



## CHAPTER 2

# DATA SCIENCE PROCESS

Kwankamon Dittakan, Ph.D.  
College of Computing  
Prince of Songkla University  
Phuket, Thailand

# CONTENT

1. Overview of the data science process
2. Setting the research goal
3. Retrieving data
4. Data preparation
5. Data exploration
6. Data modelling
7. Presentation and automation

# 1. OVERVIEW OF THE DATA SCIENCE PROCESS

The typical data science process consists of six steps:

1. Setting the research goal
2. Retrieving data
3. Data preparation
4. Data exploration
5. Data modelling
6. Presentation and automation

# 1. OVERVIEW OF THE DATA SCIENCE PROCESS

- ❖ The first step of this process is setting a *research goal*. The main purpose here is making sure all the stakeholders understand the *what*, *how*, and *why* of the project. In every serious project this will result in a project charter.
- ❖ The second phase is *data retrieval*. You want to have data available for analysis, so this step includes finding suitable data and getting access to the data from the data owner. The result is data in its raw form, which probably needs polishing and transformation before it becomes usable.

# 1. OVERVIEW OF THE DATA SCIENCE PROCESS

- ❖ Now that you have the raw data, it's time to *prepare* it. This includes transforming the data from a raw form into data that's directly usable in your models. To achieve this, you'll detect and correct different kinds of errors in the data, combine data from different data sources, and transform it. If you have successfully completed this step, you can progress to data visualization and modeling.
- ❖ The fourth step is *data exploration*. The goal of this step is to gain a deep understanding of the data. You'll look for patterns, correlations, and deviations based on visual and descriptive techniques. The insights you gain from this phase will enable you to start modelling.

# 1. OVERVIEW OF THE DATA SCIENCE PROCESS

- ❖ The fifth step is **model building** (data modelling/ model construction/ model generation). It is now that you attempt to gain the insights or make the predictions stated in your project charter. Now is the time to bring out the complicated techniques, but is often a combination of simple models tends to outperform one complicated model.
- ❖ The last step is presenting the results and automating the analysis if needed. One goal of a project is to change a process and/or make better decisions.

In reality, the process probably does not progress in a linear way from step 1 to step 6. Sometimes, we might regress and iterate between the different phases.

## 2. SETTING THE RESEARCH GOAL

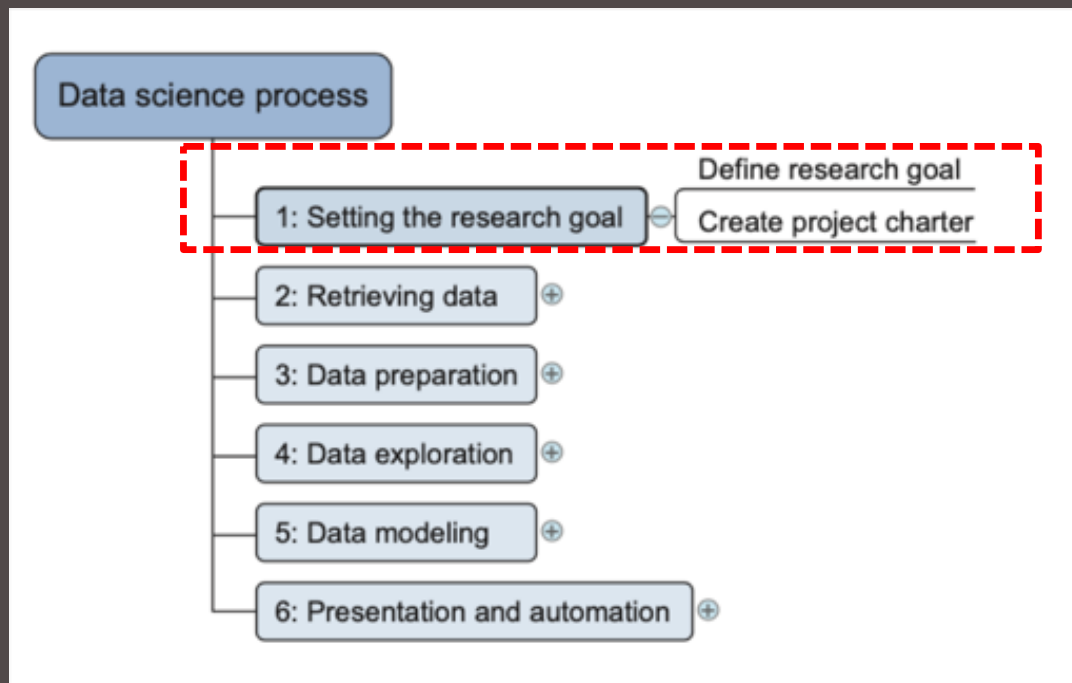
- ❖ A project starts by understanding the **what**, the **why**, and the **how** of the project.
  - ❖ What does the company expect you to do?
  - ❖ Why does management place such a value on your research?
  - ❖ Is it part of a bigger strategic picture or a “lone wolf” project originating from an opportunity someone detected?
- ❖ 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.

## 2. SETTING THE RESEARCH GOAL

- ❖ The outcome should be a clear research goal, a good understanding of the context, well-defined deliverables, and a plan of action with a timetable.
- ❖ This information is then best placed in a project charter.
- ❖ The length and formality can, of course, differ between projects and companies.
- ❖ In this early phase of the project, people skills and business acumen are more important than great technical prowess, which is why this part will often be guided by more senior personnel.



## 2. SETTING THE RESEARCH GOAL



## 2. SETTING THE RESEARCH GOAL

### 2.1 Spend time understanding the goals and context of your research

- ❖ An essential outcome is the research goal that states the purpose of the assignment in a clear and focused manner.
- ❖ Understanding the business goals and context is critical for project success.
- ❖ Continue asking questions and devising examples until grasping the exact business expectations
- ❖ Identify how the project fits in the bigger picture, appreciate how the research is going to change the business, and understand how they'll use the delivered results.
- ❖ Don't skim over this phase because many data scientists fail here.

