Data-analytics

AKI TAANILA 31.12.2020

JUHA NURMONEN 4.1.2022

Structured and unstructured data

- Structured data ≈ data which can be saved into a table or database
- Structured data can be **numerical** (e.g., price) or **categorical** (e.g., sex)

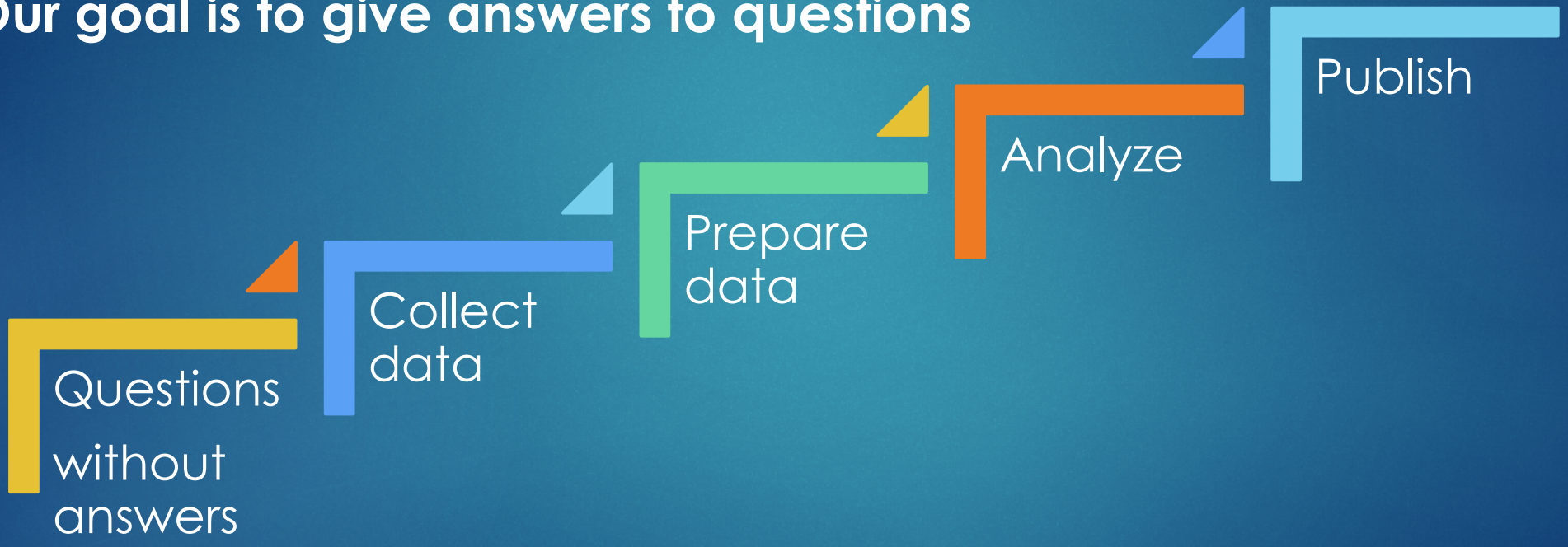
- Unstructured data can be, e.g., text, speech, pictures or videos

- **In this course we concentrate on structured data**

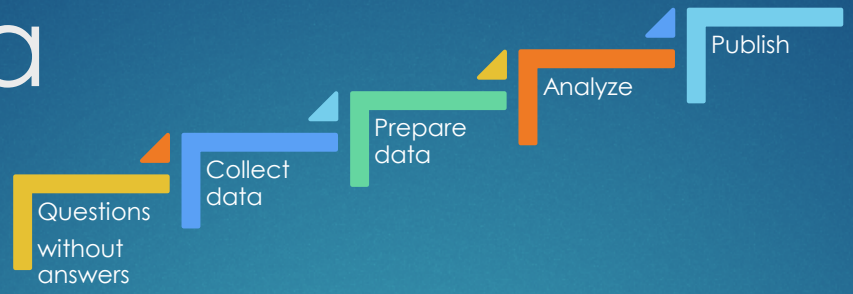
Data-analytics as a process

Data-analytics is goal-oriented activity:

Our goal is to give answers to questions



Collect data



- Is data already available, e.g., in a database, online resource, etc.
- Data can be collected using surveys, observations, sensors, etc.

File formats for stuctured data:

- Excel (.xlsx)
- Comma separated value (.csv). This is a text file where each piece of information is separated from others using the comma character. In Scandinavian text formats the comma is often replaced by the semicolon character, so that the comma can be used as decimal separator.

