RIGA TECHNICAL UNIVERSITY

Bachelor Thesis

RIGA 2023

RIGA TECHNICAL UNIVERSITY
Faculty of Information Technology and Computer Science

Institute of Applied Computer Systems

Balasuriyage Aritha Dewnith Kumarasinghe

Student of academic Bachelor Program Computer Systems

Student ID No 203AEB014

Analysis of Requirements Identification Approaches for Business Intelligence and Data Analytics

BACHELOR THESIS

Scientific adviser Dr.sc.ing., Professor
Mārīte Kirikova

RIGA 2023

RIGA TECHNICAL UNIVERSITY

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

Institute Of Applied Computer Systems

Work Preformed and Assessment Sheet of the Bachelor Thesis

The author of the graduation thesis:

Student Balasuriyage Aritha Dewnith Kumarasinghe ____________________

(signature, date)

The graduation paper thesis has been approved for the defence:

Scientific adviser:
Dr.sc.ing., Professor Mārīte Kirikova ____________________________

(signature, date)

ABSTRACT

REQUIREMENTS ELICITATION, REQUIREMENTS DEFINITION,
ANALYTICS PROJECTS

Data analytics and business intelligence have been increasing in popularity in recent years. As the popularity of analytics projects increases their complexity can also be expected to increase. This increase in complexity can lead to the misalignment between the expectation of project stakeholders and the solutions which the analysts create. Requirements engineering is a process that helps define requirements for projects based on the expectations of stakeholders. Requirements identification/elicitation is the initial stage of the requirements engineering process and consists of eliciting requirements from stakeholders to understand what they expect the result(s) of a project to be.

This thesis aims to understand the nature of requirements for data analytics and business intelligence projects and to select requirements identification/elicitation techniques that best suit data analytics and business intelligence projects.

This thesis consists of case study analyses where requirements are defined for analytics projects and the analysis of defined requirements to reveal insights into the nature of analytics project requirements. Then selected requirements elicitation methods are assessed based on analytics project attributes as well as contextual attributes. This results in the development of guidelines for the elicitation and definition of requirements for (descriptive and diagnostic) data analytics and business intelligence projects.

This bachelor thesis consists of the 66 pages, 7 figures, 12 tables, and 2 Appendixes

ANOTĀCIJA

PRASĪBU NOTEIKŠANA, PRASĪBU DEFINĪCIJA, ANALĪTISKI PROJEKTI

Datu analīzes un biznesa informācijas popularitāte pēdējos gados ir pieaugusi. Palielinoties analītikas projektu popularitātei, paredzams, ka palielināsies arī to sarežģītība. Šis sarežģītības pieaugums var izraisīt neatbilstību starp projektā ieinteresēto personu cerībām un analītiķu radītajiem risinājumiem. Prasību izstrāde ir process, kas palīdz noteikt prasības projektiem, pamatojoties uz ieinteresēto pušu cerībām. Prasību noteikšana ir prasību izstrādes procesa sākotnējais posms, un tas sastāv no ieinteresēto pušu prasību iegūšanas, lai saprastu, kāds būs projekta rezultāts.

Šī darba mērķis ir izprast datu analīzes un biznesa inteliģences projektu prasību būtību, lai izvēlētos datu analīzes un biznesa inteliģences projektiem vislabāk atbilstošās prasību noteikšanas metodes.

Šis darbs sastāv no gadījumu izpētes analīzes, kurā tiek noteiktas prasības analītikas projektiem, un definēto prasību analīzes, lai gūtu ieskatu analītikas projektu prasību būtībā. Pēc tam tiek novērtētas atlasītās prasību noteikšanas metodes, pamatojoties uz analītikas projekta atribūtiem, kā arī kontekstuālajiem atribūtiem. Tā rezultātā tiek izstrādātas pamatnostādnes (aprakstošās un diagnostikas) datu analīzes un biznesa informācijas projektu prasību noteikšanai.

TABLE OF CONTENTS

INTRODUCTION

The data analytics project, which is carried out for obtaining insights through the computer supported data analysis to provide business intelligence is different from the generic software development project as it is more service oriented than product oriented, because what a data analyst does is provide their services to clients which results in the development of insights regarding the nature or the performance of a specific object or system.

Given this more service-oriented nature of analytics and BI projects, the approach by which requirements can be defined for such projects is different. The standard approach of requirements engineering requirements is equivalent to saying that the product or system that is developed must do ‘this and that’ in a particular way defined by the stakeholders of the project. Whereas requirements within a data analytics or BI project must inform the data analyst(s) specifications regarding how to carry out the data analytics project which are more unchanging due to the consistency in the nature of the tasks that need to be completed within the analytics process (Wirth & Hipp, 2000) (define problem, acquire data, analyze data, present the results). Therefor it can be hypothesized that data analytics and BI projects will only need a consistent set of requirements that define a particular aspect of the analytics process.

