Don Duma

RECOGNIZING THE VALUE OF DATA IN BUSINESS OPERATIONS

Data analytics for business operation

Thesis
CENTRIA UNIVERSITY OF APPLIED SCIENCES
Information Technology
August 2022

ABSTRACT

Centria University of Applied Sciences	**Date** August 2022	**Author** Don Duma
Degree programme Information Technology		
Name of thesis RECOGNIZING THE VALUE OF DATA IN BUSINESS OPERATIONS. Data analytics for business operation		
Centria supervisor Jari Isohanni		**Pages** 33 + 4

The aim of this study was to demonstrate the hidden value of data that can be extracted with few commercial and open-source software tools. Any given business can collect, organize, and extract data for analysis that can be vital in making informed decisions in business operations. Subsequently, with good data analysis skills, it is possible to optimize the quality and quantity of data to be analysed in order to achieve desirable outcome. Data have become an important asset in business operations since they provide insights that could have not been gained otherwise. Finding ways to explore the riches hidden in data will profit businesses in their daily processes.

The experimental method was used in the study. Various technologies for data collections, storage, handling, and analysis were considered to achieve the outcome. Furthermore, the study focused on database technologies, data analytics tools for manipulating and modelling data in order to get clean and reliable results that can be visualised for better understanding.

The results of the study demonstrated that if businesses in different sectors pay attention to the collected data within their daily operations, beneficial insights can be gained for better decision making. Moreover, businesses should use insights from data to follow trends and megatrends and to follow different developments thus enabling them to learn from the past and predict the future.

Key words
Databases, data analysis, data cleansing, data management, data mining, data visualisation, data wrangling, external data, internal data, non-structured databases, structured databases

CONCEPT DEFINITIONS

Oracle HeatWave

HeatWave is a fully managed database service that lets developers quickly develop and deploy secure cloud native applications using the world's most popular open-source database. (Oracle 2022)

Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualisations in Python. (Matplotlib 2022)

MySQL

MySQL is a widely used relational database management system (RDBMS). (W3Schools 2022)

NumPy

NumPy is the fundamental package for scientific computing in Python. (NumPy 2022)

Pandas

Pandas is a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language. (Pandas 2022)

Power BI

Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights. (Microsoft corporation 2022)

Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. (Python 2022)

Seaborn

Seaborn is a Python data visualization library based on matplotlib. (Seaborn 2022)

SQL

SQL is a standard language for storing, manipulating, and retrieving data in databases. (W3Schools 2022)

ABSTRACT
CONCEPT DEFINITIONS
CONTENTS

FIGURES

1 INTRODUCTION

The topic of the study is recognizing the value of data in business operations and data analytics for business operation. The study will endeavour to deal with the important situation of various kind of data currently collected within businesses. The aim of the study was to demonstrate with commercial and open-source software tools how collected data can be modelled, visualised, and used to gain invaluable insights profitable to businesses.

The background of the study derives from the fact that the on-going digitalisation enabled by information technology (IT) and information systems have become an integral part of any modern businesses irrespective of the operational sector. Businesses have countless smart devices and applications installed on their premises, computer aided devices and sensors are incorporated on operating machines, websites have cookies for data collections and monitoring, smartphones and other handheld devices register activities, enterprise resource planning systems, and many more that constantly collect data.

For the purpose of this demonstration, the first set of the collected data can be categorised as internal data. Internal data includes information from sales reports, daily operations such as machine logs, personnel’s data from the human resources department, financial data, research and developments, material costs, and the records of employees’ timestamps. The second category of data collected can be identified as external data. The external data consists of information collected from websites, social media, market research, market trends, competitors, suppliers, and customers’ feedback.

Finally, the study will closely examine how the collected data are stored before being structured, processed, and analysed. Therefore, both non-structured and structured databases technologies will be addressed. Moreover, structured, and unstructured data will feature in the discussion. The recent revolution in information technology has enabled the digitalization and globalization in all sectors of business operations. Consequently, to be competitive and up-to-date, businesses are bombarded with enormous amount of data daily. Data are generated from different sources and different tools and mechanisms are put in place for their gathering.

The gathered data are worthless if not processed and analysed. Therefore, the study will also emphasise on continuous plans to maximize the importance of the collected data in order to make use of it in daily business operations. Furthermore, the study will encourage business stakeholders to see how gathered data influence decision-making processes resulting in gaining and maintaining the competitive advantages.

The aim of the study was to demonstrate that the massive amounts of amassed data should be safely stored, prioritized, and utilized to make informed business decisions. Considering the width of the topic of data collection, storage and utilization, the focal point of the study was solely on the collection of data, the categorization, management, analysis, and visualisation. The different technologies utilized to classify and process data were the main part of this study and experiment.

2 DATA STORAGE, MANAGEMENT, ANALYSIS AND VISUALISATION

The industrialization and the on-going digitalization trends have resulted in the exponential increase in volume of data being generated in companies quotidianly. According to Alghushairy & Ma (2019), data are generated by humans and machines irrespective of the industry or geolocations. Therefore, as the value of generated data increases and businesses become more dependent on the insight gained from data, reliable means of storing and archiving the data is an integral process for any operating businesses.

Alghushairy & Ma (2019) define data storage as the process of storing and archiving data in electronic storage devices that are designed and dedicated for conservation, in return, making data accessibility a priority at any time. Therefore, conventional storage devices are hardware that are used for reading and writing data through a storage medium. Storage media are physical materials for storing and retrieving data.

According to an article in Techopedia (2012), the most common data storage devices are hard drive disks, flash drives, and cloud storage while compact disks (CDs) and digital video disks DVDs) are becoming obsolete to businesses due to the incapacity in coping with the amounts and speed at which data are collected and stored.

Since data are the all-important asset for any business, it is necessary to store data in appropriate ways to support data discovery, access, and analytics. To match the demands and challenges posed by data storage, various data storage devices and technologies have been developed to increase the efficiency in data management and to enable information extraction and knowledge discovery from data. (Blumzon & Pănescu 2019.)

Eiras (2011) insists that in domain-specific fields, a lot of scientific research has been conducted to tackle the specific requirements concerning data collection, data formatting, and data storage, which have also generated beneficial feedback to computer science. This fact has encouraged world researchers to focus on new techniques to design systems with larger storage capacity, as it is the case of cloud storage solutions, and high-capacity storage systems. These systems can store the entire volume of the informational material, thereby increasing the storage capacity in comparison with two-dimensional systems that only store the information on the surface. (Eiras 2011.)

2.1 Databases

In modern society databases and database systems play a vital role. Most of the day-to-day activities involve the use of databases. Bank application, ticket reservations for any sort of applications, computerized library can be cited as instances for use of databases (Vidhya, Jeyaram & Ishwarya 2016). According to Bagui & Earp (2011), a database can be defined as a collection of related data. Insomuch as Murthy (2007) defines a database as a shared collection of logically related data, which is designed to meet the information needs of multiple users is a business.

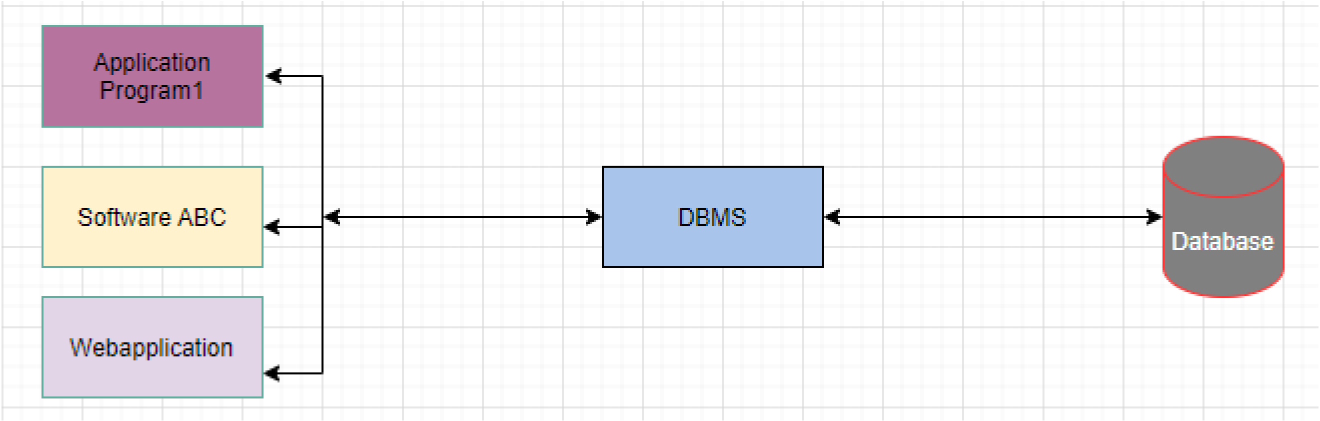


FIGURE 1. Applications and software relation to database (Connoly & Begg 2015)

Moreover, Vidhya, Jeyaram & Ishwarya (2016) define data as known facts that can be recorded that have implicit meaning. When data is processed, organized, structured, or presented in a given context it becomes information. Databases are managed using the database management systems (DBMS) which is a collection of interrelated data and a set of programs to access those data as FIGURE 1 demonstrates. The primary goal of a DBMS is to provide a way to store and retrieve database information that is both convenient and efficient.

The DBMS manages incoming data, organizes, and provides ways for the data to be modified or extracted by users or other programs. The database and DBMS collectively are known as database system. Database system is a computerized record keeping system. It is a repository or container for a collection of computerized data files. From a database, users of the system can perform or request the system to perform a variety of operations such as adding new files to the database, inserting data into existing files, retrieving, or deleting data from existing files, modifying data in existing file or removing existing files from the database. (Vidhya et al. 20016.)

Moreover, for security reasons databases have an access control method for allowing access to company's sensitive data only to those people (database users) who are allowed to access such data and to restrict access to unauthorized persons. The access control includes authentication and authorization. (Connoly & Begg 2015.)

2.1.1 Relational databases

According to MongoDB (2022), a relational database, or relational database management system (RDMS), stores information in tables. Often these tables have shared information between them, forming a relationship between tables. The name relational database is a description of this relationship between tables or schemas. A table uses columns to define the information being stored and rows for the actual data. Each table will have a column that must have unique values known as the primary key as shown in FIGURE 2. This column can then be used in other tables if relationships are to be defined between them. When one table's primary key is used in another table, this column in the second table is known as the foreign key. The most common way of interacting with relational database systems is using SQL. (MongoDB 2022.)

