Tampereen yliopisto

Mervi Jokipii

MANAGING MISSING DATA IN DATA INTEGRATION

Faculty of Information Technology and Communication Sciences
M. Sc. Thesis
May 2023

ABSTRACT

Mervi Jokipii: Managing Missing Data in Data Integration
M.Sc. Thesis
Tampere University
Master's Degree Programme in Computer Sciences
May 2023

The amount of data in the world is constantly growing at an enormous pace, especially with the expansion of the internet. Data is stored in different formats in various source systems. The goal of data integration is to provide users with unified access to heterogeneous and independent data without requiring them to understand the logic of the source systems. Users can submit queries on the mediated schema that interprets them to the source systems. The data in integration is rarely complete: it may contain incorrect or completely missing values. These missing data can be managed and enriched using various methods.

The literature review of this thesis explores data integration and its challenges, as well as the missing data mechanisms and strategies for dealing with missing data. The experimental section of this work analyses these strategies in the context of online automotive dealerships. Cars are increasingly being purchased directly from the internet or at least using the internet as a strong support in the purchasing process. Incomplete car data can lead to issues such as the car not appearing in potential buyers' search results, even resulting in the car not being sold.

The results of this work show that finding a similar car from a dataset is crucial in managing missing car data, which is not always straightforward. String matching -method is an essential part of finding a similar car, but it doesn't always give a perfectly accurate result. For this reason, the work presents a model for managing missing car data, where string matching is used only when necessary. According to the model, string matching can also be strengthened by comparing other values belonging to the same attribute group. External sources, such as pre-existing commercial databases or a company's self-built database, should also be used, if needed, to find the similar car.

Key words and terms: data integration, missing data management, imputation

TIIVISTELMÄ

Mervi Jokipii: Puuttuvan tiedon hallinta dataintegraatiossa
Pro gradu -tutkielma
Tampereen yliopisto
Tietojenkäsittelyn tutkinto-ohjelma
Toukokuu 2023

Datan määrä kasvaa yhteiskunnassamme valtavalla vauhdilla erityisesti internetin laajentuessa. Tietoa tallennetaan eri muodoissa erilaisiin lähdejärjestelmiin. Dataintegraation tavoitteena on tarjota käyttäjälle yhtenäinen pääsy heterogeeniseen ja itsenäiseen dataan ilman, että käyttäjä tuntee lähdejärjestelmien logiikkaa. Käyttäjä hakee tietoa kyselyiden avulla, joita skeemat välittävät lähdejärjestelmille. Dataintegraatiossa liikkuvassa tiedossa on usein puutteita, esimerkiksi virheellisiä tai kokonaan puuttuvia arvoja. Näitä puuttuvia tietoja on mahdollista hallita ja rikastaa erilaisten menetelmien avulla.

Tämän työn kirjallisuuskatsauksessa tutustutaan dataintegraatioon ja sen haasteisiin. Kirjallisuuskatsauksessa esitellään myös puuttuvan tiedon mekanismeja sekä erilaisia strategioita puuttuvan tiedon hallintaan. Työn kokeellisessa osiossa näitä strategioita analysoidaan autoliikkeiden verkkopalvelujen kontekstissa. Autoja ostetaan yhä enemmän suoraan internetistä tai vähintäänkin käytetään internetiä vahvasti ostoprosessin tukena. Puutteelliset auton tiedot saattavat aiheuttaa esimerkiksi sen, ettei auto osu potentiaalisen ostajan autohakuihin sivustolla, jolloin auto saattaa jäädä jopa tämän vuoksi ostamatta.

Tämän työn tuloksissa todettiin, että puuttuvan autodatan hallinnassa oleellista on löytää datajoukosta vastaava auto, mikä ei aina ole suoraviivaista. Merkkijonon vertailu on olennainen osa vastaavan auton löytymisessä, mutta se ei anna aina täydellisen oikeaa tulosta. Siitä syystä työssä esitetään puuttuvan autodatan hallintaan malli, jossa merkkijonon vertailua käytetään vain tarvittaessa. Mallin mukaan merkkijonon vertailua voidaan varmentaa myös vertailemalla muita samaan attribuuttiryhmään kuuluvia arvoja. Vastaavan auton löytämiseen käytetään tarvittaessa myös ulkoisia lähteitä, joita voivat olla valmiit kaupalliset tietokannat tai yrityksen itse rakentama täydennystietokanta.

Avainsanat: dataintegraatio, puuttuvan tiedon hallinta, tiedon rikastaminen

Contents

1 Introduction

The amount of data in the world is constantly growing, and utilizing it in various ways is vital for many organizations (Laquer, 2017). Data integration aims to combine data from different sources so that users have uniform access to it without needing to be familiar with the original data sources (Ziegler & Dittrich, 2007). Data in different sources may be in different formats, and different query methods may need to be used in databases (Doan et al., 2012). Data integration has become an essential part of software development.

Data from different sources is rarely complete for various reasons. Data may be incomplete due to errors in recording, for example (Bramer, 2007; García et al., 2015). In cases of missing data, it must be determined how to handle it. It is possible to simply ignore such incomplete data and use only complete data. Missing data can also be replaced with various methods. (Ratner, 2011)

The purpose of this thesis is to give a general understanding of data integration and missing data management to readers who have no previous experience with the subject. Literature on missing data management is often quite mathematical and probability-based. In this thesis, however, the intention is deliberately to leave out this mathematical approach and keep the explanations as high-level as possible.

In this thesis, the challenges of missing data are also researched in the context of car retail's websites and especially car data integrated to the webpages. Car sales and leasing services has taken a huge digital step in the last years like most of the business areas. Detailed information about cars on these websites mostly comes from a third-party data source. In this case as well, the data is not always complete, and individual pieces of data about the car may be missing. This missing data may sometimes be essential for the functioning of the website and thus for the car retail business. However, companies designing and developing websites cannot control the quality of the data, which is why methods must be developed for enriching missing data.

The research questions of this thesis are:

- What are the different methods to replace missing data in data integration?
- How are these methods suitable for replacing the missing data in car data integrated to the web pages of car dealerships?

This thesis begins with an overview of the theory of the topics mentioned earlier. In Chapter 2, data integration and its principles and challenges are presented. In Chapter 3, missing data, causes, and related mechanisms are discussed. Chapter 4 covers strategies related to missing data management. Chapter 5 presents the research approach of the thesis. In Chapter 6, the challenges of integrating incomplete car data on car retail's websites are presented, and ways to enrich this missing data are explored. Finally in Chapter 7, the conclusion of the research is summarized.

2 Data Integration

A large amount of data is often processed in software development. This data comes from various sources. Data is created, for example, as a result of collection or tracking or from data integration. (Laquer, 2017) Data integration is especially important for large companies that have many different sources of data, for large research projects where data is produced independently by several researchers, for the cooperation of government agencies that each have their own data sources, and for search engines that search through millions of internet data sources (Halevy et al., 2006).

The amount of processed data in the world has grown enormously during the last decades (Doan et al., 2012). The development of the Internet has made data available to everyone, and the number of different databases even within a single organization has multiplied. With this, the role of data integration has become essential the larger the organization in question. (Daraio & Glänzel, 2016; Halevy et al., 2006)

At the same time while the amount of data has increased, the area of data integration has also expanded significantly over the years. Initially, the goal of data integration was to produce a user interface that could be used to retrieve information from the company's various systems using pre-defined query models. For decades, data management was done through relational databases; the developers knew what the data looks like, how the data is stored, modified and retrieved. (Halevy et al., 2006; Laquer, 2017) Today we live in a completely different world. There are many different types of data, the amount of data has exploded and diversified and become more complicated and it is practically everywhere. (Laquer, 2017)

With the spread of the Internet, data integration faced additional challenges, such as the need to obtain data from external business partners, share data with multiple data sources, create common architectures for information sharing, or retrieve data from the numerous sources available on the Internet (Daraio & Glänzel, 2016; Golshan et al., 2017). The integration had to take into account semi-structured data, such as XML or unstructured text-based data (Golshan et al., 2017). Today, data integration is absolutely necessary as the amount of data has grown enormously and it is also more easily accessible via the internet (Halevy et al., 2006).

An important development direction of data integration has also been the inclusion of artificial intelligence (AI) in data management. Initially, AI was meant to be used to describe the content of heterogeneous data sources in data integration (Halevy et al., 2006). As the field has progressed, the role of AI grew and continues to grow as data integration systems must simultaneously account for uncertainty, be able to find facts from text-based data and learn from previous experiences (Golshan et al., 2017).

2.1 Data and Databases

Because data comes from various sources and be in different format, there might be challenges to analyse it properly. Sherman (2014) lists the five Cs of data that, by taking care of, the data will be in good shape:

- **Clean:** Most of the data is "dirty" to some extent. Dirty data contains, for example, missing values, wrong values etc.
- **Consistent:** It must be clearly known which data is correct.
- **Conformed:** The data should be analysed with respect to the same dimensions.
- **Current:** The data must be as up-to-date as possible in relation to its purpose of use.
- **Comprehensive:** All necessary data should be available.

The Organisation for Economic Co-operation and Development, OECD (2011) has also defined Quality Framework that has seven dimensions for quality:

- relevance: the extent to which data serves the intended purposes of users;
- accuracy: the level to which data provides correct estimates or descriptions of characteristics or the quantities they are designed to measure;
- credibility: level of trust that data users have in the data products, which is largely influenced by their perception of the data producer;
- timeliness: refers to the duration between the time when data becomes available and the time when the event or phenomenon it describes occurs;
- accessibility: measures how easily data can be located and retrieved
- interpretability: to how easily data can be comprehended, utilized, and analysed by users;
- coherence: refers to how logically connected and consistent data are with each other.

Data is stored in different formats in various sources, such as relational databases, flat files (e.g. CSV file), Electronic Data Interchange (EDI) and XML (Doan et al., 2012). Especially in the traditional databases the most used of these is relational database, which has proven to be an effective way to store and retrieve large amounts of data. Relational databases can combine data using the SQL language. (Kuchibhotla et al., 2009) Although relational databases are very popular and they will continue to be widely used, their weakness is their inflexibility with other systems, as this was not originally a requirement for data modelling (Reed, 2006).

JavaScript Object Notation (JSON) is a lightweight semi-structured data format language that has been replacing XML more and more over the years. Over the past years, it has emerged as the predominant data exchange format on the World Wide Web. (Lv et al., 2018) JSON is easy to read, and its structures align with concepts that software developers are familiar with, such as Arrays, Objects, and name/value pairs (Marrs, 2017).

Multi-model databases are systems that can store different types of data, such as JSON, XML, text, CSV (Laquer, 2017). Multi-model databases offer the user a single platform for different types of data, where you can search for information from different types of data with a unified query interface (Lu et al., 2018). It focuses on query functions, reduces integration challenges and eliminates migration problems (Lu & Holubová, 2019).

A dataset consists of instances with multiple attributes, each of which has a value. There are two types of data, labelled and unlabelled data. (Bramer, 2007; Serrano, 2021) In labelled data, there is a named attribute. If the named attribute has a category, its alternative values are predefined, for example course grades, and this is called classification. (Serrano, 2021) Classification is one of the most common operations in data mining. If the value is numerical and can vary freely, the numbers are continuous, and it is called regression. Unlabelled data has no named attribute. (Bramer, 2007)

2.2 Principles of Data Integration

The goal of data integration is to provide users with unified access to independent and heterogeneous data sources. Data can be located in different systems and be in different formats. (Ziegler & Dittrich, 2007) Data can be structured, semi-structured or unstructured data. Structured data encompasses relational, key/value, and graph data, while semi-structured data usually includes XML and JSON documents. Unstructured data refers to

text files that typically contain dates, numbers, and factual information. (Lu & Holubová, 2019)

According to Doan and others (2012) data integration can be viewed in small pieces as follows:

- **Queries**
  - The goal of data integration is to make queries from several different data sources.
- **Number of sources**
  - The more sources, the more challenging it is to search for information.
- **Heterogeneity**
  - Data is often retrieved from sources that have been developed separately from each other. Some may have a clear structure, such as databases, but at the other extreme there is no structure at all, such as a textual source.
- **Autonomy**
  - The sources can be managed by different organizations, in which case the data user's rights may be limited in different ways, e.g. the rights can only be to part of the data or it can only be searched at certain intervals. The owner of the data can also change the structure of the data and access to it at any time, without separately informing the user of the data.

