2 1. Teil - Theorie

2.1 1. Theoriefrage

Erkläre die 4 V`s mit deinen eigenen Worten. Vergiss dabei nicht die 4 V´s thematisch einzuordnen und an einem Beispiel zu erklären.

(Tabellearische Darstellung ist ausreichend.)

Answer:

Volume	Variety	Velocity	Veracity
The amount of data/scale of data can no longer be handled by conventional means due to huge amount of the data stored and created nowadays which are now frequently larger than terabytes and petabytes. Especially when relevant information must be brought together from different areas.	The variety of data sources and data formats require a different data analysis: structured, semi-structured and unstructured Nowadays a lot of new information is digitized. Some of them are structuralized (like date, amount, time) but most of them are unstructured (like Twitter feeds, Facebook posts, audio files, MRI images, web pages, web logs) Therefore a different approach (and tools) are needed because the unstructured data is often impossible to evaluate in relational databases.	= Speed. Timely processing of data must be ensured (the frequency of incoming data that needs to be processed). Nowadays it is more feasible to access even very large amounts of data simultaneously. Real-time processing plays a major role for many companies. A streaming application like Amazon Web Services Kinesis is an example of an application that handles the velocity of data.	= Truthfulness, correctness. Data from various sources sometimes does not arrive in the desired quality. Can we rely on the data? It can therefore not be used as intended or must be reprocessed at great expense. Data quality determines the success of big data. An example of a high veracity data set would be data from a medical experiment or trial.
Example: most of the US Companies have at least 100 Terabytes of data, 2,5 quintillions bytes are created each day*	*Example:* unstructured data like photos, text and videos	*Example*: streaming, data from the machine like modern cars sensors (monitors), 1TB of trade information during each trading session (NY stock exchange)	*Example*: customer and products ratings stored in a uniform rating scale, instead of continuous text. Amazon ratings for the products (1-5 stars)

*Used source for numbers: Infographic: The Four V's of Big Data | IBM Big Data & Analytics Hub (ibmbigdatahub.com)

2.2 2. Theoriefrage - fehlende Werte

a.) Erkläre anhand eines selbst gewählten Beispiels was fehlende Werte sind.

Answer:

Many real-world datasets may contain missing values for various reasons. Missing data are empty fields in a record. This means that the value is missing for the given rows and columns. This can occur if, for example, no information was given about the statement in the data collection. For example during collecting data for a survey from people: people with high salaries generally do not want to reveal their incomes in surveys or females generally don't want to reveal their ages.

Other example of missing data could be clinical records. Nurses may forget to record an output at a certain time point. Patients may have only one measurement of blood lactate, while the researcher is interested in exploring the impact of lactate trend on mortality outcome. Other reasons of missing values include but not limited to coding errors, faulty or nonresponses of equipment.

Nevertheless, when identifying missing values, we need to look carefully at the table and check whether columns belong together thematically.

For example during our course (USE CASE for ecar) we had an example for cars where fast charging was not included and therefore "Fast charging time" was not given. "Fast charging time" and "Fast charge function" belong together. If a car does not have the fast charge function, there can be no value in the column "Fast charging time ". Consequently, the missing data in second column are not really missing values. Here we have to think about how to fill in the values in order to be able to carry out evaluations later. If we delete the numbers, we will again lose important information.

b.) Warum muss sich bei der Datenanalyse damit beschäftigt werden?

Answer:

Missing values are missing information in the data set. If these are not identified, statements are falsified (e.g. in the case of visualisation, where the missing data are taken from the population. This changes the population each time). When identifying, it is important to consider whether there is really a missing value. A 0 is usually not a missing value, but means that the value is present, but is 0. Furthermore, it is important to note whether columns belong together in terms of content.

In machine learning for example, training a model with a dataset that has a lot of missing values can drastically impact the machine learning model's quality.

c.) Wie können fehlende Werte identifiziert werden?

Answer:

Missing values can be represented either by synonyms or by conventions. They are often encoded as NaNs, blanks or any other placeholders.

For example:

- Synonyms: ?, dummy, dummy-must-be-change, -99999
- Conventions: null, NaN (stands for: Not a Number), Na, None

In Python, the convention for a missing value is None. None comes from the Python package and is treated as a data type: object.

This means: if you execute functions like min(), max(), + etc. you get an error message! Why? Because of the data types - object cannot be added/subtracted etc. with int/float64.

Therefore you will get the following error message: "TypeError: unsupported operand type(s) for +: 'int' and 'NoneType'".

The other way to identify the missing value is to check the columns type. If column has values which by “fell” shall be an integer or float but type shows an object type, it is a hint that something is not ok.

When we try to force-change the type of the column datatypes, we could identify easily the used synonym (we become an error message with synonym).

Attachment to a, b and c: How to deal with missing values:

There is no universal way to deal with missing data. Each solution depending on the kind of problem.

For some cases it is safe to remove the data with missing values depending upon their occurrences, while in the other case removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations.

Sometimes you can drop variables if the data is missing for more than 60% observations but only if that variable is insignificant.

In my opinion, it is always better to keep data than to discard it.