## 2. SETTING THE RESEARCH GOAL

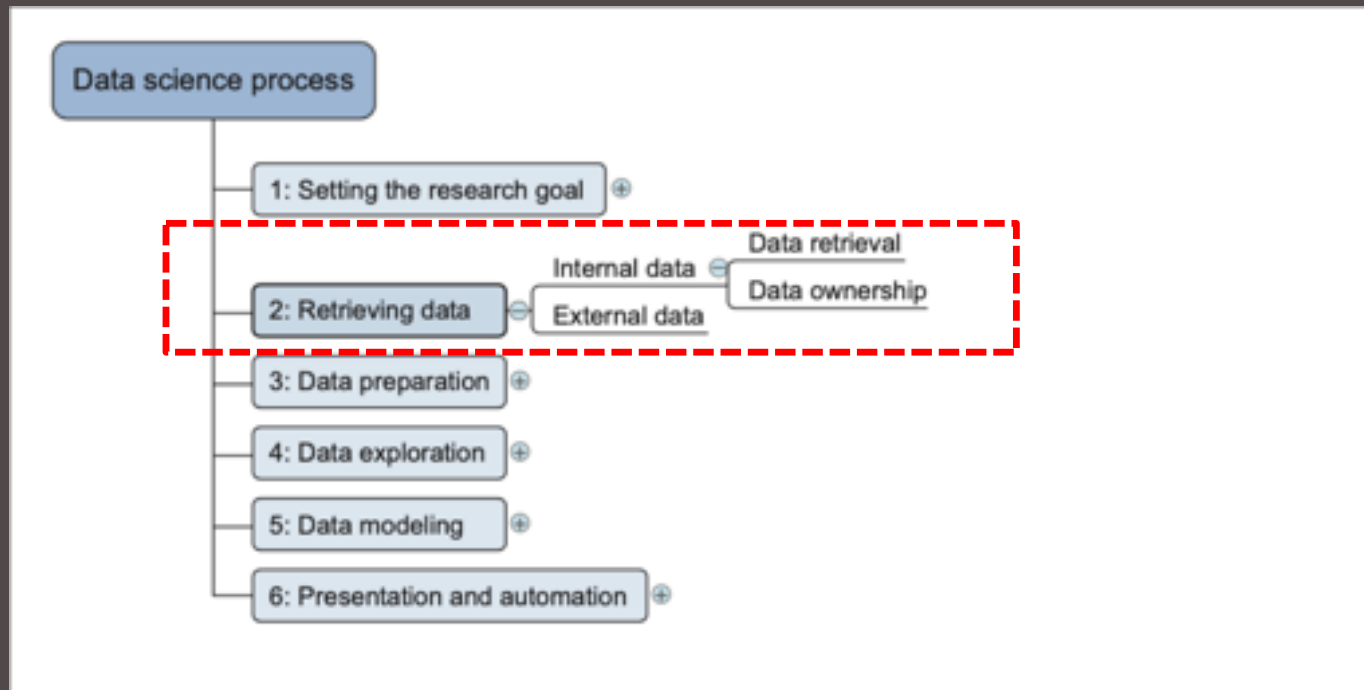
### 2.2 Create a project charter

- ❖ After we have a good understanding of the business problem, try to get a formal agreement on the deliverables.
- ❖ All this information is best collected in a project charter. For any significant project, this would be mandatory.
- ❖ Project charter requires teamwork, and the 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

## 3. RETRIEVING DATA

- ❖ Sometimes a user may need to go into the field and design a data collection process by themselves, but most of the time a user won't be involved in this step.
- ❖ Many companies will have already collected and stored the data, and what they don't have can often be bought from third parties.
- ❖ Don't be afraid to look outside the organisation for data, because more and more organisations are making even high-quality data freely available for public and commercial use.

### 3. RETRIEVING DATA



## 3. RETRIEVING DATA

- ❖ Assess the relevance and quality of **the data within the company**. This data can be stored in official data repositories such as databases, data marts, data warehouses, and data lakes
- ❖ Finding data even within the company can sometimes be a challenge. As companies grow, their data becomes scattered around many places.
- ❖ Knowledge of the data may be dispersed as people change positions and leave the company.
- ❖ Getting access to data is another difficult task. Organizations understand the value and sensitivity of data and often have policies in place so everyone has access to what they need.

## 3. RETRIEVING DATA

- ❖ If data isn't available inside the organisation, look **outside** the organization's walls.
- ❖ Many companies specialize in collecting valuable information.
- ❖ For instance, Nielsen and GFK are well known for this in the retail industry. Other companies provide data so that can be used to enrich their services and ecosystem. Such is the case with Twitter, LinkedIn, and Facebook.

## 3. RETRIEVING DATA

❖ Example of open data:

| Open data site  | Description   |
|---|---|
| Data.gov  | The home of the US Government's open data   |
| <a href="https://open-data.europa.eu/">https://open-data.europa.eu/</a> | 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  |



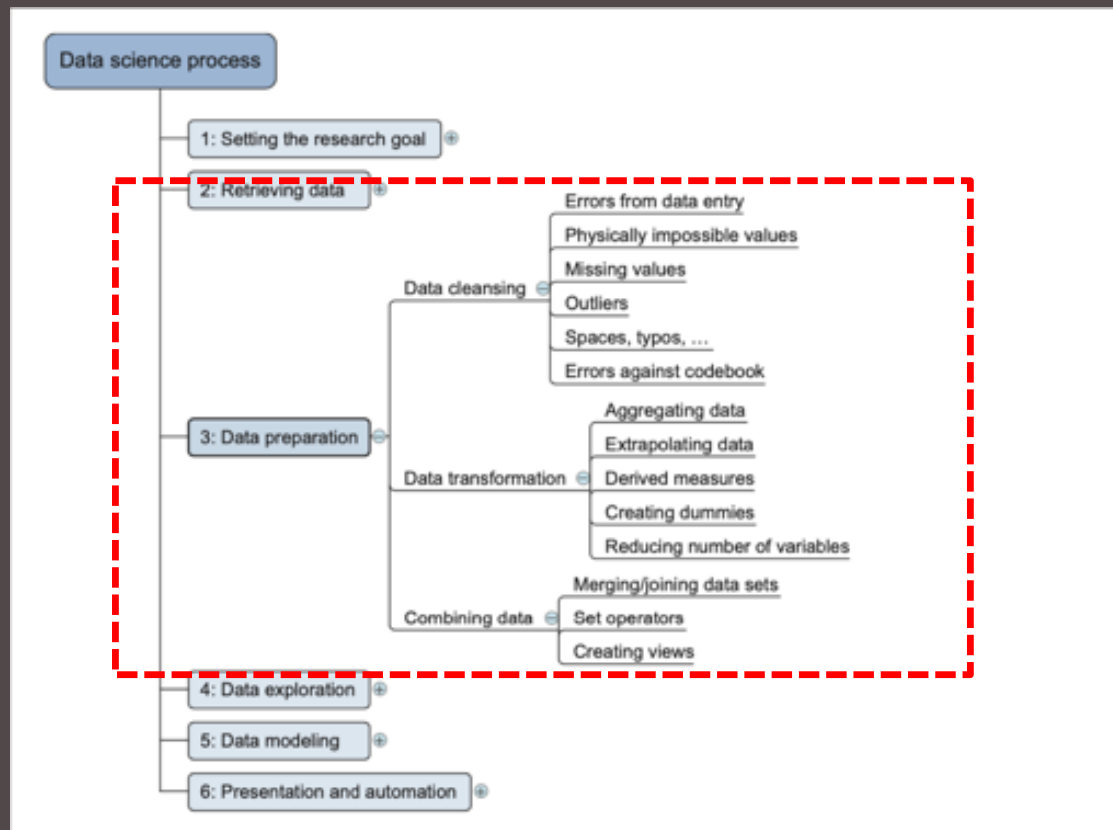
## 3. RETRIEVING DATA

- ❖ Expect to spend a good portion of your project time doing data correction and cleansing, sometimes up to 80%.
- ❖ The retrieval of data is the first time you'll inspect the data in the data science process.
- ❖ Most of the errors we'll encounter during the data-gathering phase are easy to spot, but being too careless will make us spend many hours solving data issues that could have been prevented during data import.
- ❖ During data retrieval, we check to see if the data is equal to the data in the source document and look to see if we have the right data types. When we have enough evidence that the data is similar to the data we find in the source document, then stop.

## 4. DATA PREPARATION

- ❖ Once the data retrieval has been completed, the next process is to sanitize and prepare it for use in the modelling and reporting phase.
- ❖ Doing so is tremendously important because the models will perform better and we will lose less time trying to fix strange output.
- ❖ "Garbage in equals garbage out"
- ❖ The model needs the data in a specific format, so data transformation will always come into play.
- ❖ The figure shows the most common actions to take during the data cleaning, integration, and transformation phase.