To achieve the goal of data integration, the data must be presented according to the same principles (Ziegler & Dittrich, 2007). The process consists of three main tasks: schema matching, data matching, and data fusion (Christen, 2012), although the first two should have not be considered separately in data integration (Zhao, 2007). Schema matching involves recognizing corresponding database tables, attributes, and conceptual structures from different databases that hold identical types of information (Lenzerini, 2002). Data matching involves identifying and aligning individual records across multiple databases that refer to identical real-world entities or objects. Data fusion involves combining pairs or clusters of matched records into a unified and consistent record that represents the entity accurately. (Christen, 2012)

The data integration system contains a number of data sources and a mediated schema also known as global schema (Figure 1) (Dong et al., 2009). Users can submit queries on the mediated schema, either as a structured query or a keyword query. Queries serve multiple functions such as expressing users' information requirements, defining relationships between data sources, and creating named views of the database that can be reused in

other queries. (Doan et al., 2012) The user queries the mediator schema without having to know the details of the original data source (Golshan et al., 2017). The system reformulates the query to translate it into a set of structured queries on the mediated schema. The system then uses wrappers to interact with the data sources by submitting reformulated queries, retrieving answers and sometimes also applying basic transformations to the answer. (Dong et al., 2009) The mediated schema is custom-designed for the data integration application and includes only the relevant aspects of the domain (Doan et al., 2012).

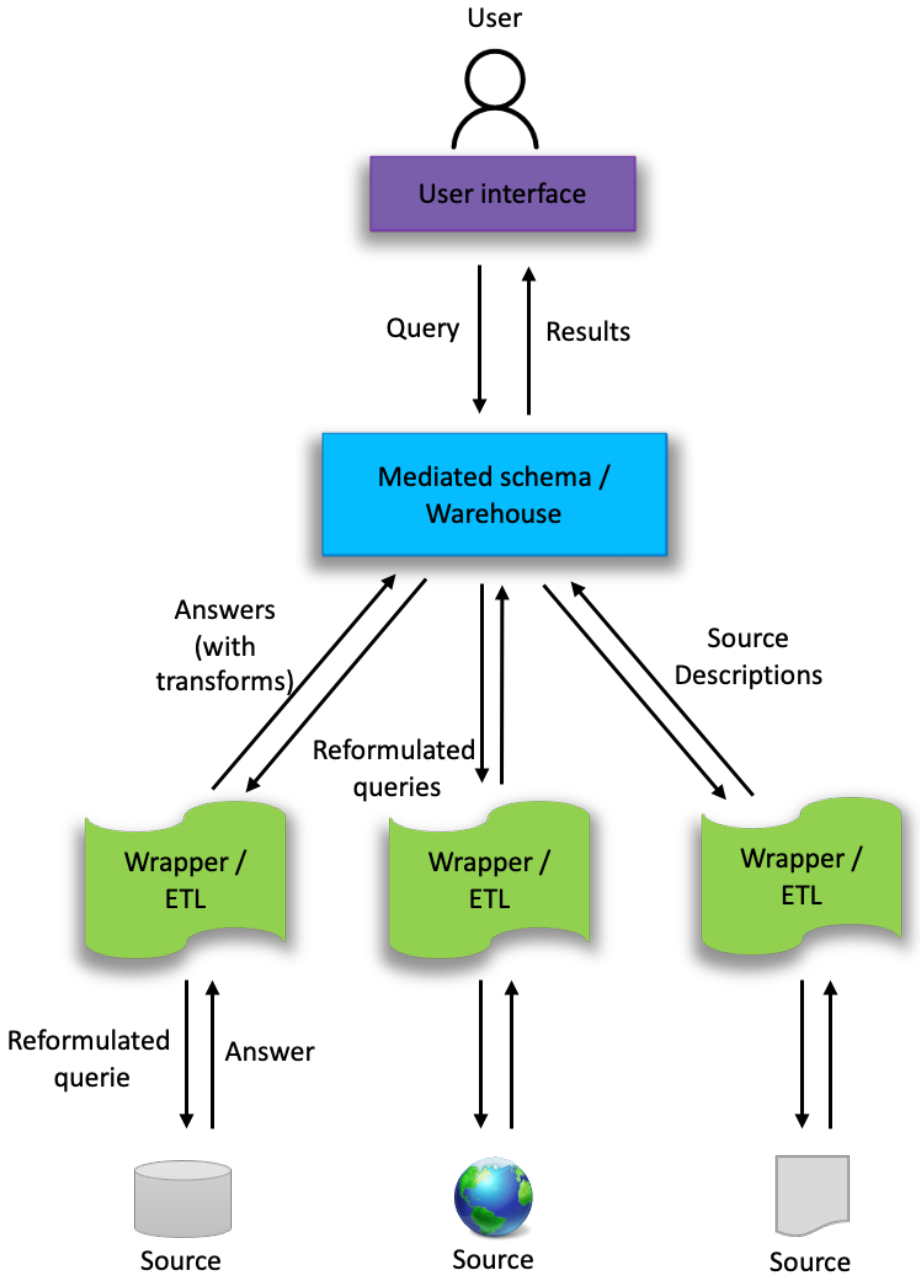

Figure 1. The data integration system.

Data sources can differ on several aspects, including the underlying data model and the types of queries supported. The critical element in developing a data integration application is the source descriptions that establish the connection between the mediated schema and the source schemas. (Halevy et al., 2006) Semantic mappings are the primary element of source descriptions as they establish the link between the mediated schema and the source schemas. Semantic mappings define the associations between attributes in the sources and the mediated schema, and the process for resolving differences in data values. (Doan et al., 2012)

Data integration systems can have different architectures, but broadly speaking all the architectures fit between two model: warehousing and virtual integration. In virtual integration, data remains in its original source and queries are made in real time when information is needed. (Doan et al., 2012) In warehousing, data is retrieved from individual sources and stored in a common database, warehouse, from which information is retrieved through queries (Lans, 2012). In the warehousing approach, mediated schemas are replaced by warehouse schemas and wrappers with extract-transform-load tool pipelines, ETLs. In contrast to wrappers, ETL often execute more complex data transformations, which can include cleansing, combining and converting values. (Doan et al., 2012) The warehouse schema includes not only the necessary source attributes, but also a physical schema that regularly retrieves data from the data sources and stores it in the warehouse (Ziegler & Dittrich, 2007). Although the architectures are different, many of the challenges discussed later are still related to both (Doan et al., 2012).

Essentially, data systems are not designed with integrations in mind. Therefore, whenever there is a need to access the system through integration, incompatible data must be combined using different reconciliation principles. (Ziegler & Dittrich, 2007) The task of data matching involves identifying structured data items that pertain to the same real-world entity (Christen, 2012). Data matching can arise in various integration scenarios. In a basic situation, there may be merging of multiple databases that have identical schemas but lack a unique global identifier and need to determine which rows are duplicates. (Christen, 2012; Doan et al., 2012) The task becomes more complex when we must combine rows from sources that have distinct schemas (Doan et al., 2012).

According to Golshan and others (2017), data integration was initially developed based on the following assumptions:

- The mediated schema is reasonably sized and can be constructed with reasonable effort.
- The data source is structured or at least semi-structured and has clearly defined schemas.
- All necessary data sources must be integrated simultaneously.
- All data integration functionalities must be available in the final data integration system.
- Most of the data in the sources is accurate and consistent or it can be brought into that form using standard methods.

These assumptions were challenged as the scale and nature of the data changed. The assumptions were based on the fact that in the beginning data integration systems processed data from a maximum of ten sources at once. (Golshan et al., 2017)

Although the goal of data integration is always to achieve a unified view of the data, the integration processes differ from each other (Ziegler & Dittrich, 2007). Figure 2 shows topics that affects the process.

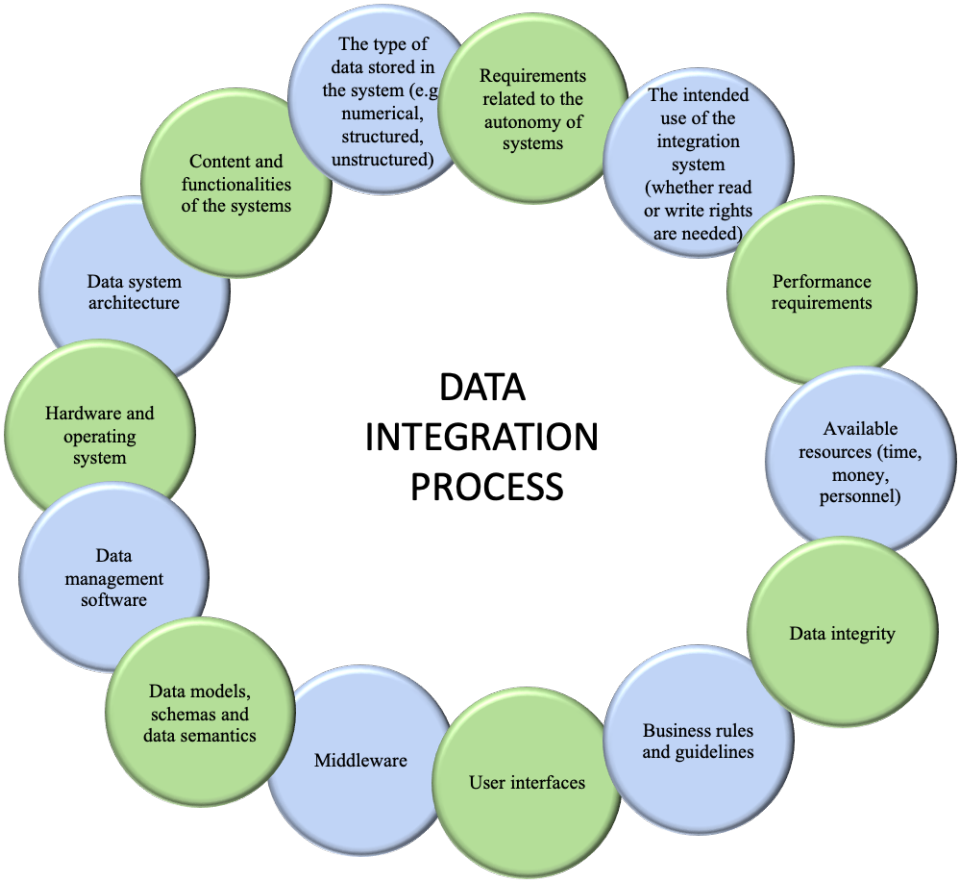

Figure 2. Topics that affect the data integration process (Ziegler & Dittrich, 2007).

2.3 Global-as-view and Local-as-view Approaches

There are two basic approaches to model the relationships between data sources and mediator schema (Lenzerini, 2002). The first is global-as-view (GAV) where the mediated schema is expressed in terms of data sources (Halevy et al., 2006; Lenzerini, 2002). The mediated schema serves as a view of the source schema, utilizing local schemas to describe intermediate schemas (Doan et al., 2012). Upon receiving a request through the global schema, the intermediate schema follows established guidelines and patterns to translate the request into an origin-specific query. The intermediate schema then sends a new query to the wrapper for processing. The wrapper examines all potential expressions and their possible combinations to fulfil the specific query. (Merieme et al., 2022)

The GAV provides a comprehensive view that is simple to design and execute since it grants control over the broker's activities (Merieme et al., 2022). The process of transforming a query from the mediated schema into a query on the data sources is conceptually uncomplicated, as the view definitions are used to unfold the query (Golshan et al., 2017).

