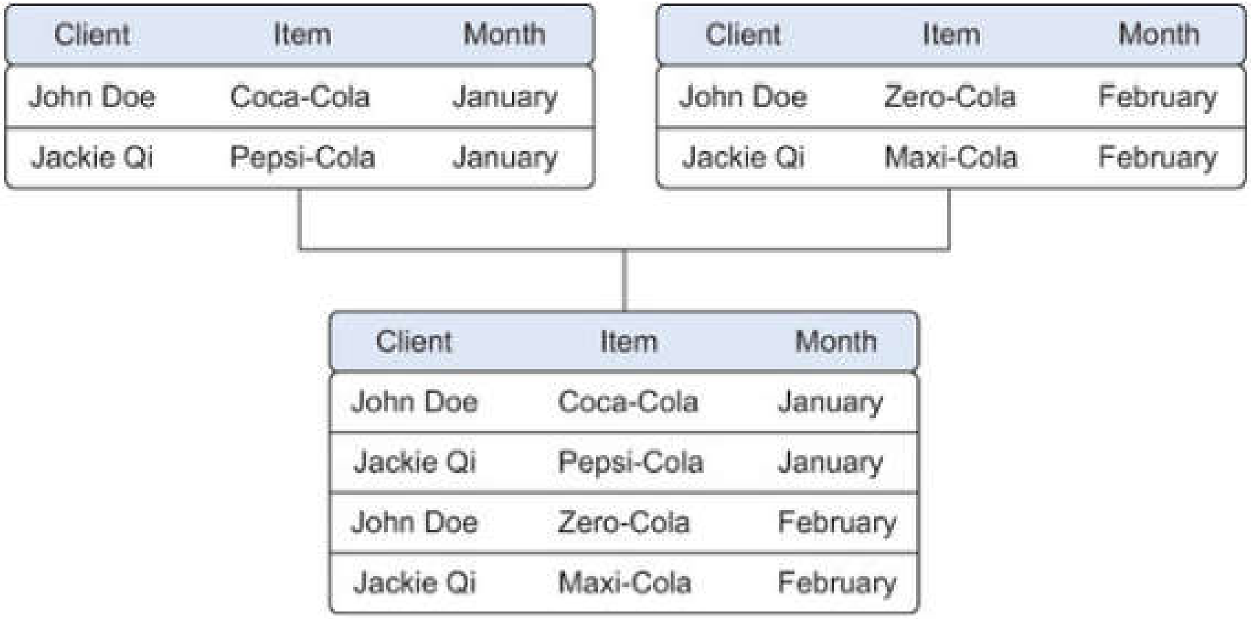
Client	Region
John Doe	NY
Jackie Qi	NC

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

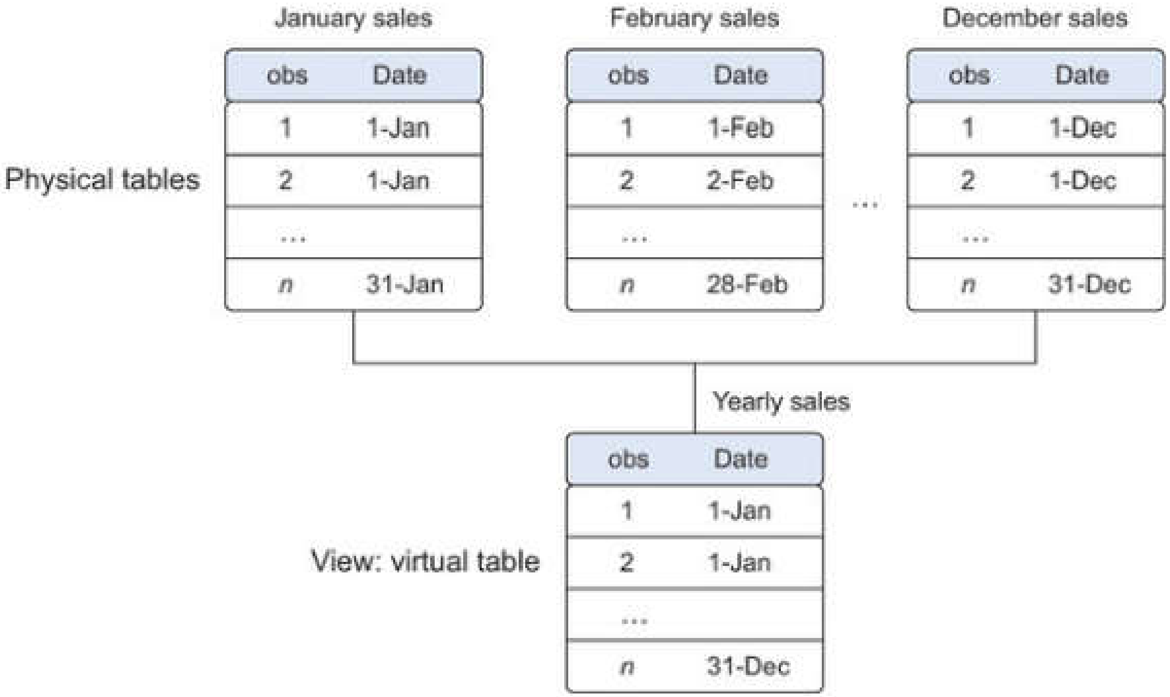
Client	Item	Month
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

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.