## 4. DATA PREPARATION



## 4. 1 DATA CLEANING

- ❖ Data cleaning is a subprocess of the data science process that focuses on removing errors in the data, so data becomes a true and consistent representation of the processes it originates from.
- ❖ By “true and consistent representation” we imply that at least two types of errors exist.
  - ❖ The first type is the interpretation error, such as when we take the value in the data for granted, like saying that a person’s age is greater than 300 years.
  - ❖ The second type of error points to inconsistencies between data sources or against your company’s standardized values.
    - ❖ An example of this class of errors is putting “Female” in one table and “F” in another when they represent the same thing: that the person is female.
    - ❖ Another example is that you use Pounds in one table and Dollars in another.
- ❖ Too many possible errors exist for this list to be exhaustive, but the next table shows an overview of the types of errors that can be detected with easy checks—the “low hanging fruit,” as it were.

## 4. 1 DATA CLEANING

| General solution  |   |
|---|---|
| Try to fix the problem early in the data acquisition chain or else fix it in the program. |   |
| Error description   | Possible solution   |
| <i>Errors pointing to false values within one data set</i>                                |   |
| 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) |
| <i>Errors pointing to inconsistencies between data sets</i>                               |   |
| 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    |

## 4. 1 DATA CLEANING

### A. DATA ENTRY ERRORS

- ❖ Data collection and data entry are error-prone processes.
- ❖ They often require human intervention, and because humans are only human, they make typos or lose their concentration for a second and introduce an error into the chain.
- ❖ But data collected by machines or computers aren't free from errors either.
- ❖ Errors can arise from human sloppiness, whereas others are due to machine or hardware failure. Examples of errors originating from machines are transmission errors or bugs in the extract, transform, and load phase (ETL).

## 4. 1 DATA CLEANING

### A. DATA ENTRY ERRORS

- ❖ For small data sets, you can check every value by hand.
- ❖ Detecting data errors when the variables you study don't have many classes can be done by tabulating the data with counts.

| Value | Count   |
|-------|---------|
| Good  | 1598647 |
| Bad   | 1354468 |
| Godø  | 15      |
| Bade  | 1       |

```
if x == "Godø":  
    x = "Good"  
if x == "Bade":  
    x = "Bad"
```

## 4. 1 DATA CLEANING

### B. REDUNDANT WHITESPACE

- ❖ Whitespaces tend to be hard to detect but cause errors as other redundant characters would.
- ❖ The cleaning during the ETL phase wasn't well-executed, and keys in one table contained whitespace at the end of a string.
- ❖ This caused a mismatch of keys such as "FR " – "FR", dropping the observations that couldn't be matched.
- ❖ If you know to watch out for them, fixing redundant whitespaces is luckily easy enough in most programming languages. They all provide string functions that will remove the leading and trailing whitespace.
  - ❖ For instance, in Python we can use the `strip()` function to remove leading and trailing spaces.



## 4. 1 DATA CLEANING

### C. FIXING CAPITAL LETTER MISMATCHES

- ❖ Capital letter mismatches are common.
- ❖ Most programming languages make a distinction between “Brazil” and “brazil”.
- ❖ In this case, you can solve the problem by applying a function that returns both strings in lowercase, such as `.lower()` in Python. `“Brazil”.lower() == “brazil”.lower()` should result in `true`.

## 4. 1 DATA CLEANING

### D. IMPOSSIBLE VALUES AND SANITY CHECKS

- ❖ Sanity checks are another valuable type of data check.
- ❖ Here we check the value against physically or theoretically impossible values such as people taller than 3 meters or someone with an age of 299 years.
- ❖ Sanity checks can be directly expressed with rules:
  - ❖  $\text{check} = 0 \leq \text{age} \leq 120$

## 4. 1 DATA CLEANING

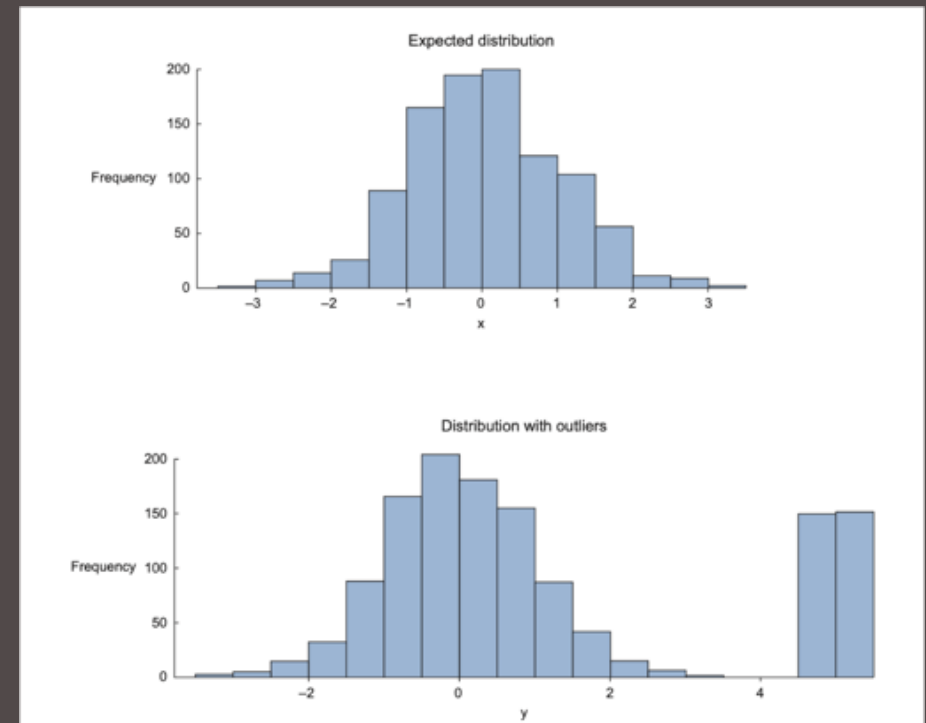
### E. OUTLIERS

- ❖ An outlier is an observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations.
- ❖ The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.

## 4. 1 DATA CLEANING

### E. OUTLIERS

- ❖ The plot on the top shows no outliers, whereas the plot on the bottom shows possible outliers on the upper side when a normal distribution is expected.
- ❖ Outliers can gravely influence your data modelling, so investigate them first.



## 4. 1 DATA CLEANING

### F. DEALING WITH MISSING VALUES

- ❖ Missing values aren't necessarily wrong, but they still need to be handled separately; certain modelling techniques can't handle missing values.
- ❖ They might be an indicator that something went wrong in your data collection or that an error happened in the ETL process.
- ❖ Common techniques data scientists use are listed in the next table

# 4. 1 DATA CLEANING

## F. DEALING WITH 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<br>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<br>You make data assumptions  |
| Modeling the value (nondependent)                            | Does not disturb the model too much   | Can lead to too much confidence in the model<br>Can artificially raise dependence among the variables<br>Harder to execute<br>You make data assumptions |

## 4. 1 DATA CLEANING

### G. DEVIATIONS FROM A CODE BOOK

- ❖ Detecting errors in larger data sets against a **codebook** or against standardized values can be done with the help of set operations.
- ❖ A codebook is a description of the data, a form of metadata. It contains things such as the number of variables per observation, the number of observations, and what each encoding within a variable means.
- ❖ For instance “0” equals “negative”, “5” stands for “very positive”. A codebook also tells the type of data you’re looking at: is it a hierarchical, graph, something else?