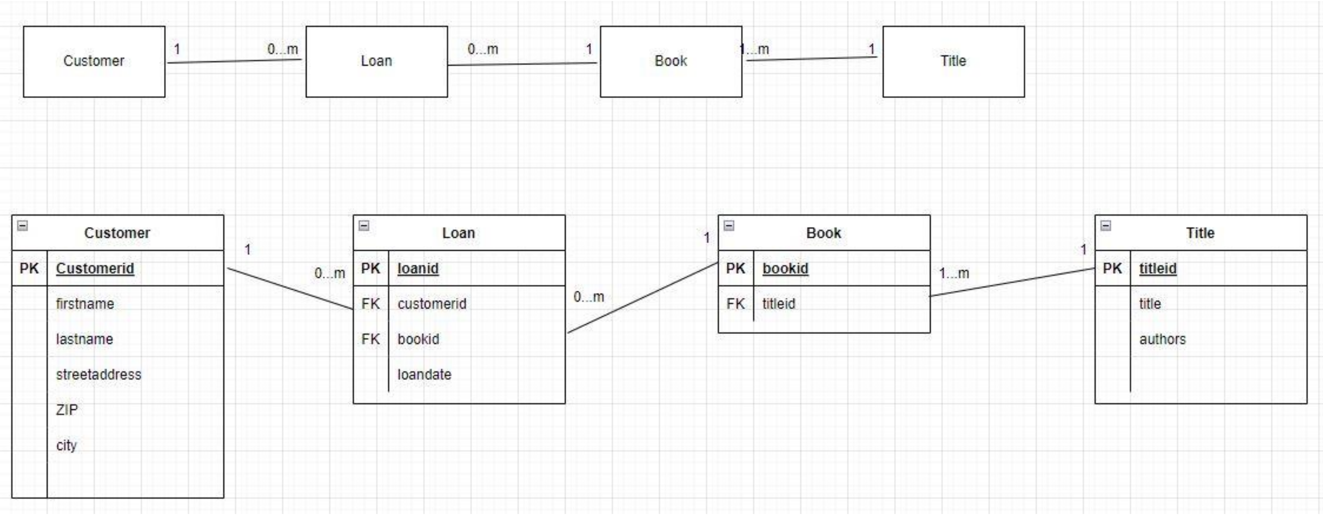


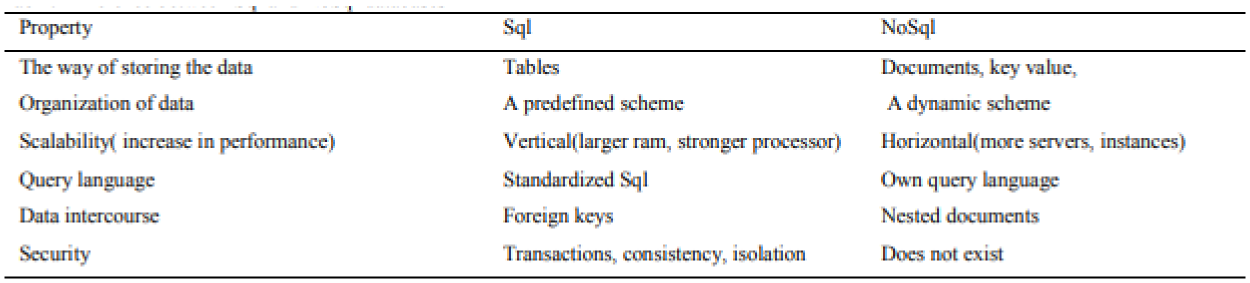
FIGURE 2. Relationships between tables or schemas in a database ((Connoly & Begg 2015)

Gupta & Agrawal (2018) describe relational databases as collection of tables having relations with data categories and constraints. Relational databases uses SQL or MySQL as the tool to access the data and is based upon ACID (atomicity, consistency, isolation, and durability) properties.

Moreover, Rai & Chettri (2018) insist that the relational database management systems (RDBMS) has several advantages such as abstraction, multiuser access, automatic optimization for searching and ACID properties enabling transaction support in extremely easy querying language. However, as the data grow exponentially as the in recent cases, scalability becomes a major issue for RDBMS leading to the introduction of non-relational databases such as NoSQL.

2.1.2 Non-Relational databases

According to Gupta & Agrawal (2018), non-relational databases provide elastic scaling and are designed using low-cost hardware. Non-relational databases are schema free, distributive, and store enormous amount of data. As described by MongoDB (2022), a non-relational database, sometimes called NoSQL (Not Only SQL), is any kind of database that does not use the tables, fields, and columns structured data concept from relational databases. Non-relational databases have been designed with the cloud in mind, making them great at horizontal scaling. Regarding the language used for querying, relational databases use the SQL and non-relational databases use the JSON query language (Rai & Chettri 2018).

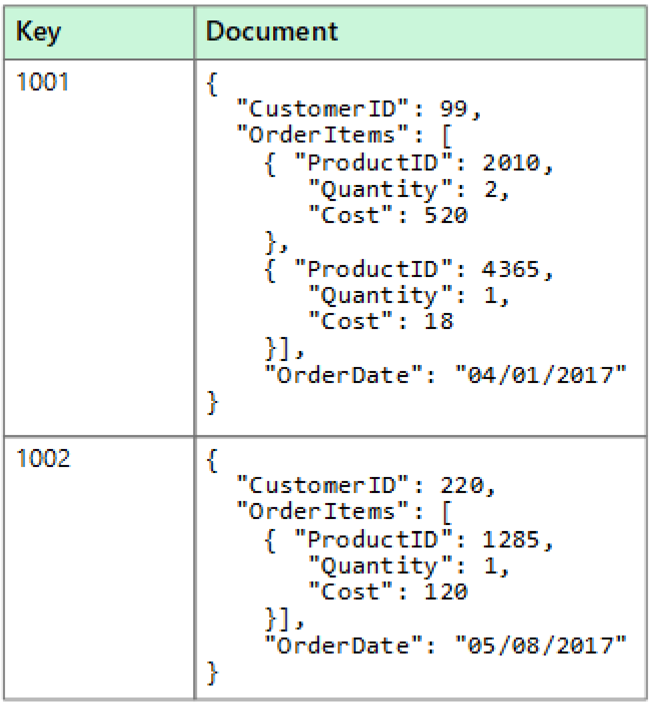


Property	Sql	NoSql
The way of storing the data	Tables	Documents, key value,
Organization of data	A predefined scheme	A dynamic scheme
Scalability(increase in performance)	Vertical(larger ram, stronger processor)	Horizontal(more servers, instances)
Query language	Standardized Sql	Own query language
Data intercourse	Foreign keys	Nested documents
Security	Transactions, consistency, isolation	Does not exist

FIGURE 3. Difference between SQL and NonSQL databases (Čerešňák & Kvet 2019)

As shown in FIGURE 3, non-relational databases are always prone to security attacks due to the unstructured nature of data and the distributed computing environment. In non-related database environment, the nodes are distributed to enable parallel computing that increases the attack surface which results in implementation of more complex procedures for security. Apart from distributed environment, data from a variety of nodes move from one node to another which leads to sharing of data and increases the risk of theft. (Gupta & Agrawal 2018.)

The major categories of non-relational or NoSQL database are document data stores, shown in FIGURE 4, columnar data stores, key or value data stores, graph data stores, time series data stores, object data stores, and external index data stores. (Microsoft corporation 2022c)



Key	Document

1001	{
	"CustomerID": 99,
	"OrderItems": [
	{ "ProductID": 2010,
	"Quantity": 2,
	"Cost": 520
	},
	{ "ProductID": 4365,
	"Quantity": 1,
	"Cost": 18
	}],
	"OrderDate": "04/01/2017"
	}

1002	{
	"CustomerID": 220,
	"OrderItems": [
	{ "ProductID": 1285,
	"Quantity": 1,
	"Cost": 120
	}],
	"OrderDate": "05/08/2017"
	}

FIGURE 4. A document data stores category of non-relational or NonSQL database. (Microsoft corporation 2022c)

2.2 Management

Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively. The goal of data management is to help people, businesses, and connected systems optimize the use of data within the bounds of policy and regulation so that they can make decisions and take actions that maximize the benefit to the business. Furthermore, a robust data management strategy is becoming more important than ever as businesses increasingly rely on intangible assets to create value. (Oracle 2022.)

The SaS website points out that data management is the practice of managing data as a valuable resource to unlock its potential for a business. Managing data effectively requires having a data strategy and reliable methods to access, integrate, cleanse, govern, store, and prepare data for analytics. In the digital world, data pours into businesses from many sources. Sources such as operational and transactional systems, scanners, sensors, smart devices, social media, video, and text. But the value of data is not based on its source, quality, or format. The value depends on what is done with data. Data management powers the processes for every successful business, across all industries. With more data and easier access to analytics comes the chance to seize more opportunities, ask more questions and solve more problems. Data management works by businesses avoiding the paradox of "garbage in, garbage out." As volumes, types and sources of data increase, the need to process data in real time expands and the urgency to manage data well remains a top priority for business success. Processes aspects such as data access, data integration, data quality, data governance, and data preparation are vital parts of data management. (SAS 2022.)

According to Munappy, Bosch, Olsson, Arpteg & Brinne (2022), data management for analysis can be defined as a process that includes collecting, processing, analysing, validating, storing, protecting, and monitoring data to ensure the consistency, accuracy, and reliability of the data. Industry products that make use of tremendous volume of digital data can successfully employ data analytics. However, real-world data need to be processed and managed before being fed as input to the analysis models. Massive and variegated data sets is challenging, and several aspects need to be considered. To ensure the high performance from data models, a set of good data management practices such as data pipelines and DataOps should be followed from data collection, through data processing and analysis, dataset preparation, and deployment of the model. Challenges like data dependency, memory management, concurrency, data inconsistency can be solved by combining database techniques and deep neural networks. (Munappy et al. 2022.)

2.2.1 Data cube (OLAP)

Data cubes allow data to be modelled and viewed in multiple dimensions within data warehouses. Data cubes are defined by dimensions and facts. Generally, dimensions are the perspectives or entities with respect to which a business wants to keep records. These dimensions allow the business to keep track of its activities and operations. Each dimension may have a table associated with it called a dimension table which further describes the dimension. A dimension table can be specified by users or by experts, or automatically generated and regulated based on data distributions. The data cubes are a metaphor for a multidimensional data storage. (Han & Kamber 2006.)

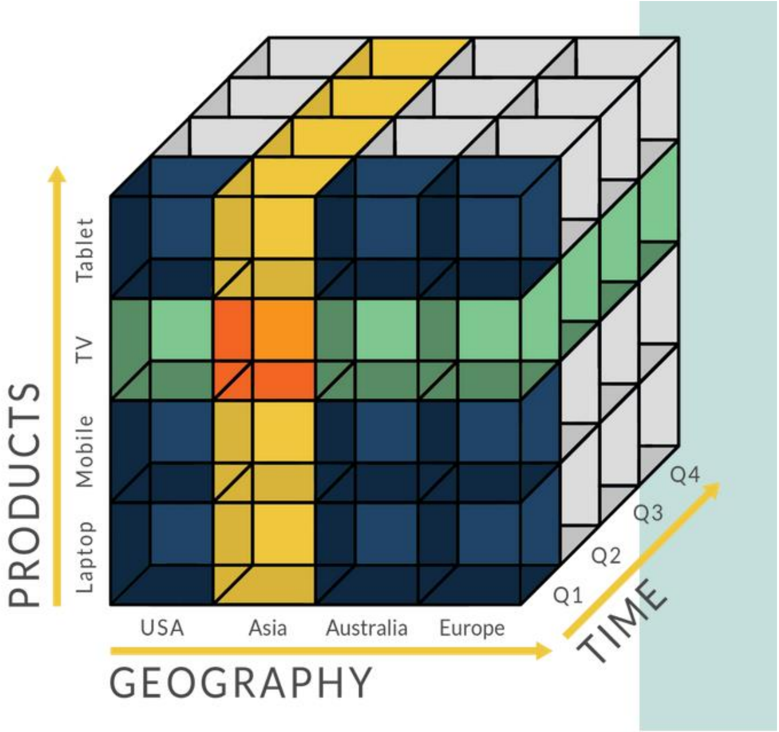


FIGURE 5. Data cube (OLAP 2022)

A data warehouse is usually modelled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum (sales amount). As a data cube provides a multidimensional view of data it allows the precomputation and fast access of summarized data. Data cubes facilitate the answering of queries as they allow the computation of aggregate data at multiple granularity levels. Traditional data cubes are typically constructed on commonly used dimensions using simple measures. (Han & Pei 2012.)

According to OLAP website, an easy definition of OLAP (Online Analytical Processing) is the technology behind many business intelligence (BI) applications. OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive uncertain scenario (budget, forecast) planning. Moreover, OLAP performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data modelling. Furthermore, OLAP is the cornerstone for many kinds of business applications for business performance management, planning, budgeting, forecasting, financial reporting, analysis, simulation models, knowledge discovery, and data warehouse reporting. OLAP enables end-users to perform ad hoc analysis of data in multiple dimensions, thereby providing the insight and understanding they need for better decision making. (OLAP 2022.)

2.2.2 Oracle MySQL HeatWave

HeatWave is a fully managed database service that lets developers quickly develop and deploy secure cloud native applications using the world’s most popular open-source database. Service such as massive-scalability, integrated, real-time query accelerator, and a fully automated in-database machine learning engine separate MySQL HeatWave from other MySQL products. This service overcomes the limitations of traditional data warehouse, analytics and machine learning environments that use periodic long-running ETL batch jobs to refresh the data. MySQL HeatWave provides a unified MySQL database platform for OLTP, OLAP and machine learning. Moreover, HeatWave is designed to enable customers to run analytics on data, which is stored in MySQL databases, without the need for ETL. This service is built on an innovative, in-memory analytics engine which is architected for scalability and performance and is optimized for Oracle Cloud Infrastructure (OCI). This results in a very performant solution for SQL analytics at a fraction of the cost compared to other industry solutions. (Oracle 2022.)

