Slide 1:

Today's theme is datasets and variables.

Slide 2:

We will talk about what a dataset is and how we should understand data.

In addition, we will talk about how we analyze data and what challenges can occur.

There are also many different types of data, and we will talk about why data types are important when visualizing a dataset.

Slide 3:

I have created three questions that you should be able to answer after this lesson.

What is a dataset?

What are objects and attributes in a visualization perspective?

What are the different data types?

The questions are designed to help you understand what is important.

My experience is that many students find this topic difficult.

My best advice is to spend a little extra time watching the lecture, do the quizzes and participate in the lab.

Slide 4:

The terms "data" and "information" are often used interchangeably, but they actually are not the same.

There are small differences between these words and their purpose.

Data is defined as individual facts, while information is the organization and interpretation of those facts.

Ultimately, you can use the two components together to identify and solve problems.

Data is defined as a collection of individual facts or statistics.

Data can come in the form of text, observations, figures, images, numbers, graphs, or symbols.

For example, data might include individual prices, weights, addresses, ages, names, temperatures, dates, or distances.

Data is a raw form of knowledge and, on its own, does not carry any significance or purpose.

put in another way, you must interpret data for it to have meaning.

Data can be simple—and may even seem useless until it is analyzed, organized, and interpreted.

Slide 5:

The Key Differences Between Data vs Information

Data is a collection of facts, while information puts those facts into context.

While data is raw and unorganized, information is organized.

Data points are individual and sometimes unrelated. Information maps out that data to provide a big-picture view of how it all fits together.

Slide 6:

Data, on its own, is meaningless. When it is analyzed and interpreted, it becomes meaningful information.

Data does not depend on information; however, information depends on data.

Data typically comes in the form of graphs, numbers, figures, or statistics.

Information is typically presented through words, language, thoughts, and ideas.

Data isn't sufficient for decision-making, but you can make decisions based on information.

Slide 7:

We can create utility based on the information we get from data.

Look at the figure on the right.

Here is a street with 200 people shown in different ways.

How 200 people look if they must drive in 177 cars,

how little space the people use without a car,

how the street will look if everyone was on a bicycle or divided into 3 buses or in 1 light rail train.

Visualization is used to show the utility of data, by using 5 different images to convey the information.

Slide 8:

When we visualize, we translate data into something that is visual.

Concrete data values become markers in a diagram, as shown on the right in a coordinate system.

In addition, information is added so that the markers make sense.

Number of delays/cancellations is displayed on the vertical position (y-axis) and time on the horizontal position (x-axis).

Slide 9:

How do we understand data?

Nathan Shredoff proposed a model for how we understand data. He defines a gradual development in our understanding based on a context.

He created a model that shows step-by-step development in understanding. He describes 4 phases "Data", "Information", "Knowledge" and "Wisdom"

Slide 10:

The first step is "data" which can be collected by doing research.

When **data** is collected in the form of raw data, or observations, it is meaningless without a context. Data can be processed; we can work with data to connect parts and gain understanding of information.

Information is about making the data easier to understand by processing/sorting raw data and presenting it in alternative ways.

When we have gained an understanding of the information, we can start to see a whole and interact with the information, then knowledge can be obtained.

Knowledge: When knowledge is obtained then the information creates utility.

Wisdom: The highest level is wisdom and is acquired knowledge (over time) and it gives us a deeper understanding and reflection.

Slide 11:

But what is a dataset?

A dataset consisting of data related to a collection of objects, where each object is described with a set of attributes often called variables. Data is often presented in a matrix with n rows and m columns.

Slide 12:

Object and attributes in a dataset

A cat can have the following characteristics, also called attributes: The characteristics/attributes of the cat: It can be, for example:

Breed, Life expectancy, Weight, Behavior, Nutrition, Price, Etc

Slide 13:

If we insert the data into a table, we can have a cat of the "Norwegian forest cat" breed. It costs 12,000. and has an expected life of 15-18 years. A Norwegian forest cat can weigh between 3-9 kilos, and it often has a calm and reserved behaviour.

Another example of a cat breed is the British shorthair. As you can see the attributes are the same.

Slide 14:

A dataset does not have to describe only physical objects. Let's look at a dataset on world happiness. There are several things that tell us about how happy a person is. For example, family status, finances, which country the person lives in and life expectancy. These factors can give an overall happiness score.

Slide 15:

Here is data from two countries - Norway and Costa Rica. Pause the video and try to answer the following questions:

What is the object in this happiness dataset?

What are the attributes?

Slide 16:

Maybe you knew the answer -

Object can be happiness and/or Country

The attributes are Family, economy, health and happiness score.

Slide 17:

Data can also come from a website.

Try pausing the video and identifying the object and attributes on H&M's website.

Slide 18:

The H&M website contains a lot of categorical data - this means data that we can divide into categories and subcategories. For example: Target group, type of clothing and size. H&M website can for example be an object with the attribute's female, male, child, baby etc. Another object could be swimwear with the attributes size, gender, material, price.

We can have many objects and attributes on a website. When we work with data analysis, it is important to be able to identify them in order to make a meaningful analysis.

Slide 19:

Where does the data come from?

Data can be the result of a direct measurement. This can be done, for example, by taking measurements of water temperature over a period of time.

Not all data can be collected by making a direct measurement.

Sometimes we have to plan how data will be collected. For example, a survey. Here, the questions must be written in advance and the participants must have time to answer the survey. The result will not occur immediately.

Data may also be collected indirectly from other sources. For example, we can measure activity, do digital tracking or collect data from Twitter.