## 4. 1 DATA CLEANING

### H. DIFFERENT UNITS OF MEASUREMENT

- ❖ When integrating two data sets, we have to pay attention to their respective units of measurement.
- ❖ An example of this would be when we study the prices of gasoline in the world.
  - ❖ To do this we gather data from different data providers.
  - ❖ Data sets can contain prices per gallon and others can contain prices per litre.
  - ❖ A simple conversion will do the trick in this case.



## 4. 1 DATA CLEANING

### I. DIFFERENT LEVELS OF AGGREGATION

- ❖ Having different levels of aggregation is similar to having different types of measurement.
- ❖ An example of this would be a data set containing data per week versus one containing data per workweek.
- ❖ This type of error is generally easy to detect, and *summarizing* (or the inverse, *expanding*) the data sets will fix it.

## 4.2 DATA INTEGRATION

- ❖ This process is mainly focused on data integration from different data sources.
- ❖ Data vary in size, type, and structure, ranging from databases and Excel files to text documents.
- ❖ There are 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.
- ❖ When we combine data, it might have the option to create a new physical table or a virtual table by creating a view. The advantage of a view is that it doesn't consume more disk space.

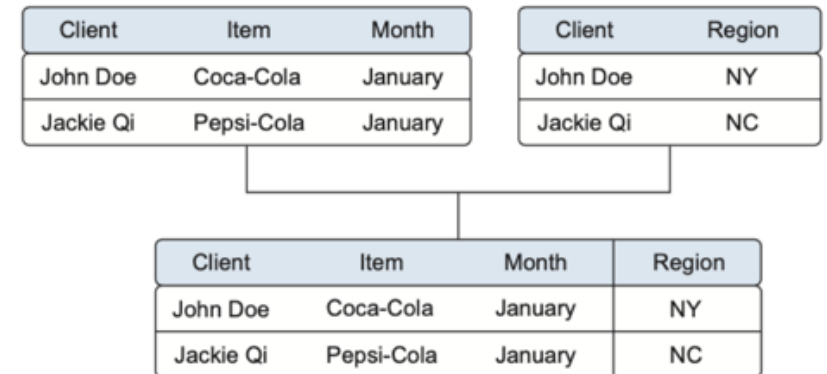
## 4.2 DATA INTEGRATION

- ❖ This process is mainly focused on data integration from different data sources.
- ❖ Data vary in size, type, and structure, ranging from databases and Excel files to text documents.
- ❖ There are 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.
- ❖ When we combine data, it might have the option to create a new physical table or a virtual table by creating a view. The advantage of a view is that it doesn't consume more disk space.

## 4.2 DATA INTEGRATION

### A. JOINING TABLES

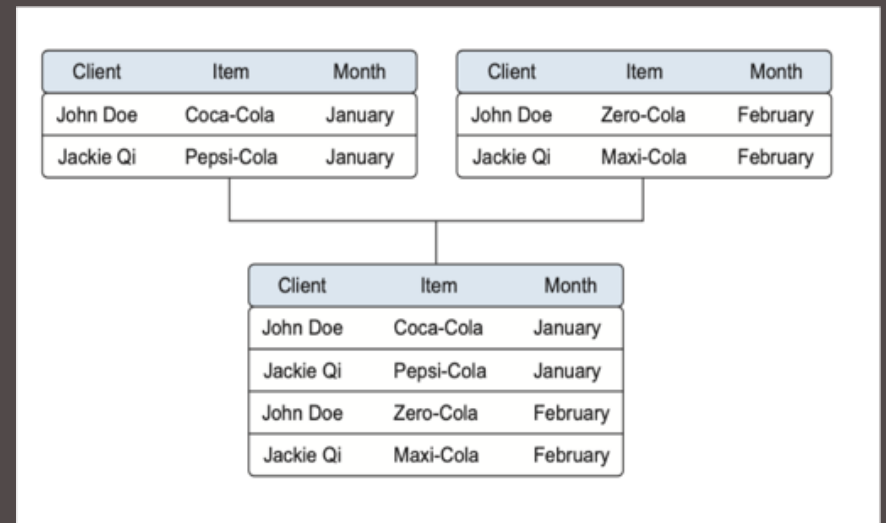
- ❖ Joining tables allows combining the information of one observation found in one table with the information that we find in another table.
- ❖ To join tables, the variables that represent the same object in both tables are applied, such as a date, a country name, or a Social Security number.
- ❖ These common fields are known as keys. When these keys also uniquely define the records in the table they are called **primary keys**.



## 4.2 DATA INTEGRATION

### B. APPENDING TABLES

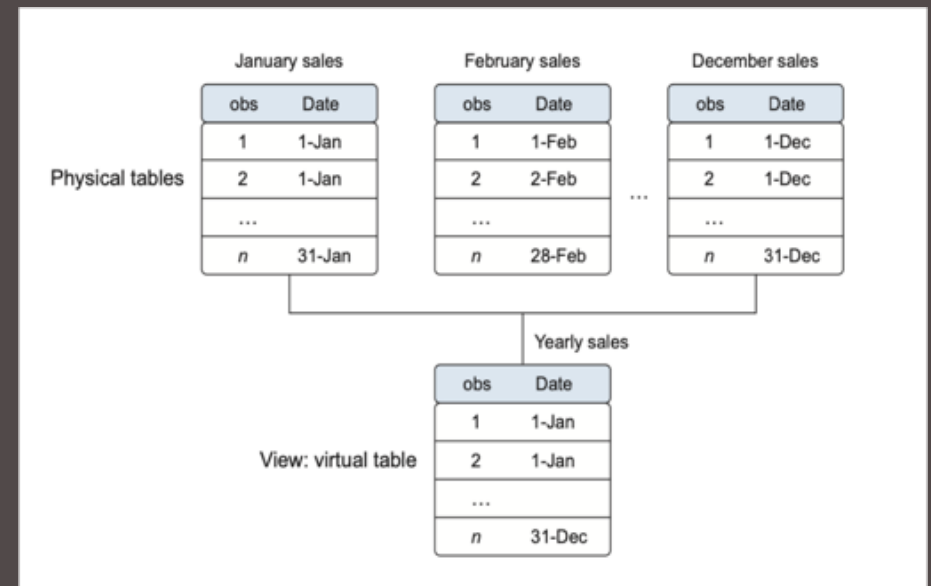
- ❖ Appending or stacking tables is effectively adding observations from one table to another table.
- ❖ From the example, the result of appending these tables is a larger one with the observations from January as well as February.
- ❖ The equivalent operation in set theory would be the union, and this is also the command in SQL, the common language of relational databases.



## 4.2 DATA INTEGRATION

### C. VIEW

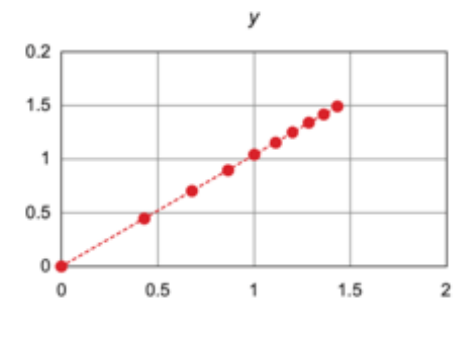
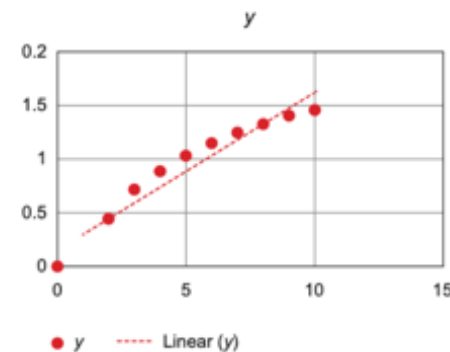
- ❖ To avoid duplication of data, we virtually combine data with views.
- ❖ For this reason, the concept of a view was invented. A view behaves as if you're working on a table, but this table is nothing but a virtual layer that combines the tables for you.
- ❖ Views do come with a drawback, however. While a table join is only performed once, the join that creates the view is recreated every time it's queried, using more processing power than a pre-calculated table would have.