2.3 Analysis

Cuesta (2013) describes data analysis as the process in which raw or unstructured data is ordered and organized, to be used in methods that help to explain the past and predict the future. Data analysis is not about the numbers, it is about making/asking questions, developing explanations, and testing hypotheses. Data Analysis is a multidisciplinary field, which combines Computer Science, Artificial Intelligence & Machine Learning, Statistics and Mathematics, and Knowledge Domain.

2.3.1 Data cleansing

According to Fakhitah & Wan (2019), data cleansing is an operation that is performed on the existing data to remove anomalies and obtain the data collection which is an accurate and unique representation of the ideal format. It involves eliminating the errors, resolving inconsistencies, and transforming the data into a uniform format. With the vast amount of data collected, manual data cleansing is almost impossible as it is time-consuming and prone to errors. Data cleansing process is complex and consists of several stages which include specifying the quality rules, detecting data error, and repairing the error.

The five phases involved in the data cleansing process are phases such as data analysis, definition of transformation workflow and mapping rule, verification, transformation, and backflow of cleaned data.

2.3.2 Data wrangling

Data wrangling refers to a variety of processes designed to transform raw data into more readily used formats. The exact methods differ from project to project depending on the data leveraged and the goal. Data wrangling can be used in operation such as merging multiple data sources into a single dataset for analysis, identifying gaps in data by filling or deleting them, deleting data that is either unnecessary or irrelevant to the project, and identifying extreme outliers in data and either explaining the discrepancies or removing them so that analysis can take place. Data wrangling can be a manual or automated process. In scenarios where datasets are exceptionally large, automated data wrangling becomes a necessity. However, in businesses that employ a full data team, a data scientist or other team member is typically responsible for data wrangling. In smaller businesses, non-data professionals are often responsible for cleaning their data before leveraging it. (Stobierski 2021.)

Endel & Piringer (2015) insist that data wrangling is not only about transforming and cleaning procedures. Many other aspects like data quality, merging of different sources, reproducible processes, and managing data provenance have to be considered. Although various tools designed for specific tasks are available, software solutions accompanying the whole process are still rare.

2.3.3 Data mining

Data mining is the process of discovering useful patterns and trends in large data sets. Moreover, data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. However, data mining can not only be achieved by automatic or semi-automatic means but also the exploration and analysis. Therefore, data mining is a discipline that must be mastered. Automation is no substitute for human input. Humans need to be actively involved at every phase of the data mining process. Rather than asking where humans fit into data mining, the question should instead inquire about how to design data mining into the very human process of problem solving. (Larose & Larose 2014.)

2.3.4 Python with libraries

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Python's high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Benefits such as Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Moreover, Python supports built in modules and packages or libraries such as NumPy, Scikit-learn, Pandas, and matplotlib, which encourage program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed. (Python.org 2022.)

According to Lee (2019) NumPy is an extension to the Python programming language that adds support for large, multidimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on the arrays. However, Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labelled” data both easy and intuitive. Pandas supports two key data structures known as Series and Data Frame.

Moreover, Python library such as Matplotlib is a Python 2D plotting library mostly used to produce publication quality charts and figures. Using matplotlib, complex charts and figures can be generated with ease. Furthermore, Scikit-learn is Python library that implements the various types of machine learning algorithms, such as classification, regression, clustering, decision tree, and more. Using Scikit-learn, implementing machine learning is now simply a matter of calling a function with the appropriate data to fit and train the model. (Lee 2019.)

2.4 Visualisation

According to IBM cloud education (2021), data visualisation is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand. Data visualisation can be utilized for a variety of purposes, and it is important to note that it is not only reserved for use by data teams. Management also leverages it to convey organizational structure and hierarchy while data analysts and data scientists use it to discover and explain patterns and trends. Microsoft corporation insists that data visualisation helps businesses turn all the granular data into easily understood, visually compelling and useful business information. (Microsoft Corporation 2022a).

2.4.1 Microsoft Power BI

Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights. The data may be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses. Power BI facilitates connections to data sources, visualise and discover what is important, and share that with stakeholders. Moreover, Microsoft Power BI has several elements that work together, starting with three basics such as a Windows desktop application called Power BI Desktop, an online SaaS (Software as a Service) service called the Power BI service, and the Power BI mobile apps for Windows, iOS, and Android devices. Beyond those three, Power BI also features two other elements, the Power BI Report Builder, for creating paginated reports to share in the Power BI service and the Power BI Report Server, an on-premises report server to publish Power BI reports, after creating them in Power BI Desktop. (Microsoft Corporation 2022b.)

2.4.2 Seaborn

Seaborn is a complementary plotting library that is based on the Matplotlib data visualisation library. Seaborn’s strength lies in its ability to make statistical graphics in Python, and it is closely integrated with the Pandas data structure. Seaborn provides high-level abstractions to that enables the building of complex visualisations for data easily. With less code, one can get more sophisticated charts with Seaborn than with Matplotlib. (Lee 2019.)

3 MAKING VALUE FROM DATA

Data are the new engine in business operations. When data are exploited and refined into information, businesses are able to do exciting and new innovations. The question is where to focus efforts to deliver business value in a world where businesses are overwhelmed with data from so many sources found from within and outside. Businesses need to have the right leadership, data strategy, technologies, and business processes to prioritize and turn complex data into meaningful information and then to ensure that insights are actually used to support decision-making. Mostly many businesses and IT leaders are focused on big data projects, neglecting the large amounts of data already residing but unexploited within their businesses. The priority should be how to exploit the data to deliver business value. (Rosslyn Analytics (2022.)

3.1 Value of data

When data are processed into sets according to context, they supply information. Moreover, data refer to raw input that when processed or arranged makes meaningful output. Information is usually the processed result of data. When data are processed into information, they become interpretable and meaningful. (Cambridge Assessment International Education 2015.)

According to Borek, Parlikad, Webb & Woodall (2013), data are the essential material to generate the information insights that a business needs to strive in today's increasingly competitive environment. Information insights allow commercial businesses to allocate resources more profitably, satisfy their customers more effectively, reduce costs, save energy and materials, and offer better products and services that consumers desire and that are priced competitively. Moreover, public authorities, governments, and local communities use analytics to minimize the financial burdens placed on taxpayers, reduce crime, optimize transportation, improve reliability and quality of utility and citizen services, and diminish environmental pollution. Furthermore, non-profit organizations are also able to take considerable advantage of information insights to increase the speed of support logistics in crisis regions, achieve a better allocation of resources to solve global challenges, and speed up medical progress in fighting diseases. Therefore, it is not an overstatement to say that advanced data analytics will drive a smarter and effective processes in business operations.

3.2 Data categorization

Data can be categorized in several ways. Data can be considered primary data, meaning that they are gathered directly from the source. Secondary data refer to the data that has been gathered and analysed by a third party. Additionally, data can be considered qualitative or quantitative. Qualitative data are measures of ‘types’ and may be represented by a name, symbol, or a number code. Quantitative data, on the other hand, are measures of values or counts and are expressed as numbers. Data collected about a categorical variable will always be qualitative and data collected about a numeric variable will be quantitative. (Australian Bureau of Statistics 2020.)

Despite the characteristics of data mentioned above, data can be considered raw (unprocessed) indicating format of the provenance. Data can also be processed and analysed if any manipulation by machines or humans has been applied. Additionally, data can be considered structured if correct structure such as rows and columns are already assigned. Data can also be unstructured. (Praveen & Chandra 2017.)

3.2.1 Unstructured and structured data

According to Praveen & Chandra (2017) all data are first gathered as raw in its state before being sorted, processed, and analysed for various usage. Typically, the data that are most readily available for use are not in a state in which they can be used easily. Such data are difficult to manipulate and typically need to be processed in some fashion before they can be used. Data which have yet to be processed are sometimes referred to as raw data, although source data are a more useful term. These data, being in their original state, are unstructured and have not yet been structured in a predefined manner. Unstructured data are typically text-heavy, like system logs, sensors' generated data, and handheld devices' data. They can also include images, videos, and audio.

Unstructured data are growing quickly due to increased use of digital applications and services. These data are valuable to businesses if analysed and interpreted correctly. They can provide rich insights that statistics and numbers just cannot detail. Additionally, unstructured data cannot be easily stored in a traditional column-row database or spreadsheet like a Microsoft Excel table. It is therefore more difficult to analyse and not easily searchable. (Grossman 2019.)

Currently, more than 80% of the data on the Internet are unstructured data (Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez & Kochut 2017). Unstructured data usually refers to information that does not reside in a relational database. In other words, the data structure of unstructured data are irregular or incomplete and there is no predefined data model. It should be noted that although some document formats like comma separated value (CSV), JavaScript object notation (JSON), and extensible markup language (XML) have some organizational properties, they usually do not have a clear predefined data model. Compared to structured data, these data are still difficult to retrieve, analyse and store. Unstructured data are easily processed by humans but are very hard for machines to understand. (Allahyari et al. 2017.)

There are some characteristics that define the unstructured data. Unstructured data have no internal identifier to let search functions recognize it. Unstructured data do not follow any semantic or rules, they are gathered in the state that they are generated. Unstructured data lack any particular format or sequence and do not possess easily identifiable structure. Therefore, due to their lack of identifiable structure, they cannot be used by computer programs easily. These data are generated by various sources in different formats and need some work to them before they can be used by businesses. (Praveen & Chandra 2017.)

Web pages produce enormous amounts of unstructured data through inbuild data collecting applications. These enormous data can be in form of images (JPEG, GIF, PNG, etc.), videos, sounds, and text files. Unstructured data can be in forms of memorandums, various reports and systems logs, and text documents. Research surveys, and customers' feedback can generate a considerable amount of unstructured data that need sorting before processing and analysing. (Allahyari et al. 2017.)

There are many advantages the unstructured data bring to businesses. Because of their lack of proper format and sequence, the data are not constrained by a fixed schema. Moreover, the gathering is very flexible due to absence of predefined schema. (Praveen & Chandra 2017.)

Regarding the unstructured data concept, unstructured data are mobile and portable. Because it is generated and stored efficiently, unstructured data prove to be flexible. Unstructured data enriche businesses’ data and enable decision-makers to work effectively and proactively. Moreover, unstructured data are scalable, their volume increase at a very high speed, and they can deal easily with the heterogeneity of provenances. These types of data have a variety of business intelligence and analytics applications. (Sivarajah, Kamal, Irani & Weerakkody 2017.)

However, as with any solution, Sivarajah et al. (2017) state that there are some disadvantages with unstructured data. It is difficult to store and manage unstructured data due to lack of schema and structure. Therefore, indexing the data is difficult and error prone due to unclear structure and lack of predefined attributes. The data are disorganized, and it consumes enormous time and resources to extract insights from them. Due to their disorganization, search results may not be very accurate if extra precautions are not taken. Since data originate from various sources, ensuring security to data is a difficult task.

In contrast to unstructured data, structured data are the data which conform to a data model. They have a well define structure, they follow a consistent order, and can be easily accessed and used by a person or a computer program. Structured data have a well-defined structure that helps in easy storage and accessibility. Data can be indexed based on text string as well as attributes. This makes search operation easy and quick. Structured data are usually stored in well-defined schemas such as databases. It is generally tabular with column and rows that clearly define its attributes. Structured data exist in a format created to be captured, stored, organized, and analysed. They are neatly organized for easy access. Consequently, structured data bring inherent benefits when dealing with high volumes of information. (Praveen & Chandra 2017.)

However, according to the article “The age of analytics, competing in a data driven world” published in 2016 by the McKinsey Global Institute, it is important to mention that despite the amount of data collected, structured data account for only about 20% of data within companies.

Structured data depend on the existence of a data model, a model of how data can be stored, processed, and accessed. Because of a data model, each field is distinct and can be accessed separately or jointly along with data from other domains. This makes structured data extremely powerful. It is possible to quickly aggregate data from various locations in the database. However, structured data records can hold unstructured data within it. In a database, data are well organised so that data definition, format and meaning is explicitly known. (Gartner 2008.)

