Lab 0
Introduction to Python

Data Mining for Computer and Systems Sciences (DAMI)

Luis Quintero

luis-eduardo@dsv.su.se

2023-08-30

DataScienceGroup

datascience.dsv.su.se

Stockholm University

Photo by Pietro Jeng on Unsplash

ILOs

I. Assess the **quality of a structured dataset**
II. Formulate **analytical questions** that can be solved with descriptive and exploratory methods
III. Recognize the types of questions that can be answered with **predictive methods**
IV. Configure **Python** to design and implement a data science project

DataScienceGroup
datascience.dsv.su.se
Stockholm University

1. Structured high-quality datasets

2. Data Mining and Predictive Analytics

3. Python

DataScienceGroup
datascience.dsv.su.se

Stockholm University

Photo by Iva Muškić from Pexels

Study case I
Bank Marketing

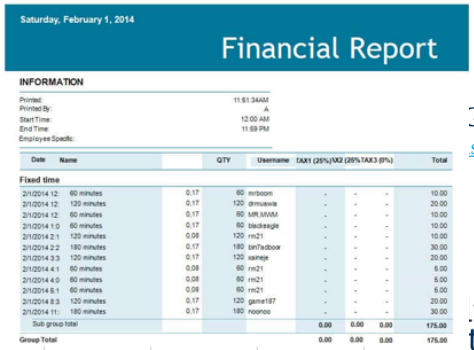
DataScienceGroup
datascience.dsv.su.se

Photo from Pexels

Bank-v1.csv

From: [Bank Marketing - UCI Machine Learning Repository]

	A	B	C	D	E	F	G	H	I	J	K	L
1	30	unemployed	married	primary	no	1787	no	no				
2	33	services	married	secondary	no	4789 euros	yes	yes	this person is the cousin of the owner			
3		management	single	tertiary	no	1350	yes	no				
4	30	management	married	tertiary	yes	1476	yes	yes		the person still owes money		
5	59	blue-collar	married	secondary	yes	0	yes	no				
6	35	management	single	tertiary		747	no	no				
7												
8	36	self-employed	married	tertiary		307	yes	no				
9	39	technician	married	secondary		147	yes	no				
10	41	entrepreneur	married	tertiary	no	221	yes	no				
11	43	services	married	primary	no	-88	yes	yes				
12	39	services		secondary	no	9374	yes	no				
13	43	admin.	married	secondary	no	264	yes	no				
14		technician		tertiary	no	1109	no	no				
15	20	student	single	secondary	no	502	no	no				
16	31	blue-collar	married	secondary	no	"€360"	yes	yes				
17	40	management	married	tertiary	yes	194	no	yes				
18	56	technician	married	secondary	yes	4073	no	no				
19	37	admin.	single	tertiary	no	2317	yes	no				
20	25	blue-collar	single	primary	no	-221	yes	no				
21	25	blue-collar	single	primary	no	-221	yes	no				



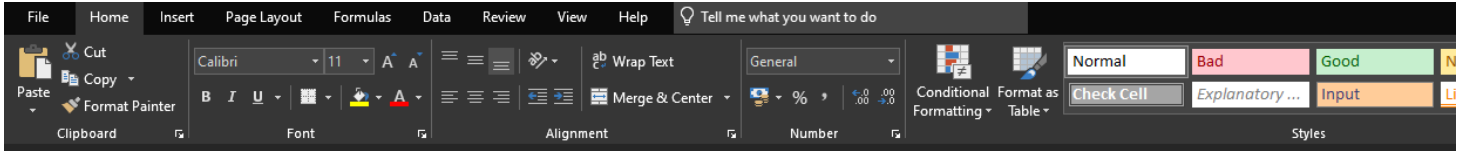
Task: Identify possible problems in the previous dataset

For example:

- Lack of **column names** with descriptions
- Some cells have **missing values**
- The **format** in the same column is inconsistent
- **Empty rows**
- **Duplicates** in the last two rows
- **Unstructured data** in random cells (imgs)

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Bank-v2.csv



	A	B	C	D	E	F	G	H
1	age	job	marital	education	default	balance	housing	loan
2	30	unemployed	married	primary	no	1787	no	no
3	33	services	married	secondary	no	4789	yes	yes
4	35	management	single	tertiary	no	1350	yes	no
5	30	management	married	tertiary	no	1476	yes	yes
6	59	blue-collar	married	secondary	no	0	yes	no
7	35	management	single	tertiary	no	747	no	no
8	36	self-employed	married	tertiary	no	307	yes	no
9	39	technician	married	secondary	no	147	yes	no
10	41	entrepreneur	married	tertiary	no	221	yes	no
11	43	services	married	primary	no	-88	yes	yes
12	39	services	married	secondary	no	9374	yes	no
13	43	admin.	married	secondary	no	264	yes	no
14	36	technician	married	tertiary	no	1109	no	no
15	20	student	single	secondary	no	502	no	no
16	31	blue-collar	married	secondary	no	360	yes	yes
17	40	management	married	tertiary	no	194	no	yes
18	56	technician	married	secondary	no	4073	no	no
19	37	admin.	single	tertiary	no	2317	yes	no
20	25	blue-collar	single	primary	no	-221	yes	no

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Structured dataset

Row, observation, sample, registry, item, subject

Columns, variables, attributes, features

age	job	marital	education	default	balance	housing	loan
30	unemployed	married	primary	no	1787	no	no
33	services	married	secondary	no	4789	yes	yes
35	management	single	tertiary	no	1350	yes	no
30	management	married	tertiary	no	1476	yes	yes
59	blue-collar	married	secondary	no	0	yes	no
35	management	single	tertiary	no	747	no	no
36	self-employed	married	tertiary	no	307	yes	no

Cell, field, value

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Bank-v2-metadata.csv

From Wikipedia: **Metadata** (or **metainformation**) is "data that provides information about other data",[1] but not the content of the data, such as the text of a message or the image itself.[2]

#	Column name	Description	Data Type	Details
1	age	age	numeric	
2	job	type of job	categorical	"admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services"
3	marital	marital status	categorical	"married", "divorced", "single"; note: "divorced" means divorced or widowed
4	education	education level	categorical	"unknown", "secondary", "primary", "tertiary"
5	default	has credit in default?	binary	"yes", "no"
6	balance	average yearly balance	numeric	in euros
7	housing	has housing loan?	binary	"yes", "no"
8	loan	has personal loan?	binary	"yes", "no"

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Types of variables / Data types

Numerical

- **Discrete** – *Only certain values are possible*
  - User ID, Age, # children, Year, Frequencies, …
- **Continuous** – *Any value within a range*
  - Height, balance $, T°, joystick values, time [ms, s]…

Categorical

- **Nominal** – *Labels without quantitative meaning*
  - Non-binary: Gender, Country, Role, Marital status, Team, …
  - **Binary**: Is Smoker? Is Active? Is Subscribed?
- **Ordinal** – *Labels with relative order/rank*
  - Education Level, Age Group, Likert Scales, any other scale…
  - **Binary**: Bad/Good? Low/High?, …

Which data type is a variable **PhoneNumber**?: 071 123 4567
The fact that it contains numbers **does not mean** it is numerical!

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Data quality

"If it cannot be guaranteed that the data is completely clean, the subsequent analysis lacks any scientific rigor and its ability to be useful for the purpose we seek."

Xavi Font - *Business Intelligence y Business Analytics (2019)*

DataScienceGroup
datascience.dsv.su.se
Stockholm University

From: Six dimensions of data quality (sbctc.edu)

6 dimensions of data quality

Accuracy
Do the data reflect reality?
E.g., age containing negative numbers

Timeliness
Are the data available when needed?
E.g., the end-of-the-month report has access to all the required metrics

Validity
Does the format follow the specific business rules?
E.g., define a date as "23/08/2023" or "23-08-23"

Completeness
Are the data complete?
E.g., the metadata describing the data is missing, or definition of the optional fields

Consistency
Do the data from different sources follow the same structure?
E.g., databases from different campuses at SU

Uniqueness
Does every instance appear only once in the dataset?
E.g., perhaps the person John William Smith and John W. Smith are the same subject.

Task: Think of 2 questions that can be answered from the structured dataframe and the process to answer them

E.g., What is the average balance of married people?
(filter marital, then mean on balance)

	A	B	C	D	E	F	G	H
1	age	job	marital	education	default	balance	housing	loan
2	30	unemployed	married	primary	no	1787	no	no
3	33	services	married	secondary	no	4789	yes	yes
4	35	management	single	tertiary	no	1350	yes	no
5	30	management	married	tertiary	no	1476	yes	yes
6	59	blue-collar	married	secondary	no	0	yes	no
7	35	management	single	tertiary	no	747	no	no
8	36	self-employed	married	tertiary	no	307	yes	no
9	39	technician	married	secondary	no	147	yes	no
10	41	entrepreneur	married	tertiary	no	221	yes	no
11	43	services	married	primary	no	-88	yes	yes
12	39	services	married	secondary	no	9374	yes	no
13	43	admin.	married	secondary	no	264	yes	no
14	36	technician	married	tertiary	no	1109	no	no
15	20	student	single	secondary	no	502	no	no
16	31	blue-collar	married	secondary	no	360	yes	yes
17	40	management	married	tertiary	no	194	no	yes
18	56	technician	married	secondary	no	4073	no	no
19	37	admin.	single	tertiary	no	2317	yes	no
20	25	blue-collar	single	primary	no	-221	yes	no

	Colname	Description	Data Type
1	age	age	numeric
2	job	type of job	categorical
3	marital	marital status	categorical
4	education	education level	categorical
5	default	has credit in default?	binary
6	balance	average yearly balance	numeric
7	housing	has housing loan?	binary
8	loan	has personal loan?	binary

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Univariate (*1 variable or column*)

- **How many managers** are in the customers list? *(filter and count on the column `job`)*
- What is the **range** of the **balance** in the bank accounts of the customers? *(min and max on the column `balance`)*
- What is the **proportion** of customers with personal **loan**? *(count customers where `loan` = "yes" / # customers)*

Bivariate (*2 variables or columns*)

- What is the **proportion** of **students** that are **married**? *(calculation on the columns `job` and `marital status`)*

Multivariate (*>2 variables or columns*)

- **Which students** under **25** y.o. have more than **1000** euros? (*job, age, balance)*
- What is the **average balance** of **married** people with minimum **secondary** education? (*pivot table on the involved columns and aggregation function)*