The drawback of GAV is that any data that is not present in any of the source schemas cannot be represented by the schema, because it relates to multiple sources (Merieme et al., 2022). Adding new sources to the existing ones is a challenging task since it is necessary to ensure that the existing sources are dependent on it. As a result, independent sources are seldom added. (Doan et al., 2012) Furthermore, adding a new source would require modifications to the mappings. Removing a data source from the global schema may also prove difficult, making it inflexible. (Merieme et al., 2022)

The local-as-view (LAV) approach involves defining the mediated schema separately from the data sources and establishing a connection by defining each source as a view of the mediated schema (Halevy et al., 2006; Ullman, 2000). The schema is formulated to remain consistent, even as certain data sources join or exit the integrated system, providing LAV with the flexibility to include or exclude sources autonomously. The key advantage of the LAV approach is the capability to register distinct resources independently of one another. (Halevy et al., 2006) On the other hand, data integration systems that implement the LAV approach are comparatively more complicated than those that employ the GAV approach. To obtain answers to queries directed at the mediated schema, they must be transformed into corresponding queries to the local schema, a technique known as query rewriting using views. (Golshan et al., 2017; Merieme et al., 2022)

The approach called global-local-as-view (GLAV) combines GAV and LAV. In GLAV a view over the data sources is defined as a view over the global schema (Golshan et al., 2017; Merieme et al., 2022). GLAV combines the expressive ability of GAV and LAV enabling source descriptions that contain frequent queries on sources (Doan et al., 2012; Merieme et al., 2022).

2.4 Challenges of Data Integration

One challenge of data integration is the varying capabilities of the systems where the data is originally located, with some systems being accessible through complex SQL queries while others can only be accessed through basic web forms (Doan et al., 2012). Another thing is the complexity of integration, as it may not always be clear what it means to integrate data or how different sets of data can operate together (Halevy et al., 2006).

The integration of structured and unstructured data has been a persistent challenge for the data integration communities, with its significance only growing over time (Dong et al., 2009). This challenge stems from the fact that data integration systems manage both structured and unstructured content and must be able to handle queries that use both keywords and structured query languages (Golshan et al., 2017).

End users are not necessarily familiar with the schemas or the system itself is too broad provide accurate search methods, in which case keyword queries may need to be used. This leads to uncertainty between the keywords used and the structural queries formed from them. (Dong et al., 2009) Also, the data itself may be unreliable or out of date. The reliability and accuracy of the data is compromised, for example, when taken from unstructured sources using information extraction techniques. (Dong et al., 2009; Halevy et al., 2006)

Another challenge is the way the data is organized in the sources. In a structured database, information is often organized according to schemas. The problem is that these schemes have been designed in different ways, even if the requirements were the same. (Doan et al., 2012) Defining a schema in advance can be difficult and time-consuming in today's world, especially when needs can also change quickly. The challenge of a relational database is not the schemas themselves, but the fact that they usually require a specific, single schema and changing this schema require a lot of work. (Laquer, 2017)

The semantic mapping between the data source and the schema is not necessarily accurate and may be based on some degree of estimation (Dong et al., 2009). There may also be differences in the way the data is presented, even if the information is otherwise identical. A typical example in databases is, for example, storing a person's name. In some systems, the full name is stored in the same field, while in others there may be two fields; one for the surname and one for the first name. (Doan et al., 2012) The users may not have enough knowledge and skill to make accurate mappings, or the users do not understand the area enough, and therefore do not even know what the correct mappings are. There may also be so much source information that it is impossible to maintain accurate mappings. (Dong et al., 2009)

Then there are also social and administrative reasons. Data integration often relies to people collaborating and sharing data (Halevy et al., 2006). The original owner of the source system may want to restrict access to the information for many different reasons (Doan et al., 2012). Data integration involves finding right data, but access to data may be limited based on the user's role (Halevy et al., 2006). Restrictions are made e.g. because the data search loads the system. It may also be that someone wants to withhold information to maintain power within the organization. Naturally, today’s tightened requirements for the processing of personal data also require precise restrictions on data rights. (Doan et al., 2012)

According to Dong and others (2009), there are three types of uncertainty associated with data integration:

- Uncertain data
- Uncertain queries
- Uncertain scheme mappings.

Dong and others (2009) introduced an architecture for a data integration system that takes into account uncertainties related to integration and it differs from the traditional integration system with four points:

- Data model is based on probability and these probabilities must be attached to each tuple in data processing and to the schema mappings. These probabilities are used to prioritize the answers of the mapping.
- The system first formulates the keyword query into several alternative structural queries, after which it tries to infer different structural elements from them to form a more precise query for the data sources.

- Instead of trying to find all the answers, the goal is to find a certain number of the best answers and prioritize them efficiently.
- The processing of queries must be more flexible.

3 Missing Data

In the previous chapter, the principles of data integration were introduced. However, data that moves in integration is rarely complete.

3.1 Causes and Definition of Missing Data

It may contain erroneous values or entirely missing values (Sherman, 2014). A typical dataset often lacks information for certain variables or cases (Allison, 2002). There may be various reasons for the missing values, for example:

- errors in manual data entry
- failure or malfunction of the data storage device
- incorrect measurements
- in case of forms, fields have been added to the form after some of the forms have already been filled out
- data that could not be obtained (for example personal information) (Bramer, 2007; García et al., 2015).

Graham (2012) categorizes missing values into two different types: item nonresponse and wave nonresponse. These categories emerge especially in survey research. Item nonresponse happens when a survey respondent doesn't answer to some questions or does not complete certain segments of the survey despite filling out some portions of it. (Graham, 2012) For instance, when income is asked in surveys, a significant portion of respondents usually decline to respond. In self-administered surveys, individuals may inadvertently overlook or forget to answer some questions, and even experienced interviewers may sometimes neglect to ask certain questions. Additionally, certain questions may not apply to certain respondents, such as asking unmarried individuals to rate their marriage. (Allison, 2002)

Wave nonresponse pertains to longitudinal studies where the same group of individuals is measured at two or more points in time (waves). Wave nonresponse means that a respondent does not complete the entire measure or survey. (Graham, 2012) For example, some respondents may die or relocate before the next wave of interviews (Allison, 2002).

The best solution would always be to get the source data corrected, but this is rarely possible. In software development the data may come, for example, from the systems of another organization(s), whose data integrity cannot be influenced. Incomplete data can cause, for example:

- Loss of efficiency
- Complications in data processing and analysis
- Misleading results due to differences between complete and missing data (García et al., 2015).

Processing missing data practically always takes time (Kaiser, 2014). Methods and algorithms related to data processing are rarely able to directly process missing information, so such data usually requires data pre-processing, where these deficiencies are corrected in one way or another (García et al., 2015; Kaiser, 2014). However, the missing values are not necessarily completely independent and separate values, but often have a statistical connection with other corresponding existing values (Lakshminarayan et al., 1999).

Little & Rubin (2014) defines the missing data as follows:

> *"Missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value."*

Graham (2012) simplified the same thing as

> *"Missingness is the state of being missing."*

and stated that

> *"The value is either missing or it is not."*.

3.2 Missing Data Mechanisms

Little & Rubin (2014) stated that the mechanisms of missingness are critical since the properties of missing data techniques are heavily influenced by the nature of the dependencies in these mechanisms. Especially important is the fact whether the absence of variables is linked to the underlying values of the variables in the dataset.

Graham & Donaldson (1993) divides missing data mechanisms into two categories: accessible and inaccessible. Accessible means that the cause of the missingness is a variable that has been measured for all cases and is therefore available for analysis. In case of

inaccessible, the cause of the missingness has not been measured for every case or is otherwise unavailable for analysis.

Little and others (2022) classifies mechanisms of missing data into three categories:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Not missing at random (NMAR) or Missing not at random (MNAR).

The missing data mechanism is classified as MCAR when the probability of a missing value is independent of both the observed data and the value that is missing (Carpenter & Kenward, 2013; Kaiser, 2014). Graham (2012) describes MCAR like this:

> *"Cases with data for a variable, and cases with missing data for a variable, are each random samples of the total. This situation is achieved if the cause of missingness is a completely random process such as flipping a coin."*

Or as stated by Horton & Kleinman (2007) missingness does not result from any known or unknown factor. If this assumption holds for all variables, the group of individuals with complete data can be regarded as a randomly selected subset of observations from the original set (Allison, 2002).

An example of such missing data is the loss of a laboratory sample during the process (Kaiser, 2014). Even though MCAR is a strong assumption, it can be justifiable in some circumstances, particularly when data are missing due to the research design. Such designs may be used, for example, when a variable is very expensive to measure. In this case, costly variable is measured only for a random subset, which leads that the data is missing completely at random for the remaining sample. (Allison, 2002)

Missing at random (MAR) occurs when the probability of missing values is related to the observed data, but not the missing value itself (Allison, 2002). Data that is incomplete due to structural reasons is classified as MAR (Kaiser, 2014). This does not imply that the probability of observing a variable on an individual is independent of the variable's value. On the contrary, in the context of MAR, the likelihood of observing a variable is influenced by its value. (Carpenter & Kenward, 2013) But as Allison (2002) noted, it is not possible to verify if the MAR condition is met, because the missing data values are unknown, which prevents us from comparing complete and incomplete data values to determine whether they systematically differ on that variable.

Graham (2012) gives a reading speed as an example of the MAR. With the long time-limited survey, faster users will fill the whole survey, while slower ones may not answer every question especially at the end of survey. Nonetheless, it is possible to evaluate reading speed early on in the survey when most participants are expected to provide responses. As a result, any biases that may stem from reading speed can be managed by integrating the reading speed variable into the model for analysing missing data.

When the mechanism behind missing data is not MCAR or MAR, it is referred to as Missing Not At Random (MNAR) or Not Missing At Random (NMAR) (Carpenter & Kenward, 2013). MNAR occurs when the absence of missing values is related to the missing values themselves. One way to address this issue is to revisit the primary data source and trying to either find the missing value or find a mechanism to infer it. (Kaiser, 2014) While in some situation MNAR may be more plausible than MAR, conducting an analysis under MNAR is significantly more challenging (Carpenter & Kenward, 2013). Graham & Donaldson (1993) called this type of missingness as inaccessible missingness, because the reason for the missingness has not been quantified and, as a result, is not usable for analysing. An example of NMAR is the measure of income in survey research. People with higher incomes are more likely to leave the income answer blank. (Graham, 2012)

One more way to categorize missing data mechanisms is ignorable and nonignorable. Mechanism is ignorable if

1. Data is MAR and
2. The parameters that regulate the missing data process are independent of the parameters that require estimation (Allison, 2002; Zhou et al., 2014).

Ignorability basically implies that modelling the missing data mechanism as a component of the estimation process is unnecessary. Nonetheless, specific techniques are required to effectively utilize the data. (Allison, 2002; Buuren, 2012)

If the data is not MAR, the missing data mechanism is referred as nonignorable. In such situations, it is typically necessary to model the missing data mechanism to acquire precise estimates of the relevant parameters. (Allison, 2002; Zhou et al., 2014)

4 Strategies for Dealing with Missing Data

As Allison (2002) stated,

> *"The only really good solution to the missing data problem is not to have any".*

But in real life this is rarely possible. Incomplete data analysis typically involves methods where missing values are either completely ignored or replaced with other values inferred from the available data (Voillet et al., 2016). Graham (2012) stated that handling missing data involves making assumptions regarding plausible values of the missing data. The choice between different methods depends on the nature and quantity of the available data, the intended use of the data, the users of the data, as well as the extent of missing values (Lakshminarayan et al., 1999).