Once a consistent set of requirements have been identified and validated a generic (to analytics projects) requirements framework will be created that specifies the requirements that must be defined for all analytics projects. Followed by the creation of guidelines for the selection of requirements elicitation/identification (which is the first phase of the analytics project) techniques that will facilitate the definition of requirements within the context of a data analytics or BI project based on contextual attributes that relate to all projects (Carrizo, Dieste et el., 2014).

The goal of this thesis is to develop guidelines that provides insight into what requirements need to be defined as well as how to elicit/identify these requirements within the context of a data analytics or business intelligence project.

The goal of the thesis leads to the main research question:

What are the requirements that need to be defined for the successful completion of a data analytics or business intelligence project and what requirements elicitation techniques are best suited for the definition of these requirements?

To accomplish this goal and to answer the research question, the following tasks are defined:

- to define the scope of the thesis through the literature review,
- to define attributes that relate to the analytics project,
- to analyze analytics project cases to understand the nature of requirements for business intelligence and analytics solutions, and to identify and validate the requirements that need to be defined for all projects,
- to identify which requirements elicitation techniques will be most suitable for the eliciting each of the defined requirements based on the properties of said requirement,
- to define contextual project attributes that will be used to evaluate the effectiveness of selected requirements elicitation techniques,
- to evaluate requirements elicitation techniques based on the defined contextual project attributes,
- demonstrate the utility of the guidelines within a project context.

The structure of the thesis is as follows:

- Chapter 1: Definition of the scope of thesis based on a review of available literature.
- Chapter 2: Definition of research method used in the thesis in relation to the tasks that are undertaken within the context of this thesis.
- Chapter 3: The definition of project attributes that relate to analytics projects, to provide an overview of what attributes relate to analytics projects.
- Chapter 4: Analysis of case study through a manual text analysis of research articles that relate to analytics project cases, with the goal of defining recurrent requirements that relate to analytics projects.
- Chapter 5: Development of an analytics project model based on the defined analytics project attributes and generic requirements. Resulting in generic requirements framework.
- Chapter 6: Selection of or the definition of requirements elicitation techniques that can be utilized in analytics projects, to define the

requirements within a data analytics project. Resulting in guidelines for the selection of requirements elicitation techniques.

- Chapter 7: Application of developed guidelines within a project case to demonstrate its utility.
- Section 8: Conclusion drawn from the tasks undertaken within the thesis as well as what future research that can be done.
- Appendix 1: Case study analyses of descriptive analytics projects
- Appendix 2: Case study analyses of diagnostic analytics projects

1. DEFINITION OF THE SCOPE OF THE THESIS

This chapter defines the scope of the thesis by:

1. introducing concepts of data analytics and business intelligence,
2. explaining the analytics process and its phases,
3. defining what requirements are within the context of data analytics,
4. describing the difference between requirements elicitation and requirements identification, and finally
5. describing the changes to the requirements elicitation process in relation to the modern workplace.

1.1. Data Analytics and Business Intelligence (BI)

Data analytics is the field of study that relates to systematically analyzing a real-world system by using mathematical and statistical techniques on data that represents the system in question. The processing of data analytics consists of acquiring data, applying mathematical and statistical (data analysis) techniques on the acquired data, and conveying the results of this analysis with the overarching goal of creating value (in the form of insights) through the analysis of the acquired data.

Given the broadness of the data analytics as a field of study different concepts such as data analysis and business intelligence are derivates of data analytics; with data analysis being a sub process within the data analytics process and business intelligence being data analytics that has been tailored analyzing and presenting business related (i.e., financial, and economic) data.

1.2. The analytics process

This thesis considers four phases that constitute a data analytics project. They are as follows:

1. Initiation – This phase of the analytics project is the starting point and can consist of project stakeholders defining the goal(s) of the project, creating a project plan, as well as defining requirements.
2. Acquisition – This phase of the analytics project will be undertaking the acquisition of data from various data sources and transforming it into a format that allows for the application of data analytics techniques. It should be that data

cleaning will be outside of the scope of this thesis due to higher emphasis being placed on the analysis.

3. Analysis – This phase consists of the application of data analysis techniques on the acquired data with the goal of gaining insights with regard to the system to which the acquired refers to.
4. Presentation – This phase consists of conveying the insights produced by the analysis phase.