DataScienceGroup
datascience.dsv.su.se
Stockholm University

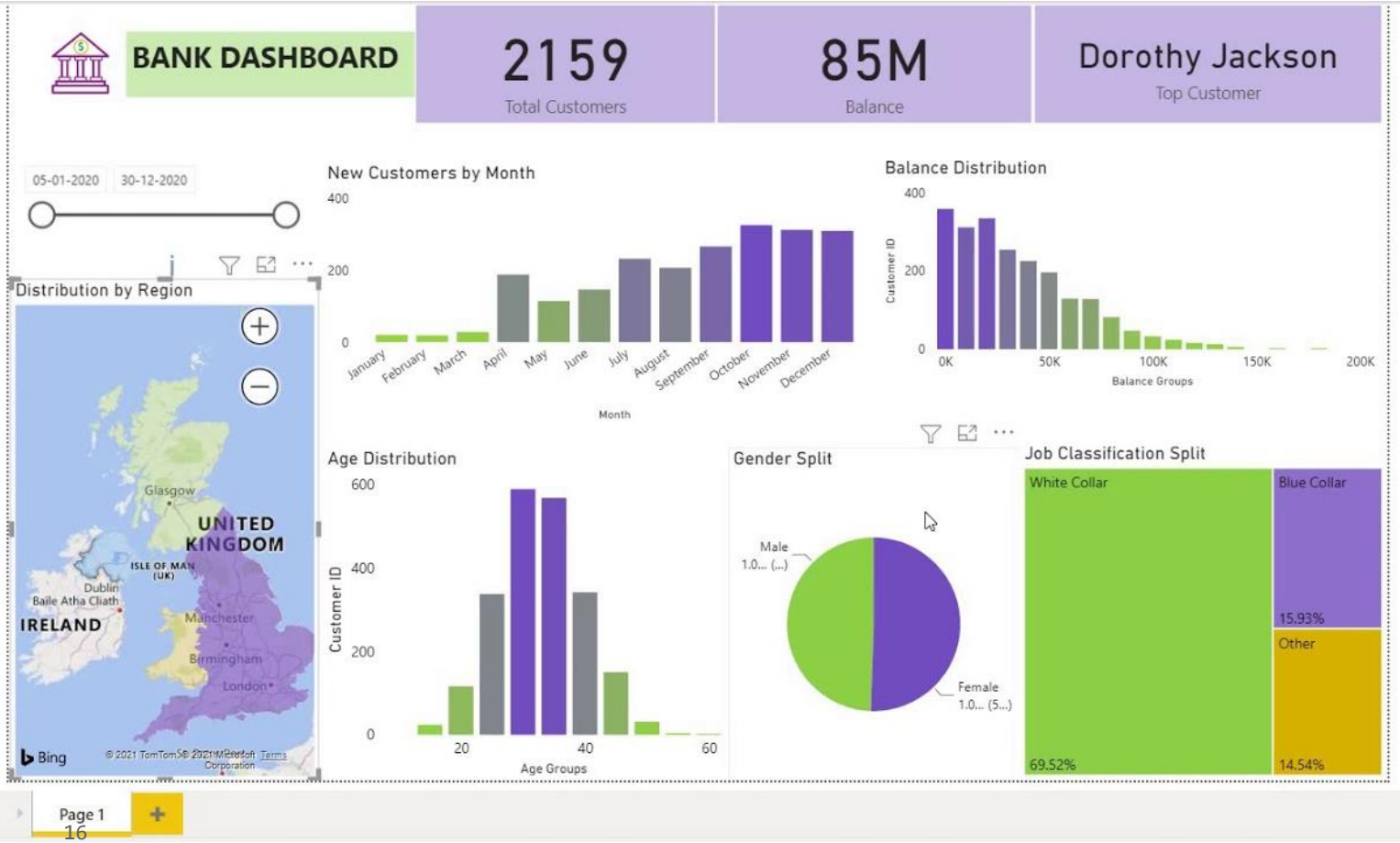
Formulating analytical questions

Every analytical question leads to a certain **analytical approach** and specific involved **variables/features/columns**.

The value of the data can be discovered asking the right questions!
More **specific** questions lead to more valuable and processable answers.

DataScienceGroup
datascience.dsv.su.se
Stockholm University

From: https://www.youtube.com/watch?v=gJhWWeJTIww

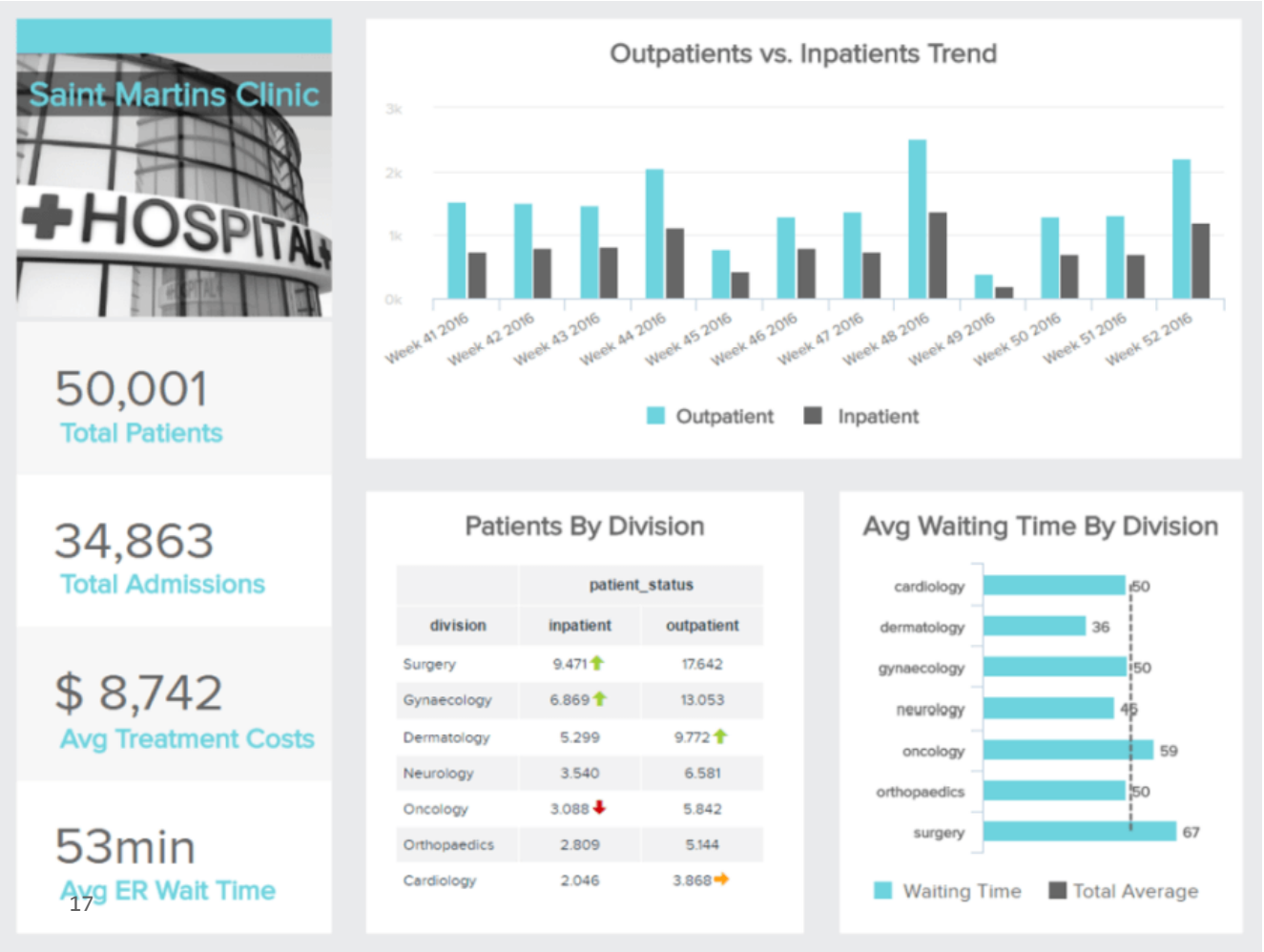


Which questions can be answered from this dashboard?

DataScienceGroup
datascience.dsv.su.se

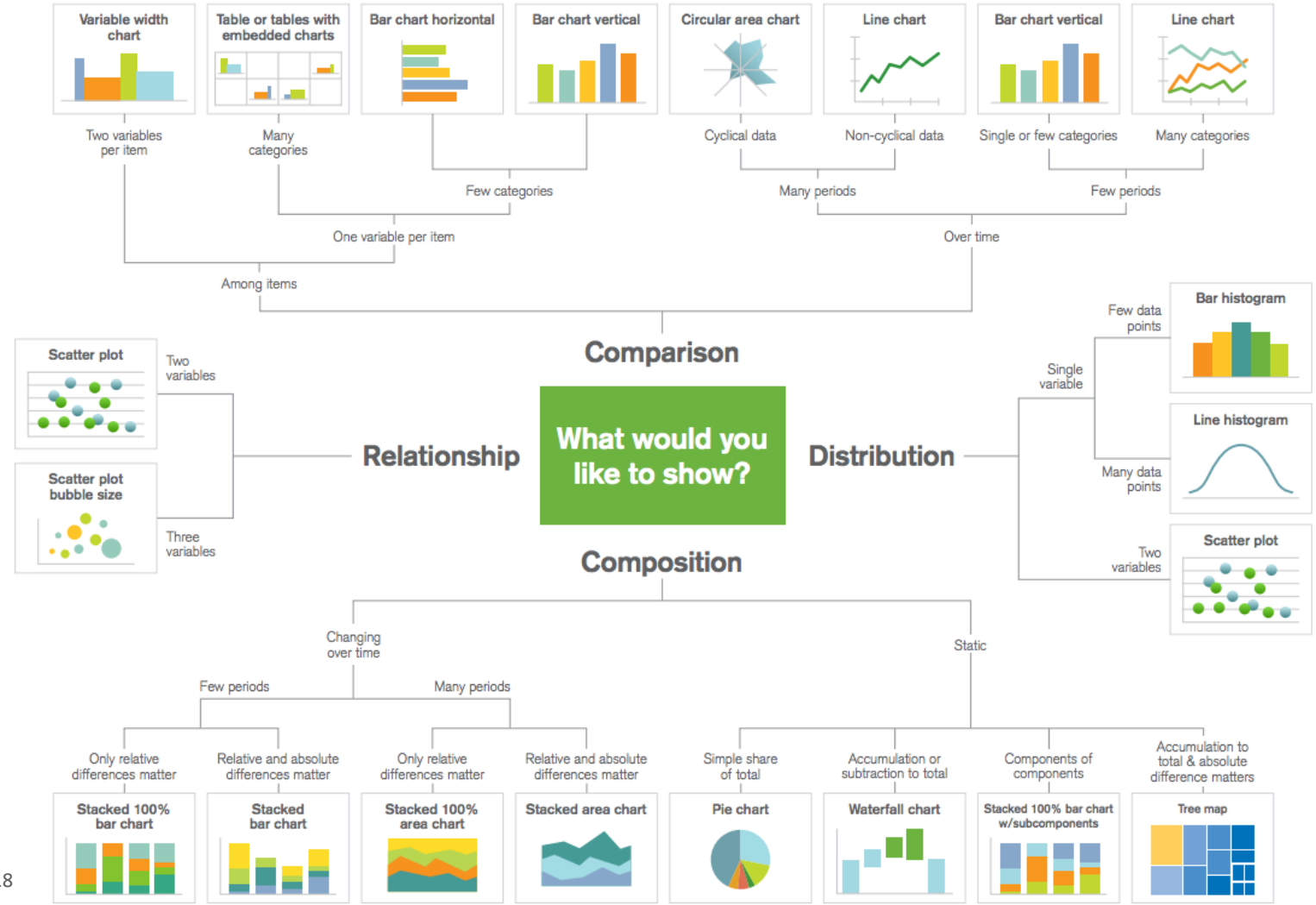
Stockholm University

From: What Is A Data Dashboard? Definition, Meaning & Examples (datapine.com)



Which questions can be answered from this dashboard?