Traditional commonly used strategies for dealing missing data are, for example, complete-case Analysis and different single imputation approaches. Commonly held belief is that missing data solutions usually yield satisfactory results at best, even when the quantity of missing data is moderate and the missing data assumption is satisfied. (Ratner, 2011) Bramer (2007) claimed that none of these methods is more reliable than the others for all possible datasets. A key issue with all the traditional imputation methods is that conclusions drawn from the imputed data fail to consider the uncertainty associated with imputation (Little & Rubin, 2014). There are also newer approaches like multiple imputation and maximum likelihood that try to overcome the weaknesses of the previously mentioned methods (Allison, 2002).

4.1 Complete-case Analysis

When dealing with missing data, one common approach is to simply discard incomplete cases, a strategy referred to as complete-case analysis, listwise deletion or casewise deletion (Allison, 2002). As all instances with missing data have been eliminated, there is no longer a need to address the issue of missing data. This approach is easy to implement and may work well when there is a lot of data, but relatively small amount of missing data. (Mockus, 2008; Zhu, 2014) However, this approach may cause significant biases and may not always be effective, especially when making inferences for subgroups (Little & Rubin, 2014).

Assuming the data is MCAR, the reduced sample will represent a random subset of the initial sample, and estimates obtained using listwise deletion will be unbiased. However, if the data is MAR, listwise deletion may produce biased estimates. (Allison, 2002; Little et al., 2022)

One advantage of complete-case analysis is its simplicity. It allows for standard statistical analysis to be applied without modification, and univariate statistics can be compared across variables because they are all calculated on a common sample base of units (Little & Rubin, 2014; Ratner, 2011). It also does not require special computational methods (Allison, 2002) and it does not cause incorrect data to enter the source (which happens in many other approaches) (Bramer, 2007).

However, the approach also has significant disadvantages, primarily the potential loss of information that results from removing incomplete cases (Little et al., 2022). This loss of information can lead to a decrease in precision and introduce bias, especially when the missing data mechanism is not MCAR. The degree of bias and precision loss is influenced by the proportion of complete cases, the distribution of missing data, and the degree of dissimilarity between complete and incomplete cases. (Little & Rubin, 2014)

Despite its limitations, listwise deletion is not necessary a bad method for handling missing data (Allison, 2002; Little et al., 2022). While it does not use all available information, it can still provide valid inferences when the data are MCAR (Allison, 2002). Little and others (2022) stated that for some regression problems listwise deletion is optimal, and multiple imputation is actually less efficient. Alternatives to listwise deletion, such as imputation methods, may provide better results, but they also require more complex computational methods and assumptions about the missing data mechanism (Allison, 2002).

4.2 Pairwise Deletion

Pairwise deletion, which is also referred to as available case analysis, is a strategy for managing missing data that focuses on computing summary statistics for available cases rather than excluding them entirely from analysis (Shi et al., 2020). This approach is applicable to a range of linear models, such as linear regression, factor analysis, and intricate structural equation models. The idea behind pairwise deletion is to calculate the summary statistics using all the cases that are available, rather than only those that have complete data. (Marsh, 1998)

This allows more information to be utilized, potentially making pairwise deletion more efficient than listwise deletion. However, the implementation of pairwise deletion is not straightforward, and there are ambiguities in how to compute the summary statistics. (Allison, 2002) The variations do not lead to significant differences in estimators' properties, but they can lead to biased estimates and test statistics, particularly when missingness is not missing completely at random (MCAR) (Shi et al., 2020).

One significant limitation of pairwise deletion is the inability to accurately estimating standard errors. Estimation of standard errors involves identifying the sample size, which is not obvious with pairwise deletion. (Shi et al., 2020) Additionally, pairwise deletion may produce covariance or correlation matrices that lack positive definiteness, making regression computations impossible. These difficulties, along with its sensitivity to departures from MCAR, limit pairwise deletion's general recommendation as an alternative to listwise deletion. (Graham, 2012)

In practice, pairwise deletion is not recommended for parameter estimation, as better parameter estimates can be obtained using other methods. Therefore, despite its potential advantages in using all available data, pairwise deletion has several limitations and may not be the optimal approach for handling missing data. (Graham, 2012; Marsh, 1998; Shi et al., 2020)

4.3 Single Imputation

Many methods of handling missing values fall into a category called imputation (Allison, 2002). The process of imputation can be described as any technique that fills in missing data to produce a complete dataset. The goal of imputing missing data is to restore or reduce the loss of information resulting from incomplete data. (Ratner, 2011) The fundamental concept underlying imputation is to replace each missing value in a dataset with a reasonable estimate or guess, and then proceed with data analysis as if no data was missing (Allison, 2002).

Imputation has several advantages, including being flexible and producing a comprehensive dataset that can be analysed using standard methods and software. This practical utility of applying preferred technique or software can be of significant value to data users. (Little & Rubin, 2014) Compared to listwise deletion, imputation is potentially a more

efficient method since it does not sacrifice any units. By preserving the complete sample, imputation can prevent the loss of power caused by a reduced sample size. It can maintain high precision, if the observed data contains valuable information for forecasting the missing values and this information is used in the imputation process. (Schafer & Graham, 2002) Moreover, imputation can prove to be advantageous in cases where data is analysed by multiple individuals or entities. This is because performing imputation before all analyses helps guarantee that all entities are considering the same set of units, which, in turn, makes it easier to compare results. (Rässler et al., 2013)

But like all the missing value approaches, it has disadvantages. One of the pitfalls of imputation is that it can be challenging to implement well, especially in multivariate settings (Schafer & Graham, 2002). Treating imputed data as complete data results in underestimated standard errors and overestimated test statistics (Allison, 2002). These methods do not account for imputation uncertainty (Little & Rubin, 2014). Therefore, it is crucial to choose the appropriate imputation method based on the observed data, and to ensure that the imputation method used is reliable and accurate. If the assumptions made during imputation are incorrect, this can lead to biased or unreliable results. Such substituted values should not be treated the same way as complete data. (Kaiser, 2014)

Little & Rubin (2014) stated that it's important to generate a method for creating a predictive distribution for the imputation based on the observed data and they divided these methods into two categories:

- Explicit modelling: The predictive distribution is based on a formal statistical model and therefore, the assumptions are explicit. Methods like mean imputation, and regression imputation.
- Implicit modelling: The emphasis is on an algorithm, which could suggest the existence of an underlying model. While the assumptions are implicit, they still require thorough evaluation to ensure their reasonable. Methods like hot deck imputation and cold deck imputation.

Last Observation Carried Forward (LOCF)

The Last Observation Carried Forward (LOCF) method is a straightforward approach for dealing with missing values. It replaces every missing value with the corresponding last observed value (Lachin, 2016; Mavridis et al., 2019) and is commonly used in longitudinal studies of continuous outcomes under the assumption of Missing Completely at Random (Lachin, 2016). The LOCF method assumes that the outcome remains the same after

the last observed value, and thus, no time effect exists since the last observation (Zhu, 2014). Due to its simplicity and ease of implementation, LOCF has become a popular method for handling missing data problems especially in clinical trials (Mavridis et al., 2019). Unlike the complete case (CC) method, LOCF does not reduce the sample size (Zhu, 2014). But Lachin (2016) stated that LOCF can only be considered unbiased if the missing data is purely random and the data used for LOCF imputation has an identical distribution to the missing data. Lachin (2016) continued saying that as it is impossible to demonstrate that the distributions are completely identical, all LOCF analyses remain questionable.

Mean Imputation

Mean imputation is a method of dealing with missing values by replacing them with the mean value of the cases that have data available on that variable (Allison, 2002; Ratner, 2011). This method can only be used for numerical values. Often, the most common attribute value is used in conjunction with this method when a numerical value is accompanied by a symbolic attribute. (Kaiser, 2014)

Both Allison (2002) and Graham (2012) strongly advised that this method should be avoided as it produces biased estimates of variances and covariances there is no easy way to estimate standard errors.

Regression Imputation

When utilizing regression imputation, the missing values in a dataset are replaced with predicted values that are based on a regression analysis conducted on the existing variables for that particular data point (Little & Rubin, 2014; Mohideen et al., 2021; Ratner, 2011). If variable X has missing data for some cases and Y's are the matching variables, then X is regressed on the Y's using complete-case analysis dataset (Allison, 2002). Little & Rubin (2014) stated that mean imputation can be seen as a variant of regression imputation where the predictor variables are represented by dummy indicator variables for the imputed mean values.

The most common attribute imputation

In this method, the missing value is replaced by the most common value of the attribute. (Grzymala-Busse & Hu, 2001) For example, if out of a hundred attribute values, 80 have

the value X, 12 have the value Y, and 8 have the value Z, it might make sense to replace the missing value with the value X. However, this may not be a reliable approach in cases where the occurrence percentages of values are very close, such as 35%, 35%, and 30%. (Bramer, 2007)

Hot Deck Imputation

Hot deck imputation replaces a missing attribute value in a record by drawing a value from a distribution estimated using the available data in the sample (Little & Rubin, 2014). Essentially, this method works by finding a similar case with a known variable value and using that value to fill in the missing value in the current case (Lakshminarayan et al., 1999). Hot deck imputation is frequently utilized especially in survey practice and may entail complex schemas for selecting similar units to carry out the imputation process (Little & Rubin, 2014).

One of the advantages of this method is that it imputes actual values, which leads to more realistic outcomes. Additionally, it doesn't rely on strict parametric assumptions and can integrate covariate information. A weakness is that finding suitable matches between existing and missing values is essential and it is more challenging in smaller samples compared to larger ones. (Andridge & Little, 2010)

Cold Deck Imputation

Cold deck imputation is similar to hot deck imputation, but differs in one way: it replaces a missing value for a given attribute from an external source and not from the current data sample (Lakshminarayan et al., 1999; Little & Rubin, 2014).

Closest Fit

Closest fit approach replaces the missing value with an existing value of the same attribute that resembles the case of the missing value as closely as possible (Grzymala-Busse et al., 2005). The problem with this approach can be, for example, if the value is replaced with the value of an instance that happens to be an outlier in the data. This can be at least partly replaced by using several closest cases. (Kaiser, 2014)

k-Nearest Neighbour

While the previous method searched for only one closest instance, here multiple instances are sought. The k-nearest neighbour method is commonly used for object classification, where the k-nearest data points to the object are searched for, and through this, the nearest class to the object is determined. (Kaiser, 2014; Sessa & Syed, 2017) The disadvantage of this method is that the algorithm goes through the entire dataset, which is why it cannot necessarily be used in very large amounts of data (Sessa & Syed, 2017).

4.4 Multiple Imputation

Multiple imputation (MI) was first proposed by Rubin (1977) to handle nonresponse problems in large surveys so that data users could analyse completed dataset. The method takes into account the uncertainty associated with disturbances in the results caused by missing data (Lakshminarayan et al., 1999). Nowadays multiple imputation is considered as one of the most prevalent and adaptable statistical methods for addressing missing data issues and is extensively utilized across various fields of study (He et al., 2021).

MI involves imputing a missing value multiple times to generate multiple data sets (Zhu, 2014). The fundamental steps of MI are:

1. Predictive distribution for the missing values is estimated based on the observed values in the dataset;
2. The missing values are replaced with randomly selected draws from the predictive distribution, which are not the means of the distribution;
3. The second step is repeated X times (where $X > 0$) to create X datasets, each of which contains different sets of draws for the missing values. (Little et al., 2022)

This results in multiple completed datasets, each with both original and imputed values (He et al., 2021). By generating draws from the predictive distribution, imputation introduces variance into the estimates across MI datasets, thereby enabling the appropriate evaluation of imputation uncertainty. After imputing the missing data, the subsequent steps in the MI procedure are not significantly more complex than those involved in a single imputation approach. (Little et al., 2022)

One thing to be considered when using MI is how many imputations is needed (Zhu, 2014). Rubin (1987) claimed that good inferences can be done for only 3 - 5 imputed data

sets. Schafer (1997) stated that over ten imputations is needed when the fraction of missing information is very large. Bodner (2008) recommended that the number of imputed datasets should be proportional to the percentage of individuals with missing data. (Allison, 2002) stated that good imputation methods use all information related to missing cases.

