6th Seminar – Requirements Analysis, Explorative Data Analysis

1 Exploratory data analysis of server performance

We measured the following performance metrics on a server:

Time of measure [ms]	500	600	700	800	900
Requests processed in the last 100 ms [request]	11	12	21	18	20
Average serving time in the last 100 ms [ms]	15	20	21	25	27
CPU utilization of the last 100 ms [%]	12	13	16	17	19
HDD I/O utilization of the last 100 ms [%]	55	63	87	61	73

a. Display the number of processed requests and the CPU utilization on a scatterplot diagram! Interpret the diagram!

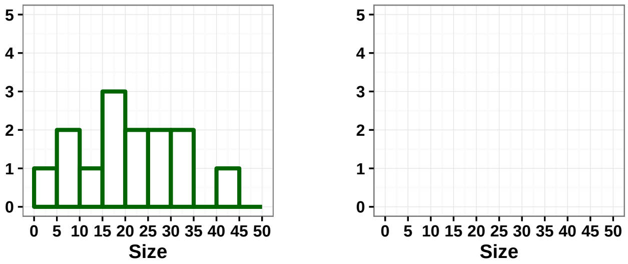
b. What is the server's throughput at the time of the first measure? What is the average and median throughput based on these 5 measurements? What belongs to the 40% quantile?

c. Can we assume some kind of causal relationship between some of the metrics?

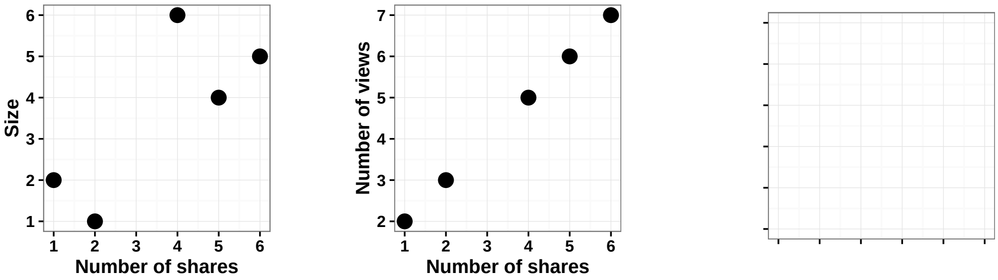
2 Picture gallery – data analysis

In our online picture gallery users can search and display pictures that match some search phrase.

a. We displayed the distribution of the album's size on the following histogram. We would like a histogram with twice the bin size of the original one, since in order to organize our storage more efficiently we only need to know how many albums we have with picture sizes below 10, between 10 and 20, and so on. Create this new histogram!



b. We chose 5 albums and displayed their size and number of views in relation to their number of shares on two scatterplot diagrams. Is it true that the bigger is the album, the more view it has? Answer this question with a third scatterplot diagram that shows the albums' number of views in relation to their size!



c. We would like to know the popularity of the albums so we calculated the average and median of the number of views based on the scatterplot diagram. Can we always use this method based on a scatterplot diagram? How will these values change if we upload a new album that was viewed 40 times?

3 Sensor network (previous exam exercise) – data analysis

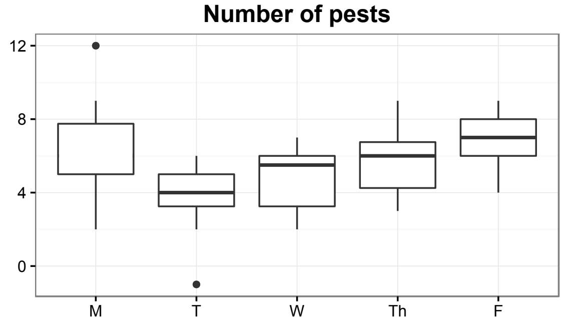
We have an agriculture sensor network that helps us to track the states of our open-field, glasshouse and foil tent areas based on some measured values (temperature, humidity, luminous intensity, wind speed, detected pests, etc.).

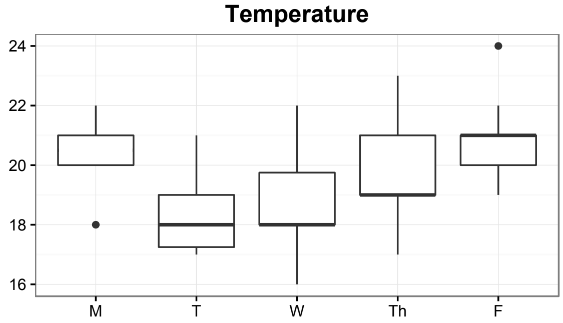
Date	Temp. [°C]	Hum. [%]	Pests [piece]
2015. 05. 04. 08:00	18	66,00	3
2015. 05. 04. 09:00	20	65,75	6
2015. 05. 04. 10:00	20	65,75	8
2015. 05. 04. 11:00	20	65,50	9
2015. 05. 04. 12:00	20	65,50	5
2015. 05. 04. 13:00	21	65,00	12
2015. 05. 04. 14:00	21	64,70	5
2015. 05. 04. 15:00	21	64,70	6
2015. 05. 04. 16:00	21	64,60	7
2015. 05. 04. 17:00	22	64,00	2