Within the context of this thesis two of the four different types of data analytics(Cote, 2021) will be considered, namely:

1. Descriptive analytics: This type of data analytics seeks to answer questions that follow the format "What happened?". Current or historical data is analyzed in order to unearth trends or relationships within the data. The end result of descriptive analytics projects are usually data visualizations (mostly in the form of reports) that show the underlying trends or relationships within a given data set.
2. Diagnostic analytics: This type of data analytics seeks to answer questions that follow the format "Why did something happen?". The underlying trends or relationships within data is analyzed in order to come up with a hypothesis, which is then proven or disproven using a (history-oriented) hypothesis testing.

The two other types of data analytics projects are "Prescriptive" and "Predictive" analytics and these two types of data analytics much more complex than the aforementioned descriptive and diagnostic analytics. With this in minds, predictive and prescriptive analytics will be outside of the scope of the thesis with the expectation that these two types will be considered in further research.

1.3. Requirements for Data Analytics Projects

The *ISO/IEC/IEEE International Standard - Systems and software engineering -- Life cycle processes -- Requirements engineering*, in ISO/IEC/IEEE 29148:2018(E) defines a requirement as a 'statement that translates or expresses a need and its associated constraints and conditions.

Within the context of an analytics project a need would refer to what insight the stakeholder wishes to gain about a specific system and the constraints associated with that need are:

1. What are the sources of data that relate to the specific system,
2. What analysis techniques can be used on said data, and
3. What methods(visualizations) can be used to present the insights needed.

Given the complexity of constraints (mentioned above) that relate to user requirements this thesis will consider the constraints as additional requirements for the analytics projects and seek to define them as such.

1.4. Requirements Elicitation vs. Requirements Identification.

Within a data analytics project several different sources can be utilized to define requirements. Within the context of this thesis the following sources will be considered:

1. Stakeholders: who define the goal of the project, the data sources that can be utilized within the project, what software can be utilized within the project and specifications regarding the presenting results of the project.
2. Data Files: which are files that will be utilized within the project and therefore can be used to specify the data extraction process.
3. Databases: which like data files can be used to specify the data extraction process.
4. Software(s)/tools: the software that can be utilized within the project and will constrain factors such as:
    a. what extraction techniques can be utilized using the software,
    b. what analysis techniques can be utilized using the software, and
    c. what visualizations (graphs, plots) can be achieved based on the functionally of the software.

The main differentiation that can be made within these sources is whether they are human(stakeholder) or not (software, data files, database, software). Within this thesis that requirement elicitation will relate only to techniques that 'elicit' requirements from stakeholders and requirements identification techniques will relate to techniques that can be used to 'identify' requirements using nonhuman resources.

1.5. Changes to the requirements elicitation process

With the changes in the paradigm of the modern workplace caused by the COVID-19 pandemic (Ozimek,2020) as well as the advances in communication technology, there needs to be a rethinking of the approach towards requirements elicitation from stakeholders. A great example of this change is reflected in the increased popularity of using video conferencing software for meetings, meaning that for using interviews for requirements elicitation there is no spatial constrain to be considered given that anyone around the world you can interviewed provided, that they have time and a semi-stable internet connection. Taking this into consideration some of the generic requirement's elicitation approaches will be redefined taking modern technologies and workplace practices into account. Although not as the main emphasis, this thesis will consider the digitalization of the requirements elicitation process as a part of its focus.

2. RESEARCH METHODS

This chapter of the thesis defines the steps how the research within the thesis will be carried out:

1. project attributes that relate to the analytics project will be defined.
2. a case study analysis will be conducted and the results of the case study analyses will be used to create a requirements guidelines.
3. the selection of requirements elicitation techniques and their evaluation will be undertaken to provide elicitation techniques guidelines.
4. the result of the thesis will be applied within a project case.

2.1. Definition of Analytics Project Attributes

The definition of analytics project attributes will be done based on the authors conceptualization of an analytics project and within this state the defined attributes and their potential values will be purely theoretical. These theoretical attributes will form the hypothesis that X is an attribute that relates to analytics project and can take the value Y. This hypothesis will be validated based on literature sources that explicitly mentions the attribute or its value in relation to data analytics.

2.2. Case Study Analysis

To accomplish the goal of this thesis, the first steps that need to be taken are the definition of requirements for analytics projects that have already been completed. Doing so will provide further insight into what requirements need to be defined for an analytics project from a more practical point of view.

To achieve this, a case study analysis will be undertaken through the text analysis of research articles that are published as a result of analytics projects. Five case studies will be analyzed for each, descriptive and diagnostic, type of analytics, and for each of the studies the requirements will be defined for one of the phases of the analytics project (Initiation, Acquisition, Analysis, Presentation).