When using Multiple Imputation (MI) to handle missing data, there are certain assumptions and limitations to keep in mind, such as:

1. Missing data mechanism must be MAR;
2. The imputation model should align with the analysis model;
3. The imputation algorithm needs to take into account the variables that are associated with the missingness of the data (as well as any related variables) (Zhu, 2014).

MI has two significant advantages, including the ability to utilize complete-data methods for data analysis and the incorporation of random errors in the imputation process. MI can be employed with any model and data type without the need for specialized software. (Allison, 2002; Zhu, 2014) Furthermore, MI enhances the efficiency of the estimates by minimizing standard errors (Little & Rubin, 2014).

Little & Rubin (2014) stated that only drawback of MI over single imputation is the additional effort required to generate the imputations and analyse the results, as well as the need for more data storage. Nonetheless, with the current state of computing technology, the storage demands are often insignificant, and the analysis process is not significantly more challenging since it essentially involves carrying out the same task X times instead of once. Allison (2002) added that MI is easy to do in the wrong way, but using good software to do the imputations reduces this risk. Allison (2002) also stated that MI produces different estimates in every use which can lead to weird situations where different researchers obtain different results from the same data while using the same methods.

4.5 Maximum Likelihood

The basic concept behind Maximum Likelihood (ML) estimation is to select estimates that maximize the likelihood of observing what has been observed. This necessitates a formula that can express the probability of data as a function of both the data and the unknown parameters. (Allison, 2002; Newman, 2014) In cases where observations are

independent, the overall probability for the sample can be obtained by multiplying the individual likelihoods for each observation (Allison, 2002). ML offers a high degree of flexibility and explicitly articulate the fundamental model assumptions, enabling them to be evaluated (Ibrahim et al., 2005). Although ML is a significant improvement over conventional methods for handling missing data, it has certain limitations. While ML theory and software are readily available for linear and log-linear models, they are generally not available for more complex models. (Allison, 2002)

An essential difference between MI and likelihood approaches is that in likelihood methods, the missing values are addressed within the model-fitting process, whereas in MI, they are addressed prior to the analysis (Schafer & Graham, 2002). Maximum likelihood or multiple imputation can be performed even when the assumption is that the data is not missing at random. However, obtaining accurate results can be challenging, because these methods are highly sensitive to assumptions regarding the missingness mechanism or the distributions of variables with missing data. (Allison, 2002)

5 Research Approach

As in many other business areas, the sales and leasing services of cars have also become increasingly digital. Potential buyers often first examine the range of cars available on dealers' websites before physically going to the store. Nowadays, more and more cars are also bought or leased directly online without ever visiting a physical store.

5.1 Background

According to a study by Paytrail (2019), the share of online purchases for cars and other vehicles in Finland was 8% of all product purchases made online in 2019 (Figure 3).

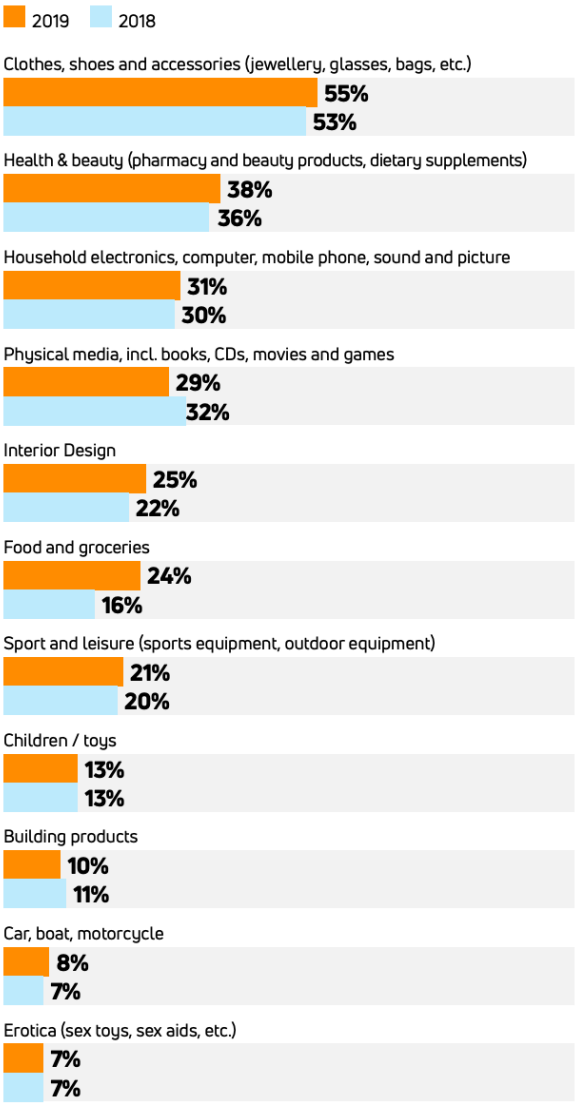

Figure 3. Share of consumers who buy physical goods online (Paytrail, 2019).

Danske Bank (2018) reported already a few years ago that 16 percent of Finns would be willing to buy a car entirely online and almost 90 percent used the internet as their source of information before their last car purchase. Large car dealerships, such as K-Auto and Saka, have a strong focus on digitalization in their strategy and development plans (Kesko, 2023; Saka, 2023). In Finland, there is also online-only players such as Beely, which offers leasing cars.

5.2 The Purpose of the Research

The aim of this thesis was to investigate how to replace the missing values of car data integrated into the websites of companies selling or leasing cars. The case company had started or was about to start internal company projects on the topic, because the missing data caused challenges in the web page developing.

The research questions of this thesis are:

- What are the different methods to replace missing data in data integration?
- How are these methods suitable for replacing the missing data in car data integrated to the web pages of car dealerships?

The purpose was to compare and find ways for the case company to manage missing data in data integration, using literature and discussions in work groups. Through literature, the goal was to find as many different ways as possible to manage the missing data and then analyse these methods from the perspective of car data. At the same time as the research was being conducted, the example company had an internal project underway to find solutions specifically for missing electric car data. The writer of this thesis participated in this internal project. In addition, the writer had several free-form discussions with different people in the case company, such as developers and product owners, to clarify the data integration process as a whole and to understand the extent of the various challenges related to missing data.

5.3 The Case Company

The case company is Crasman Oy, which produces various digital services for large and small companies in Finland, such as Intersport, Messukeskus and Familon. One of the

company's large customer segments are also car dealerships, where the goals of this thesis are focused. Crasman was founded in 1996 and offers services in the following areas:

- Strategy and Consulting
- Design
- Development and Technologies
- Digital Marketing and Content Services (Crasman, 2023).

Crasman's goal is to act as the customer's digital partner. Crasman employs a little over 100 people and has offices in three different locations in Finland: Helsinki, Tampere and Joensuu. (Crasman, 2023) The case company was chosen because the writer of this thesis works there as a web developer and has been participating in the company’s internal project related to managing missing data.

6 Managing Missing Data in Car Data Integration

Car dealers' websites usually display all the cars that the company has for sale. The features and other information of the cars are listed precisely on the pages. According to an article by Autovista24 (2021), it is particularly important in the online car market that all the details are described accurately to avoid unsatisfied customers.

6.1 Data Integration for Car Data

The car dealership usually receives the basic information about cars from an external provider, whose system is also used to handle the entire car purchase process. Companies producing such services in Finland include, for example, Alma Ajo (Websales) ja Netwheels (GT-X) (Alma Ajo, 2023; Netwheels, 2023). The original source of car data is often the Finnish Transport and Communications Agency's, Traficom's, registration data, which is then enriched by the before mentioned companies. Traficom offers this car data as open data that is available to everyone (Liikenne- ja viestintävirasto, 2023). Appendix 1 contains a table of all the information produced by Traficom.

The car data contains both continuous and categorical values. Continuous values include, for example, model year, power and engine capacity. Categorical values include, for example, body type, fuel type and transmission. Table 1 on the next page shows more extensive examples of the attributes of a single car and their values, and whether they are continuous or categorical.

In addition to the basic information, the car salesperson can enter additional information for the car, such as information related to the car's equipment and of course price. This additional information is entered into the same system where the car data is located.

Attribute	Example values	Datatype
Make	BMW Peugeot Volkswagen	Categorical
Model	X3 408 Golf	Categorical
Year	2020	Continuous
Body type	SUV Sedan Hatchback	Categorical
Fuel type	Gasoline Diesel Electric Plug-in-hybrid	Categorical
Drive type	All wheel drive Rear wheel drive	Categorical
Transmission	Automatic Manual	Categorical
CO2 emissions	158 g/km	Continuous
Power	180 hp	Continuous
Fuel consumption	5,6 l/100km	Continuous
Acceleration	7.2 sec	Continuous
Doors	5	Categorical
Seats	5	Categorical
Engine capacity	1998 cm3	Continuous
Length	4713 mm	Continuous
Width	1827 mm	Continuous
Height	1445 mm	Continuous
Weight	2150 kg	Continuous
Electrical range	435	Continuous
Battery capacity	78,1 kWh	Continuous

Table 1. Examples of car data.

A company developing car dealer's website integrates data into its own database. Since the data is stored in its own database and is not retrieved from the original source every time it's used, the architecture of the data integration is called warehousing. At the same time data is integrated, the data is modified and enriched in different ways. Data is retrieved as semi-structured JSON format. Updated data is retrieved from the external system regularly, for example a few times a day. Figure 4 shows the architecture of data integration in car dealer's website.

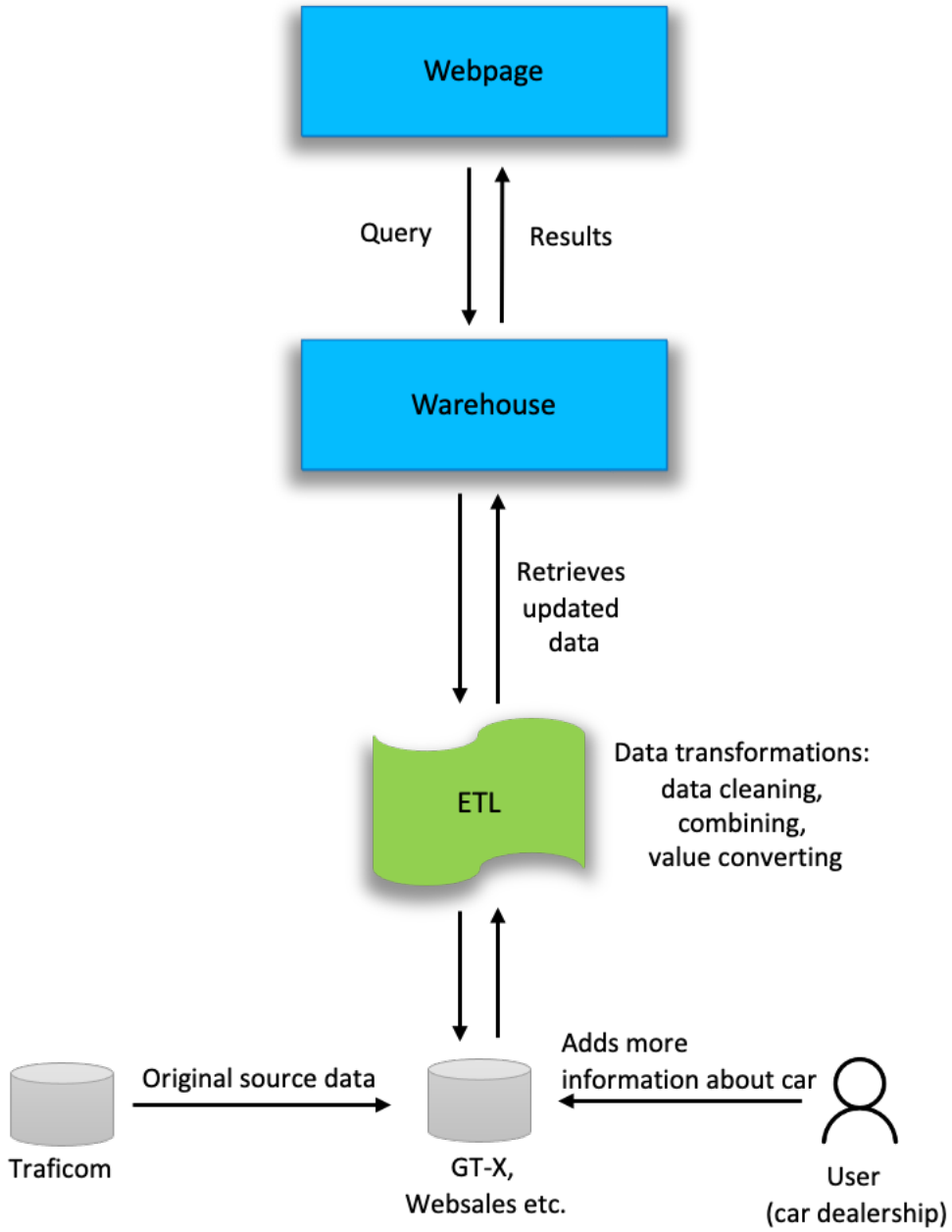

Figure 4. Architecture of data integration in car dealer's website.