Structured data reside in fixed fields within a record or file. Similar entities are grouped together to form relations or classes, and entities in the same group have same attributes thus, enabling easy to access and query. As a result, data elements are addressable, efficient to process and analyse. (Tandy, Ceolin & Stephan 2016.)

Since all data have a provenance, structured data are not an exception. Structured data come from structured query language (SQL) databases, spreadsheets such as Microsoft Excel, smart manufacturing systems, online forms, sensors such as global positioning system (GPS) or radio frequency identification (RFID) tags, systems' logs and journals, handheld devices, retail, and ecommerce. Considering the nature of structured data, there are advantages that cannot be ignored. Structured data make operations, such as updating and deleting, easy due to its well-structured form. Data can be captured and stored securely, and the scalability is not a problem in case there is incrementation of data. Data analysis is quicker because data is already structured and ready for processing. (Gartner 2008.)

3.2.2 Internal and external data

Internal data is data retrieved from inside the company to make decisions for successful business operations (Arthur 2013). These data are important to determine whether the strategies the company is currently using are successful or if modifications are necessary. As the word internal indicates, data are internal if a company generates, owns, and controls them. There are many benefits to why businesses are interested in the internal data. Internal data can benefit businesses that want to improve efficiency and productivity and businesses that are struggling to be profitable. (Baud, Franchot & Roncalli 2002.)

Internal data are the engine that helps decision makers run and optimize their daily business operations. These data are reliable because the sources of provenance are accurate, thus verification is not always necessary. While internal data are data generated from within the company, they cover areas such as operations, maintenance, personnel (human resources), sales, marketing, and finance. Each area provides a unique perspective, yet the data connect the departments. Furthermore, internal data are extremely important for the success of any business because they are readily available for process and analysis whenever there is such a need. The ability to make quick decisions is enhanced by rapid access to data. Another advantage of internal data is that they show a very clear trajectory of the business without any dependency on outside resources. (Arthur 2013.)

In the other hand, external data are the data generated outside the organization and therefore, the company neither owns nor controls them. These data can range from economic trends, market research, consumers or customers’ behaviour and feedback, and government regulations within an industry. To collect data from outside the sources, different tools and methods such as website data collectors, line of business applications, questionnaire surveys, and various market research are utilized. External data help businesses' decision makers better understand their customer base and the competitive landscape. Businesses need a clear view of what is happening outside to make truly insightful business decisions. (Schatsky, Camhi & Muraskin 2019.)

According to Schatsky et al. (2019), businesses are increasingly seeking better insights by collecting and analysing third-party data. This is not an exception for businesses in industries including financial services, logistics, technology, health care, retail, and the public sector. Most businesses are using external data to gain new insights that can help increase efficiency and revenue.

Mining external data for insights is increasingly important, therefore, businesses know they can gain valuable insights by analysing the data they generate from their operations. However, because internally generated information can leave gaps and show a one-sided picture of operations, businesses are increasingly moving to incorporate new, non-traditional, and external sources of data into their analyses. These data can include almost anything. Collecting and analysing external data enable businesses to foresee and prepare for risks and opportunities which they could otherwise easily have missed with inputs limited to data generated from internal operations, customers, and first-tier suppliers. With globalisation businesses increasingly operate as part of networks consisting of business partners such as suppliers, resellers, channel partners, regulators, and other stakeholders. These networks are often globally distributed and potentially affected by financial, political, and/or environmental factors. (Arthur 2013.)

Furthermore, collecting and analysing external data illuminate how factors such as shifting market trends and behaviours, competitor initiatives, or geopolitical events can affect their business. External data sources are helping businesses to personalize marketing offers, improve human capital decisions, gain new revenue streams by launching new products or services, enhance risk visibility and mitigation, and better anticipate shifts in demand for their products and services. (Wilson 2019.)

3.3 Data analytics for business operations

Data analytics can be used to find alternative ways of assessing business issues. Analysing huge amount of data is complex for a variety of reasons. Traditionally data analysis has considered discrete data that can be handled using well-established and sophisticated quantitative techniques, such as data feeds from sensors. Processing such data is straightforward and can easily be automated. The data generated through social media present a greater challenge. Data are unstructured and come in a range of formats, often with multimedia content and threads of previous textual dialogues embedded within it. Analysing such qualitative textual data requires specific skills. Analysts are working on tools to achieve this, but the ability of software to analyse text remains rather limited. (Wamba, Ngai, Riggins & Akter 2017.)

According to Laursen & Thorlund (2010), Business analytics can be defined as delivering the right decision support to the right people at the right time. Business analytics gives companies, the business user, data, information, or knowledge, which can be acted upon or not. Business analytics is about improving the business’ s basis for decision making, its operational processes, and the competitiveness obtained when a business is in possession of relevant facts and knows how to use them. Laursen & Thorlund insist that business analytics activities must always be based on the business-driven environment, with the management specifying or creating one single information strategy, which must be subject to the company’ s overall business strategy (vision, mission, and objectives). Business analytics is a holistic and hierarchical discipline, stretching from business strategies to sourcing from operational data sources. The business-driven environment must assume full ownership and manage the process. The technically oriented environment must support the process with infrastructure, data delivery, and the necessary application functionality.

Business analytics is a support process. It can be seen as a chain that is only as strong as its weakest link. If, for instance, the analyst cannot derive the right information from data, then all other activities are of no use. The same is true if right data are not delivered the analysts, or if the business users choose not to act based on the new knowledge. (Laursen & Thorlund 2010.)

Finally, operational analytics allows business leaders to identify trends and patterns that inform decision-making, drive optimal operational performance, and cut costs. Predictive and prescriptive operational analytics provides answers to questions such as, if the business can predict outages with data from connected devices and if it is possible to identify areas that need maintenance before a problem occurs. Moreover, operational analytics empowers executives to make strategic decisions that improve the business’s overall performance. It also gives leaders the tools to transform their wealth of customer, operational, and product data into valuable insights that lead to agile decision-making and financial success. (Ohio University 2020.)

3.4 Data analytics for decision making

Data analytics initiatives are critical for transforming traditional organizational decision making into data-driven decision making. The use of data analytics can help businesses collect and analyse data and make decisions and predictions, which can provide useful guidance for further decision making. Using data analytics helps businesses identify, share, and analyse data resources such as production information, logistics information, and price information, as well as encourage them to develop matching data analytics capabilities. Moreover, data analysis capabilities may fully exploit the value of data and provide businesses with insights that are helpful for optimizing allocation, product traceability, operation planning, decision making, and implementation. (Lei, Lin, Ouyang & Luo 2022.)

Lei et al (2022) observe that decision-making quality refers to the correctness and accuracy of decisions, which is evaluated by decision effectiveness and decision efficiency in the process of decision making. Decision effectiveness focuses on the accuracy, precision, and reliability of decision results, whereas decision efficiency considers the time, cost, and other aspects of the resources involved.

Since data are ever present at all stages of the industrial chain, this has transformed the way that businesses make decisions, enabling businesses to quickly identify opportunities and problems, shorten the process of decision making, and improve decision-making quality. Data analytics provide businesses with accurate production information and improve decision-making quality through intelligent prediction functions. Moreover, the application of data analytics in business processes can assist businesses in quickly transferring market and customer information and performing real-time analysis and insights to support decision making. According to this conception, efficient decisions can help businesses control costs, ensure product quality, and improve customer satisfaction. (Lei et al (2022.)

4 SOLUTION IMPLEMENTATION AND DEMONSTRATION

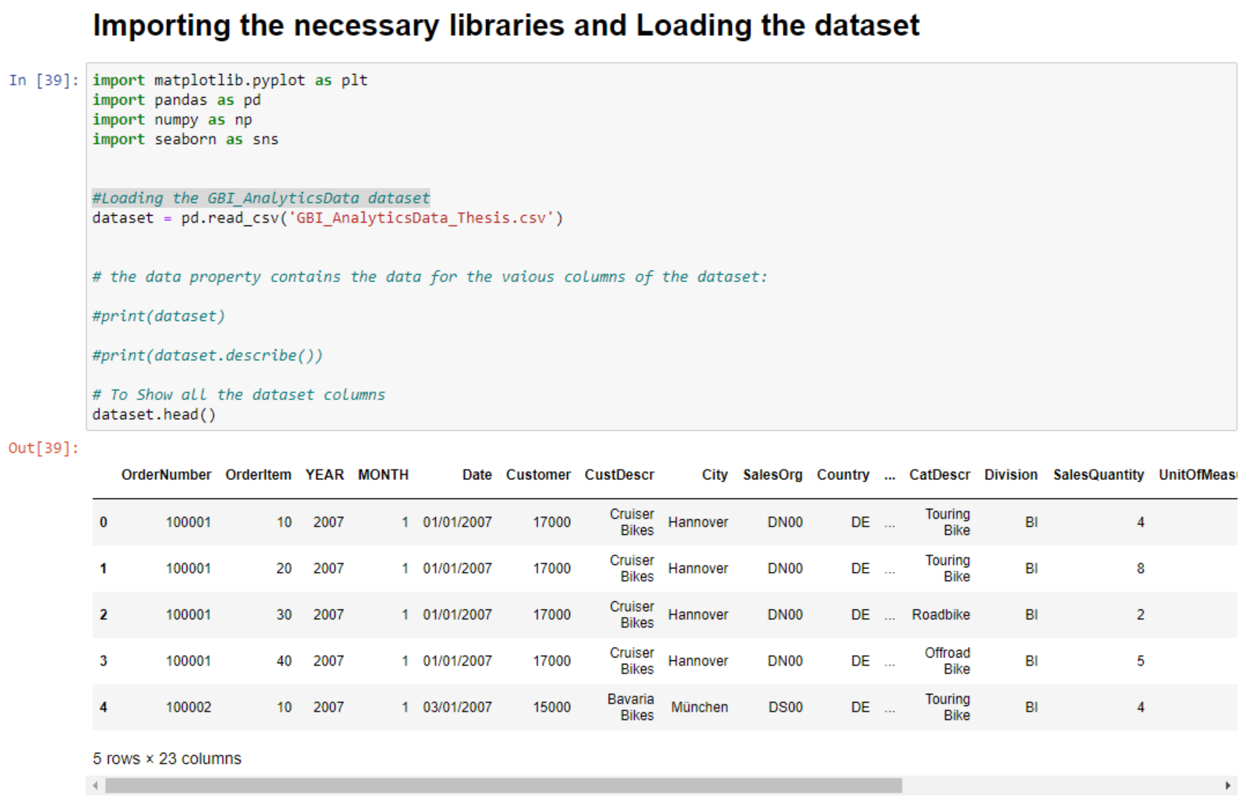
The main idea of this demonstration is to show how a dataset can be manipulated and analysed in order to extract vital information for business operations. Companies collect data from various sources and if good technological techniques are applied for exploitation and analysis, vital information can be extracted for business enablement. For the purposes of this case, the four steps of data analysis (Cleansing, modelling, analysing, and interpreting) will be used instead of five because a dataset was ready.

4.1 About the dataset

The Global Bike Inc. (G.B.I) dataset, which has exclusively been created for SAP UA global curricula used by SAP University alliances was used for this purpose. The dataset is comma separated values format representing the sales report for years 2007 to 2016. It is a reliable structured dataset which is commonly used in SAP analysis courses. Before the transformation, the dataset has 132759 rows and 23 columns. The dataset file name is GBI_AnalyticsData_Thesis.csv which is locally stored in a disk.

4.2 Importing the necessary Python libraries and loading the dataset

Python together with libraries such as Matplotlib, Seaborn, Pandas, and NumPy were at the heart of the exercise. The demonstration was conducted using the integrated development environment (IDE) Jupiter notebook. The process started by importing the necessary libraries and loading the dataset GBI_analyticsData_Theis.csv which is locally stored on a computer as shown in FIGURE 6.