Slide 20:

Now we will look at typical challenges when analyzing a dataset.

Slide 21:

This is a dataset collected during the Roskilde festival - perhaps some of you already know this festival - but it is northern Europe's largest music festival and is and takes place every year in Denmark.

This is a dataset that shows sales of food and drink in the "Meyers" food stand.

Let's try to analyze this raw data

Pause the video and try to identify possible challenges that may complicate the analysis of the dataset

---

We have no information about how the data is collected and therefore we have to make our own assumptions.

Perhaps you also noticed that there were several empty fields in the table.

For example, under the payment category. Cards are shown, but what do the empty fields represent? Perhaps they represent payments made with cash? but We don't know for sure.

Rows 5 to 10 are missing more information. We do not know which Item or item category has been purchased and the price looks quite strange.

Under the category time, we can also see that the time is registered in two different ways. When data is collected, it must always be done in the same way.

When we visualize, we also must assess how many variables we include. The information from the dataset can tell several stories.

If we wanted to find out which products sell the most and the least in Meyer's food stand, we could select certain product groups and make a comparison.

When selecting data, it is very important not to exclude relevant data because this will make the visualization misleading.

For example, if we only select 5 best-selling products, but the data set shows that there are 7 best-selling products. It would be wrong to exclude 2 products.

All these challenges make it difficult to visualize the dataset. We can make assumptions, but we must never visualize anything other than what the dataset shows in order not to mislead the audience. The more uncertainties there are, the bigger the chance of making a wrong visualization.

Slide 22:

On the last slide we talked about typical challenges - here they are summarized.

Slide 23:

Data sets from the Roskilde festival were an example of raw data.

We can also derive data - On the right side, a figure of different car prices is shown.

Each marker symbolizes a car.

The first row shows all markers, and the second row only shows the lowest price and highest price.

Between the two values is a box .... it shows that the remaining cars and prices are within the box's limits. The price limits are slightly below 2000 pounds and 1700 pounds.

The purpose of this visualization is to show the cheapest and most expensive car price, but without removing all other prices, - but at the same time make it easy for the audience to quickly get an overview of the prices.

Slide 24:

Another way to show car prices is, for example, on Finn

A user wants to buy a car and goes to Finn's website. The goal is to make it easy for the user to find the right type of car based on several criteria.

Slide 25:

When we talk about data, it can be divided into two main groups. Quantitative and qualitative data.

Slide 26:

Quantitative data is the value of data, in the form of counts or numbers.

It can be used for mathematical calculations and statistical analysis to make real-life decisions.

“How much did that laptop cost?” it is a question that will collect quantitative data.

Quantitative data is best visualized in graphs and charts.

Pounds or kilograms for weight, dollars for cost, are examples of quantitative units of measurement.

Slide 27:

Qualitative data is information that cannot be counted, measured, or easily expressed using numbers.

It can be collected from text, audio, and images.

It can be shared through data visualization tools, such as word clouds, concept maps, timelines, and infographics.

Qualitative data analysis tries to answer questions about what actions people take and what motivates them to take those actions.

Slide 28:

As the figure shows, we have subgroups within both types of data.

Categorical, nominal and ordinal data belong to qualitative data.

Numerical, interval and ratio data belong to quantitative data.

Slide 29:

Numerical data is characterized by the fact that we can make calculations. For example, add numbers, divide, calculate an average.

Slide 30:

In Interval data, data is measured along a scale where each point is equidistant from one another. Interval data holds no true zero. This means that we can have values way above and below 0. For example, temperature.

Slide 31:

Ratio Data can have the same values as interval data, but Ratio data has a defined zero point. It is possible not to have a negative weight or you can not be -4 years old. Zero is the lowest point.

Slide 32:

Let's look at qualitative data.

Slide 33:

Categorical data is characterized by the fact that data can be given a category value.

For example, students are a category and teacher are a category.

The categories have no mathematical meaning. We cannot multiply the two groups and get a meaningful result.

Slide 34:

Nominal data is used as a "label". The order of the labels is not significant. Examples of labels are "single" and "married".

Slide 35:

Ordinal data - Then the order of the data is important.

The picture shows how hot the different chilies are.

Another example of ordinal data is a pain scale.

Pain can be described as minimal, moderate, severe and unbearable pain.

Side 36:

We also have other types of data.

Slide 37:

Binary data is data whose unit can take on only two possible states.

E.g., 0 or 1 or dead or alive.

Slide 38:

Continuous data is a type of numerical data that refers to the unspecified number of possible measurements between two realistic points.

Continuous data is all about accuracy, therefore often written in decimal numbers. All numbers in an interval can be used.

E.g.,

The weight of newborn babies.

The daily wind speed

The temperature of a freezer.

Slide 39:

Discrete data is a numerical type of data that includes whole, concrete numbers with specific and fixed data values determined by counting.

E.g., Number of children in a family. We can count how many children there are in a family. It is often a relatively stable number.

Slide 40:

Here is a summary so you can see the differences between discrete and continuous data.

Slide 41:

Here is an overview:

Interval and ratio data contain more information than ordinal data, which contains more information than nominal data.

We can go from continuous data to discrete to ordinal to nominal - but never the other way around.

Side 42:

Why are data types relevant for data visualization?

This is important because there are different ways to visualize data based on data type. We will talk more about that next week.

Slide 43:

Next week we will talk about data representation. Then I will review the first exam's assignment. In the physical class there will be an opportunity to ask questions.

See you next week.