Prepare data

- Combining data collected from different sources
- Renaming variables (columns)
- Removing duplicates
- Checking time formats, units, etc.
- Recognising unusual values and errors
- Approach to missing values
- Calculating new variables, classification, recoding
- Familiarize yourself with the data: sorting, filtering

Analyze

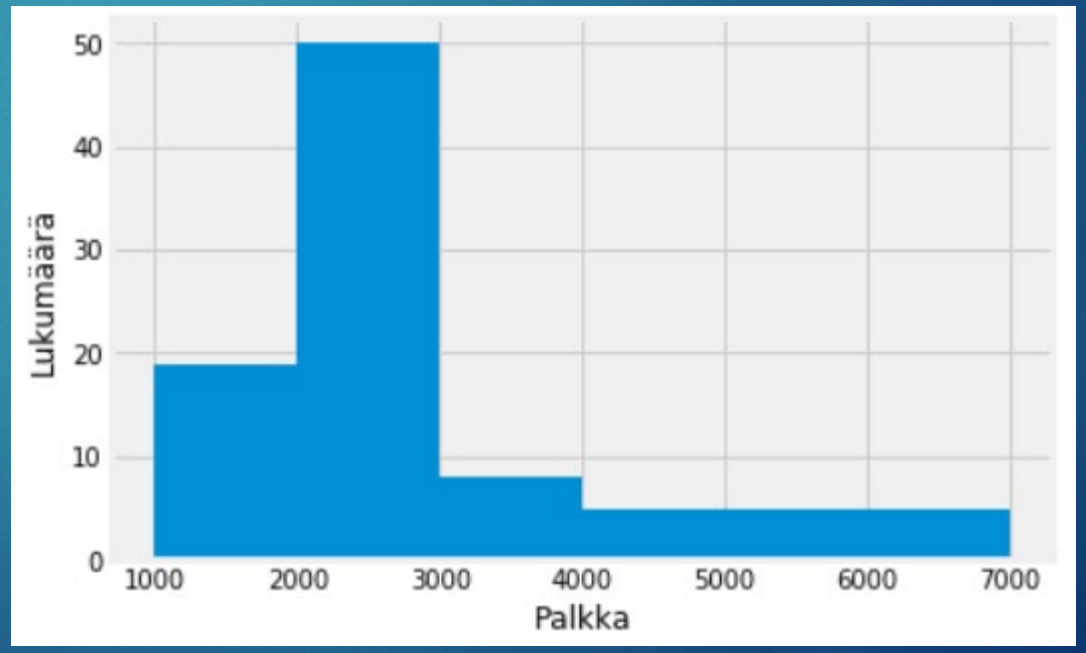
Analyzing part of data analytics can be divided into four sectors:

Descriptive analytics

- Frequency table
- Classified distribution
- Statistical numbers

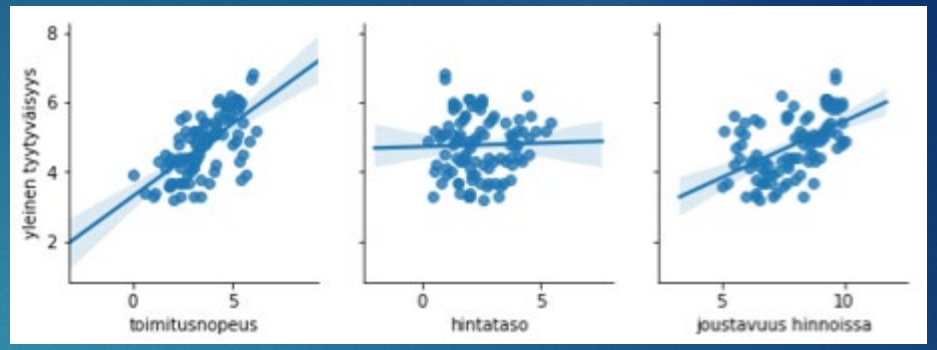
	lkm	%
Peruskoulu	27	33.3 %
2. aste	30	37.0 %
Korkeakoulu	22	27.2 %
Ylempi korkeakoulu	2	2.5 %
Yhteensä	81	100.0 %