DataScienceGroup
datascience.dsv.su.se
Stockholm University



DataScienceGroup
datascience.dsv.su.se
Stockholm University

1. Structured high-quality datasets

2.Data Mining and Predictive Analytics

3. Python

DataScienceGroup
datascience.dsv.su.se

Stockholm University

Photo by Iva Muškić from Pexels

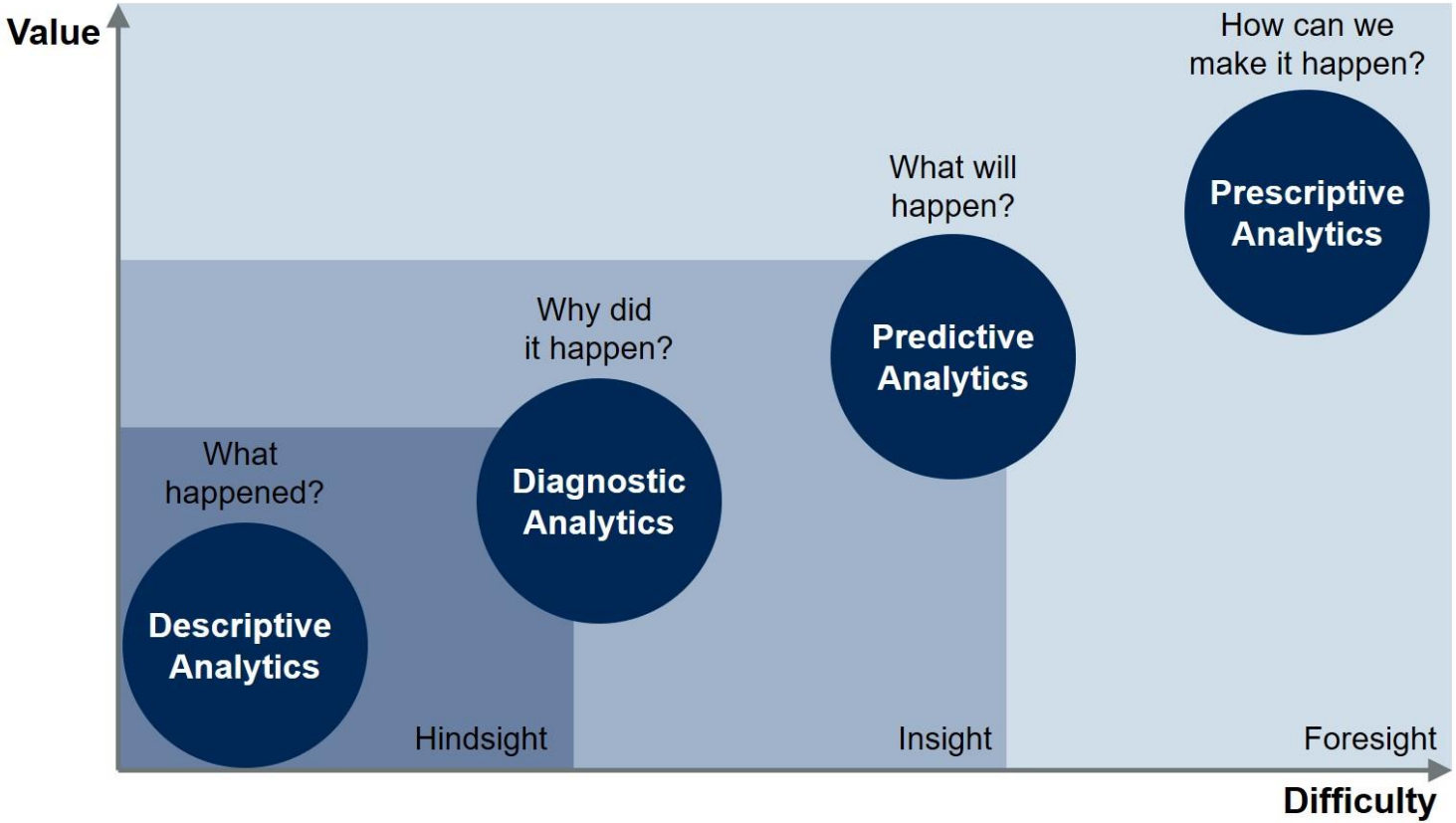
- ✓ Structured data
- ✓ Descriptive metadata
- ✓ Clean data
- ✓ Well-defined questions
- ✓ Descriptive analytical approaches

How is data mining and machine learning different from data analytics in Excel?

DataScienceGroup
datascience.dsv.su.se
Stockholm University

From: blogs.gartner.com

Four types of data analytics



DataScienceGroup
datascience.dsv.su.se
Stockholm University

Exploratory and Descriptive Analytics

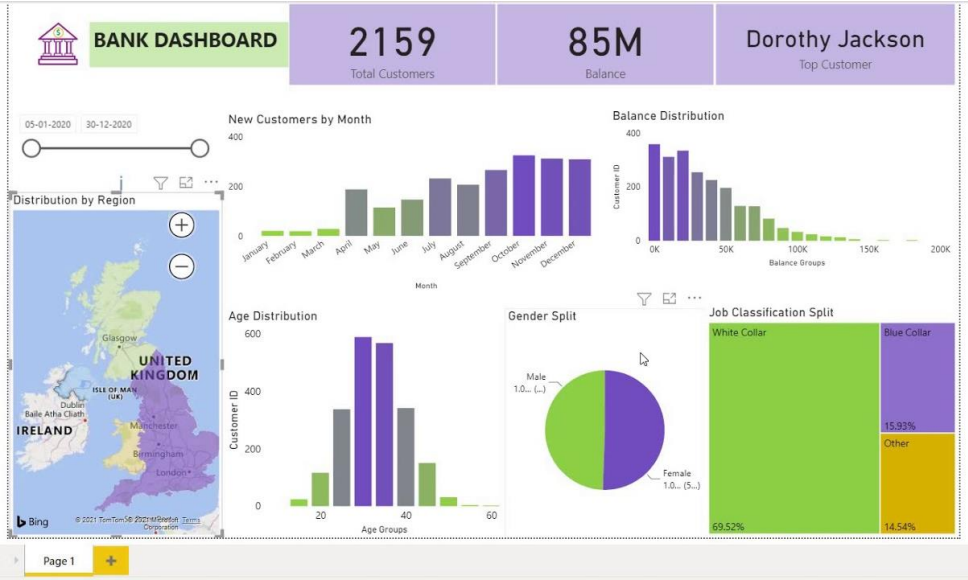
It is introductory, retrospective and answers to the question: **What happened?**

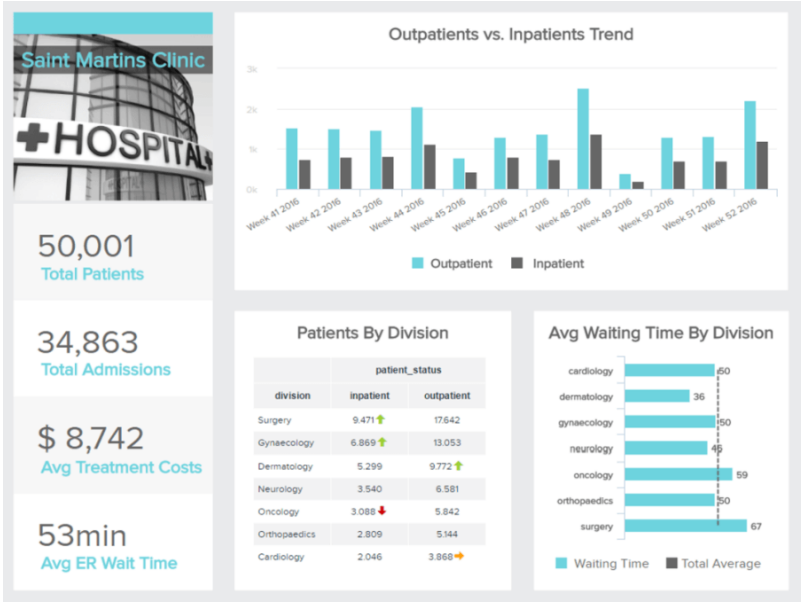
Most data analyses in a business are descriptive, which makes it the most common type.

EDAs are usually visualized in simple reports, control panels, or KPI dashboards in software visualization frameworks.

DataScienceGroup
datascience.dsv.su.se
Stockholm University

EDAs are usually enough to answer simple questions

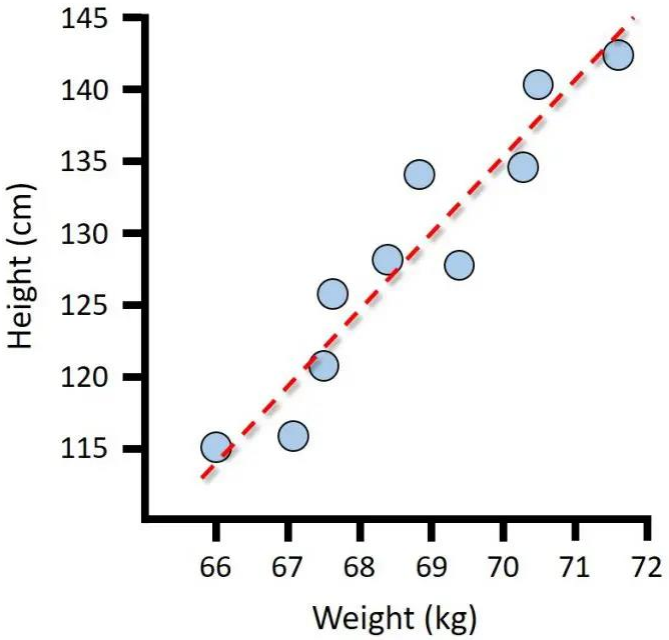




DataScienceGroup
datascience.dsv.su.se

Stockholm University

Diagnostic analytics



Answers to the question:
Why did something happened?

Some approaches:

- Correlation analysis
- Inferential statistics
- Clustering

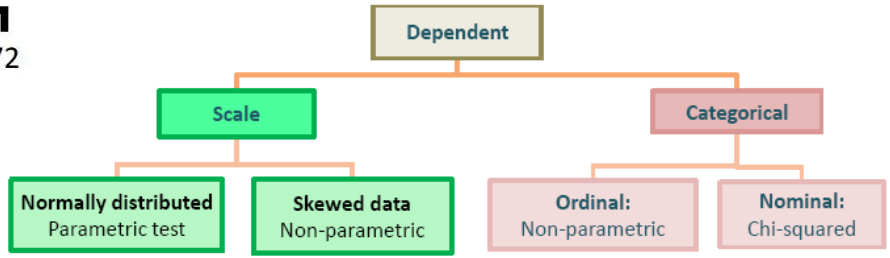


Image from: tutorsquickguidetostatistics.pdf (statstutor.ac.uk)

DataScienceGroup
datascience.dsv.su.se
Stockholm University

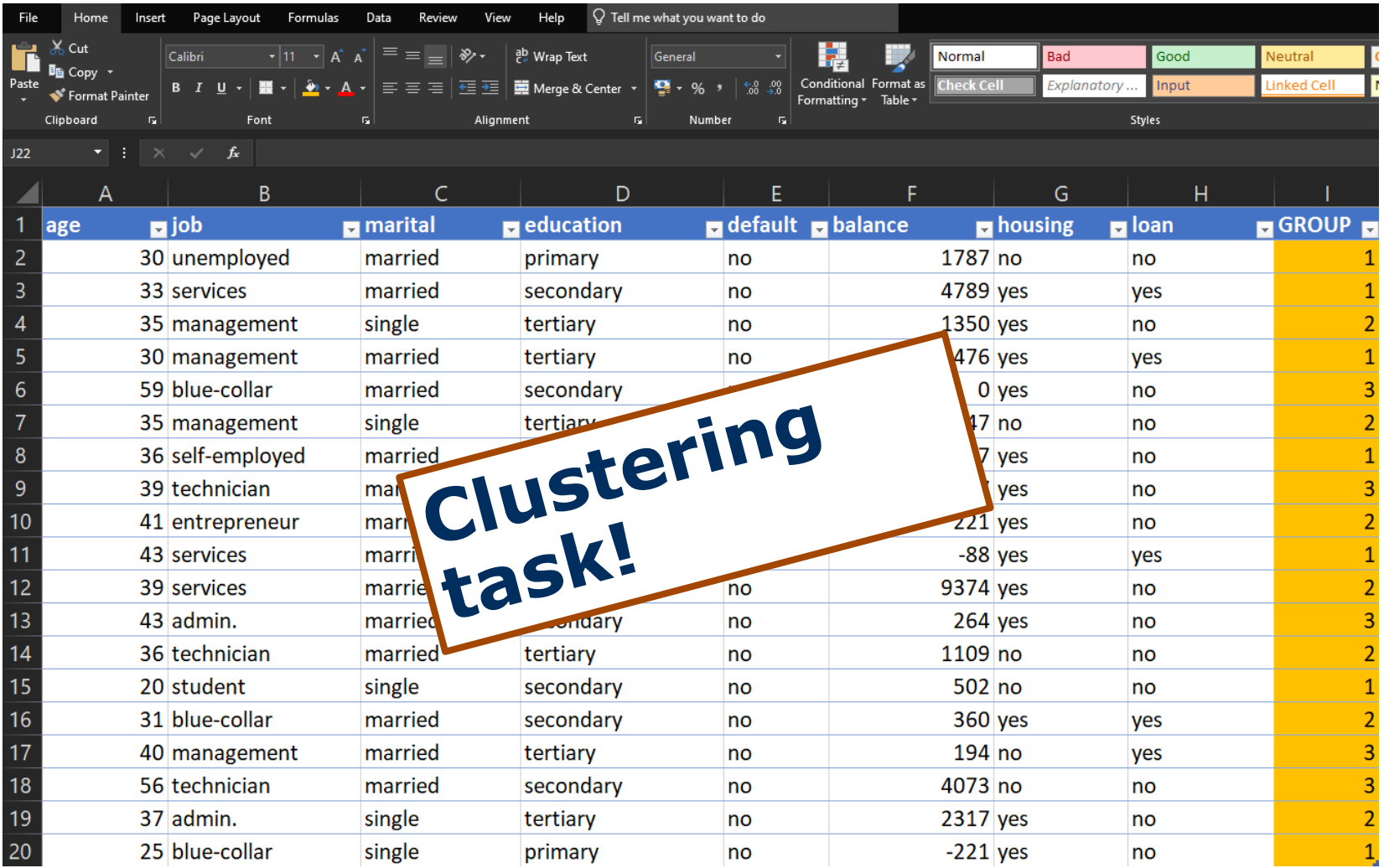
But you can ask more complicated questions from the data…

Task: How many groups of similar customers can be created from the bank dataset?