## 4.3 DATA TRANSFORMATION

- ❖ Certain models require their data to be in a certain shape.
- ❖ Once the data is cleansed and integrated, this is the next task to be performed: transforming the data so it takes a suitable form for data modelling.
- ❖ Data scientists use special methods to reduce the number of variables but retain the maximum amount of data.

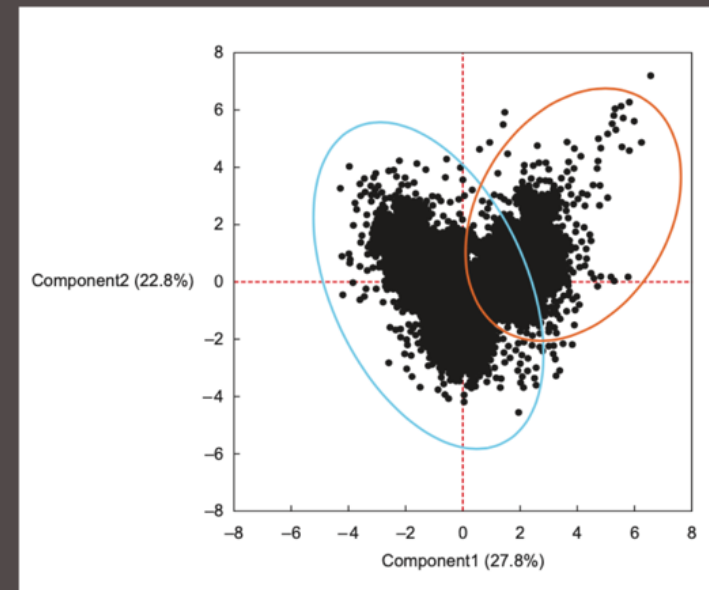
| x      | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |
|--------|------|------|------|------|------|------|------|------|------|------|
| log(x) | 0.00 | 0.43 | 0.68 | 0.86 | 1.00 | 1.11 | 1.21 | 1.29 | 1.37 | 1.43 |
| y      | 0.00 | 0.44 | 0.69 | 0.87 | 1.02 | 1.11 | 1.24 | 1.32 | 1.38 | 1.46 |



## 4.3 DATA TRANSFORMATION

### A. REDUCING THE NUMBER OF VARIABLES

- ❖ Sometimes we 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 the model makes the model difficult to handle, and certain techniques don't perform well.
- ❖ From the example figure, principal components analysis (PCA) is applied.






## 4.3 DATA TRANSFORMATION

### B. TURNING VARIABLES INTO DUMMIES

- ❖ 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.
- ❖ An example is turning one column named Weekdays into the columns Monday through Sunday.
- ❖ Turning variables into dummies is a technique that's used in modeling and is popular with, but not exclusive to, economists.

| 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    |

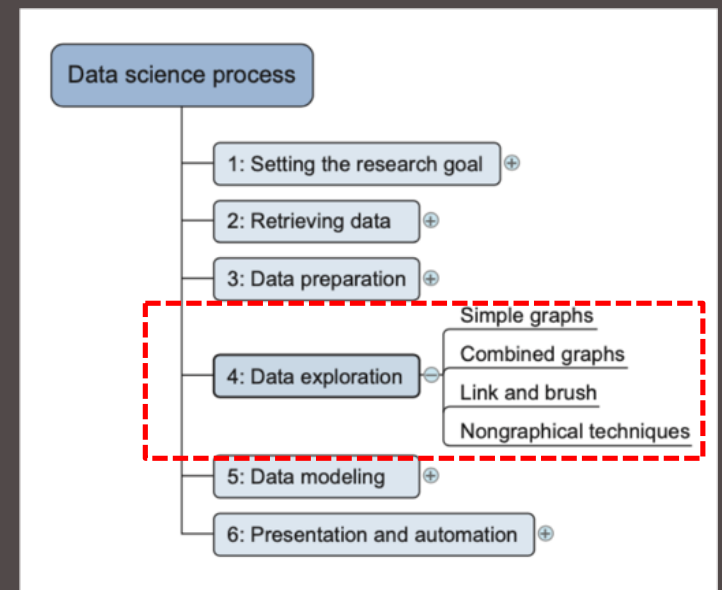


A diagram showing a vertical line from the 'Gender' column of the top table branching into two arrows labeled 'M' and 'F', pointing to the 'Male' and 'Female' columns of the bottom table respectively.

| 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      |

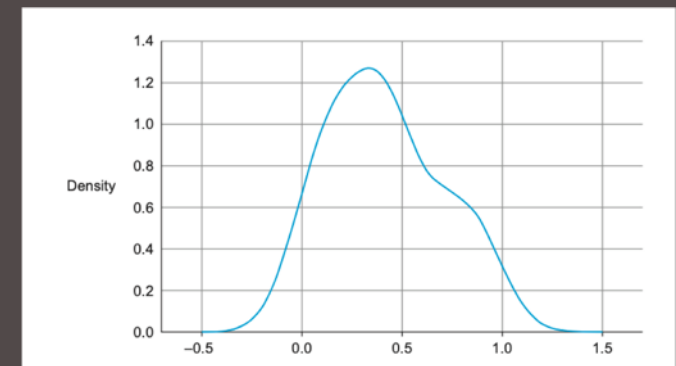
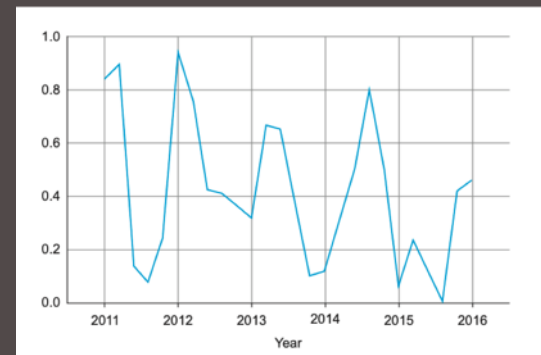
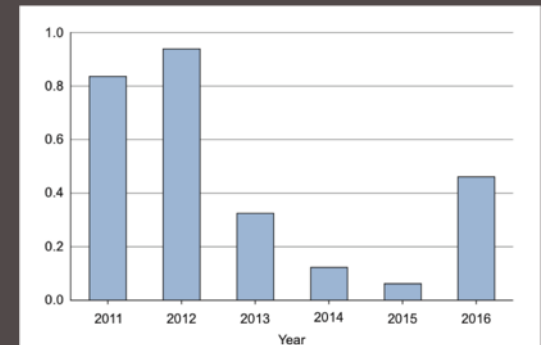
## 5. DATA EXPLORATION

- ❖ During exploratory data analysis we take a deep dive into the data.
- ❖ Information becomes much easier to grasp when shown in a picture, therefore using graphical techniques to gain an understanding of the data and the interactions between variables.
- ❖ The goal isn't to clean the data, but it's common that sometime still discover anomalies we missed before, forcing us to take a step back and fix them.