It is not really a complex integration process, as data is only retrieved from one source at once. As is stated in the previous sections, the data is rarely perfect, and in this case, too, the data is occasionally incomplete in terms of some essential attributes. In practice, missing data can be divided into two groups:

1. Occasionally missing individual data
2. Data that is always missing.

Incomplete data belonging to the first group is due, for example, to an error that has occurred at some point in the manual data entry, either because the data has not entered at all or because the information was entered in the wrong format, causing the external system to fail to integrate it into its own system. At this case, since missing values does not depend on existing or missing values, it can be concluded that the missing mechanism is Missing Completely At Random, MCAR. Therefore, a certain value may be present in one car but missing in another for some reason. Potential buyers often search for cars on websites using various filters to narrow down the extensive selection of cars offered by dealerships. If the car lacks some information that the user wants the car to have, such a car will not appear in the search results at all. For example, if the buyer wants a car with at least 150 horsepower, but the car's information does not include horsepower at all, the car will not appear in the search results.

The second group currently includes critical values especially related to electric cars. As can be seen from the data listing in Appendix 1 from Traficom, it does not contain any data specific to electric cars, such as the car's range or battery size. In this case, the missing data mechanism is Not Missing At Random, NMAR, because it is always certain attributes whose values are missing. As Kaiser (2014) defined NMAR, absence of missing values is related to the missing values themselves. Since the data do not come directly from any system, the car salesperson must enter it manually into the system. In addition to the fact that such a process takes up resources, it is also prone to various errors when the information is entered manually.

As Allison (2002) stated “The only really good solution to the missing data problem is not to have any”. However, the developer of the web pages cannot fix the original data because it comes from an external source. The company must come up with a way to enrich the data in terms of missing information.

6.2 Analysing Strategies for Dealing with Missing Car Data

It is important to note that the different ways of handling missing values presented in the literature and in the previous sections of this thesis are often based on situations where the data is used in various studies and analyses. In these cases, calculations, distributions and trends are derived from all the data, so replacing a single value with an approximate value does not necessarily affect the final result. In the case of car data, the single attribute

value is used as is, which is why it should be exactly correct. As stated earlier, a potential customer may make their purchase decision based on the information on the dealership's website, in which case the lack of information may lead to the buyer not ending up with the car in question.

Complete-case analysis would drop the entire data of the car, which of course in this case is a completely excluded option. It would also be possible to ignore the missing value and continue using the data as it is, which would resemble pairwise deletion to some extent. But as previously noted, this would result in the car not being found when using search filters that include the attribute with the missing value.

Imputation methods that rely on adding approximate values to missing values are also problematic in this case. Such methods include, for example, mean imputation, regression imputation and the most common attribute imputation. Using approximate values is easier in a situation where summaries are made of data, where a single value may not have that much weight. In this case, however, the values must be exactly correct, and approximate values are not acceptable. For example, consider a car with a missing fuel consumption rate. In this case, the missing consumption would be, for example, 6.9 liters/100km, and as a result of the imputation, the value would be 7.1. Once again, the car might be excluded from search results if the user is searching for cars with a fuel consumption rate below 7 liters/100 km. The small deviation in the value may be irrelevant to one buyer, but significant to another. Since buyers have different preferences and values for car features, it is impossible to determine definitively which approximate values are acceptable and which are not, and to what extent.

Additionally, calculating the values of the same attributes across the entire population of cars would result in significant bias, because the data for different car makes and models differ greatly from each other. All of the above also applies to methods such as multiple imputation and maximum likelihood, because the values are not exact in these cases either.

On the other hand, if approximate values were used, they might cause more distortion in the values in the future. If the same attribute's value were missing again from a car and an attempt was made to replace it based on existing data, the initially imputed value would be included, which could further distort the actual value.

Hot deck imputation, Closest fit and k-Nearest Neighbor methods actually give the same and correct result in the case of car data, because they are all based on finding similar cases in the data. If a similar case can be found flawlessly, its attribute values are the same as those in the missing data, and it does not matter whether one or more of these similar cases are found, because one is enough. When using hot deck imputation, one thing to decide is whether the use of the same similar case should be numerically limited. In this case, the cars in the database vary as new ones are released and old ones are sold. Therefore, there is no need to make a restriction. In addition, the base assumption is that the existing car data is correct, which is why the data can be replaced according to one case. There may also be so few cars of the same model and year in the database that it is impossible to find more than one match.

If a similar case cannot be found in the same data set, there should be a possibility to search for a similar case from external sources. In this case, the method of the choice is cold deck imputation. This method should also be used if certain source data is always missing, such as the previously mentioned data of electric cars.

The method of Last Observation Carried Forward would be suitable in practice due to the merits of the immutability of car data, if this is implemented in such a way that the last entered value is taken from the similar case car. However, the method is primarily designed for longitudinal studies, and in this case entering exactly the latest value does not bring any improvements to data compared, for example, to hot imputation method, because the values of similar cars remain the same all the time. In fact, it might also be more difficult to implement, because all entered data needs to have timestamp in this case. Another problem is that if no similar car is found in the dataset, this method cannot be applied.

In conclusion, the most critical aspect of replacing missing values in car data, is to find a similar case, which will be covered in the next section.

6.3 Defining Similar Case

The key attributes for finding a similar case are make, model, year and model specification of the car. Matching make, model and year is fairly easy, because these attribute values usually come in a consistent format from the source systems, or they can be easily transformed into a consistent format with the help of ETLs. The situation is more difficult

with the model specification. It is a text field that car salesperson can modify freely. In this case, the model specification may have been written in different ways. In addition, the model specification may contain a lot of other description related to the car, such as equipment information. Therefore, the best way to match two model specifications is through string matching, which can lead to relatively reliable results if a suitable algorithm is found for the matching. The problem of string matching involves identifying strings that pertain to the same real-world entity (Doan et al., 2012) and it's used often in data integration (Huang & Madey, 2004). However, finding a match in this way is not 100% certain, unless the model specifications happen to be identical. Table 2 shows examples related to model specifications.

Make	Model	Model description
BMW	320	F31 Touring 320d A xDrive Business Luxury (HUD, Koukku, Sporttipenkit, Nahat, YMS!)
BMW	320	G20 Sedan 320i A xDrive Business M Sport // Koukku / HiFi / Tutkat **BPS takuu 24kk**
BMW	320	F31 Touring 320d A xDrive Edition M Sport **Prof Navi, Kangas/Alcantara, Koukku, LED**
Peugeot	308	Allure e-HDi 115 FAP
Peugeot	308	SW Active PureTech 130 Automaatti
Peugeot	308	GT Hybrid First Edition 225 EAT8 // Navi / Tehdas takuu voimassa / Peruutuskamera / 225hv
Volkswagen	Golf	Golf 1.6 FSI Hatchback
Volkswagen	Golf	Variant GTD 2,0 TDI 135 kW (184 hv)
Volkswagen	Golf	GTE Plug-In Hybrid 110 kW (150 hv) DSG-aut
Tesla	Model 3	Long Range Dual Motor (MY22) // Vetokoukku / Lasikatto / Connectivity Paketti / Autopilot
Tesla	Model 3	Performance Dual Motor AWD (MY20)
Tesla	Model 3	Dual-Motor Long Range 77kWh

Table 2. Examples of model specifications.

The requirement for the level of accuracy in string matching depends on the attribute that is missing. Many attributes need very accurate match in model specification. These kinds of attributes are, for example, engine capacity, power and fuel consumption. These attributes varies, if the model isn't exactly same. Attributes such as body type or physical dimensions, like length and weight, are possible to match with string matching of less certainty. These are kind of dimensions that do not change, for example, if the car's drive type, engine capacity or power varies. In fact, with many cars model specification is not even needed to compare these attributes. Only car models that have different body types with the same model name do need the model specification in matching. These are cars like BMW 3-series, Peugeot 308 and Volkswagen Golf that have sedan, minivan and sometimes hatchback models with same model name. And if car has many body types, there is usually only one specific word that needs to be match in model specification, for example Touring, Variant, SW.

Since finding similar cases is a little bit different, whether car model has different body types or not, it might be a good starting point to have this information in the database. It would be quite easy to create database that holds those car models that have multiple body types. Then when the similar car is searched, this database can be used when deciding whether finding the similar car needs string matching or not. Also the attributes should be grouped according to their need for weaker or stronger string matching.

The challenge of using string matching is that it is an approximate method that may not give a completely accurate answer. One approach to gaining certainty in finding the correct case is to compare other corresponding attributes of the car to each other. This kind of attribute group could be, for example:

- CO_2 emissions
- Power
- Fuel consumption
- Acceleration

Example: a car has a missing value with one of these attributes and similar car is found through string matching with some uncertainty. If both cars have same values with other three existing values, then missing value can be replaced with corresponding value of similar car without uncertainty.

Figure 5 presents the process of finding similar car as a flow chart. This process requires two steps before implementation:

1. Defining attributes that need always full string matching of model description, if value is missing, and creating database for the attributes.
2. Defining car models that have multiple body types and creating database for them.

This process starts by finding cars with same make, model and year. Then there is checking if the missing attribute belongs to the group that not necessarily need full string matching of model description. If it does belong to this group then there is a check, if car model with missing data has many different body types. Then the replacing of missing value is done either with or without string matching.