Image from: Prasad Patil | Towards Data Science

How do you group them? Age, job, or balance,… ?

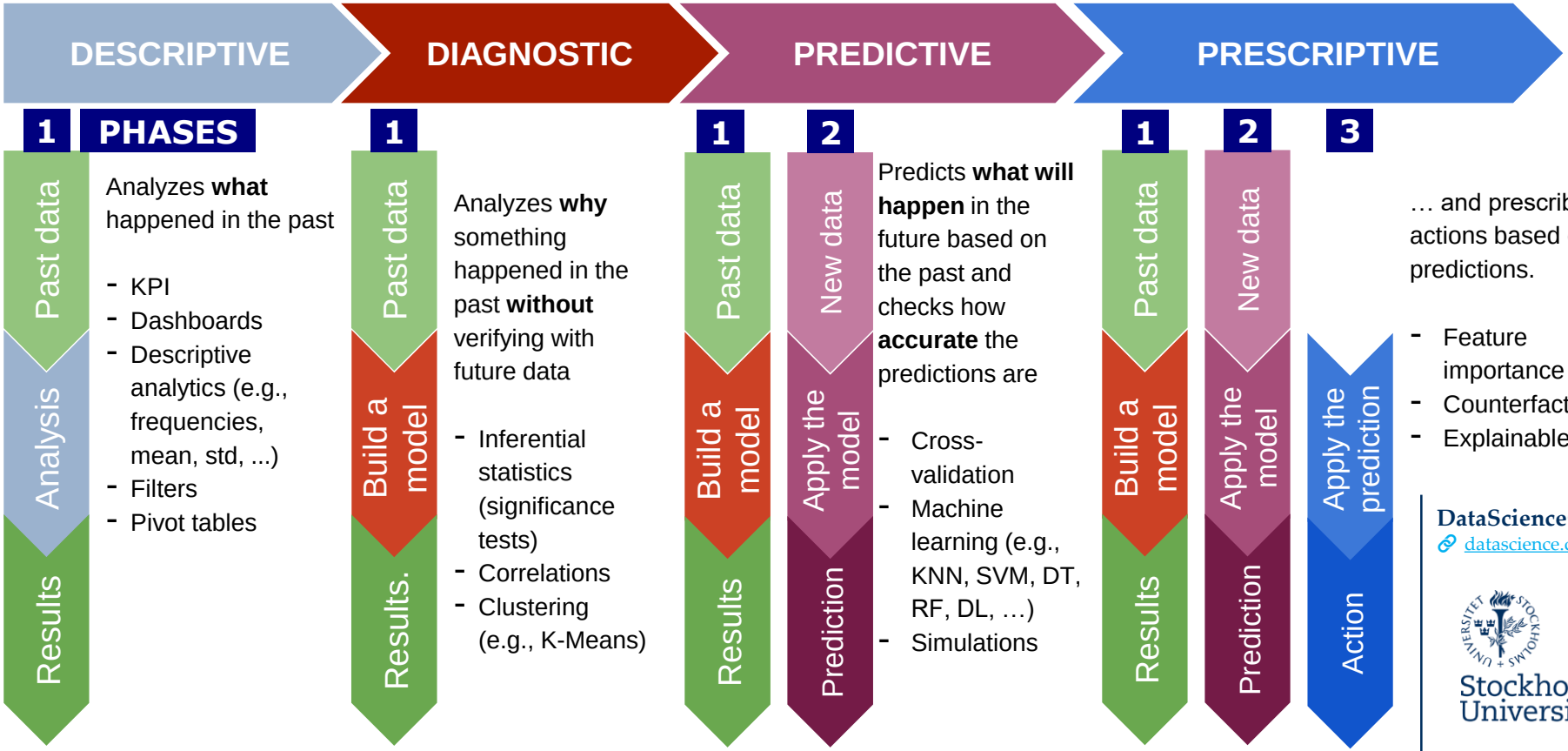
DataScienceGroup
datascience.dsv.su.se
Stockholm University



	A	B	C	D	E	F	G	H	I
1	age	job	marital	education	default	balance	housing	loan	GROUP
2	30	unemployed	married	primary	no	1787	no	no	1
3	33	services	married	secondary	no	4789	yes	yes	1
4	35	management	single	tertiary	no	1350	yes	no	2
5	30	management	married	tertiary	no	476	yes	yes	1
6	59	blue-collar	married	secondary		0	yes	no	3
7	35	management	single	tertiary		47	no	no	2
8	36	self-employed	married			7	yes	no	1
9	39	technician	mar				yes	no	3
10	41	entrepreneur	marr			221	yes	no	2
11	43	services	marri			-88	yes	yes	1
12	39	services	marrie		no	9374	yes	no	2
13	43	admin.	married	ondary	no	264	yes	no	3
14	36	technician	married	tertiary	no	1109	no	no	2
15	20	student	single	secondary	no	502	no	no	1
16	31	blue-collar	married	secondary	no	360	yes	yes	2
17	40	management	married	tertiary	no	194	no	yes	3
18	56	technician	married	secondary	no	4073	no	no	3
19	37	admin.	single	tertiary	no	2317	yes	no	2
20	25	blue-collar	single	primary	no	-221	yes	no	1

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Types of data analytics



DataScienceGroup
datascience.dsv.su.se

Stockholm University

Study case II

Plant classification

DataScienceGroup
datascience.dsv.su.se

Stockholm University

Photo from Pexels

Non-structured datasets

Edgar Anderson went on an expedition to quantify the morphologic variations of the iris flowers of three species. The data was collected in the Gaspé Peninsula in southern Canada “all from the same pasture, picked up on the same day and measured at the same time by the same person with the same apparatus”.

Iris Versicolor

Iris Setosa

Iris Virginica

DataScienceGroup
datascience.dsv.su.se
Stockholm University

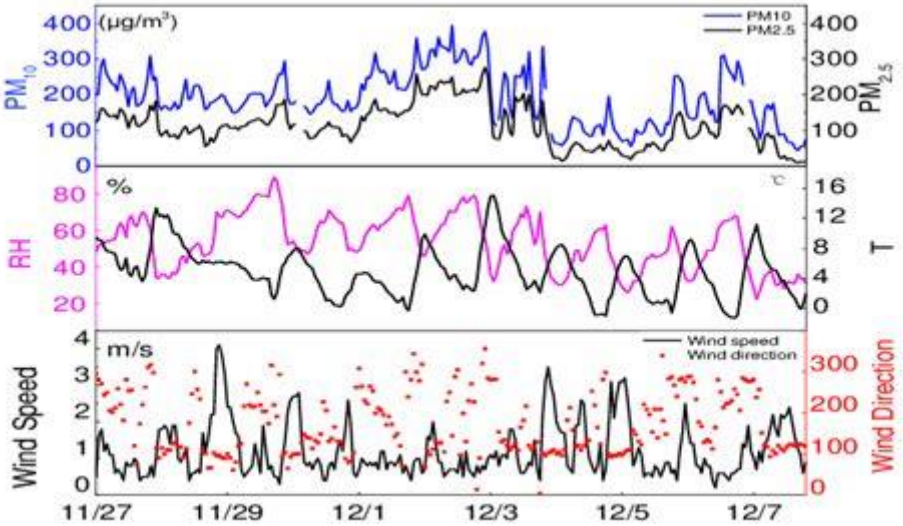
Data Modalities

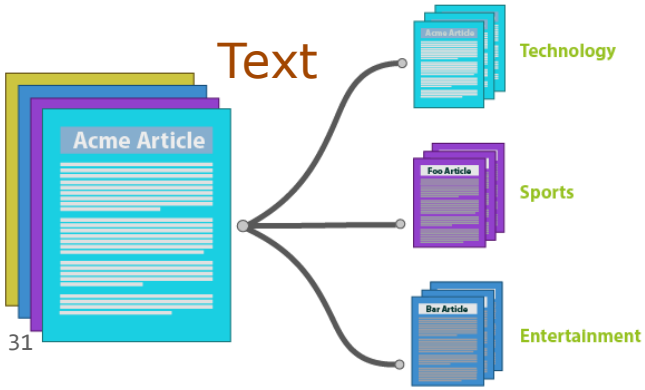
	age (n)	job (c)	marital (c)	education (c)	balance (n)	housing (c)
0	30	unemployed	married	primary	1787	no
1	33	services	married	secondary	4789	yes
2	35	management	single	tertiary	1350	yes
3	30	management	married	tertiary	1476	yes
4	59	blue-collar	married	secondary	0	yes
5	35	management	single	tertiary	747	no

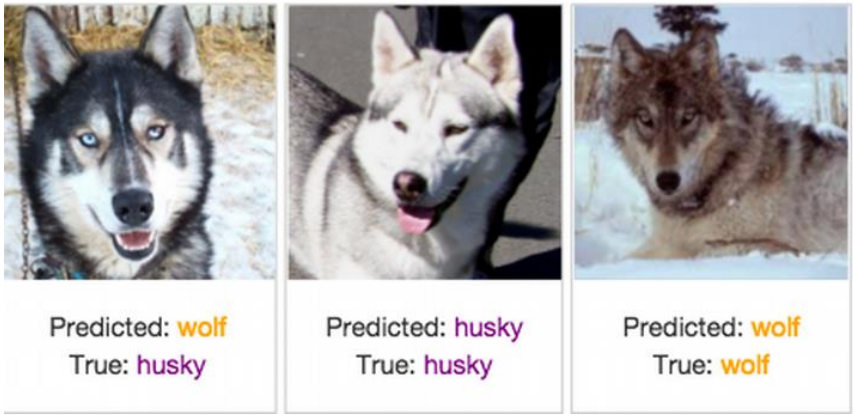
Tabular/Structured

(we will use this modality during the labs!)