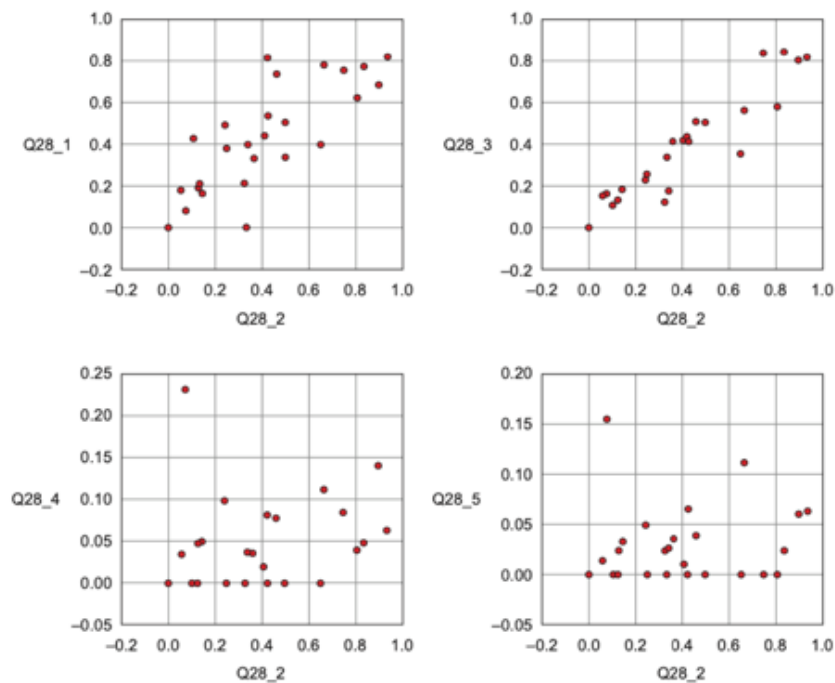


## 5. DATA EXPLORATION

- ❖ The visualization techniques using in this phase range from simple line graphs or histograms to more complex diagrams such as Sankey and network graphs.
- ❖ Sometimes it's useful to compose a composite graph from **simple graphs** to get even more insight into the data.
- ❖ Other times the graphs can be animated or made interactive to make it easier and, let's admit it, way more fun.

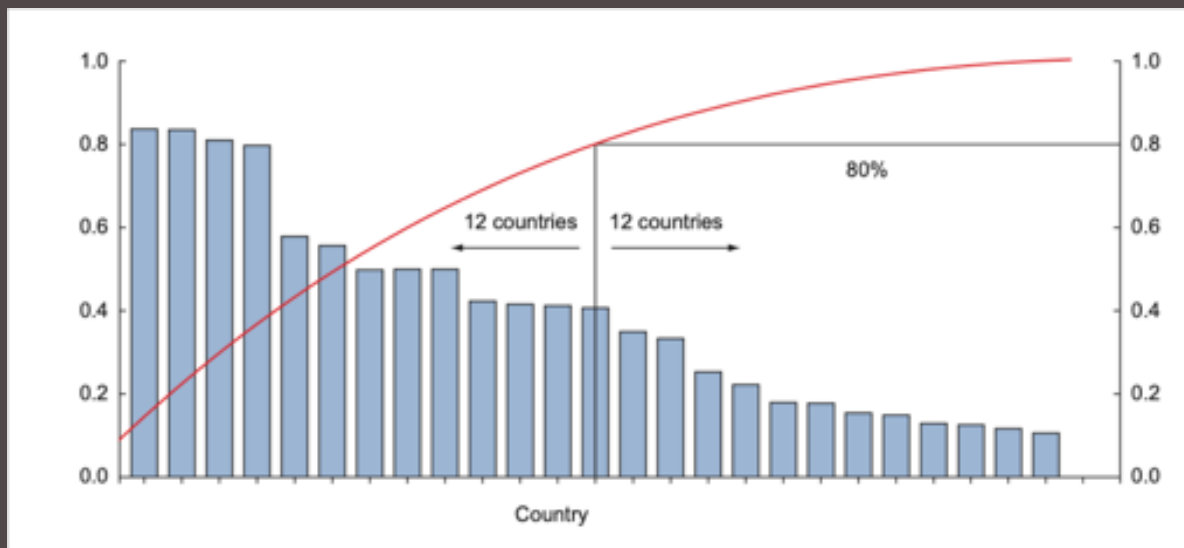


## 5. DATA EXPLORATION



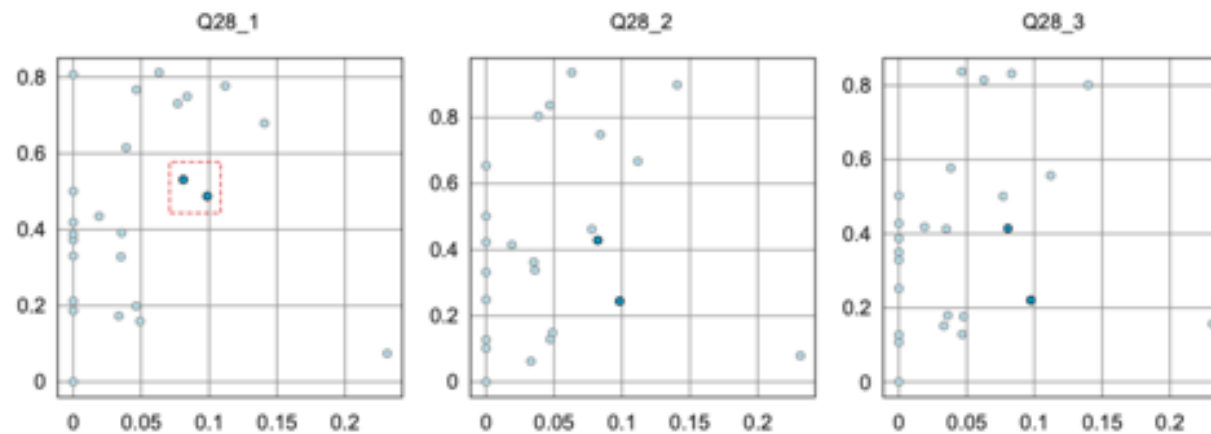
Drawing **multiple plots** together can help we understand the structure of the data over multiple variables.

## 5. DATA EXPLORATION



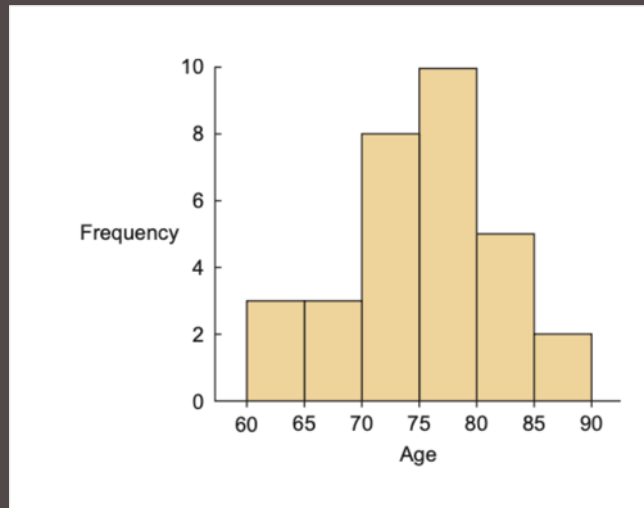
- ❖ A **Pareto diagram** is a combination of the values and a cumulative distribution. It's easy to see from this diagram that the first 50% of the countries contain slightly less than 80% of the total amount.
- ❖ If this graph represented customer buying power and we sell expensive products, we probably don't need to spend our marketing budget in every country; we could start with the first 50%.