January sales

obs	Date
1	1-Jan
2	1-Jan
…	
n	31-Jan

February sales

obs	Date
1	1-Feb
2	2-Feb
…	
n	28-Feb

December sales

obs	Date
1	1-Dec
2	1-Dec
…	
n	31-Dec

Yearly sales

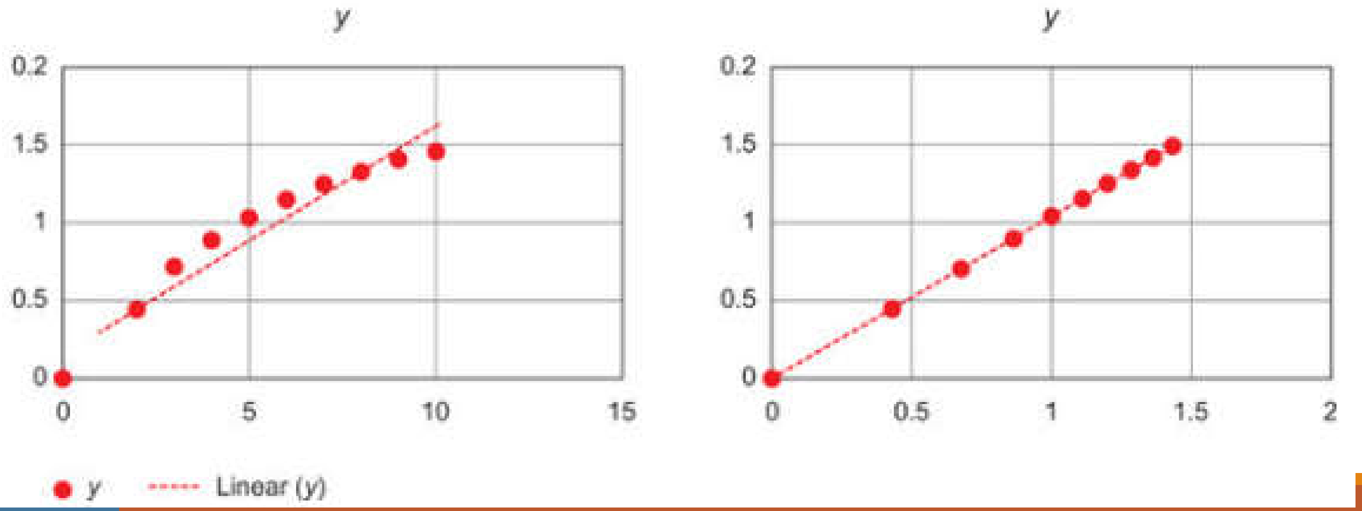
obs	Date
1	1-Jan
2	1-Jan
…	
n	31-Dec

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.

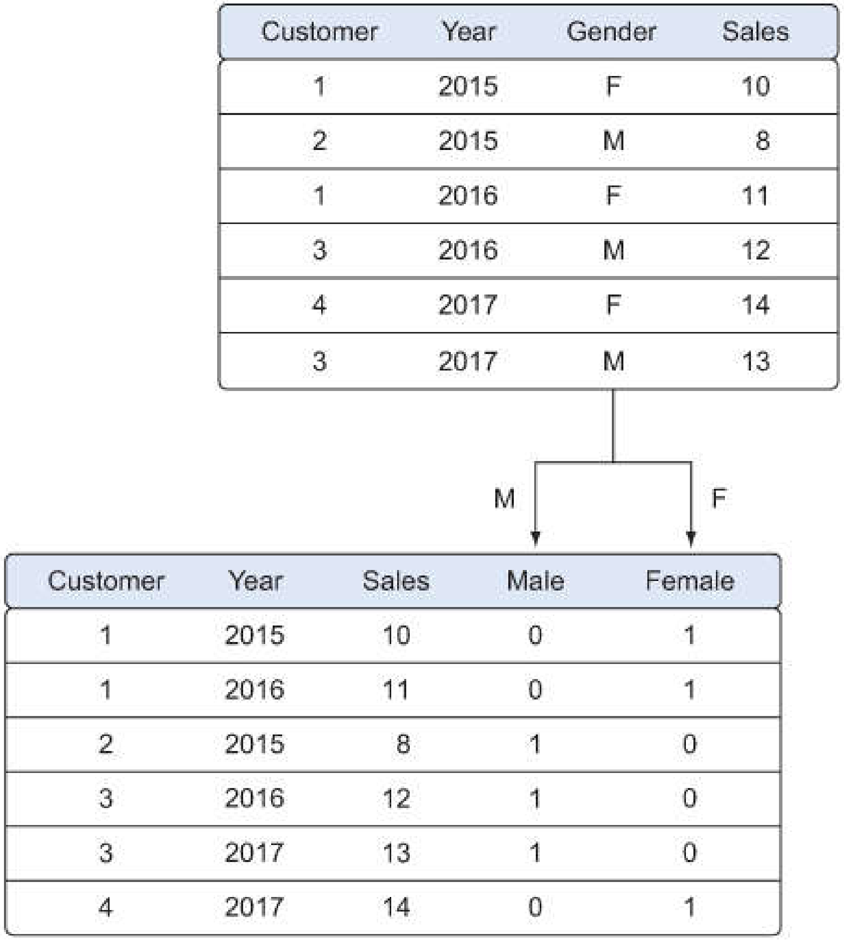
Dummy variables can only take two values: true(1) or false(0).

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