Time series





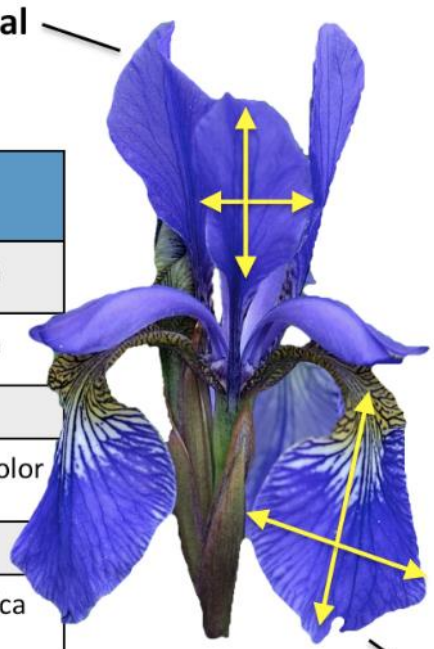


Images/Video

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



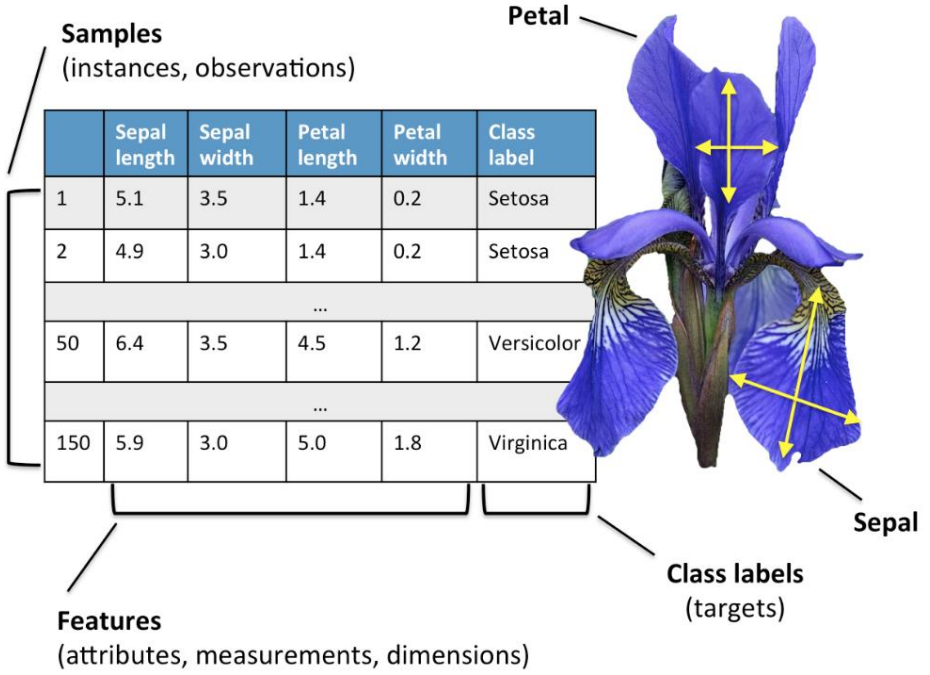
Features
(attributes, measurements, dimensions)

Class labels
(targets)

Image from: (sebastianraschka.com)

DataScienceGroup
datascience.dsv.su.se
Stockholm University

	A	B	C	D	E
1	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
52	7	3.2	4.7	1.4	versicolor
53	6.4	3.2	4.5	1.5	versicolor
54	6.9	3.1	4.9	1.5	versicolor
55	5.5	2.3	4	1.3	versicolor
56	6.5	2.8	4.6	1.5	versicolor
57	5.7	2.8	4.5	1.3	versicolor
58	6.3	3.3	4.7	1.6	versicolor
59	4.9	2.4	3.3	1	versicolor
60	6.6	2.9	4.6	1.3	versicolor
61	5.2	2.7	3.9	1.4	versicolor
102	6.3	3.3	6	2.5	virginica
103	5.8	2.7	5.1	1.9	virginica
104	7.1	3	5.9	2.1	virginica
105	6.3	2.9	5.6	1.8	virginica
106	6.5	3	5.8	2.2	virginica
107	7.6	3	6.6	2.1	virginica
108	4.9	2.5	4.5	1.7	virginica
109	7.3	2.9	6.3	1.8	virginica
110	6.7	2.5	5.8	1.8	virginica
111	7.2	3.6	6.1	2.5	virginica



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Task: Which questions can be asked from this dataset?

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Predictive Task

Based on the 150 samples collected previously in the iris dataset.

Create a model that allows us to determine whether a **new** iris flower is either:

- setosa
- virginica
- versicolor

	A	B	C	D	E
1	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
52	7	3.2	4.7	1.4	versicolor
53	6.4	3.2	4.5	1.5	versicolor
54	6.9	3.1	4.9	1.5	versicolor
55	5.5	2.3	4	1.3	versicolor
56	6.5	2.8	4.6	1.5	versicolor
102	6.3	3.3	6	2.5	virginica
103	5.8	2.7	5.1	1.9	virginica
104	7.1	3	5.9	2.1	virginica
105	6.3	2.9	5.6	1.8	virginica
151	5.9	3	5.1	1.8	virginica
152	5.2	3.1	3.5	1.4	?
153	6.5	3.5	1.6	1.3	?
154	5	3	4	0.5	?

new samples!!

DataScienceGroup
datascience.dsv.su.se

Stockholm University

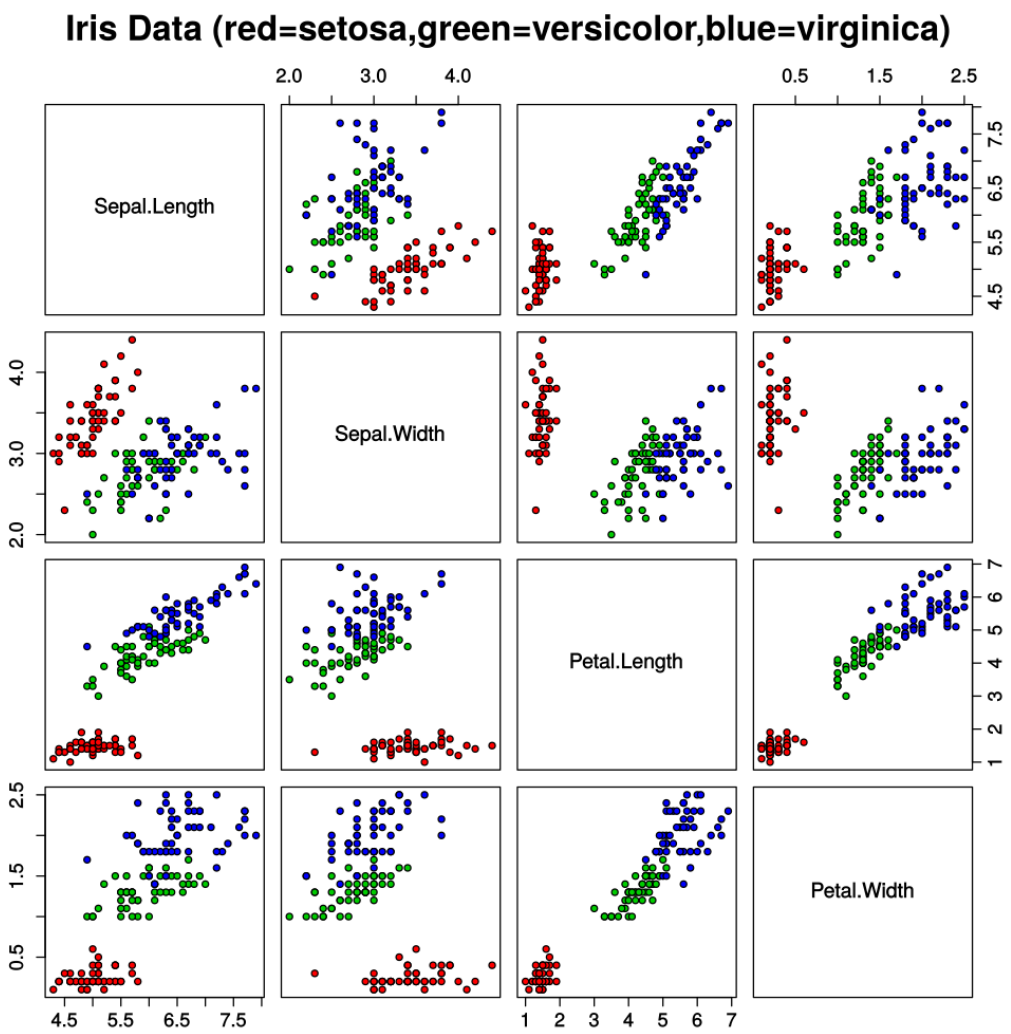


Image from: Iris dataset scatterplot - Iris flower data set - Wikipedia

A linear model

The dataset is so famous because it easily unveils predictive variables visually.

A pair plot let us see which pair of variables are more likely to separate the three species **accurately** with a **linear model**.

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Predictive analytics

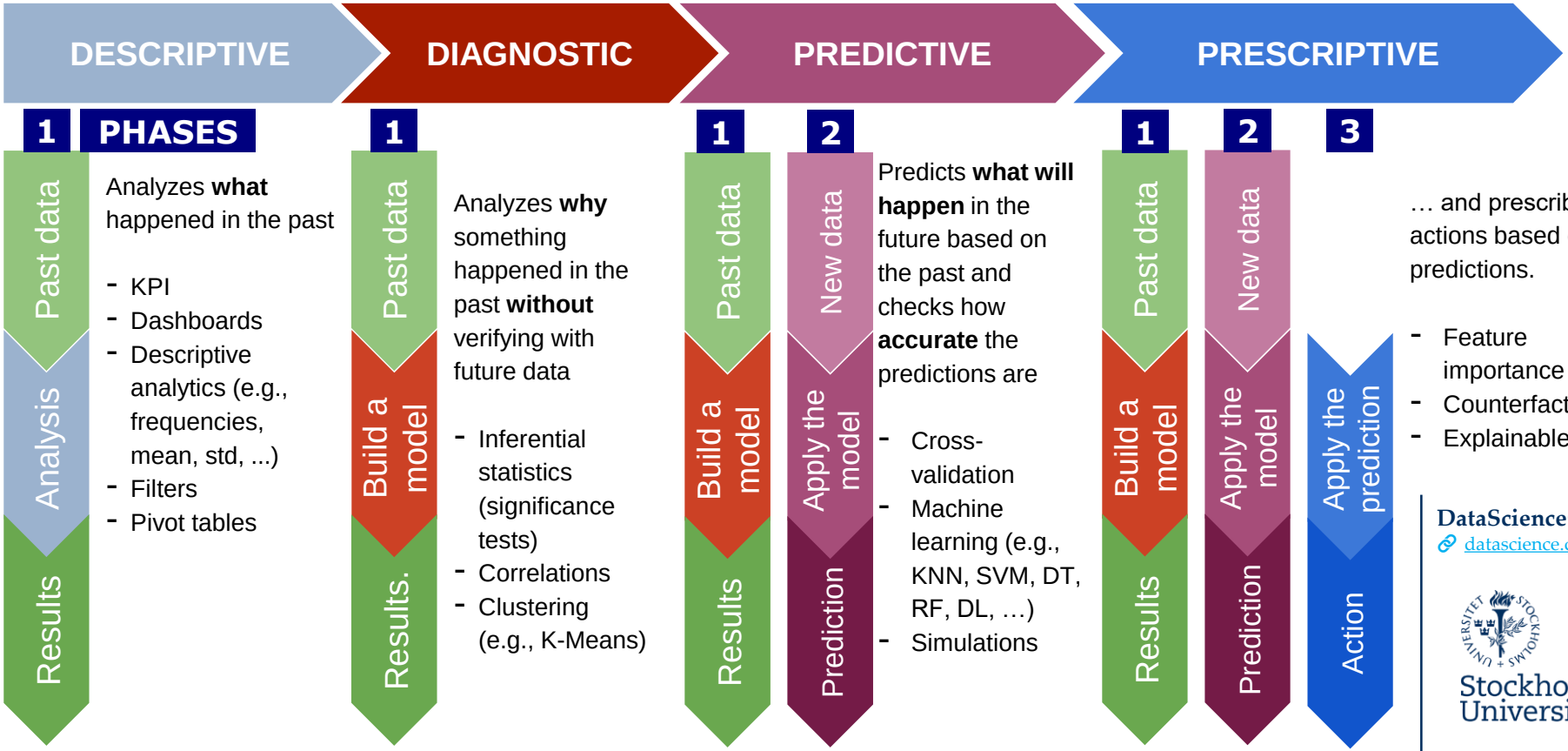
Generally applies statistical algorithms and **machine learning** techniques to answer questions related to **future values based on historical data**.