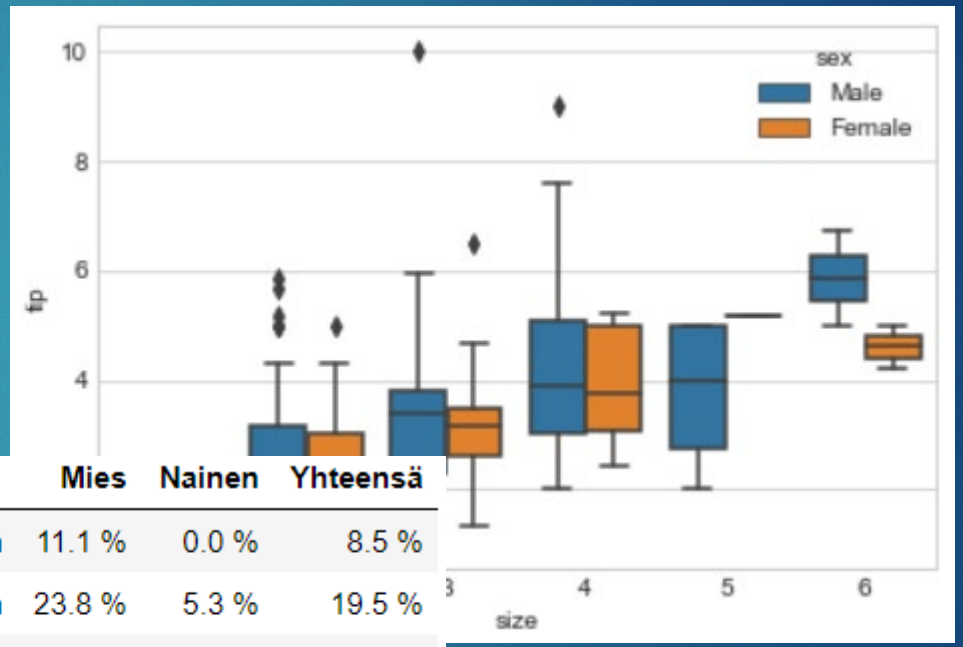
	ikä	palveluv	palkka
count	82.00	80.00	82.00
mean	37.95	12.18	2563.88
std	9.77	8.81	849.35
min	20.00	0.00	1521.00
25%	31.00	3.75	2027.00
50%	37.50	12.50	2320.00
75%	44.00	18.25	2808.00
max	61.00	36.00	6278.00



Diagnostics analytics

- Cross-tabulation
- Comparing statistical numbers in different groups
- Correlation coefficient
- Scatter chart





	Mies	Nainen	Yhteensä
Erittäin tyytymätön	11.1 %	0.0 %	8.5 %
Tyytymätön	23.8 %	5.3 %	19.5 %
Ei tyytymätön eikä tyytyväinen	36.5 %	36.8 %	36.6 %
Tyytyväinen	23.8 %	42.1 %	28.0 %
Erittäin tyytyväinen	4.8 %	15.8 %	7.3 %

Predictive analytics

- Time series forecasting
- Regression models (predicting numerical variable)
- Classification models (predicting categorical variable)

Prescriptive analytics

- Prescriptive analytics is the area of data analytics where direct operating instructions are given
- For example: What should we do in order to increase the sales?

Publish

Publishing method of results of data analytics depends on the purpose of use of the results. Suitable methods could be:

- Slide show
- Report
- Dashboard
- Input for a computer program

Data analytics tools

A FEW EXTRACTS

Excel

- Data tab on the Excel Ribbon contains user-friendly tools for retrieving data from several sources (databases, online sources, etc.). Same tools are used in PowerBI.
- Excel is suitable for preparing data, descriptive and simple diagnostics analytics.
- Excel is not well suited to big data.