Through this text analysis requirements will be defined for a said analytics project (case study) denoted by “Pro requirement” followed by a number that takes the decimal format where the integer denotes the case study number and the number following the decimal point denoting the project requirement number. For example, Pro requirement 4.2 would be the second project requirement elicited through the analysis of case study number four. These project requirements will contain text (within quotation marks) from the relevant research articles to validate the fact that this requirement does relate to this project and is not something that is fabricated.

Once a project requirement is defined a generic requirement will be derived from said requirement. These generic requirements will be defined in such a way that it relates to the analytics project and act as a ‘requirement for an analytics project requirement’. This will be done under the assumption that these generic requirements will apply to all analytics of a specific type (descriptive or diagnostic) and this fact will be validated by the defining of project requirements that fulfill these generic requirements in all subsequent case study analyses following the case study in which the generic one was defined. The generic requirements will be denoted in a similar manner to project requirements but instead of “Pro requirement”, “Gen requirement” will be used.

When validating a generic requirement by applying it to a subsequent case study a project requirement that fulfils said generic requirements will be defined. The original number used to define the generic requirement will still be used alongside the term “Pro requirement”. Additionally, the font colors will be changed to blue if the requirement can be defined; and red if the requirement cannot be defined. The more times project requirements are defined to fulfil generic requirements a stronger case can be made for their validity.

This part of the thesis can be summarized as the reverse engineering of requirements from the end results of an analytics project, i.e., the research articles produced from the case study. These requirements are then used to derive requirements that relate to all data analytics projects. Figure 2.1 shows a graphical representation of this process.

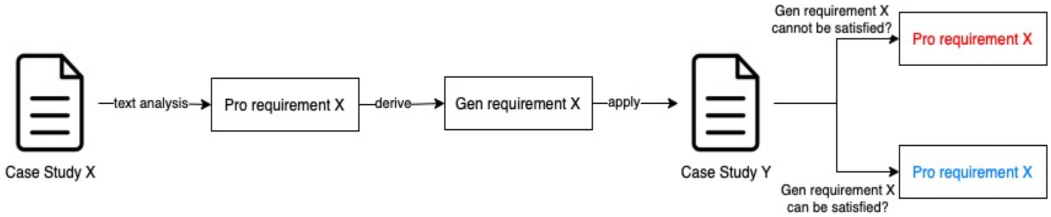


Fig. 2.1. Generic requirements definition and validation process

Additionally, there can be an instance where some segments of the direct quotation from the research article must be cut for the definition of a requirement; this will be done using ~~strikethrough~~ to denote that this part of the quotation is invalid. And comments that relate to the generic requirements will be denoted using a "*" symbol followed by the comment.

It should be mentioned that the case study analysis will be considered as part of the analysis stage of the thesis development process and the generated results of this analysis will be part of the solution phase of the thesis.

Further, the thesis generic requirements that are defined through the case study analysis will be analyzed. This analysis will consist of looking at the generic requirements that were defined from each of the projects and how many times said requirement was validated where the more the requirement was validated the more credibility it has.

Additionally, a model of analytics project attributes based on the results of Section 3 will be used to show which generic requirements can be used to define which analytics project attributes.

As a result of analyzing, the generic requirements will demonstrate which requirements need to be defined for all analytics projects and producing a list of generic requirements which will act as guidelines for requirements that can be applied to all analytics projects.

2.3. Accumulation and evaluation of different requirements elicitation techniques

2.3.1. Selection of requirements elicitation techniques that apply to analytics projects

Requirements elicitation techniques that are appropriate within the context of an analytics project are selected based on their perceived utility as expressed in literature sources.

For each requirements elicitation technique, a small explanation is provided on why this requirement elicitation techniques can be considered as effective within the context of this thesis, and certain alteration regarding how the method is carried out will also be defined in order to facilitate application of the techniques within a data analytics project.

2.3.2. Selection of contextual attributes that aid in the selection of requirements elicitation technique

Certain contextual attributes can be considered as very important for the selection requirements elicitations techniques (Carrizo, Dieste et el., 2014). These contextual attributes consider factors that relate to the elicitor of the requirement the stakeholder and therefor are independent of the type of project being considered. Therefore, the results of (Carrizo, Dieste et el., 2017) will be used to define the most important contextual attributes which relate to requirements elicitation technique selection.

Once the contextual attributes have been defined the selected requirements elicitation techniques will be evaluated based on how effective they would be given a specific attribute value. The results of this evaluation will be helpful for an analyst to choose between the recommended requirements elicitation techniques based on contextual attributes.

This evaluation will result in a table that will serve as guidelines for the selection of requirements elicitation techniques based on the contextual attributes.