E.g.,

- K-nearest Neighbors (KNN)
- Decision Trees (DT)
- Random Forest (RF)
- Support Vector Machines (SVM)
- Deep Learning (DL)

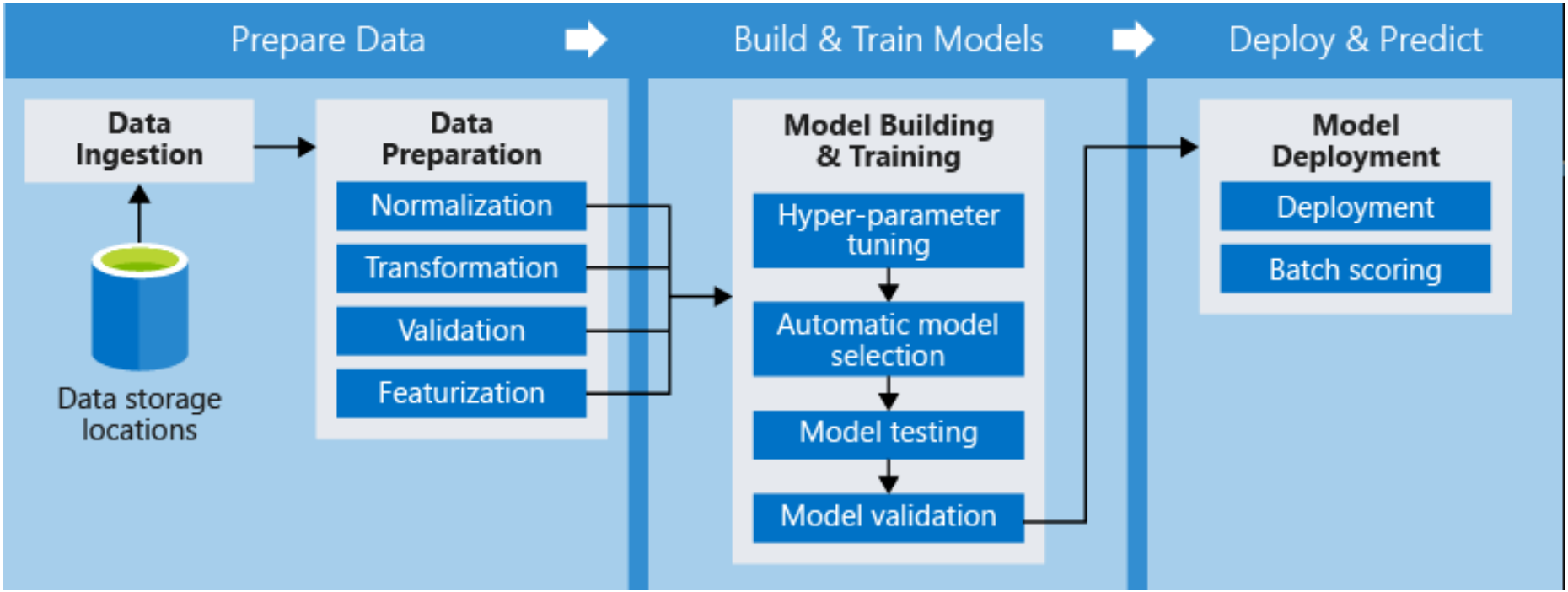
DataScienceGroup
datascience.dsv.su.se
Stockholm University

Types of data analytics



DataScienceGroup
datascience.dsv.su.se

Stockholm University



We will work with preprocessed datasets to understand how to prepare the data, build and fine tune models, and deploy them on applications.

DataScienceGroup
datascience.dsv.su.se
Stockholm University

During the rest of the labs

Understanding the **practicalities** of new analytical approaches that answer specific type of questions:

- Which are the frequent patterns and associations from a large dataset? (**rule mining**)
- Which groups of observations are similar to each other? How many groups?(**clustering**)
- What would be the value of a specific column in a new observation based on prior experience in the same problem? (**classification**)

DataScienceGroup
datascience.dsv.su.se
Stockholm University

1. Structured high-quality datasets
2. Data Mining and Predictive Analytics

3.Python

DataScienceGroup
datascience.dsv.su.se

Stockholm University

Photo by Iva Muškić from Pexels

All the files for the Lab session should be downloaded once.

Lab 0: How to start programming in Python?

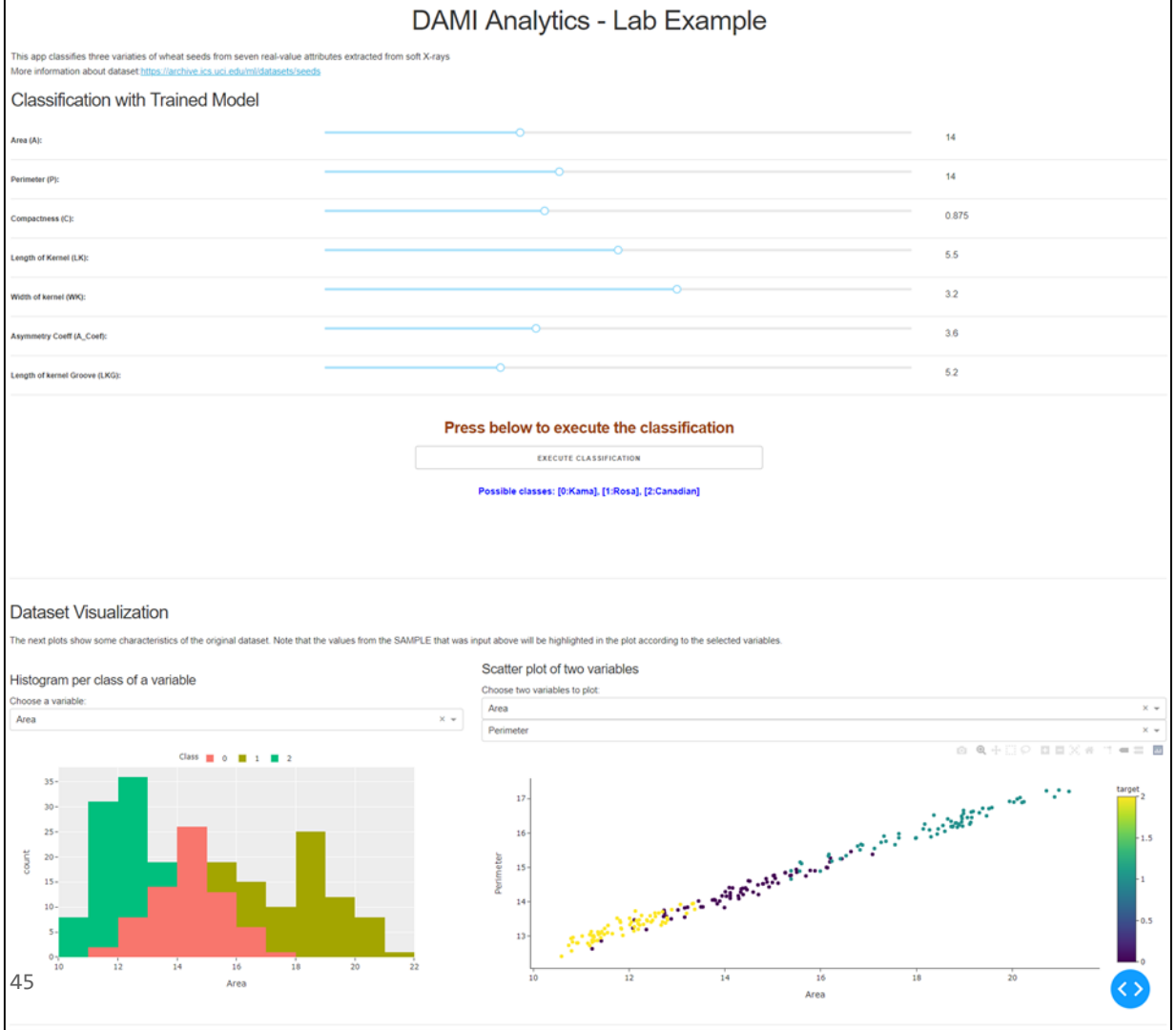
Lab 1: How to understand and process my original dataset to create a machine learning model?

Lab 2-3: Which type of models can I train based on the characteristics of my data?

Lab 4: Having multiple models, how can I evaluate which one might perform better on unseen data?

Lab 5: How can I reuse my trained model in a production-ready environment?

DataScienceGroup
datascience.dsv.su.se
Stockholm University



The interactive web platform allows manipulation of input variables and data visualization of the input compared to original dataset.

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Tools

Python programming language

Why not Excel?

It has limitations regarding:

- Manipulation of large datasets
- Reproducibility of analysis pipelines
- Not compatible to train machine learning models

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Python Ecosystem for Data Science

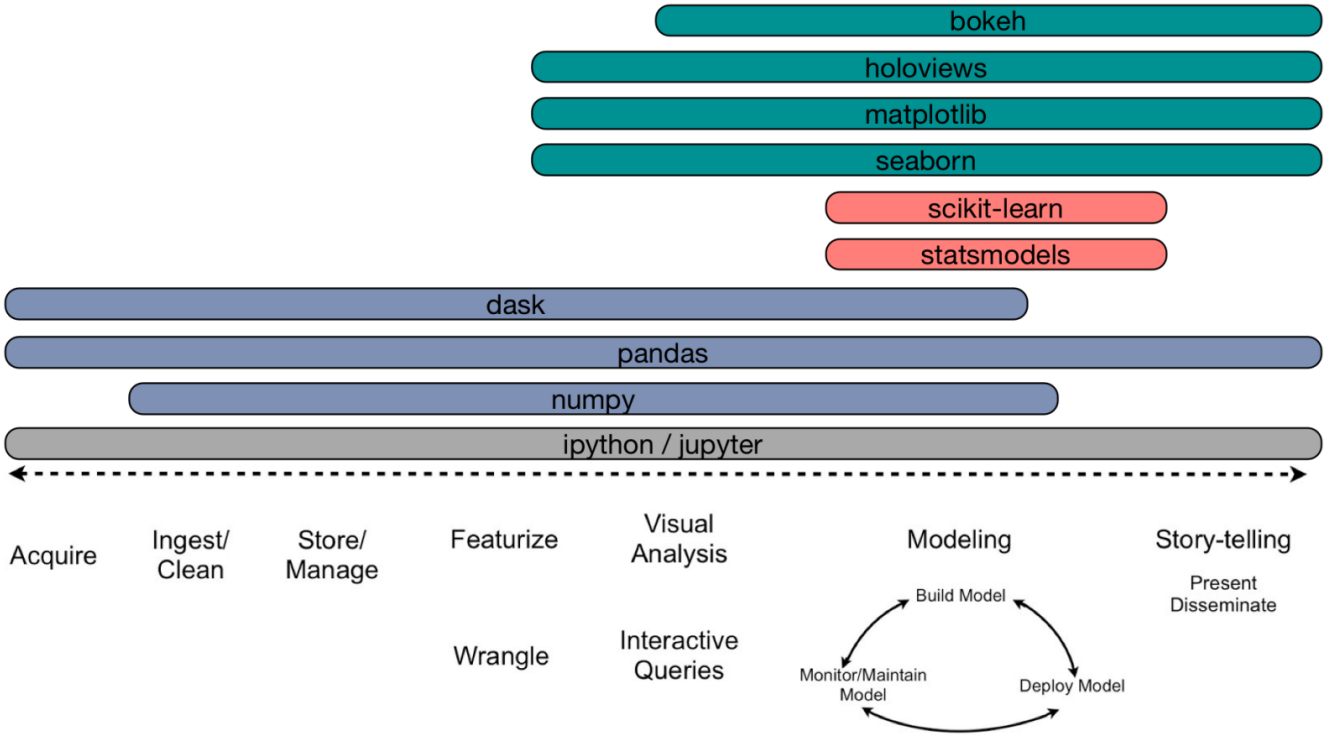
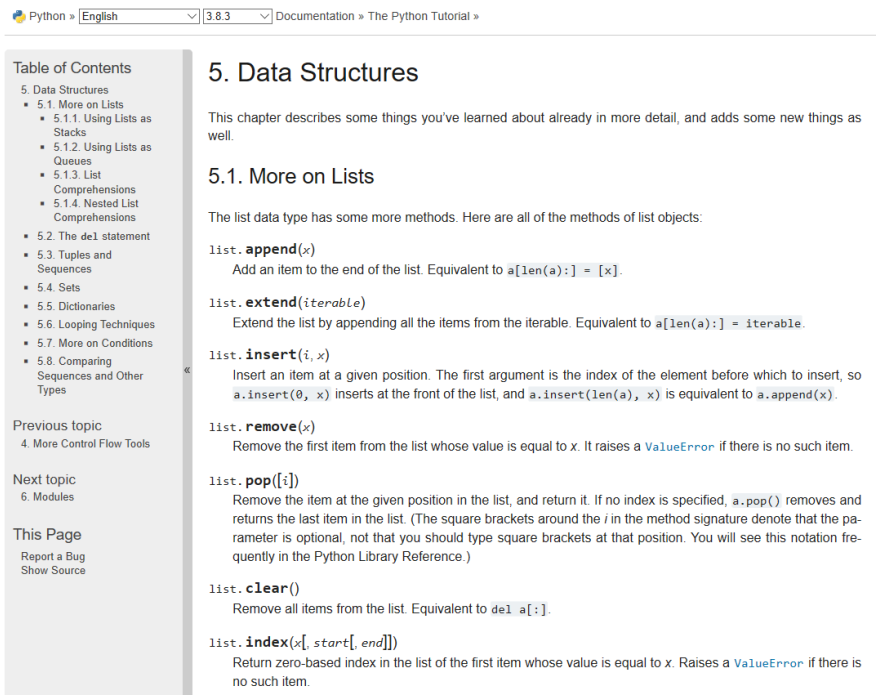


Image from Christian Staudt.