## 5. DATA EXPLORATION



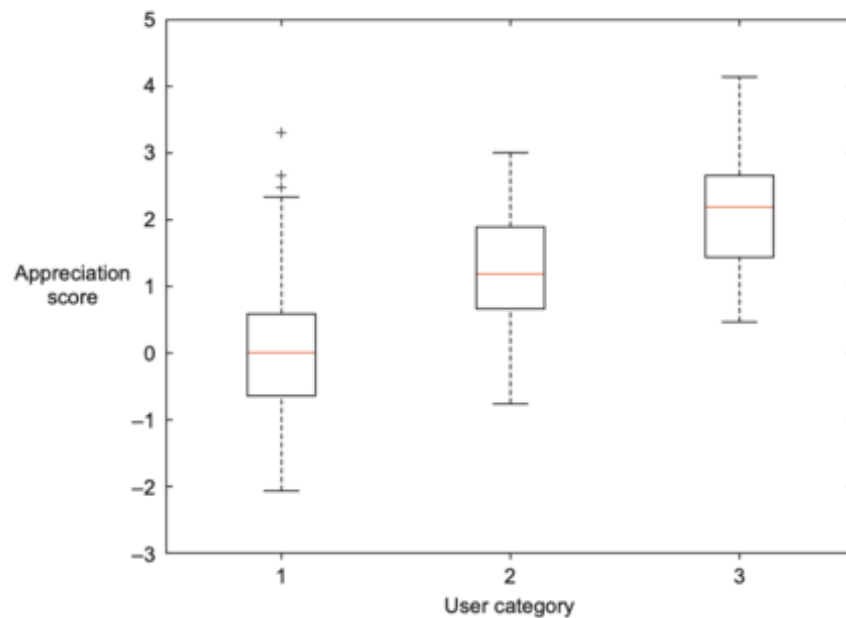
Link and brush allows we to select observations in one plot and highlight the same observations in the other plots.

## 5. DATA EXPLORATION



- ❖ In a **histogram** a variable is cut into discrete categories and the number of occurrences in each category are summed up and shown in the graph.
- ❖ Example histogram: the number of people in the age-groups of 5-year intervals

## 5. DATA EXPLORATION



- ❖ The **boxplot**, on the other hand, doesn't show how many observations are present but does offer an impression of the distribution within categories.
- ❖ It can show the maximum, minimum, median, and other characterizing measures at the same time.
- ❖ Example boxplot: each user category has a distribution of the appreciation each has for a certain picture on a photography website.



## 6. DATA MODELLING

- ❖ With clean data in place and a good understanding of the content, we're ready to build models with the goal of making better predictions, classifying objects, or gaining an understanding of the system that we're modeling.
- ❖ This phase is much more focused than the exploratory analysis step, because we know what we're looking for and what we want the outcome to be.
- ❖ Building a model is an iterative process. The way to build the model depends on whether we go with **classic statistics** or the somewhat more recent **machine learning** school, and the type of technique we want to use. Either way, most models consist of the following main steps:
  1. Selection of a modeling technique and variables to enter in the model
  2. Execution of the model
  3. Diagnosis and model comparison

## 6.1 MODEL AND DATA SELECTION

- ❖ The data selection is used to select the variables needed in the model and a modelling technique.
- ❖ The exploratory analysis should already give a fair idea of what variables will help to construct a good model.
- ❖ Many modeling techniques are available, and choosing the right model for a problem requires judgment.
- ❖ The model performance should be considered and whether the project meets all the requirements to use the model, as well as other factors:
  - ❖ Must the model be moved to a production environment and, if so, would it be easy to implement?
  - ❖ How difficult is the maintenance on the model: how long will it remain relevant if left untouched?
  - ❖ Does the model need to be easy to explain?

## 6.2 MODEL EXECUTION

- ❖ Most programming languages, such as Python, already have libraries such as StatsModels or Scikit-learn.
- ❖ These packages use several of the most popular techniques.
- ❖ Coding a model is a nontrivial task in most cases, so having these libraries available can speed up the process.
- ❖ The following code, it's fairly easy to use linear regression with StatsModels or Scikit-learn.

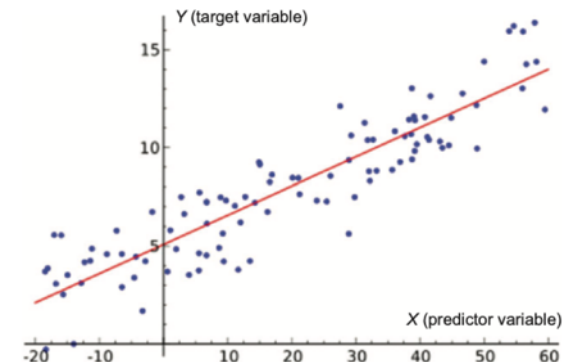
```
import statsmodels.api as sm
import numpy as np
predictors = np.random.random(1000).reshape(500,2)
target = predictors.dot(np.array([0.4, 0.6])) + np.random.random(500)
lmRegModel = sm.OLS(target,predictors)
result = lmRegModel.fit()
result.summary()
```

Shows model  
fit statistics.

Fits linear  
regression  
on data.

Imports required  
Python modules.

Creates random data for  
predictors (x-values) and  
semi-random data for  
the target (y-values) of the  
model. We use predictors as  
input to create the target so  
we infer a correlation here.



## 6.2 MODEL EXECUTION

|                   |                  |                     |           |
|-------------------|------------------|---------------------|-----------|
| Dep. Variable:    | y                | R-squared:          | 0.893     |
| Model:            | OLS              | Adj. R-squared:     | 0.893     |
| Method:           | Least Squares    | F-statistic:        | 2088.     |
| Date:             | Fri, 30 Oct 2015 | Prob (F-statistic): | 7.13e-243 |
| Time:             | 12:44:31         | Log-Likelihood:     | -176.74   |
| No. Observations: | 500              | AIC:                | 357.5     |
| Df Residuals:     | 498              | BIC:                | 365.9     |
| Df Model:         | 2                |                     |           |
| Covariance Type:  | nonrobust        |                     |           |

Model fit: higher is better but too high is suspicious.

p-value to show whether a predictor variable has a significant influence on the target. Lower is better and <0.05 is often considered "significant."

|    | coef   | std err | t      | P> t  | [95.0% Conf. Int.] |
|----|--------|---------|--------|-------|--------------------|
| x1 | 0.7658 | 0.040   | 19.130 | 0.000 | 0.687 0.844        |
| x2 | 1.1252 | 0.039   | 28.603 | 0.000 | 1.048 1.202        |

|                |        |                   |         |
|----------------|--------|-------------------|---------|
| Omnibus:       | 34.269 | Durbin-Watson:    | 1.943   |
| Prob(Omnibus): | 0.000  | Jarque-Bera (JB): | 13.480  |
| Skew:          | -0.125 | Prob(JB):         | 0.00118 |
| Kurtosis:      | 2.235  | Cond. No.         | 2.51    |

Linear equation coefficients.  
 $y = 0.7658x_1 + 1.1252x_2$ .