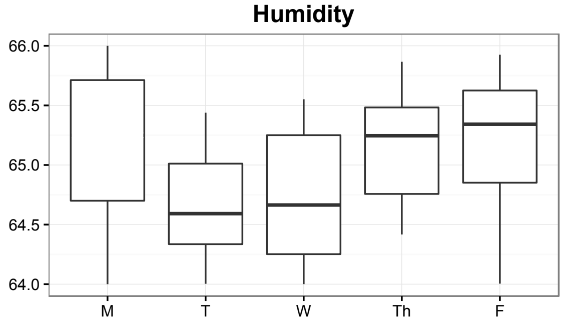
a. Unfortunately the middle values (median) of Monday, May 4th are missing from the figures. Draw them based on the data in the table!

b. Interpret the diagrams: which variable's/variables' first quartile is strictly monotonic in time?

c. (Extra task.) We would like to compare the temperature values and pest numbers of Monday in a parallel coordinates diagram.







4 Sensor network (previous exam exercise) – perf. analysis ($*$)

(Performance analysis exercises related to Exercise 3.) The different types of sensors provide data from a 100 meters radius around their location. The sensors forward their timestamped data to the central server through a radio communication-based network. The central server processes the requests then archives them to a storage unit. Our organization installed 4500 sensors and each one sends one measurement data in every minute. The system can successfully handle this load. The radio communication network can forward 100 measurement data every second. The central server's CPU is idle (not doing anything) in 75% of the time. Writing a measurement data to the storage unit takes 8 ms.

a. How many measurement data in a second is the current throughput of the system?

b. What is the throughput, maximum throughput and utilization of the radio network, CPU and storage?

c. How many more sensors can we install (to improve the measurement accuracy) without upgrading our infrastructure? Assume linear scaling!

d. The radio network uses smart encoding, so more than one sensor can forward data at the same time. How many sensors are forwarding data at the same time (overlapping) over the network currently and during maximum load, if a forwarding takes 40 ms?

5 Requirement analysis of train protection system

We are designing a train protection system. The main goal of the system is to prevent the collision of trains. The key to building a proper system is a requirement specification of good quality, since the test cases and other control mechanisms will be built based on these requirements.

Table 1: Train protection system requirements (partial)

R1	**Safety**	Trains mustn't collide on the supervised track system.
R2	**Operation**	It must be ensured that the trains reach their destination.
R3	**Optimality**	The travel time of trains must be minimized.
R4	**Track sections' supervision**	The track must be divided into sections and maximum one train can be on a section.
R5	**Dividing into sections**	The track must be divided into sections.
R6	**Occupancy**	Maximum one train can be on a section.
R7	**Detecting occupancy**	We have to detect somehow whether there is a train on a section or not.
R8	**Fault tolerance**	We have to be prepared for the malfunction of components.
R9	**Occupancy sensors**	Occupancy of a section must be detected in a redundant manner, based on multiple types of sensors.
R10	**Rail sensor**	Rail sensors must be installed in every section that signal whether there is a train on the section or not.
R11	**Camera system**	Cameras must be installed on the sections where it is possible for observation purposes.
R12	**Positioning**	Trains must continuously signal their positions towards the central control unit.
R13	**GPS subsystem**	The trains must be equipped with a GPS subsystem.
R14	**Wireless connection**	It must be ensured that the trains can provide their positions to the central control unit via a wireless network.
R15	**Train control**	It must be ensured that a train can be stopped before driving onto an occupied section.
R16	**Stopping trains**	The central control unit must be able to immediately stop a train.
R17	**Support of train types**	The system must support every type of trains capable of travelling on rails.
R18	**Unmodifiable trains**	We mustn't use methods that require to change the control system of trains.

a. Gather the participants that are involved (or affected) when building a system like this (so called *stakeholders*), that is, they can make demands for the system in form of requirements!

b. After identifying the stakeholders we gathered their requirements, a part of which can be seen in Table 1. Construct a graph that shows the dependencies between the requirements! Draw a directed arc in the graph from A to B if *(1)* requirement A is part of requirement B (*composition*), *(2)* requirement A *refines* requirement B, or *(3)* requirement A can be *derived* from requirement B. The exact relations between the requirements are not important, only that there is a relation.

c. From the above requirements which are functional requirements? What type of extra-functional requirements can we find in the table (safety, performance, reliability, etc.)?

d. Are the gathered requirements consistent? If not, then show an example of a contradiction.

e. From the gathered requirements give examples for directly verifiable requirements!

Optional Excercise

6 Social website

We operate a social web company. Due to its recent rising popularity, response times have increased greatly. The business goal is to have 1500 simultaneous user requests served with less than 4 seconds of response time in average.

a. What minimal throughput should the service infrastructure be designed for, if delays outside our infrastructure (network traffic latency, HTML rendering on the client side) can be estimated as 1 second?

b. According to measurements, an average user request in the redesigned web site takes 20 ms CPU time on the web server, and occupies the database server for 12,5 ms. Currently we have 15 web servers to handle the requests, while the database is replicated to 5 machines. Assuming linear scalability, how much additional units of each kind of server should we buy to meet the above goal?

c. ($*$) Calculate the utilization of each kind of server in the extended system. If the goal is to push the average utilization of the servers below 50% even during peak hours, do we need to scale out further?

d. Let’s consider only 2 webservers and 3 database servers. Create state-based models about the resources in the infrastructure that model the availability of the resources (available or in use). What design decisions do we face? What are the pros and cons of the choices?