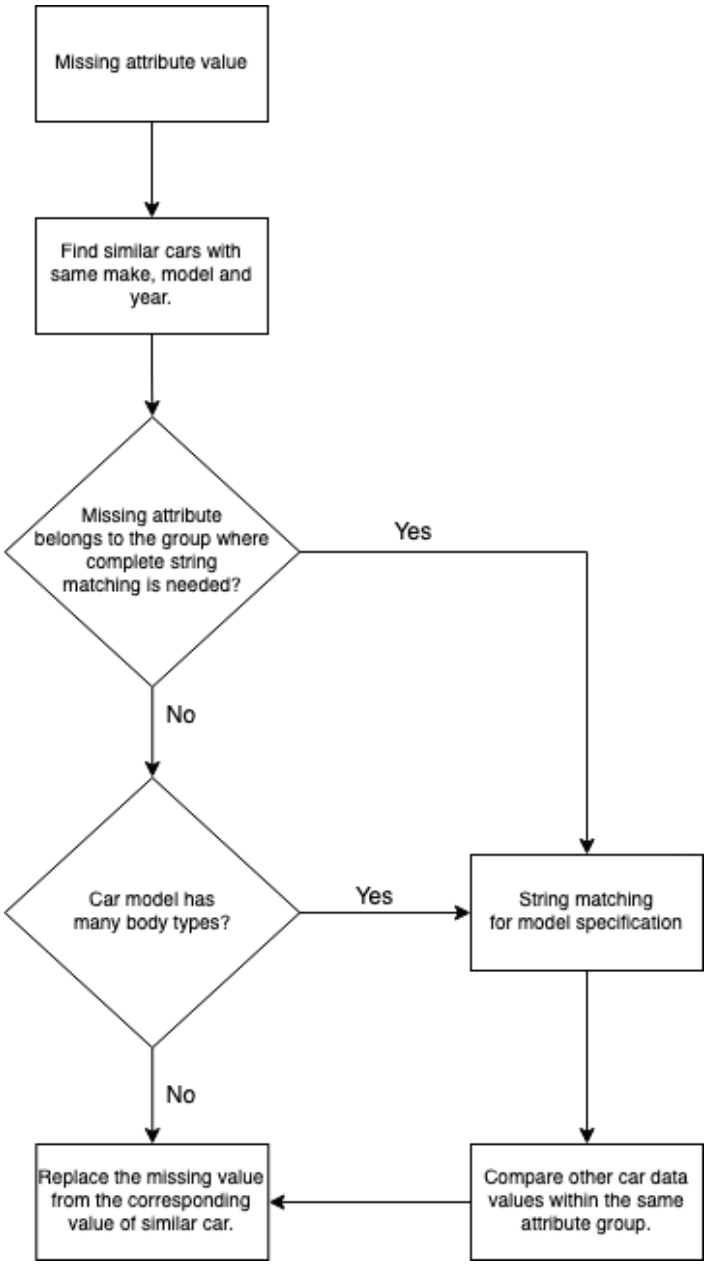

Figure 5. A process of finding a similar car.

If a similar case cannot be found from the same dataset, an external data source should be available as a backup. It is relatively safe to assume that no car is unique in terms of this basic information and a similar car can be found in one way or another. Of course, there are exceptions, for example among particularly expensive or rare cars, but in these cases the responsibility for entering the data can be transferred to car seller, because these are extremely rare cases. Process in Figure 5 can be used within the same dataset and external source data.

External data sources can be divided into two categories. Supplementary data can be obtained from an external provider that supplies ready-made car data. Such are, for example, Auto-Data.net (Auto-Data.net, 2023) and EV Database (EV Database, 2023), which specializes in electric car data. Another option, or rather a parallel option, is to create a company's own database of at least some of the car's attributes, from which missing information could be retrieved. This could happen automatically, for example, so that whenever some data is missing from the car, it would be added to the company's database when replaced. In this way, it could be used by all the company's customers, and as the database grows, it would reduce the need for integrating supplementary data from another provider's data. Car data has the advantage that it does not change for older cars, so once the information is added to the database, it does not need to be updated anymore (assuming the data is correct).

7 Conclusion

Missing data and its management are a permanent part of data integration, because it is impossible to completely eliminate the causes of missing data. As data integration becomes more complex, managing missing data also becomes more challenging. But there are also still simple data integration processes for which there is no direct solution, even though they have been extensively researched. Many integration processes have their own needs and basic requirements, which is why a company that uses integration must use resources to find a solution that is just right for them.

While conducting this thesis, a large number of different methods for managing missing data were found in the literature, not all of which were possible or even reasonable to include in the work. These methods and the literature presenting and analyzing them were strongly characterized by one assumption. They were developed for situations where the data itself is used to conduct various types of studies. In this case, the data is treated as a single large set and conclusions are drawn from their various distributions and averages, etc., where the significance of individual values is relatively low.

In this thesis, the purpose was to find ways to replace missing data in a situation where values were used as such, not as a set. Because of this, many imputation methods based on approximate values found in the literature proved to be useless. In the case of car data, the most essential thing in finding the right value is to find a similar case in the data, either from the same data set or from another dataset. The value of the corresponding attribute in the matching case can be used as is to replace the missing data. So the best imputation method is actually hot deck imputation or, in case no similar case can be found in same dataset, cold deck imputation.

An assumption was made regarding car data that if it exists, it is always correct. However, the research did not consider what would happen if an incorrect value was originally entered for the car and used to replace a missing value in a similar case. In situations where there are several similar cars, it may be safer to use values from multiple cars to ensure that the value is definitely correct. It is also necessary to consider whether it is sufficient to have only one similar case for a car. If only one was found in the same dataset, then it would be possible to find more similar cases from other data sources.

Although the company developing web pages does not have access to the original data, it would be worth considering trying to improve the quality of the original data. Currently,

empty information in an attribute does not prevent the addition of a car to the website, but it would be good to have an automatic notification sent to the car dealer, who should complete the data before publishing the car on the website. Regarding the model specification, it would be good to develop common labeling methods to obtain more reliable results from string matching. However, agreeing on common practices never brings full certainty to uniform labeling methods, because manual input processes are always prone to errors. Also, the people who record the information may change, and the agreed practices may not always be passed on. Similar error notifications as with missing data could be considered for this, so if the specification is not in accordance with the agreement, the car seller will receive a message about it.

However, as a starting point, completing the data through similar cases produces a relatively good result in car data, and incorrect situations are mostly individual cases. For this reason, the website developer should consider how far it is really worth taking the management of missing data.

8 References

Allison, P. D., 2002. Missing Data. Quantitative Applications in the Social Sciences. Thousand Oaks: SAGE Publications.

Alma Ajo, 2023. Websales. [Online] Available at: *https://almaajo.fi/websales/* [Accessed 19 April 2023].

Andridge, R. R. & Little, R. J. A., 2010. A Review of Hot Deck Imputation for Survey Non-response. International Statistical Review, 78(1), pp. 40-64.

Auto-Data.net, 2023. Wiki Automotive Catalog. [Online] Available at: *https://www.auto-data.net/en/* [Accessed 13 April 2023].

Autovista24, 2021. Does online-only used-car remarketing stand a chance?. [Online] Available at: *https://autovista24.autovistagroup.com/news/does-online-only-used-car-remarketing-stand-a-chance/* [Accessed 10 April 2023].

Bodner, T., 2008. What Improves with Increased Missing Data Imputations?. Structural Equation Modeling: A Multidisciplinary Journal, Volume 15(4), pp. 651-675.

Bramer, M., 2007. Principles of Data Mining. 1st ed. London: Springer London.

Buuren, S. v., 2012. Flexible Imputation of Missing Data. Boca Raton: Chapman & Hall.

Carpenter, J. & Kenward, M., 2013. Multiple Imputation and Its Application. Hoboken: John Wiley & Sons.

Christen, P., 2012. Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. 1st ed. Berlin: Springer.

Crasman, 2023. Crasman Oy. [Online] Available at: *https://www.crasman.fi/* [Accessed 15 April 2023].

Danske Bank, 2018. Yhä useampi ostaa auton suoraan verkosta. [Online] Available at: *https://danskebank.fi/sinulle/artikkelit/2018/09/yha-useampi-ostaa-auton-suoraan-verkosta* [Accessed 10 April 2023].

Daraio, C. & Glänzel, W., 2016. Grand challenges in data integration—state of the art and future perspectives: an introduction. Scientometrics, Volume 108(1), pp. 391-400.

Doan, A., Halevy, A. & Ives, Z., 2012. Principles of Data Integration. San Francisco: Elsevier Science & Technology.

Dong, X. L., Halevy, A. & Yu, C., 2009. Data integration with uncertainty. The VLDB journal, Volume 18, pp. 469-500.

EV Database, 2023. Electric Vehicle Database. [Online] Available at: *https://ev-database.org* [Accessed 13 April 2023].

García, S., Luengo , J. & Herrera, F., 2015. Data Preprocessing in Data Mining. New York: Springer.

Golshan, B., Halevy, A., Mihaila, G. & Tan, W.-C., 2017. Data Integration: After the Teenage Years. Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Volume 127745, pp. 101-106.

Graham, J. W. & Donaldson, S., 1993. Evaluating Interventions With Differential Attrition: The Importance of Nonresponse Mechanisms and Use of Follow-Up Data. Journal of Applied Psychology, Volume 78, pp. 119-128.

Graham, J. W., 2012. Missing data: Analysis and design. New York: Springer.

Grzymala-Busse, J. W. & Hu, M., 2001. A Comparison of Several Approaches to Missing Attribute Values in Data Mining. Lecture Notes in Computer Science, Volume 2005, pp. 378-385.

Grzymala-Busse, J. W., Grzymala-Busse, W. J. & Goodwin, L. K., 2005. A Closest Fit Approach to Missing Attribute Values in Preterm Birth Data. Lawrence: Lecture Notes in Artificial Intelligence, Volume 1711, pp. 405-413.

Halevy, A., Rajaraman, A. & Ordille, J., 2006. Data Integration: The Teenage Years. VLDB 2006 - Proceedings of the 32nd International Conference on Very Large Data Bases, pp. 9-16.

He, Y., Zhang, G. & Hsu, C.-H., 2021. Multiple Imputation of Missing Data in Practice: Basic Theory and Analysis Strategies. Milton: CSC Press.

Horton, N. J. & Kleinman, K. P., 2007. Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American statistician., Volume 61(1), pp. 79-90.

Huang, Y. & Madey, G., 2004. Web Data Integration Using Approximate String Join. Proceedings of the 13th International World Wide Web Conference on Alternate Track, Papers and Posters, pp. 364-365.

Ibrahim, J., Chen, M.-H., Lipsitz, S. & Herring, A., 2005. Missing-Data Methods for Generalized Linear Models: A Comparative Review. Journal of the American Statistical Association, Volume 100(469), pp. 332-346.

Kaiser, J., 2014. Dealing with Missing Values in Data, Volume 5, pp. p.42-51.

Kesko, 2023. Autokaupan strategia. [Online] Available at: *https://www.kesko.fi/sijoittaja/strategia/toimialojen-strategiat/autokaupan-strategia/* [Accessed 10 April 2023].

Kuchibhotla, H. N., Dunn, D. & Brown, D., 2009. Data Integration Issues in IT Organizations and a need to map different data formats to store them in relational databases. Proceedings of the 41st Southeastern Symposium on System Theory, pp. 1-6.

Lachin, J. M., 2016. Fallacies of last observation carried forward analyses. Clinical Trials, Volume 13(2), pp. 161-168.

Lakshminarayan, K., Harp, S. A. & Samad, T., 1999. Imputation of missing data in industrial databases. Applied Intelligence, Volume 11(3), pp. 259-275.

Lans, R. F., 2012. Data virtualization for business intelligence architectures revolutionizing data integration for data warehouses. 1st ed. Amsterdam: Elsevier/MK.

Laquer, E., 2017. Defining Data-Driven Software Development. Sebastopol: O'Reilly Media, Inc.

Lenzerini, M., 2002. Data Integration: A Theoretical Perspective. Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 233-246.

Liikenne- ja viestintävirasto, 2023. Avoin data. [Online] Available at: *https://tieto.traficom.fi/fi/tietotraficom/avoin-data?toggle=Ajoneuvojen%20avoin%20data* [Accessed 11 April 2023].

Little, R. J. A. & Rubin, D. B., 2014. Statistical analysis with missing data. 3rd ed. Wiley.

Little, R. J., Carpenter, J. R. & Lee, K. J., 2022. A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis, Inverse Probability Weighting, and Multiple Imputation. Sociological Methods & Research.