Importing the necessary libraries and Loading the dataset

```
In [39]: import matplotlib.pyplot as plt
         import pandas as pd
         import numpy as np
         import seaborn as sns


         #Loading the GBI_AnalyticsData dataset
         dataset = pd.read_csv('GBI_AnalyticsData_Thesis.csv')


         # the data property contains the data for the vaious columns of the dataset:

         #print(dataset)

         #print(dataset.describe())

         # To Show all the dataset columns
         dataset.head()
```

Out[39]:

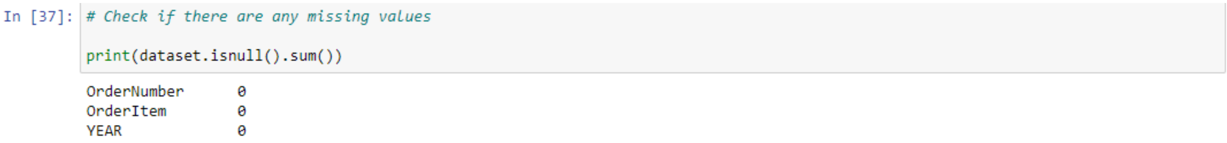
	OrderNumber	OrderItem	YEAR	MONTH	Date	Customer	CustDescr	City	SalesOrg	Country	...	CatDescr	Division	SalesQuantity	UnitOfMeas
0	100001	10	2007	1	01/01/2007	17000	Cruiser Bikes	Hannover	DN00	DE	...	Touring Bike	BI	4	
1	100001	20	2007	1	01/01/2007	17000	Cruiser Bikes	Hannover	DN00	DE	...	Touring Bike	BI	8	
2	100001	30	2007	1	01/01/2007	17000	Cruiser Bikes	Hannover	DN00	DE	...	Roadbike	BI	2	
3	100001	40	2007	1	01/01/2007	17000	Cruiser Bikes	Hannover	DN00	DE	...	Offroad Bike	BI	5	
4	100002	10	2007	1	03/01/2007	15000	Bavaria Bikes	München	DS00	DE	...	Touring Bike	BI	4	

5 rows × 23 columns

FIGURE 6. Library import and dataset loading

4.3 Preparing the dataset (Cleansing and transforming)

After the dataset was loaded and examined, it was time to do some data cleansing using the known methods such as checking for null or NAN rows or values within the dataset as shown below in FIGURE 7. In this case there were null or NAN rows because the dataset was already modelled by SAP.



```
In [37]: # Check if there are any missing values

         print(dataset.isnull().sum())

         OrderNumber      0
         OrderItem        0
         YEAR             0
```

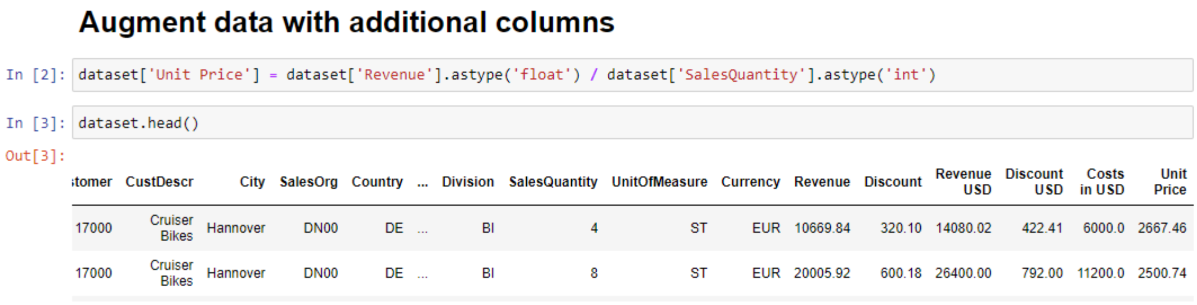
FIGURE 7. Checking the null values

To see all the duplicate rows within the dataset, the following command was used with the argument for keep set to false and the result of the duplicate () function as the index into data frame.

print(dataset.duplicated(keep=False))

The keep argument allows to indicate duplicates The keep argument allows you to specify how to indicate duplicates as follows, the default is 'first': All duplicates are marked as True except for the first occurrence. 'Last': All duplicates are marked as True except for the except for the last occurrence. False: All duplicates are marked as true.

After the dataset was examined and cleansed, there was a need for a new column for the purpose of the analysis ahead. The column Unit Price was added to the dataset by dividing the Revenue by Sales Quantity as the FIGURE 8 indicates.



Augment data with additional columns

```
In [2]: dataset['Unit Price'] = dataset['Revenue'].astype('float') / dataset['SalesQuantity'].astype('int')

In [3]: dataset.head()

Out[3]:
```

ıtomer	CustDescr	City	SalesOrg	Country	...	Division	SalesQuantity	UnitOfMeasure	Currency	Revenue	Discount	Revenue USD	Discount USD	Costs in USD	Unit Price
17000	Cruiser Bikes	Hannover	DN00	DE	...	BI	4	ST	EUR	10669.84	320.10	14080.02	422.41	6000.0	2667.46
17000	Cruiser Bikes	Hannover	DN00	DE	...	BI	8	ST	EUR	20005.92	600.18	26400.00	792.00	11200.0	2500.74

FIGURE 8. Adding columns to a dataset

4.4 Data analysis and visualisation with Python

After preparing the dataset for analysis the demonstration was conducted by answering to four business related questions. Moreover, various plotting methods (techniques) were used to visualise the results. The following questions were asked, and answers and interpretations (explanations) were also provided.

Question *1: What was the best month for sales? How much was earned that month?*

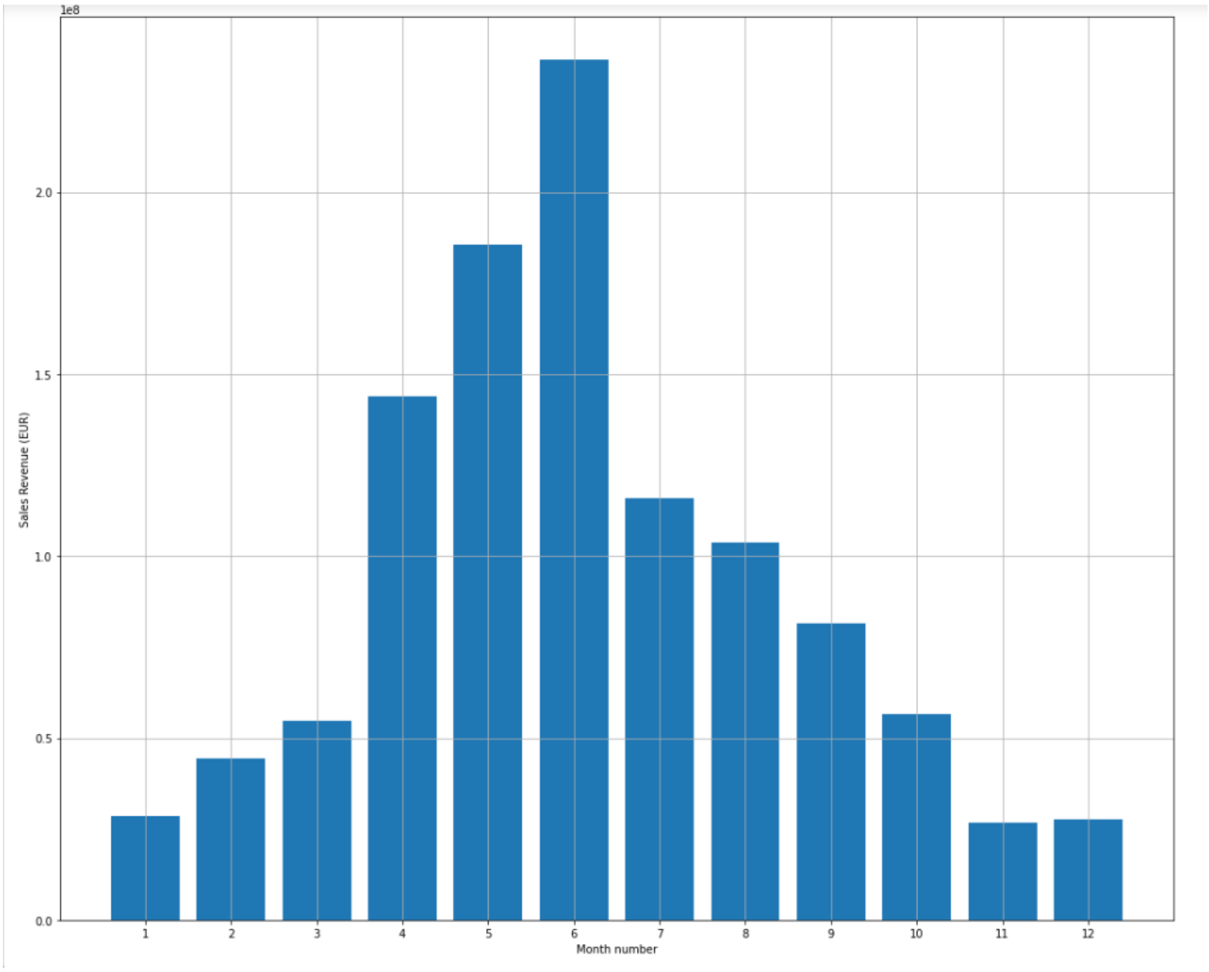


FIGURE 9. The results of the first question

According to the results of revenue generated in FIGURE 9, the best month for sales is June with revenue of 2.363534e+08 and worst month is November with 2.682781e+07 in revenue. The trend seems to be that more bicycles and accessories are sold during the warmer months compared to colder months. Weather has the big impact on the sales of bicycles and accessories. Warmer weather means more biking and colder weather means less biking. The sales start to pick up in spring and decrease in autumn.

Question 2: *What city had the highest number of sales?*

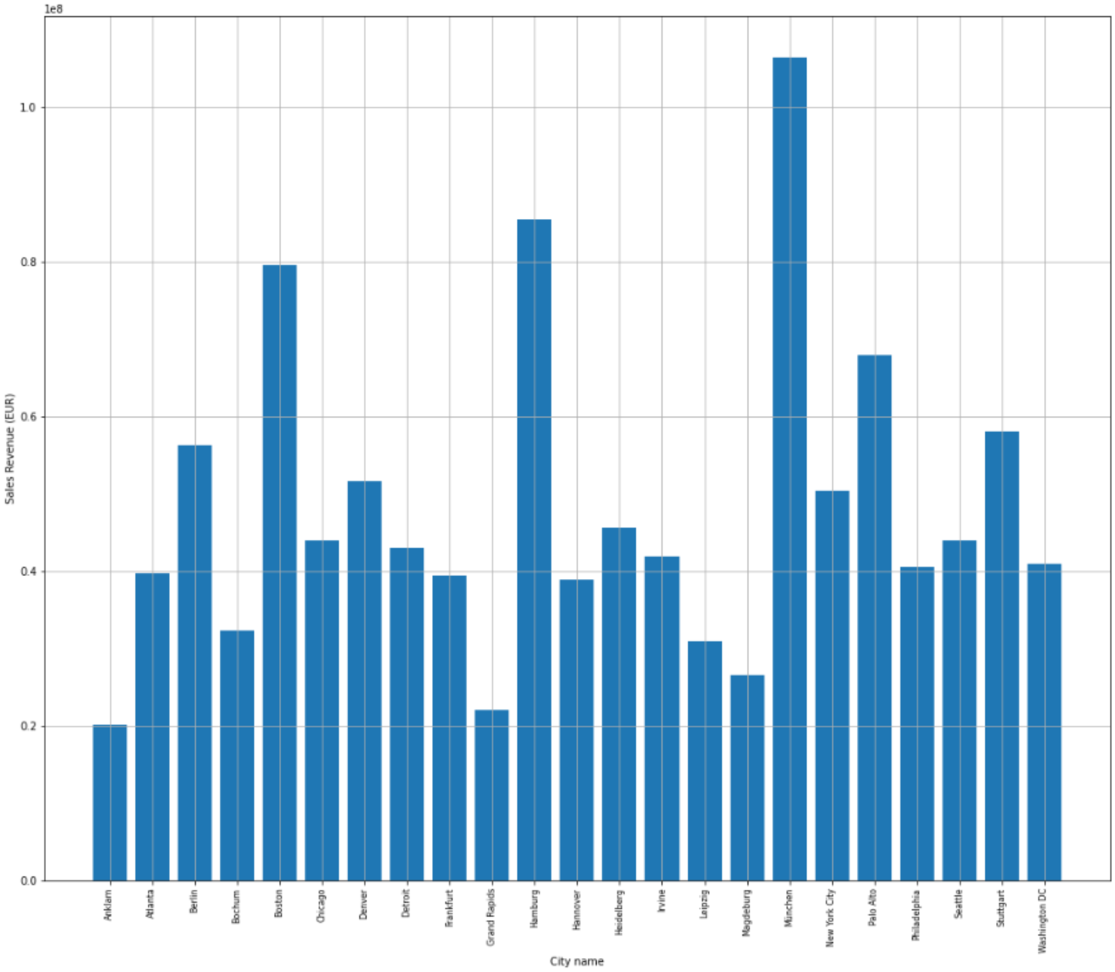
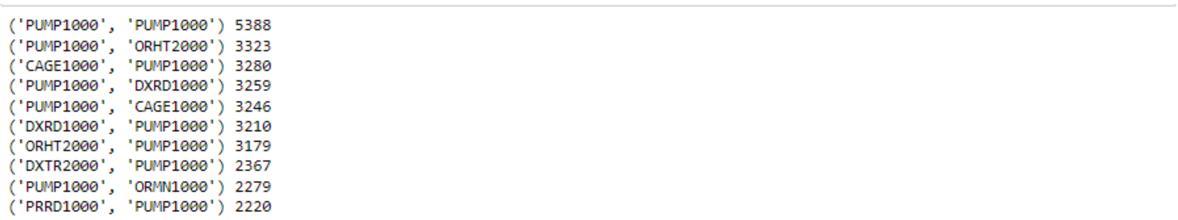


FIGURE 10. The results of the second question

The results in FIGURE 10 indicate that the city of Munich has better revenue in general and Anklam has the worst. This could be because of the culture in the city and the size as well. The city of Munich is more populated than the small city of Anklam.