Programming languages

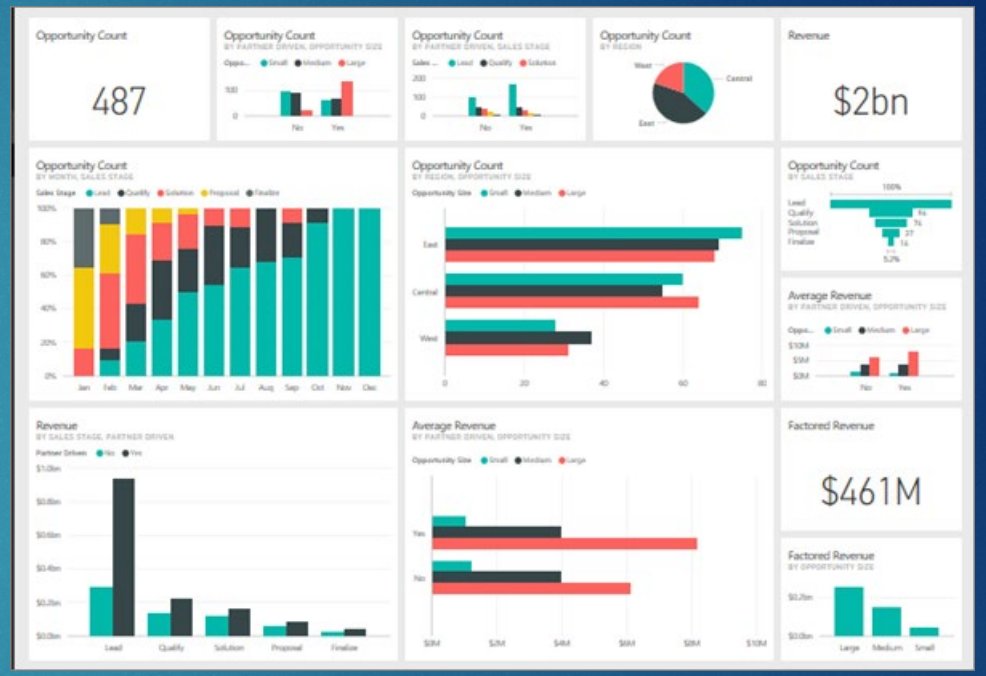
- Python is a popular programming language in data analytics.
- Python is also a general-purpose programming language.
- Anaconda Distribution is a free package which contains everything that data analytics requires.

- In addition to Python, R is a popular data analytics programming language
- Unlike Python, R is not a general-purpose programming language
- More information about the R programming language is at https://www.r-project.org/

- SQL is a standardized query language with which informations searches from databases can be performed

BI (Business Intelligence) tools

- Business intelligence described for decision making purposes
- Information is visualized as Dashboard
- Tools used are, e.g., PowerBI Desktop, Tableau, Qlik, …



PowerBI Desktop

- PowerBI Desktop is a free software, but publishing and creating interactive dashboards are subject to a charge
- There is a lot of material of good quality about using PowerBI Desktop online:
- Microsoft PowerBI learning material front page https://powerbi.microsoft.com/en-us/learning/
- Guided learning https://docs.microsoft.com/en-us/power-bi/guided-learning/
- Examples: https://docs.microsoft.com/en-us/power-bi/sample-datasets

Data mining tools

- Data mining contains information searching, combining and cleaning up data from different sources, predictive analytics and machine learning models
- There are several application software for data mining
- Orange https://orange.biolab.si/ is a free software implemented with Python (installation can be managed using Anaconda Navigatorin)
- Knime https://www.knime.com/ is a free software implemented with Java
- Rapid Miner https://rapidminer.com/ is a popular software implemented with Java

Other tools

- Statistical software, e.g. SPSS
- Large enterprise software systems, e.g. SAP and SAS

Why Python?

- Python covers all areas of data analytics
- You learn coding using one of the most popular programming languages at the moment
- Python is fast and performs well also with big amount of data
- Once written code can always be run again -- code can be utilized for several data
- Code also gives you a kind of documentation for your analysis
- Using code data analytics can be automated
- Python works in the same way on **Windows**, **Linux** and **MacOS** operating systems
- Python is free

Course execution

Course contents

- Preparing data
- Descriptive analytics
- Diagnostics analytics
- Time series
- Time series forecasting
- Regression models
- Classification models

- Python programming language is used as main tool and as programming environment Jupyter notebook, included in Anaconda package, is used

Course execution

- In assignments you analyze your own data or data provided
- **Essential: data analytics is used in order to answer questions. Pose your data good questions!**
- Information about open source data (Finnish page) https://tilastoapu.wordpress.com/avoin-data/
- In assignments you learn studying and preparing data, and also descriptive, diagnostic and predictive analytics
- Solutions to assignments are returned as Jupyter notebook format
- Returned Jupyter notebooks must include Python code and calculated results as well as **comments, explanations and documentation**

Course evaluation

Four assignments:

- Studying data, preparing data and descriptive analytics 0—5 points
- Diagnostic analytics 0—5 points
- Time series and time series forecasting 0—5 points
- Models (regression and classification) 0—5 points

- Each assignment has to be passed with a minimum grade of one point
- One passes the course if each assignment has grade one or more
- The course overall evaluation is the average of points of assignments