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Documentation



Python » English 3.8.3 Documentation » The Python Tutorial »

Table of Contents

5. Data Structures
- 5.1. More on Lists
  - 5.1.1. Using Lists as Stacks
  - 5.1.2. Using Lists as Queues
  - 5.1.3. List Comprehensions
  - 5.1.4. Nested List Comprehensions
- 5.2. The del statement
- 5.3. Tuples and Sequences
- 5.4. Sets
- 5.5. Dictionaries
- 5.6. Looping Techniques
- 5.7. More on Conditions
- 5.8. Comparing Sequences and Other Types

Previous topic
4. More Control Flow Tools

Next topic
6. Modules

This Page
Report a Bug
Show Source

5. Data Structures

This chapter describes some things you've learned about already in more detail, and adds some new things as well.

5.1. More on Lists

The list data type has some more methods. Here are all of the methods of list objects:

`list.append(x)`
Add an item to the end of the list. Equivalent to `a[len(a):] = [x]`.

`list.extend(iterable)`
Extend the list by appending all the items from the iterable. Equivalent to `a[len(a):] = iterable`.

`list.insert(i, x)`
Insert an item at a given position. The first argument is the index of the element before which to insert, so `a.insert(0, x)` inserts at the front of the list, and `a.insert(len(a), x)` is equivalent to `a.append(x)`.

`list.remove(x)`
Remove the first item from the list whose value is equal to *x*. It raises a `ValueError` if there is no such item.

`list.pop([i])`
Remove the item at the given position in the list, and return it. If no index is specified, `a.pop()` removes and returns the last item in the list. (The square brackets around the *i* in the method signature denote that the parameter is optional, not that you should type square brackets at that position. You will see this notation frequently in the Python Library Reference.)

`list.clear()`
Remove all items from the list. Equivalent to `del a[:]`.

`list.index(x[, start[, end]])`
Return zero-based index in the list of the first item whose value is equal to *x*. Raises a `ValueError` if there is no such item.

- Usually we program having at arm's distance the documentation of the packages we plan to use.

Python:	https://docs.python.org/3/contents.html
Numpy:	https://numpy.org/doc/
Pandas:	https://pandas.pydata.org/docs/reference/
Sk-learn:	https://scikit-learn.org/
Matplotlib:	https://matplotlib.org/stable/api/
Seaborn:	https://seaborn.pydata.org/examples/

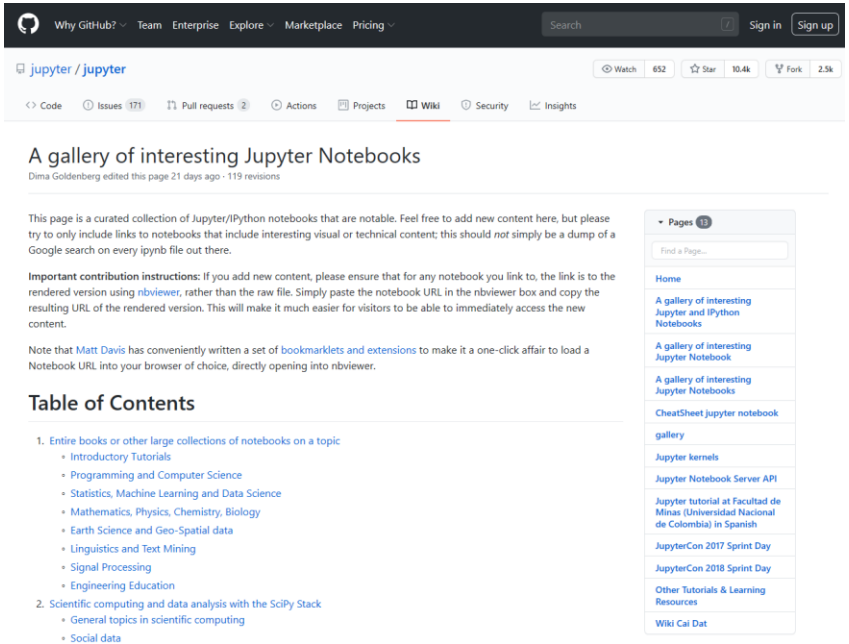
DataScienceGroup
datascience.dsv.su.se
Stockholm University

Relevant Python Packages/Modules

- Python packages/modules
  - Numpy (*vectorized operations*)
  - Pandas (*dataframe operations*)
  - Scikit-learn (*machine learning*)
  - Matplotlib (*data visualization*)
  - Seaborn (*easier data visualization*)
- Jupyter notebooks
  - Provide interactive workflow for DS
  - Combines text, math, and code in a single document

DataScienceGroup
datascience.dsv.su.se
Stockholm University

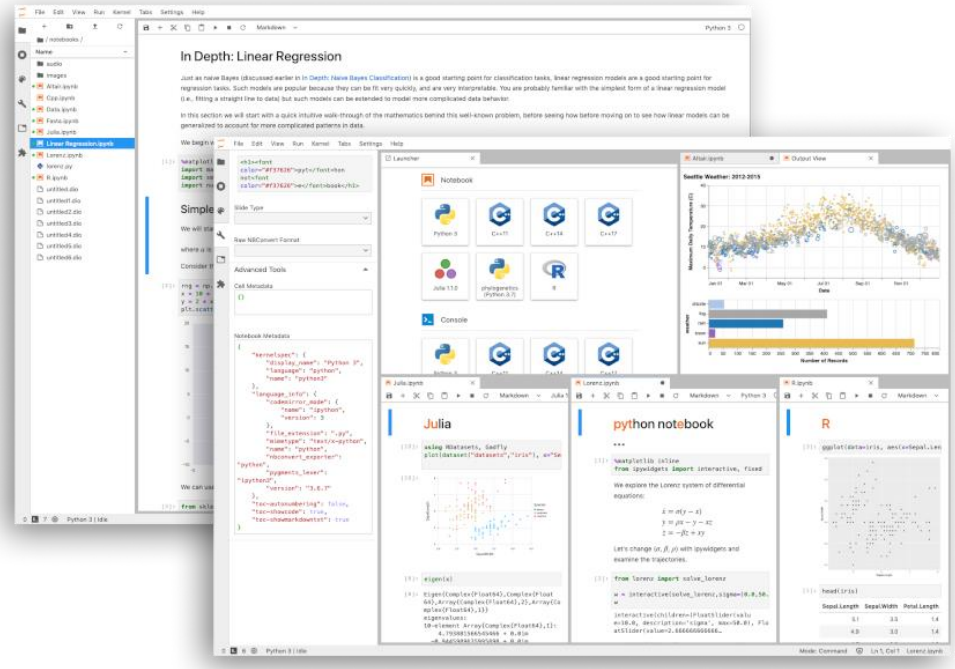
Example of Jupyter notebooks



https://github.com/jupyter/jupyter/wiki

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Python and Jupyter notebooks



Installation alternatives

1. Google Colab
2. Python and Browser
3. Python and IDE
4. Anaconda Toolkit
5. Miniconda Toolkit

Image from Project Jupyter.

DataScienceGroup
datascience.dsv.su.se
Stockholm University

1) Python and IDE

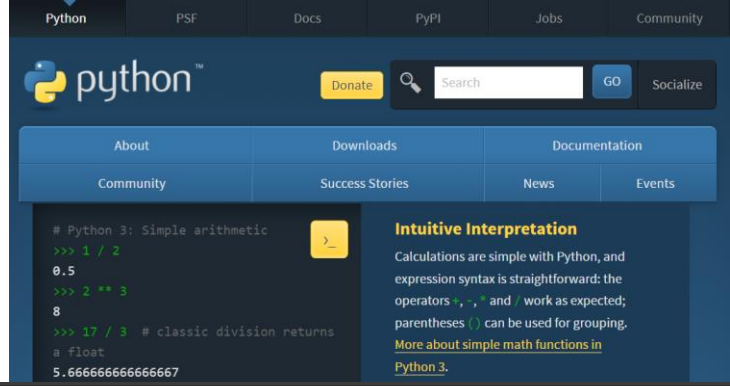
- Installing Python

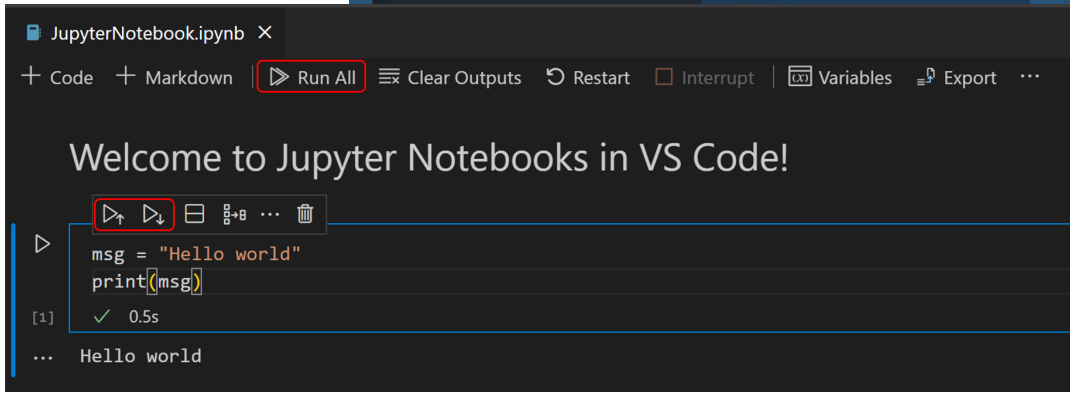
https://www.python.org/

- Choosing an IDE compatible with Jupyter notebooks (e.g., VS Code)

https://code.visualstudio.com/

- Local development
- Runs on your computer's resources



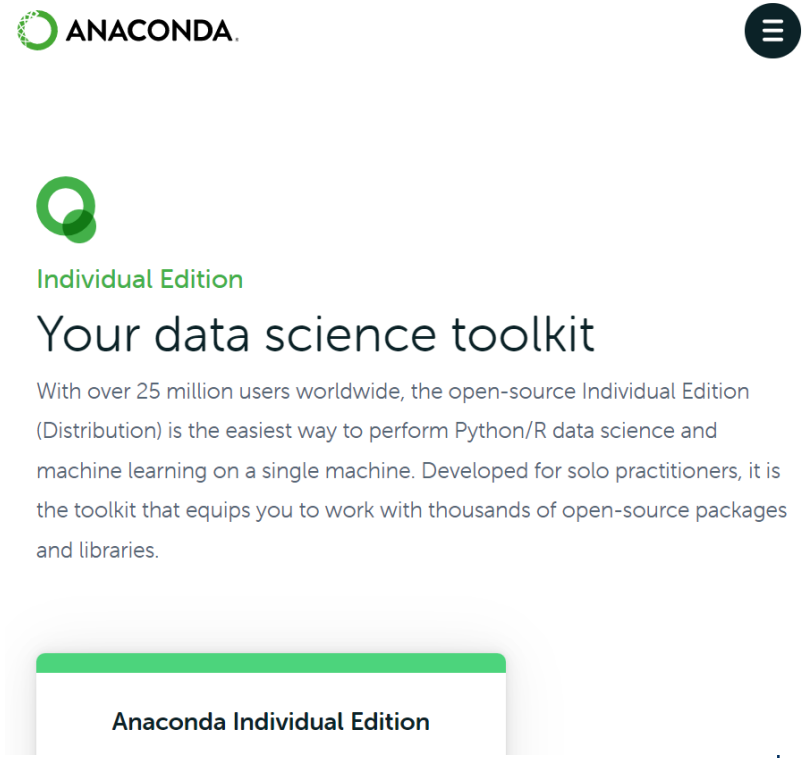


DataScienceGroup
datascience.dsv.su.se
Stockholm University

2) Python with Anaconda

- A toolkit that includes software to develop Data Science projects
- It includes Python, also the R programming language
- It installs more than 1.500 packages
- Requires around 3GB of space