2.4. The application of the developed guidelines with a project case

The evaluation of the results of this thesis will be done using by applying the requirements and elicitation techniques guidelines within an RTUs real estate analytics project that relate by contacting a researcher who will take part in the analytics project.

3. DEFINITION OF ANALYTICS PROJECT ATTRIBUTES

Table 3.1 below shows accumulated analytic project attributes that relate to analytics projects, the potential values that these attributes can take, the literature sources that discuss the application of said potential values (it should be noted that the columns with "-" means that these potential attributes were defined by the author), and the last column contains comments regarding these attributes.

Table 3.1

List of Analytics Project Attributes

Project Phase	Attributes	Potential(example) Values of Attributes	Literature source	Comments on the Attribute
Initation	Project Context/Field	Industrial Process	(Runkler, 2020)	This attribute of the analytics project refers to the problem domain within which the data analytics project is being carried out.
		Business		
		Biomedical		
	Object/System of intrest	Organisational unit	-	This attribute defines what object(s) or system(s) the analytics project is concerned with.
		Energy Systems	-	
		Computer systems	-	
	Analytics type	Descriptive analytics	(Cote, 2021)	Defining the type of analytics that is being carried out within the project is very important, especially for the analysis phase of the project since the goal of the project is very dependant on the type of analytics.
		Diagnostic analytics		
		Prescriptive analytics		
		Predictive analytics		
	Project Documentation	Project Plan	-	The documentation that is provided can change the amount of detail an analyst is
		Data dictionary	(Gradwell,1988)	

Project Phase	Attributes	Potential(example) Values of Attributes	Literature source	Comments on the Attribute
		Data Flow Diagram	(Olayan, Patu et al., 2013)	provided on things that relate to company data, available resources and expected results.
Acquisition	Data type used	Numeric data	-	The types of data that are used with any analytics project can be placed into one of these categories, regardless of the format. The type of data used within a project can affect factors such as the type of analytics method that will be used as well as the possible visualisation methods.
		Textual data	-	
		Graphical data	-	
	Dataset(s) used	Preexisting dataset	-	Whether or not the datasets used in the analytics project are already present or need creating changes the work required within the data acquisition phase. The types of data that are used with any analytics project can be placed into one of these categories, regardless of the format. The type of data used within a project can affect factors such as the type of analytics method that will be used as well as the possible visualisation methods.
		Created dataset	-	
	Data source(s)	System(s)	-	The source from which data is collected affects the data extraction methods that are required as well as how external factors such as data privacy laws affect the analytics project.
		Data stores(s)	-	
		Personal(humans)	-	
	Dataset cardinality	Single-source	-	This attribute within an analytics project defines whether or not the analytics project uses a singular data source or multiple sources. Depending on this variable the complexity of the extraction process changes. This is derived from the mathematical term 'cardinality of sets'.
		Multi-source	-	

Project Phase	Attributes	Potential(example) Values of Attributes	Literature source	Comments on the Attribute
	Data extraction technique	Web data extraction techniques i.e Roadrunner	(Salah, Okush et el., 2019)	The technique used to extract data is dependent on the form of data that is extracted and who or what said data is extracted from. This makes the data extraction process a large part of the data analytics project.
		ETL(Extract,Load, Transfer)	(Kadadi, Agrawal et al.,2014)	
		Questionnaire		
	Data Integration	NOSQL Algorithm	(Kadadi, Agrawal et al.,2014)	Data integration is dependent on the cardinality of the dataset. If there are multiple sources of data then data integration might be necessary inorder to combine them into a singular format.
Analysis	Analysis technique/ method/methodology	Correlation Analysis	(Runkler, 2020)	The technique of analysis being done impacts most aspects of the analytics project and vice versa. Therefore the selection of the techniques used in the project and the definition of technique used is essential for the success of the project.
		Regression Analysis		
		Text Analysis	(Gururajan, Clark et al., 2014)	
		Model based Analysis	(Yang, Wang, 2020)	
		Clustering	(Runkler, 2020)	
		Statistical Analysis	(Hakami, Pramanik et al., 2022)	
		Visual Analysis	(Wu, Niu, et al., 2023)	
		Financial Analysis	(Borodin, Mityushina et al., 2021)	
	Analysis tools/software	Pandas(library) in Python 3(Programming Language)	(Unpingco,2021)	The tools that are used to carry out the analysis change the nature of the analytics project similar to the selected analysis technique. A good demonstration of this is how the software will affect "Data types used" because different tools have different inbuilt data types, as well
		R(programming language)	(Ghahramani & Prokofieva, 2021)	