Question 3: *What products are most often sold together?*



```
('PUMP1000', 'PUMP1000') 5388
('PUMP1000', 'ORHT2000') 3323
('CAGE1000', 'PUMP1000') 3280
('PUMP1000', 'DXRD1000') 3259
('PUMP1000', 'CAGE1000') 3246
('DXRD1000', 'PUMP1000') 3210
('ORHT2000', 'PUMP1000') 3179
('DXTR2000', 'PUMP1000') 2367
('PUMP1000', 'ORMN1000') 2279
('PRRD1000', 'PUMP1000') 2220
```

FIGURE 11. The products sold together

With the information provided in FIGURE 11, the products that can be sold together for promotions can be noticed. Product such as PUMP1000 is commonly sold with ORTH2000 for example. And the water bottles (CAGE1000) are mostly bought together with the pumps (PUM1000). The sales personnel can use this information for advertising and for promotional purposes.

Question 4: *What product sold the most? why was it sold the most?*

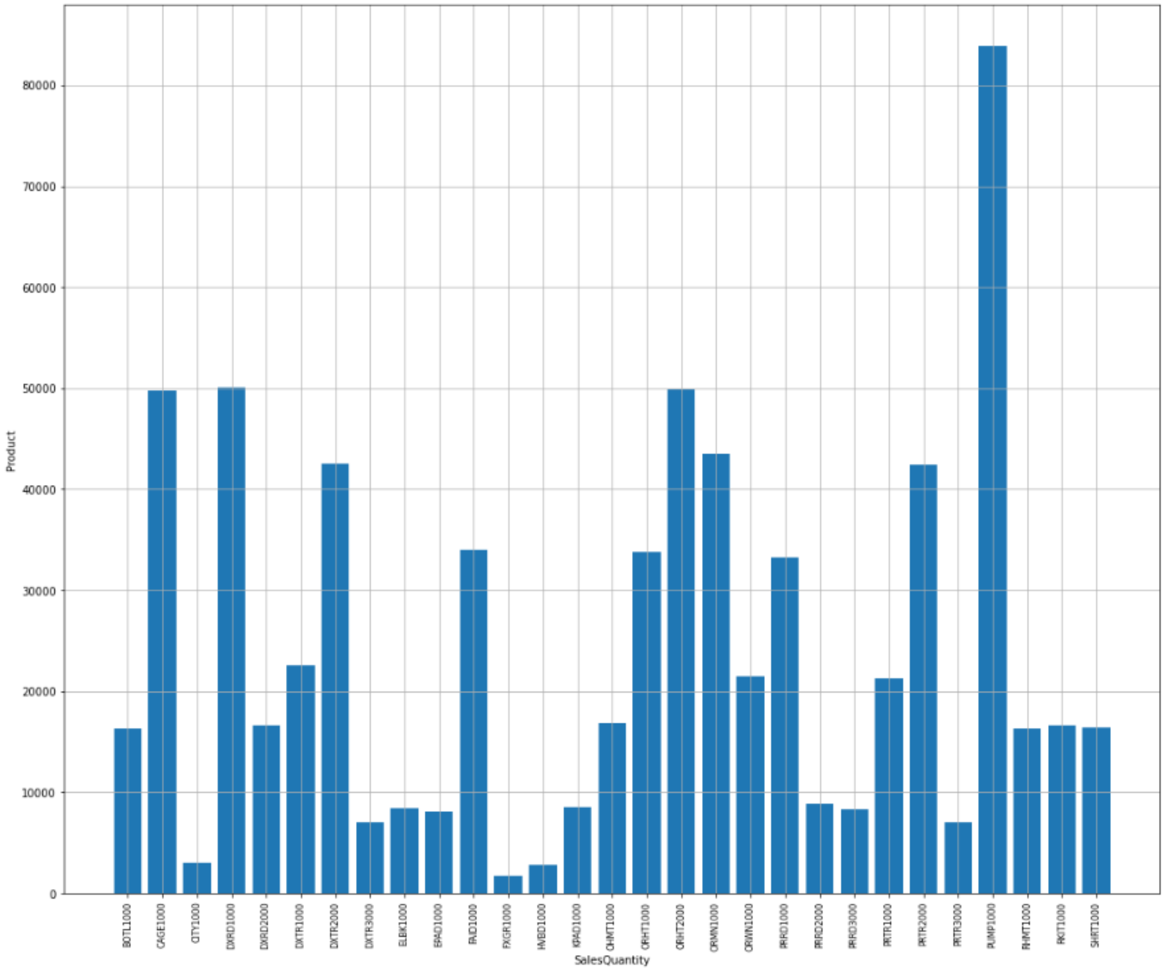


FIGURE 12. The most sold products

From the results shown in FIGURE 12, the pumps (PUMP1000) were the most sold, followed by The CAGE1000, the ORHT200 Men's Off-Road Bike, and the DXRD1000 Road Bike Alu Shimano.

To prove why some products sold most because of prices, the FIGURE 13 shows that there is a correlation between the price and quantity ordered, lesser the prices, higher the demand. In some cases, road bikes are in higher demands though they are expensive.

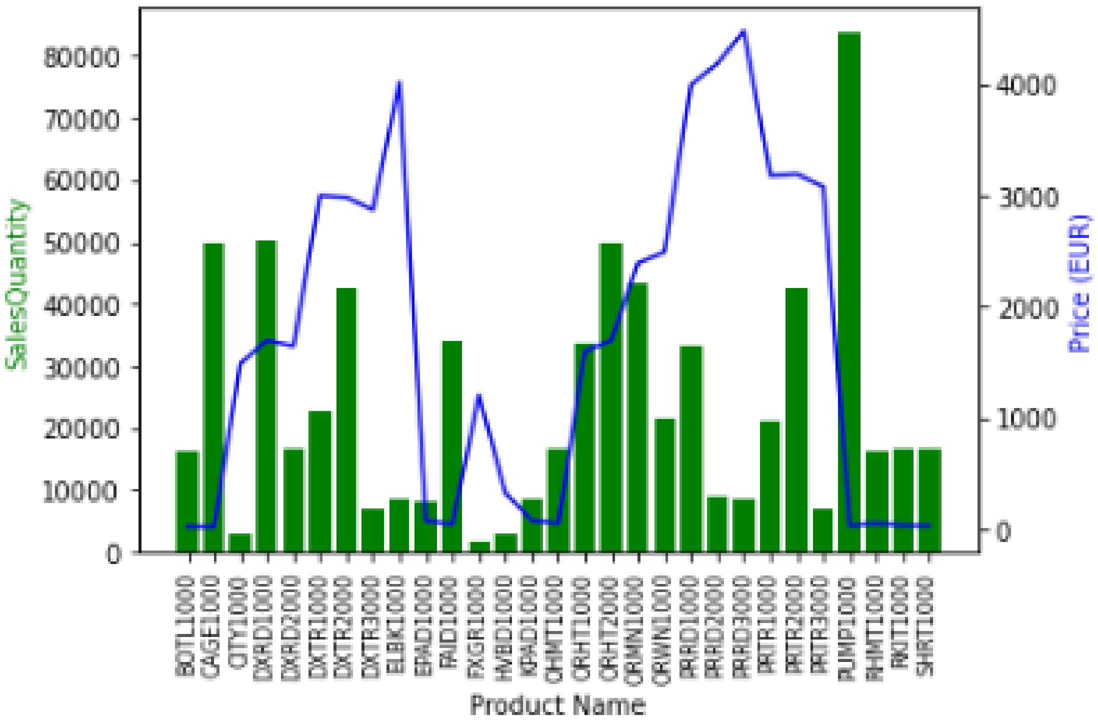
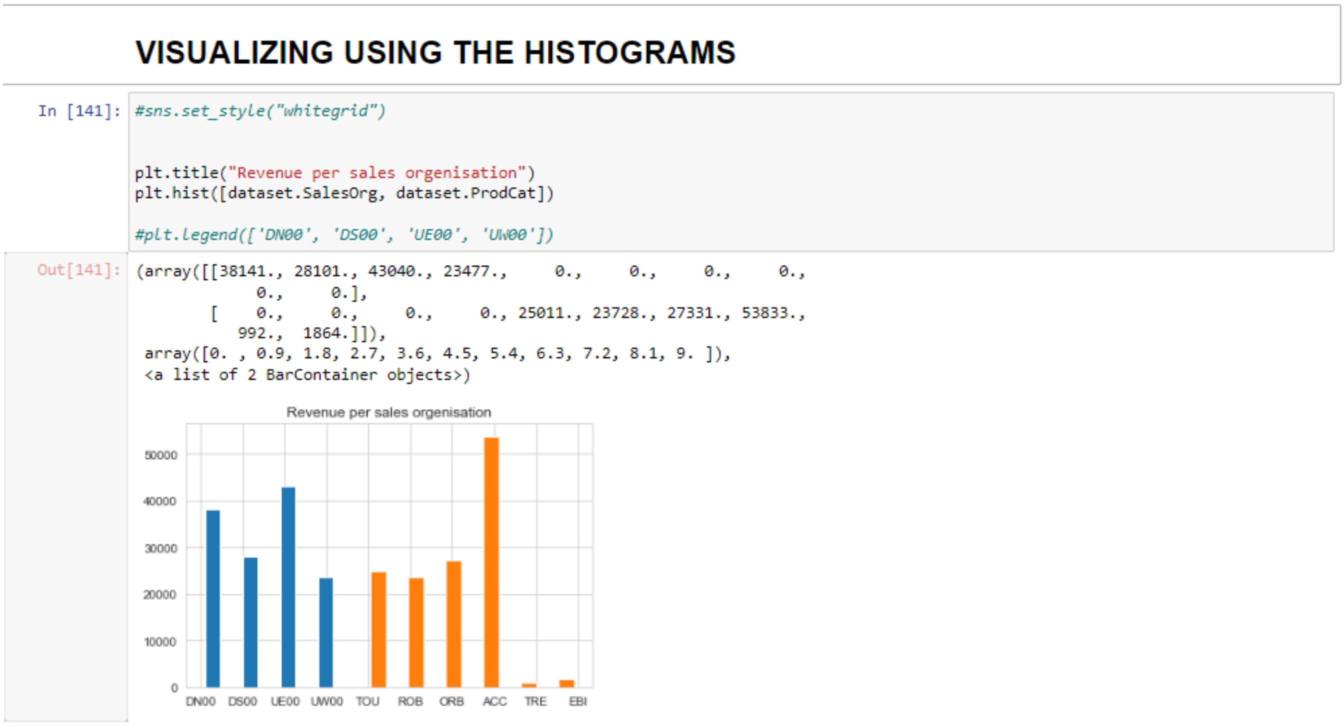


FIGURE 13. Correlation between price and demand

The Python programming language allows different libraries to perform same tasks. In examples above, the Matplotlib module was mostly used for plotting. However, the Seaborn library enables plotting for good looking charts and diagrams as shown in FIGURE 14.