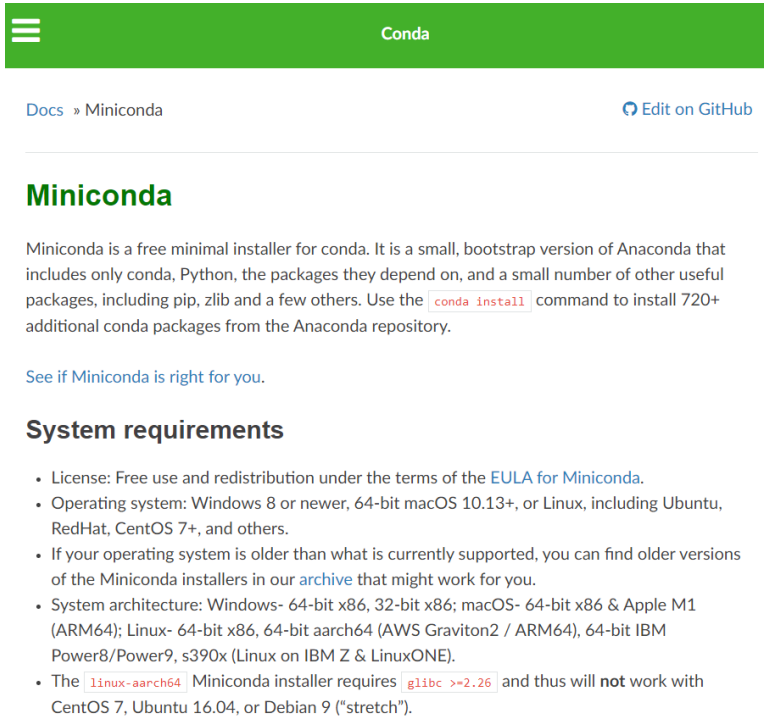
https://www.anaconda.com/



3) Python with Miniconda

- A smaller version of Anaconda (~60MB)
- Install manually the libraries as required.

https://docs.conda.io/en/latest/miniconda.html



Conda

Docs » Miniconda

Edit on GitHub

Miniconda

Miniconda is a free minimal installer for conda. It is a small, bootstrap version of Anaconda that includes only conda, Python, the packages they depend on, and a small number of other useful packages, including pip, zlib and a few others. Use the `conda install` command to install 720+ additional conda packages from the Anaconda repository.

See if Miniconda is right for you.

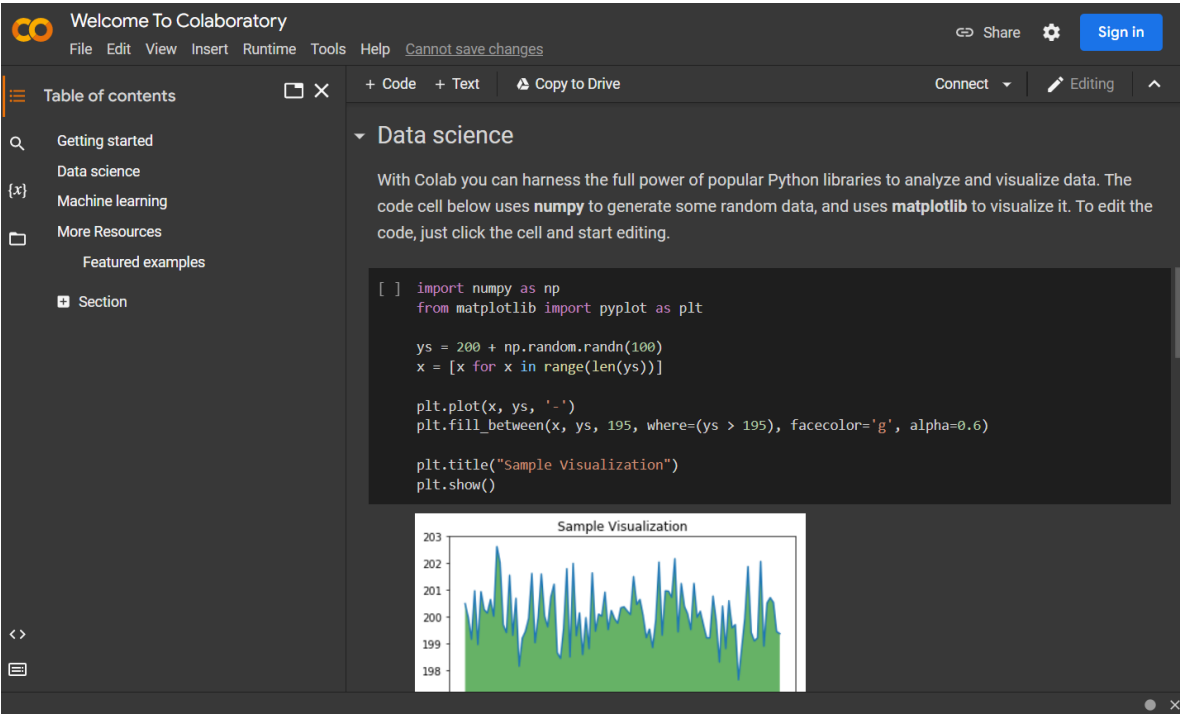
System requirements

- License: Free use and redistribution under the terms of the EULA for Miniconda.
- Operating system: Windows 8 or newer, 64-bit macOS 10.13+, or Linux, including Ubuntu, RedHat, CentOS 7+, and others.
- If your operating system is older than what is currently supported, you can find older versions of the Miniconda installers in our archive that might work for you.
- System architecture: Windows- 64-bit x86, 32-bit x86; macOS- 64-bit x86 & Apple M1 (ARM64); Linux- 64-bit x86, 64-bit aarch64 (AWS Graviton2 / ARM64), 64-bit IBM Power8/Power9, s390x (Linux on IBM Z & LinuxONE).
- The `linux-aarch64` Miniconda installer requires `glibc >=2.26` and thus will **not** work with CentOS 7, Ubuntu 16.04, or Debian 9 (“stretch”).

DataScienceGroup
datascience.dsv.su.se

Stockholm University

4) Google Colaboratory



- Hosted online
- Runs on Google's servers
- Requires uploading your data to the cloud
- Runs only for a limited time

Access from your browser:
https://colab.research.google.com/

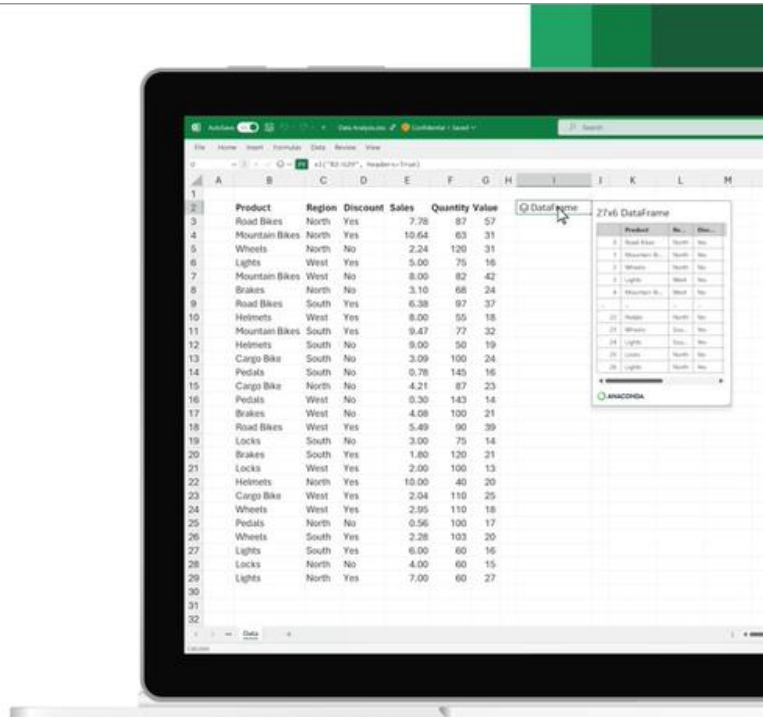
DataScienceGroup
datascience.dsv.su.se

Stockholm University

Python in Excel (microsoft.com)

Microsoft 365

Python
in Excel



DataScienceGroup
datascience.dsv.su.se
Stockholm University

Checklist for upcoming Lab 1

- ❑ Install Python >3.9 in your personal computers
- ❑ Install these packages with **pip**
  (or **conda** if using Anaconda)
  - ❑ `jupyter`
  - ❑ `numpy`
  - ❑ `pandas`
  - ❑ `matplotlib`
- ❑ Open and run the Jupyter notebook Lab 0 available from the iLearn website
- ❑ **Optional**: Practice the basic syntax of Python

Python prerequisites for the labs

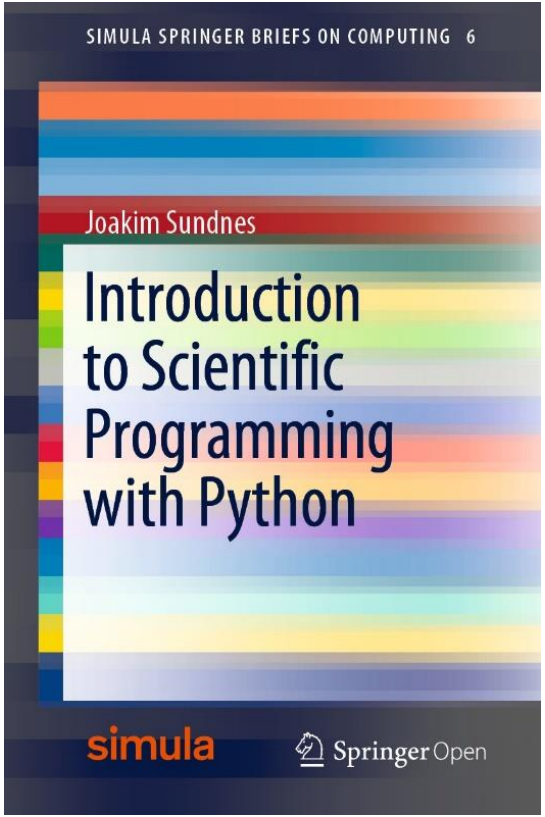
(videos will be posted on iLearn explaining these concepts)

1. Pythonic syntax
   - Interpreter, Math Operations, Variables, Function Call
2. Data Types and Data Structures
   - Variables, List, Tuples, String, **Dictionary!**
3. Package Management (PIP)
   - numpy, pandas, jupyter, scikit-learn, scipy
   - Working with existing open-source projects
4. Conditionals and Loops with Jupyter
5. Definition of Functions

DataScienceGroup
datascience.dsv.su.se
Stockholm University

Resources

- [Introduction to Scientific Programming with Python | SpringerLink]
- [Beginners Guide / Programmers - Python Wiki]



DataScienceGroup
datascience.dsv.su.se
Stockholm University

ILOs Recap

I. Assess the quality of a structured dataset
II. Formulate analytical questions that can be solved with descriptive and exploratory methods
III. Recognize the types of questions that can be answered with predictive methods
IV. Configure Python to design and implement a data science project

Luis Quintero


luis-eduardo@dsv.su.se


https://luisqtr.com

DataScienceGroup
datascience.dsv.su.se

Stockholm University