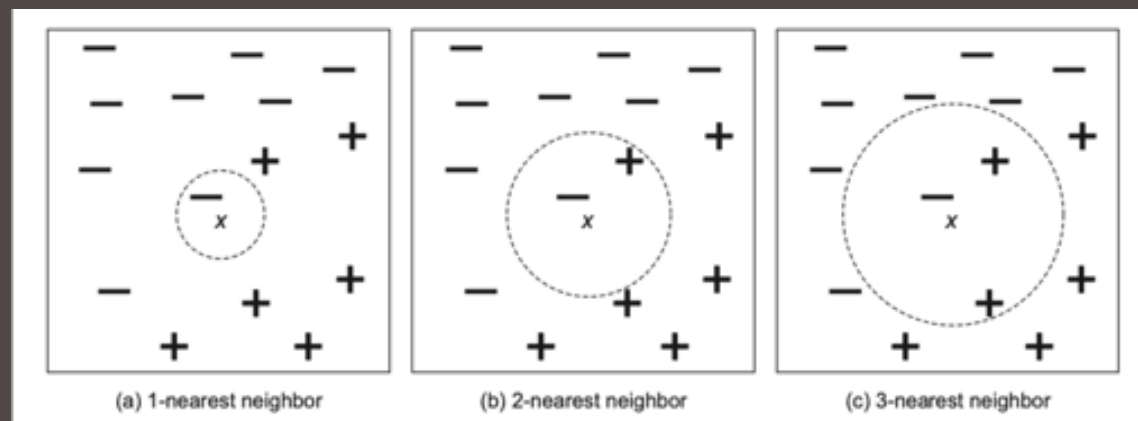
Figure 2.23 Linear regression model information output

Linear regression model  
information output

## 6.2 MODEL EXECUTION

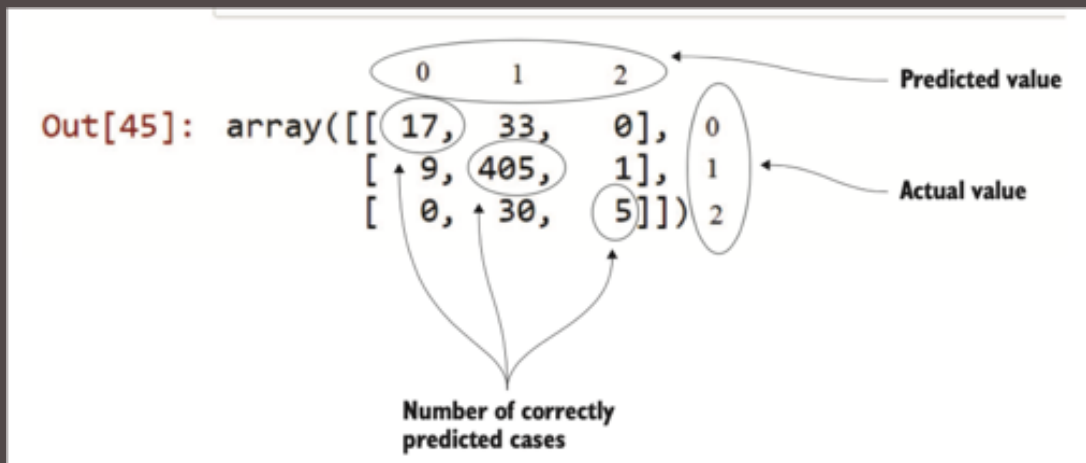
Linear regression works if we want to predict a value, but what if we want to classify something? Then we go to classification models, one of classification techniques is k-nearest neighbors.

K-nearest neighbors looks at labeled points nearby an unlabeled point and, based on this, makes a prediction of what the label should be.



K-nearest neighbor techniques look at the k-nearest point to make a prediction.

## 6.2 MODEL EXECUTION



- ❖ Classification output:
- ❖ Confusion matrix: it shows how many cases were correctly classified and incorrectly classified by comparing the prediction with the real values.

\*\* Remark: the classes (0,1,2) were added in the figure for clarification.

## 6.3 MODEL DIAGNOSTICS AND MODEL COMPARISON

- ❖ One there are multiple models have been generated from which you then choose the best one based on multiple criteria.
- ❖ The model should work on unseen data.
- ❖ You use only a fraction of your data to estimate the model and the other part, the holdout sample, is kept out of the equation.
- ❖ The model is then unleashed on the unseen data and error measures are calculated to evaluate it.
- ❖ One of the error measure is the mean square error.
- ❖ Mean square error is a simple measure: check for every prediction how far it was from the truth, square this error, and add up the error of every prediction.

## 6.3 MODEL DIAGNOSTICS AND MODEL COMPARISON

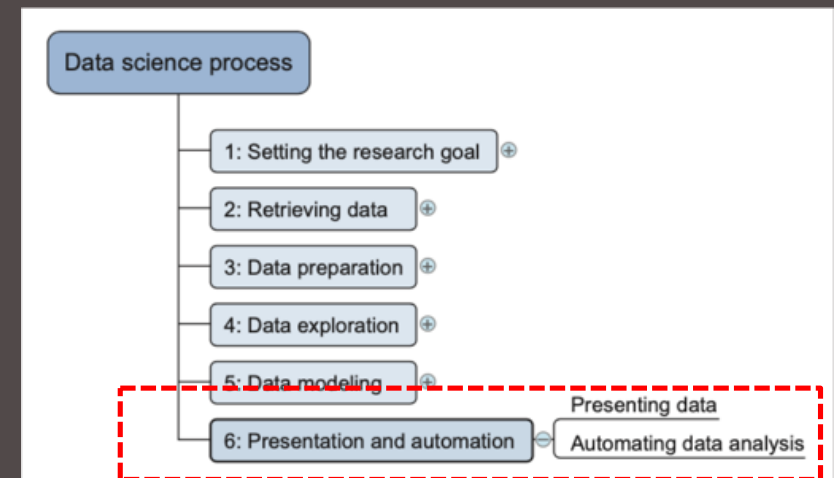
|           | <i>n</i> | Size | Price | Predicted<br>model 1 | Predicted<br>model 2 | Error<br>model 1 | Error<br>model 2 |
|-----------|----------|------|-------|----------------------|----------------------|------------------|------------------|
| 80% train | 1        | 10   | 3     |                      |                      |                  |                  |
|           | 2        | 15   | 5     |                      |                      |                  |                  |
|           | 3        | 18   | 6     |                      |                      |                  |                  |
|           | 4        | 14   | 5     |                      |                      |                  |                  |
|           | ...      | ...  |       |                      |                      |                  |                  |
|           | 800      | 9    | 3     |                      |                      |                  |                  |
|           | 801      | 12   | 4     | 12                   | 10                   | 0                | 2                |
|           | 802      | 13   | 4     | 12                   | 10                   | 1                | 3                |
| 20% test  | ...      |      |       |                      |                      |                  |                  |
|           | 999      | 21   | 7     | 21                   | 10                   | 0                | 11               |
|           | 1000     | 10   | 4     | 12                   | 10                   | -2               | 0                |
| Total     |          |      |       |                      |                      | 5861             | 110225           |

A holdout sample helps us compare models and ensures that we can generalize results to data that the model has not yet seen.



# 7. PRESENTATION AND AUTOMATION

- ❖ Once we successfully analyzed the data and built a well-performing model, it's time to present the findings.
- ❖ This is an exciting part because we have to explain what we found to the stakeholders.



## 7. PRESENTATION AND AUTOMATION

- ❖ Sometimes people get so excited about your work that we need to repeat it over and over again because they value the predictions of the models or the insights that we produced.
- ❖ For this reason, we need to **automate** the models.
- ❖ This doesn't always mean that we have to redo all of the analysis all the time.
- ❖ Sometimes it's sufficient that we implement only the model scoring; other times we might build an application that automatically updates reports, Excel spreadsheets, or PowerPoint presentations.
- ❖ The last stage of the data science process is where the soft skills will be most useful, and extremely important.