VISUALIZING USING THE HISTOGRAMS

In [141]:
```
#sns.set_style("whitegrid")


plt.title("Revenue per sales orgenisation")
plt.hist([dataset.SalesOrg, dataset.ProdCat])

#plt.legend(['DN00', 'DS00', 'UE00', 'UW00'])
```

Out[141]:
```
(array([[38141., 28101., 43040., 23477.,     0.,     0.,     0.,     0.,
             0.,     0.],
        [    0.,     0.,     0.,     0., 25011., 23728., 27331., 53833.,
           992.,  1864.]]),
 array([0. , 0.9, 1.8, 2.7, 3.6, 4.5, 5.4, 6.3, 7.2, 8.1, 9. ]),
 <a list of 2 BarContainer objects>)
```

FIGURE 14. Visualisation with Seaborn library

4.5 Data analysis and visualisation with Power BI

With the same dataset Microsoft Power BI was used to do some data transformation and analysis in the end. Moreover, Power BI has an option for getting data from various sources as shown in FIGURE 15 below. In this case the Test/CSV was the correct option since the dataset was in .csv format.

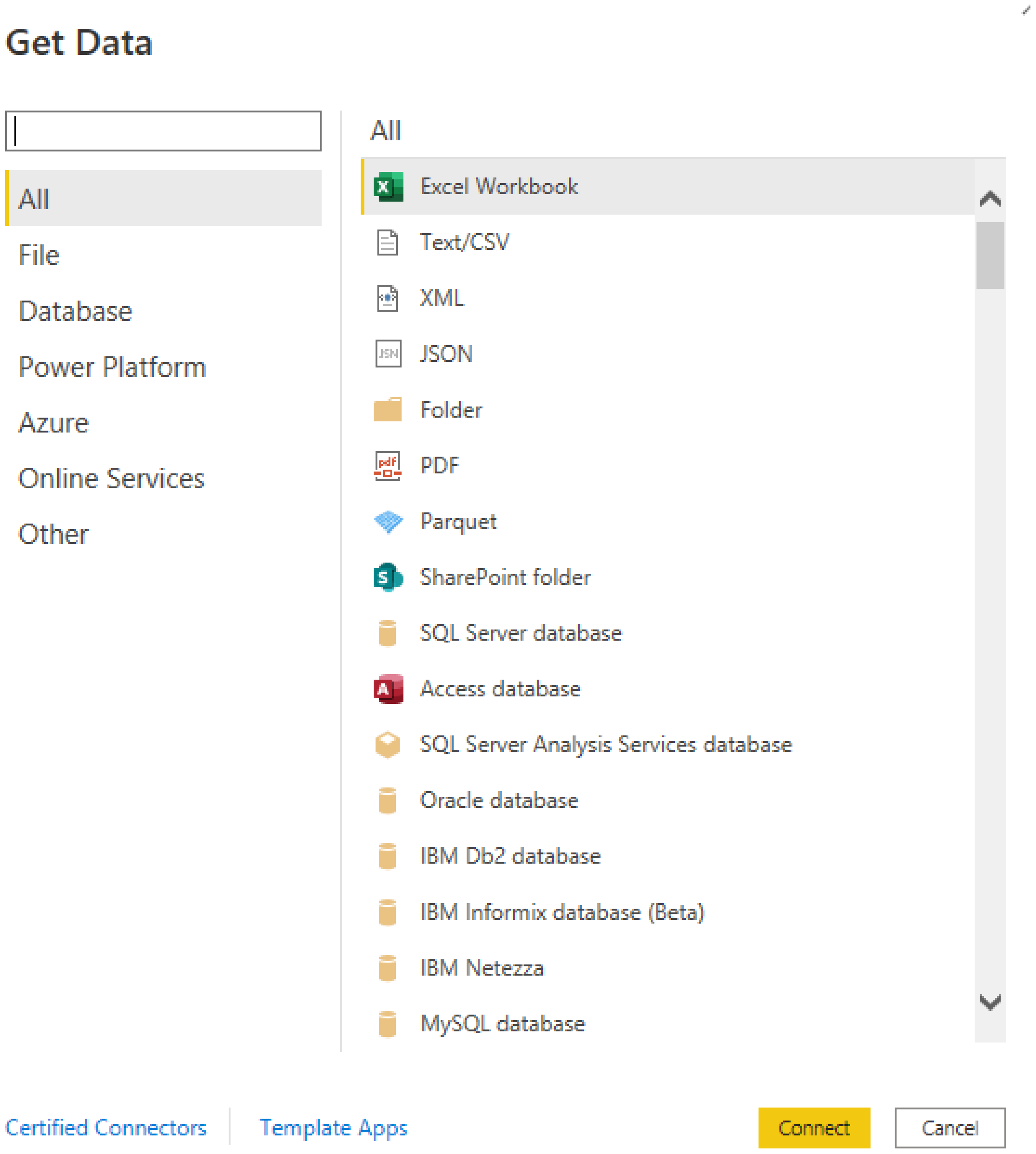


FIGURE 15. Some of data sources supported by Power BI

Without any programming efforts it is possible to analyse data with the help of tools such as Power BI. Below is an example of two bar charts produced with Power BI. The green chart represents the revenue by month in year 2016. The results reflect those produces with Python with June being the best month. The colourful chart represents the revenue by sales organizations for each year from 2007 to 2016. It is clear to see that the USA East sales organization performs better than USA West. The Germany North performs slightly better than Germany as shown in FIGURE 16.

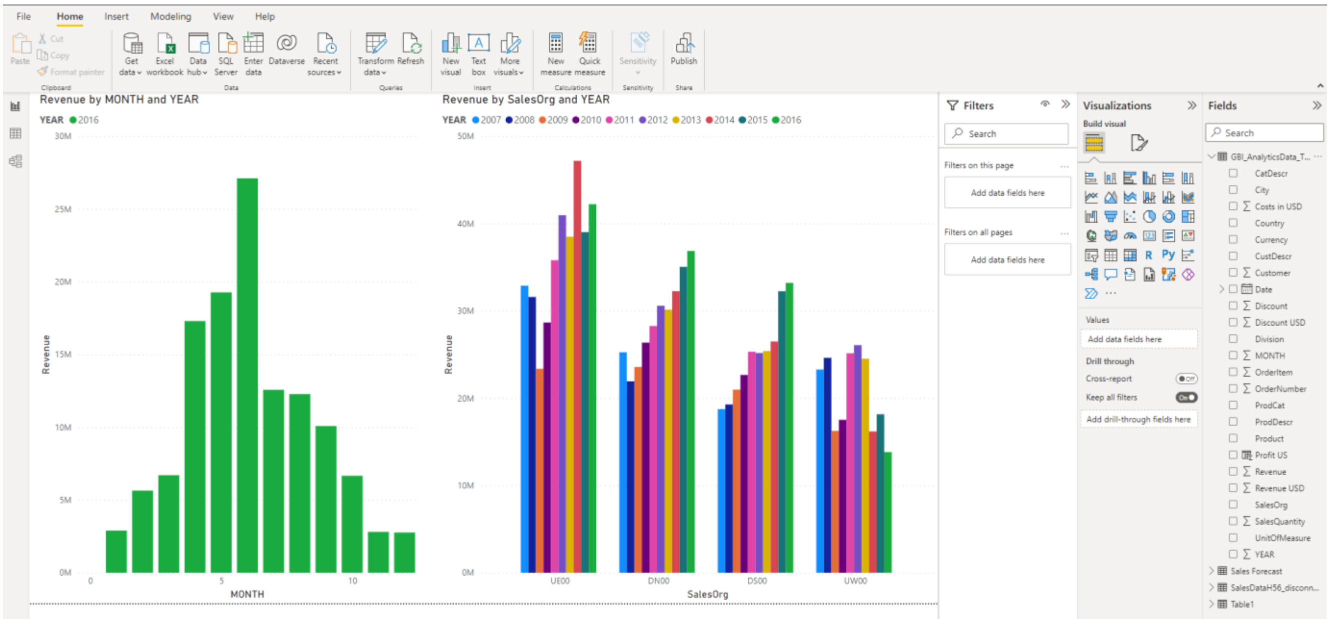


FIGURE 16. Data visualisation with Power BI

Moreover, for better visibility and interactive performance, Power BI has the option for producing dashboards with simple clicks once the reports are made. FIGURE 17 shows a dashboard that was produced with GBI_AnalyticsData_Thesis.csv dataset.

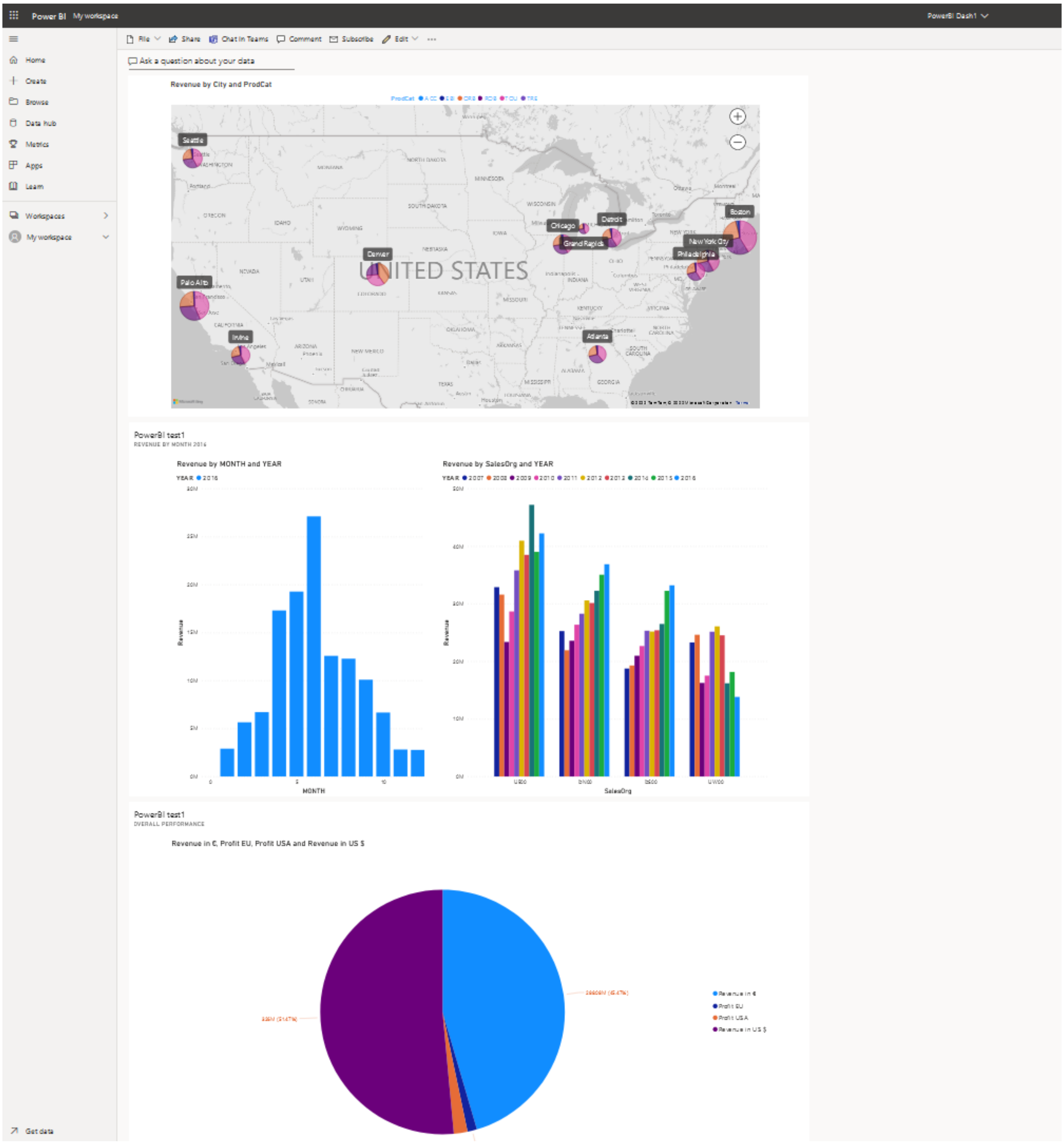


FIGURE 17. A dashboard of sales data produced with Power BI

5 CONCLUSIONS

Raw data can be found in any kind of companies and organizations. This is inspite of the industry, organizations in public or private sector, technology sector or social sector, for profit or non-profit organizations. Data has become the well spring of business operations at least for those who have learned to explore the hidden treasures discovered after data analysis and interpretation.