		Teradata	(Chawda & Thakur, 2016)	as rules regarding how those data types can be manipulated.
		Hadoop	(Praveena & Bharathi, 2017)	
Presenting of results	Report Specification	Results formating	-	Most if not all analytics projects produce a report that is given to the clients as a project artefact. Therefore formatting this report in the most acceptable way is vital to client satisfaction.
		Catergorisation of results	-	
	Graphical results	Line graph	-	The visualizations used within a data analytics project are not the most vital characteristic but they determine how well the finding are presented, therefore it is the culmination of the project. It is dependent on the characters within the acquisition and analysis phase.
		Bar Chart	-	
		Histogram	-	
	Interative results	Tableau dashboard	(Praveena & Bharathi, 2017)	Through the use of various software interactive tools such as BI dashboards can be produced as a result of the analytics, one of the most popular visualisation tools is Tableau.

4. ANALYSIS OF CASE STUDIES

This chapter of the thesis contains the analysis of case studies that relate to descriptive and diagnostic analytics (organized by the stages of analytics project to which the requirements refer to) and the analysis of the results produced by the case study analyses. The results of the analyses results is positioned as guidelines that represent what requirements need to be defined for analytics projects. In this chapter only the analysis of one case is demonstrated for each type of analytics. Other case analyses are available in Appendix 1 (for descriptive analytics) and Appendix 2 (for diagnostic analytics).

4.1. Descriptive analytics case study analysis

4.1.1. Results of the analysis of the first descriptive analytics case study

Research article: Students' perceptions of a community health advocacy skills building activity: A descriptive analysis (Hardin-Fanning, Hartson et al., 2023)

Initiation:

Pro requirement 1.1: This analytics project must "explore students' perceptions of the benefits of a discussion activity about a controversial health issue, and to describe the impact of the opportunities to form valid arguments using empirical evidence on students' perceptions of their ability to be advocates"

Gen requirement 1.1: An analytics project must have a clearly defined goal.

Pro requirement 1.2: The methods used in this project will consist of "students were invited to provide feedback on their perceptions of activity benefits. Descriptive analyses were conducted."

Gen requirement 1.2: The analytics project must have clearly defined methods that will be used in order to achieve the mentioned goal.

*The main point of emphasis being what type of data analytics will be required inorder to achieve the goal.

Acquisition:

Pro requirement 1.3: This project will use "post assignment survey (Appendix B) included questions asking how much the activity helped the student learn the following advocacy skills: (1) form a valid argument using scientific evidence; (2) use credible sources when forming opinions; and (3) begin to see themselves as advocates for improving the health of individuals and communities."

Gen requirement 1.3: The data analytics project must have a source(s) of data and how it will be collected.

Analysis:

Pro requirement 1.4: The project will carry out descriptive analysis by using "Descriptive statistics".

Gen requirement 1.4: The specific method(s) that will be used to carry out the analysis will be selected.

Pro requirement 1.5: The project will use IBMs "SPSS" software to conduct descriptive statistics.

Gen requirement 1.5: The software(s) or tools that will be used to carry out the analysis must be explicitly mentioned.

Presentation:

Pro requirement 1.6: The insights provided by the data analytics project will be presented in the form of a bar chart showing the frequency distribution for each of the responses by each category of students (graduate or undergraduate).

Gen requirement 1.6: If and how the findings of the data analysis must be graphically presented should be defined.

Note: The results of the remaining four case study analyses are included in Appendix 1.

4.1.2. Definition of Generic requirements that relate to descriptive analytics projects

A list of the defined generic requirements that relate to each phase of the descriptive analytics project is provided below, these will be referred to as Gen(des) requirements since a distinction between generic and project requirements is not necessary but specifying which style of analytics project this generic requirement refers to is needed. So 'des' is used to denote that the requirement refers to the descriptive analytics. The numbers of requirements trace back to the case studies where the first figure refers to a case study where the requirement was derived from, and a second number refers to the number of a derived requirement in that case study (the details are available in Appendix 1 and at the end of this sub-section). The generic requirements are derived from the analysis of four case studies and rechecked by one additional case study.

Initiation

1. Gen(des) requirement 1.1: An analytics project must have a clearly defined goal.
2. Gen(des) requirement 1.2: The analytics project must have the clearly defined methods that will be used in order to achieve the mentioned goal.
3. Gen(des) requirement 3.1: The level of detail and technicality required when describing the methodology used within the project must be defined based on the knowledge level of the client(s).