Lu, J. & Holubová, I., 2019. Multi-model Databases: A New Journey to Handle the Variety of Data. ACM Computing Surveys, Volume 52, pp. 1-38.

Lu, J., Liu, Z. H., Xu, P. & Zhang, C., 2018. UDBMS: Road to Unification for Multi-model Data Management. Advances in Conceptual Modeling, Volume 11158, pp. 285-294.

Lv, T., Yan, P. & He, W., 2018. Survey on JSON Data Modelling. Journal of Physics: Conference Series, Volume 1069(1), pp. 1-5.

Marrs, T., 2017. JSON at work: practical data integration for the web. 1st ed. Beijing: O'Reilly.

Marsh, H. W., 1998. Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. Structural Equation Modeling, Volume 5(1), pp. 22-36.

Mavridis, D. et al., 2019. Allowing for uncertainty due to missing and LOCF imputed outcomes in meta-analysis. Statistics in Medicine, Volume 38(5), pp. 720-737.

Merieme, E. A., Mohamed, A., Ali, C. & Fakhri, Y., 2022. A survey on the challenges of data integration. The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings.

Mockus, A., 2008. Missing Data in Software Engineering. Guide to Advanced Empirical Software Engineering, pp. 185-200.

Mohideen, D. F. M., Raj, J. S. S. & Raj, R. S., 2021. Regression Imputation and Optimized Gaussian Naïve Bayes Algorithm for an Enhanced Diabetes Mellitus Prediction Model. Brazilian Archives of Biology and Technology, Volume 64, pp. 1-16.

Netwheels, 2023. GT-X. [Online] Available at: *https://www.netwheels.fi/tuotteet/gt-x/* [Accessed 19 April 2023].

Newman, D. A., 2014. Missing Data: Five Practical Guidelines. Organizational Research Methods, Volume 17(4), pp. 372-411.

OECD, 2011. Quality Framework for OECD Statistical Activities. [Online] Available at: *https://www.oecd.org/sdd/qualityframeworkforoecdstatisticalactivities.htm* [Accessed 19 3 2023].

Paytrail, 2019. Nordic e-commerce 2019. [Online] Available at: *https://www.paytrail.com/hubfs/Nordic-Ecommerce-2019.pdf* [Accessed 10 April 2023].

Ratner, B., 2011. Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. 2nd ed. Baton Rouge: CRC Press LLC.

Reed, D., 2006. Take a good look. Data Strategy, Volume 2(4), pp. 24-29.

Rubin, D. B., 1977. Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. Journal of the American Statistical Association, Volume 72(359), pp. 538-543.

Rubin, D. B., 1987. Multiple imputation for nonresponse in surveys. New York: Wiley.

Rässler, S., Rubin, D. B. & Zell, E. R., 2013. Imputation. Wiley interdisciplinary reviews. Computational Statistics, Volume 5(1), pp. 20-29.

Saka, 2023. Autokaupan murros heijastuu digitalisaatioon ja asiakaskokemukseen – "Data on meille polttoainetta". [Online] Available at: *https://saka.fi/fi/yritys/ajankohtaista/autokaupan-murros-heijastuu-digitalisaatioon-ja-asiakaskokemukseen/* [Accessed 10 April 2023].

Schafer, J. L. & Graham, J. W., 2002. Missing Data: Our View of the State of the Art. Psychological Methods, Volume 7(2), pp. 147-177.

Schafer, J., 1997. Analysis of incomplete multivariate data. London: Chapman & Hall.

Serrano, L. G., 2021. Grokking Machine Learning. Shelter Island, New York: Manning Publications.

Sessa, J. & Syed, D., 2017. Techniques to Deal with Missing Data. International Conference on Electronic Devices, Systems, and Applications, pp. 1-4.

Sherman, R., 2014. Business intelligence guidebook: from data integration to analytics. Waltham, MA: Morgan Kaufmann.

Shi, D., Lee, T., Fairchild, A. J. & Maydeu-Olivares, A., 2020. Fitting Ordinal Factor Analysis Models With Missing Data: A Comparison Between Pairwise Deletion and

Multiple Imputation. Educational and Psychological Measurement, Volume 80(1), pp. 41-66.

Ullman, J. D., 2000. Information integration using logical views. Theoretical Computer Science, Volume 239(2), pp. 189-210.

Voillet, V., Besse, P., Liaubet, L., San Cristobal, M., González, I., 2016. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. BMC Bioinformatics, Volume 17(1).

Zhao, H., 2007. Semantic matching across heterogeneous data sources. Communications of the ACM, 50(1), pp. 45-50.

Zhou, X.-h., Ding, X., Liu, D. & Zhou, C., 2014. Applied missing data analysis in the health sciences. 1st ed. Hoboken: John Wiley & Sons Inc.

Zhu, X., 2014. Comparison of Four Methods for Handing Missing Data in Longitudinal Data Analysis through a Simulation Study. Open Journal of Statistics, Volume 4, pp. 933-944.

Ziegler, P. & Dittrich, K. R., 2007. Data Integration – Problems, Approaches, and Perspectives. Conceptual Modelling in Information Systems Engineering, pp. 39-58.

Appendix 1: Open Data for Vehicles - Traficom

Aineistokuvaus - Ajoneuvojen tekniset tiedot			
28.1.2021			
tieto	**kuvaus**	**Beskrivning**	**Description in brief**
ajoneuvoluokka	Ajoneuvon direktiivin mukainen luokittelu.	Fordons klassificering enligt direktivet	Vehicle classification according to the directive.
ensirekisterointipvm	Ajoneuvon ensirekisteröintipäivä.	Datum för första registrering	Date of first registration
ajoneuvoryhma	Ajoneuvoluokkaa tarkempi luokittelu ajoneuvoille	Mer exakt klassificering för fordon	More accurate classification for vehicles
ajoneuvonkaytto	Ajoneuvon käyttötieto.	Information om fordonsanvändning	Vehicle usage information
variantti	Ajoneuvon variantin yksilöivä tunniste	Unik identifierare för fordonets variant	Unique identifier of the vehicle variant
versio	Ajoneuvon version yksilöivä tunniste	Unik identifierare för fordonsversionen	Unique identifier of the vehicle version
kayttoonottopvm	Ajoneuvon käyttöönottopäivä	ibruktagningsdatumet	The date on which the vehicle was introduced into use
vari	Ajoneuvon väri.	Fordonets färg.	Colour of the vehicle
ovienLukumaara	Ajoneuvon ovien lukumäärä.	Antal dörrar i fordonet	Number of doors in the vehicle
korityyppi	Ajoneuvon korityyppi.	Fordonstyp	The body type of vehicle.
ohjaamotyyppi	Ajoneuvon ohjaamotyyppi.	Fordonshyttstyp	The cab type of vehicle
istumapaikkojenLkm	Istumapaikkojen lukumäärä.	Antal platser.	Number of seats
omamassa	Ajoneuvon mitattu koko omamassa kilogrammoina.	Vikt / massa	Mass
teknSuurSallKokmassa	Teknisesti suurin sallittu kokonaismassa kilogrammoina (valmistajan sallima).	Tekniskt tillåten maximal vikt i kg (tillåten av tillverkaren).	Technically permissible maximum mass in kilograms (permitted by the manufacturer).
tieliikSuurSallKokmassa	Tieliikenteessä suurin sallittu kokonaismassa kiloina.	Vid vägtransport, den högsta tillåtna lastmassan i kg.	In road transport, the maximum permissible laden mass in kilograms.
ajonKokPituus	Ajoneuvon kokonaispituus millimetreinä.	Den totala längden på ett fordon i millimeter.	The total length of a vehicle in millimetres.
ajonLeveys	Ajoneuvon leveys (mm)	Fordonsbredd i millimeter	Vehicle width in millimetres
ajonKorkeus	Ajoneuvon korkeus (mm)	Fordonets höjd i millimeter	The height of a vehicle in millimetres.
Kayttovoima	Ajoneuvon käyttövoima.	Fordonets drivkraft	Driving power of the vehicle
iskutilavuus	Moottorin iskutilavuus kuutiosenttimetreinä (cm^3).	Kapacitet / slagvolym / cylindervolym	Engine capacity / cylinder capacity
suurinNettoteho	Moottorin suurin nettoteho kilowatteina (kW).	Maximal nettomotoreffekt i kilowatt (kW)	The maximum net engine power in kilowatts (kW)
sylintereidenLkm	Sylintereiden lukumäärä.	Antalet cylindrar.	The number of cylinders.
ahdin	Kyllä/ei tieto ahtimen olemassaolosta.	Turbokompressor (Ja/Nej)	Supercharged (True/False)
sähköhybridi	Onko ajoneuvo sähköhybridi	Indikation (Ja/Nej) om fordonet är ett eldrivet hybridfordon.	Indication (Y/N) of whether the vehicle is an electrically powered hybrid vehicle.
sähköhybridin luokka	Sähkökäyttöisen hybridiajoneuvon luokka	Kategori av hybridelektrisk fordon	Category of hybrid electric vehicles
merkkiSelvakielinen	Ajoneuvon merkki selväkielisenä	Fordonets märke	Vehicle make in plain text
mallimerkinta	Ajoneuvon mallimerkintä.	Fordonets modell	Vehicle model
vaihteisto	Vaihteiston tyyppi.	transmission/växellåda	transmission / gearbox
vaihteidenLkm	Eteenpäin vievien vaihteiden lukumäärä.	Antal växlar framåt.	Number of forward gears.
kaupallinenNimi	Valmistajan ilmoittama kaupallinen nimi.	Handelsnamn från tillverkaren.	Trade name given by the manufacturer.
voimanvalJaTehostamistapa	Ajoneuvon jarrujen voimanvälitys ja tehostamistapa.	Fordonsbromsöverföring och boost-läge.	Vehicle brake transmission and boost mode.
tyyppihyvaksyntanro	Ajoneuvon tyypin yksilöivä tunniste.	typgodkännandenummer	type-approval number
yksittaisKayttovoima	Ajoneuvon käyttövoima.	Fordonets drivkraft. Körkraft-specifika data på fordonet, med start från frågetyp 6xx (EURO VI-data).	Driving power of the vehicle. Driving power -specific data on the vehicle, starting from query type 6xx (EURO VI data).
kunta	Ajoneuvon todennäköisin käyttökunta	Troligen kommun där fordonet används	Probably the municipality where the vehicle is used
NEDC_CO2	Ajoneuvon hiilidioksidipäästö (CO_2) grammoina. (NEDC)	Fordonets koldioxidutsläpp (CO2) i gram. (NEDC)	Vehicle carbon dioxide (CO2) emissions in grams. (NEDC)
NEDC2_CO2	Ajoneuvon hiilidioksidipäästö (CO_2) grammoina. (NEDC)	Fordonets koldioxidutsläpp (CO2) i gram. (NEDC)	Vehicle carbon dioxide (CO2) emissions in grams. (NEDC)
WLTP_CO2	Ajoneuvon hiilidioksidipäästö (CO_2) grammoina. (WLTP)	Fordonets koldioxidutsläpp (CO2) i gram. (WLTP)	Vehicle carbon dioxide (CO2) emissions in grams. (WLTP)
WLTP2_CO2	Ajoneuvon hiilidioksidipäästö (CO_2) grammoina. (WLTP)	Fordonets koldioxidutsläpp (CO2) i gram. (WLTP)	Vehicle carbon dioxide (CO2) emissions in grams. (WLTP)
matkamittarilukema	Viimeisin katsastuksessa todettu matkamittarin lukema	Den senaste mätaravläsningen hittades under inspektionen	Last odometer reading found during the inspection
valmistenumero2	10 ensimmäistä numeroa valmistenumerosta henkilöautoilta.	De första tio siffrorna i serienumret för personbilar.	The first 10 digits of the serial number for passenger cars.
jarnro	Juokseva numerointi	Sekventiell numrering	Sequential numbering