Companies spend enormous amount of money on gathering data, some even pay money to companies that have already collected, transformed, analysed data in order to find insights leading to gaining and maintaining competitive advantage. Nowadays it is inevitable for any business to successfully operate without continually analysing the information gained from their financial reports, human resources department, sales reports, marketing reports, operational reports, and machine reports. To be successful, companies are employing skilful data scientists, data analysts, and business intelligence professionals in order to utilise the insights to their advantages. Failing to catch the moment can be detrimental to the point of company's survival. Missing information can be a missed opportunity if data analysis and results interpretation is not taken seriously.

The case study proved that with some programming skills it is possible to collect data, store data, manipulate data and do analysis in order to gain insights that can be vital for companies' daily operations. Python programming language has gained popularity because of its ease of use and fast support community which makes it a best choice for data analysis.

Moreover, as proven in the case study, commercial tools such as Power BI make it even easier to collect and analyse data for both the technically and non-technically minded. Visualisation features provided by both the commercial and non-commercial tools make the results easily readable and understandable. With this said, companies need skilful data analysts that can interpret and explain some of the mysteries hidden in data.

Finally, to better achieve decision effectiveness and decision efficiency, businesses should make full use of data analytics tools to accelerate the transformation from traditional to data-driven decision making. Businesses should employ data analytics tools to integrate databases, detect market changes, identify competitive opportunities, reorganize organizational resources, and finally make high-quality decisions.

REFERENCES

Alghushairy, O. & Ma, X. 2019. *Data storage.* Available at: https://www.researchgate.net/publication/335754159_Data_Storage. Accessed 21.5.2022.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. & Kochut, K. 2017. *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. Available at: https://www.researchgate.net/publication/318336890_A_Brief_Survey_of_Text_Mining_Classification_Clustering_and_Extraction_Techniques. Accessed 01.06.2022.

Arthur, L. 2013. *Big data marketing: Engage your customers more effectively and drive value*. Available at: ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=1426518. Accessed 3.6.2022.

Australian Bureau of Statistics 2020. *Statistical language: Quantitative and Qualitative data*. Available at: https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language+-+quantitative+and+qualitative+data.. Accessed 29.5.2022.

Bagui, S. & Earp, R. 2011. *Database Design Using Entity-Relationship Diagrams.* Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=1449552. Accessed 11.05.2022.

Baud, N., Franchot, A. & Roncalli, T. 2002. *Internal Data, External Data and Consortium Data: How to mix them for measuring operational risk*. Available at: https://www.researchgate.net/publication/251242277_Internal_Data_External_Data_and_Consortium_Data_-_How_to_Mix_Them_for_Measuring_Operational_Risk. Accessed 5.6.2022.

Blumzon, C. F. I. & Pănescu, A. T. 2019. *Good research practice in non-clinical pharmacology and biomedicine*. Available at: https://www.researchgate.net/publication/337691364_Data_Storage. Accessed 30.5.2022.

Borek, A., Parlikad, A. K., Webb, J. & Woodall, P. 2013. *Total Information Risk Management : Maximizing the Value of Data and Information Assets*. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=1386476. Accessed 27.05.2022.

Cambridge Assessment International Education 2015. *Data, information and knowledge*. Available at: https://www.cambridgeinternational.org/Images/285017-data-information-and-knowledge.pdf. Accessed 17.5.2022.

Čerešňák, R., Kvet M.201. *Comparison of query performance in relational a non-relation databases* Available at: https://www-sciencedirect-com.ezproxy.centria.fi/science/article/pii/S2352146519301887. Accessed 08.08.2022.

Cuesta, H. 2013. *Practical Data Analysis*. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=1507840. Accessed 19.05.2022.

Connolly, T. & Begg, C. 2015 *Database Systems, A Practical Approach to Design, Implementation, and Management. sixth Edition, global Edition.*

Eiras, J. R. 2011. *Data collection and storage*. Available at: https://ebookcentral-proquest-com.ezproxy.centria.fi/lib/cop-ebooks/reader.action?docID=3021663&query=EIRAS. Accessed 22.5.2022.

Endel, F. & Piringer, H. 2015. *Data Wrangling: Making data useful again*. Available at: https://www.sciencedirect.com/science/article/pii/S2405896315001986. Accessed 19.05.2022.

Fakhitah, R. & Wan, M. N. W. Z. 2019. *A Review on data Cleansing Methods for big data*. Available at: https://www.sciencedirect.com/science/article/pii/S1877050919318885. Accessed on 19.05.2022.

Gartner, T. 2008. *Kernels for structured data, World Scientific Publishing Company*. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=1193194. Accessed 3.6.2022.

Grossman, R. L. 2019. *Data lakes, clouds, and commons: a review of platforms for analysing and sharing Genomic data*. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0168952518302257. Accessed 28.5.2022.

Gupta, N. & Agrawal, R. 2018. *Advances in computers, A Deep Dive into NoSQL Databases: The Use Cases and Applications*. Available at: https://www-sciencedirect-com.ezproxy.centria.fi/topics/computer-science/relational-database. Accessed 13.05.2022.

Han, J. & Pei, J. 2012. *Data mining (Third edition): Data warehousing and online analytical pro-cessing*. Available at: https://www.sciencedirect.com/topics/computer-science/data-cube. Accessed 30.5.2022.

Han, J. & Kamber, M. 2006. *Data mining: Concepts and techniques (second edition)*. Available at: https://books.google.fi/books?id=AfL0tYzOrEC&pg=PA189&dq=Data+cube&hl=en&sa=X&ved=2ahUKEwic4sX17sPxAhWDHXcKHemPAcIQ6AEwAHoECAYQAg#v=onepage&q=data%20cubes&f=false. Accessed 19.5.2022.

IBM cloud Education 2021. *Data visualization*. Available at: https://www.ibm.com/cloud/learn/data-visualization. Accessed 23.05.2022.

Larose, D. T. & Larose, C. D. 2014. *Discovering Knowledge in Data : An Introduction to Data Mining*. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=1699137. Accessed 20.05.2022.

Laursen, G. H. N. & Thorlund, J. 2010. *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting*. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=547101. Accessed 30.05.2022.

Lee, W-M. 2019. *Python Machine Learning*. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=5747364. Accessed 23.05.2022.

Lei, L., Lin, J., Ouyang, Y. & Luo, Xin 2022. *Evaluating the impact of big data analytics usage on the decision-making quality of organizations*. Available at: https://www-sciencedirect-com.ezproxy.centria.fi/science/article/pii/S0040162521007861. Accessed 30.05.2022.

McKinsey Global Institute 2016. *The age of analytics: Competing in a data-driven world*. Available at: https://www.mckinsey.com/~/media/mckinsey/industries/public%20and%20social%20sector/our%20insights/the%20age%20of%20analytics%20competing%20in%20a%20data%20driven%20world/mgi-the-age-of-analytics-full-report.pdf.. Accessed 7.6.2022.

Microsoft corporation 2022c. *Non relational data and NoSQL*. Available at: https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/non-relational-data. Accessed 08.08.2022.
Microsoft corporation 2022a. *What is data visualization?* Available at: https://powerbi.microsoft.com/en-us/data-visualization/. Accessed 23.05.2022.

Microsoft corporation 2022b. *What is Power BI?* Available at: https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview. Accessed 23.05.2022.

MongoDB 2022. *Relational vs. non-relational databases*. Available at: https://www.mongodb.com/compare/relational-vs-non-relational-databases. Accessed 12.5.2022.

Munappy, A. R., Bosch, J., Olsson, H. H., Arpteg, A. & Brinne, B. 2022. *Data management for production quality deep learning models: Challenges and solutions*. Available at: https://www-sciencedirect-com.ezproxy.centria.fi/science/article/pii/S0164121222000905. Accessed 13.05.2022

Murthy, C.S.V. 2007. *Database Management Design*. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=3011229. Accessed 11.05.2022.

Ohio university 2020. *4 Uses for Analytics in Business*. Available at: https://onlinemasters.ohio.edu/blog/4-uses-for-analytics-in-business/#:~:text=Data%2Ddriven%20analytics%20helps%20businesses,decision%2Dmaking%20and%20financial%20success. Accessed 30.05.2022.

OLAP 2022. *What is OLAP*. Available at: https://olap.com/olap-definition/. Accessed 19.05.2022.

Oracle 2022. *What is data management.* Available at: https://www.oracle.com/database/what-is-data-management/. Accessed 13.05.2022.

Praveen, S. & Chandra, U. 2017. *Influence of structured, semi-structured, unstructured data on various data models*. Available at: https://www.researchgate.net/profile/Umesh-Chandra-8/publication/344363081_Influence_of_Structured_Semi-_Structured_Unstructured_data_on_various_data_models/links/5f6c6ee7a6fdcc0086386767/Influence-of-Structured-Semi-Structured-Unstructured-data-on-various-data-models.pdf. Accessed 28.5.2022.

Python 2022. *What is Python? Executive summary*. Available at: https://www.python.org/doc/essays/blurb/. Accessed 23.05.2022.

Rai, R. & Chettri, P. 2018. *Advances in Computers, A Deep Dive into NoSQL Databases: The Use Cases and Applications*. available at: https://www-sciencedirect-com.ezproxy.centria.fi/topics/computer-science/relational-database. Accessed 13.05.2022.

Rosslyn Analytics 2022. *Data: The Art of the Possible How to create value from your different data sources*. Available at: https://www.shrm.org/ResourcesAndTools/hr-topics/behavioral-competencies/Documents/RA_Data_TheArtOfPossible.pdf. Accessed 27.05.2022.

SaS 2022. *Data Management What It Is and Why It Matters*. Available at: https://www.sas.com/en_us/insights/data-management/data-management.html#:~:text=Data%20management%20is%20the%20practice,and%20prepare%20data%20for%20analytics. Accessed 13.05.2022.

Schatsky, D., Camhi, J. & Muraskin, C. 2019. *How third-party information can enhance data analytics*. Available at: https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/smart-analytics-with-external-data.html. Accessed 16.6.2022.

Sivarajah, S., Kamal, M. M., Irani, Z. & Weerakkody, V. 2017. *Critical analysis of big data challenges and analytical methods.* Available at: https://www.sciencedirect.com/science/article/pii/S014829631630488X. Accessed 15.6.2022.

Stobierski, T. 2021. *Data Wrangling: What It Is & Why It's Important*. Available at: https://online.hbs.edu/blog/post/data-wrangling. Accessed 19.05.2022.

Tandy, J., Ceolin, D. & Stephan, E. 2016. *CSV on the web: Use cases and requirements*. Available at: https://www.w3.org/TR/csvw-ucr/. Accessed 16.6.2022.

Techopedia 2012. *Data storage*. Available at: https://www.techopedia.com/definition/23342/data-storage. Accessed 16.5.2022.

Vidhya, V., Jeyaram, G. & Ishwarya, K.R. 2016. *Database Management Systems*. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=5248352. Accessed 11.05.2022.

Wamba, S. F., Ngai E. W.T., Riggins F. & Akter S 2017. *Big data and business analytics adoption and use : a step toward transforming operations and production management?* Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=4810481. Accessed 30.05.2022.

Wilson, E. 2019. *Forecaster's & Planner's Guide to Data*. Available at: https://demand-planning.com/2019/08/26/forecasting-data-types/. Accessed 10.7.2022.