Acquisition

4. Gen(des) requirement 1.3: The analytics project must have a source(s) of data and how it will be collected.
5. Gen(des) requirement 2.1: The permissions regarding the use of incomplete data sets must be defined within the context of the analytics project.
6. Gen(des) requirement 3.2: How data was collected from the data source must be defined for an analytics project.
7. Gen(des) requirement 4.1: The data contained within the data source must be defined as well as which of that data will be used for the data analytics.
8. Gen(des) requirement 4.2: Specification regarding the ETL (Extract Transform Load) must be defined for analytics projects.
9. Gen(des) requirement 4.3: How data warehousing is carried out in the analytics project, then the specifications regarding the data warehouse must be specified.

Analysis

10. Gen(des) requirement 1.4: The specific method(s) that will be used to carry out the analysis will be selected.
11. Gen(des) requirement 1.5: The software(s) or tools that will be used to carry out the analysis must be explicitly defined.
12. Des requirement 3.3: Practices that relate to the reliability of analysis results must be defined for the analytics project.

Presentation

13. Gen(des) requirement 1.6: How the findings of the analysis must be graphically presented should be defined.

14. Gen(des) requirement 4.4: How the findings of the analysis can be used must be defined.

Table 4.1 shows the requirements that have been elicited from each case study analysis. A total of 14 requirements have been elicited, with the first case study providing the highest number of generic requirements (6 requirements), and case study five being just used to validate the 14 requirements. The average number of requirements elicited within the first four case studies is 3.33, and the standard deviation is 2.08 whereas if you consider all five case studies then the mean standard deviation becomes 2.80 and 2.38 respectively. Given that the standard deviation and the average requirements are very similar it can be concluded that analytics projects requirements have a lot of variation between them.

Table 4.1

Generic Requirements that have been Elicited from Each Case Study Analysis

Case study 1	1.1	1.2	1.3	1.4	1.5	1.6
Case study 2	2.1					
Case study 3	3.1	3.2	3.3			
Case study 4	4.1	4.2	4.3	4.4		
Case study 5	n/a	n/a	n/a	n/a	n/a	n/a

Table 4.2 shows the results of the validation of the generic requirements from subsequent four case study (CS) analyses with one additional case study (CS 5).

Table 4.2

Validation of Generic Requirements that have been Elicited from Each Case Study Analysis

Gen (des)	1.1	1.2	1.3	1.4	1.5	1.6	2.1	3.1	3.2	3.3	4.1	4.2	4.3	4.4
CS 1	Y	Y	Y	Y	Y	Y								
CS 2	Y	Y	Y	Y	Y	Y	Y							
CS 3	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y				
CS 4	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
CS 5	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	N	N	N	Y

Legend:

🟨 Initiation Phase 🟩 Acquisition Phase 🟦 Analysis Phase 🟥 Presentation Phase

Y The requirement was validated successfully

N The requirement was not validated

The first seven generic requirements that were defined through the analysis of the first two cases studies could be defined for all the case studies. This can be used to infer that that Des requirement 1.1 to 1.6, and 2.1 are requirements that define factors within an analytics project which need to be defined for all analytics projects.

Gen(des) requirement 3.1 although not vital within the context of a research paper is of importance when developing business intelligence or analytics solutions given that the final user of the solution is defined (client(s)) and therefore it is much easier and more important to understand the level of expertise of that individual and the level of detail that is required when presenting the results.

Gen(des) requirements 4.1-4.3 all deal with the data acquisition part of case study 4; this increase in requirements is due to the complexity of the data acquisition methods used to formulate the solution. It is true that all analytics projects must have four phases, but the complexity of the methods used for these phases varies from project to project. This must be a considered when developing guidelines for analysts to elicit requirements and can be addressed by either focusing on removing more technical requirements when eliciting requirements from stakeholders that do not have an extensive level of technical knowledge or by having the elicitor make a judgment regarding the need for stakeholder input regarding the more complex aspects of the analytics project and thereby defining the more technical requirements through introspection.

Gen(des) requirement 3.3 relates to defining factors that relate to the viability of the analytics solution. It would be in the best interest of an analysts developing an analytics project to explicitly state whether there was factor discovered that might affect the viability of the analytics solution; the same way most research articles have a deceleration of competing interest. This will help improve the trust that the client has with the analytics solution.

All descriptive generic requirements appear to have been validated by the fifth case study sufficiently except for Gen(des) requirements 4.1-4.3 which are only necessary for projects that have a more complex data acquisition phase where data integration and warehousing methods been to be explicitly defined.

4.2. Diagnostic analytics case study analysis

4.2.1. Results of the analysis of the first diagnostic analytics case study

Research Article: Diagnostic Analysis for outlier detection in big data analysis (Ridzuan & Zainon, 2022)

Initiation:

Pro requirement 1.1: This project seeks to "addressed the concept of data quality diagnosis to identify the outlier presented in the dataset"

Gen requirement 1.1: The analytics project must have a clearly defined goal.

Pro requirement 1.2: "big data", "Data quality", and "Outlier" were defined within the context of the data analytics project.

Gen requirement 1.2: The analyst be aware of the level of expertise of the client and define the key terminology within the context of the data analytics project accordingly.

Pro requirement 1.3: This analytics project must undertake "Data quality diagnosis was run on the dataset to understand the data and identify errors that appeared in the dataset".

Gen requirement 1.3: The analytics project must have a clearly defined system/object on which the diagnostic analysis is carried out.

Pro requirement 1.4: This project will use "outlier" which is "evaluated by comparing them with the general distribution of the values inside the column" in order to evaluate the dataset..

Gen requirement 1.4: The analytics project must have a quantitative metric(s) that is used to evaluate the system.

Acquisition:

Pro requirement 1.5: The “Global Food Prices Dataset” will be obtained from “Humanitarian Data Exchange”.

Gen requirement 1.5: The analytics project must have defined which data sets will be used and where these data sets will be acquired.

Pro requirement 1.6: The dataset “contains 1048576 records and 17 column listings which consist of the following attributes; Country Id, Country Name, State Id, State Name, Market Id, Market Name, Food Id, Food Name, Currency Id, Currency Name, Type Id, Type Name, UnitMetric Id, UnitMetric Name, Month, Year, price and Commodity Source”.

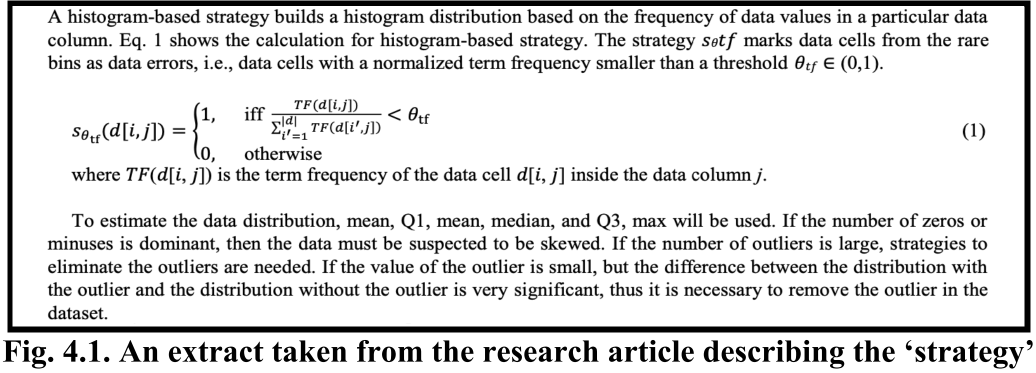
Gen requirement 1.6:The properties to the dataset that is used within the analytics project must be defined.

Analysis:

Pro requirement 1.7: In order to identify the “the outlier” “a histogram-based strategy is chosen”.

Gen requirement 1.7: The ‘strategy’ that will be employed to carry out the diagnostic analysis must be defined.

Pro requirement 1.8: The specifics regarding the histogram based strategy for this project is defined as (Fig. 4.1.):



A histogram-based strategy builds a histogram distribution based on the frequency of data values in a particular data column. Eq. 1 shows the calculation for histogram-based strategy. The strategy $s_\theta tf$ marks data cells from the rare bins as data errors, i.e., data cells with a normalized term frequency smaller than a threshold $\theta_{tf} \in (0,1)$.

$$s_{\theta_{\text{tf}}}(d[i,j]) = \begin{cases} 1, & \text{iff } \frac{TF(d[i,j])}{\sum_{i'=1}^{|d|} TF(d[i',j])} < \theta_{\text{tf}} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $TF(d[i,j])$ is the term frequency of the data cell $d[i,j]$ inside the data column j.

To estimate the data distribution, mean, Q1, mean, median, and Q3, max will be used. If the number of zeros or minuses is dominant, then the data must be suspected to be skewed. If the number of outliers is large, strategies to eliminate the outliers are needed. If the value of the outlier is small, but the difference between the distribution with the outlier and the distribution without the outlier is very significant, thus it is necessary to remove the outlier in the dataset.

Fig. 4.1. An extract taken from the research article describing the ‘strategy’ (adopted from (Ridzuan & Zainon, 2022))

Gen requirement 1.8: An analytics project must have an in depth definition of the strategy that will be used to carry out the analytics. Which includes specific equations that will be used and what the variables within said equation are.

Presentation:

Pro requirement 1.9: This analytics project must result in a “visualization of the price in the dataset with the outlier and without the outlier” showing that the “presence of the outlier provides a significant difference in the graph”.

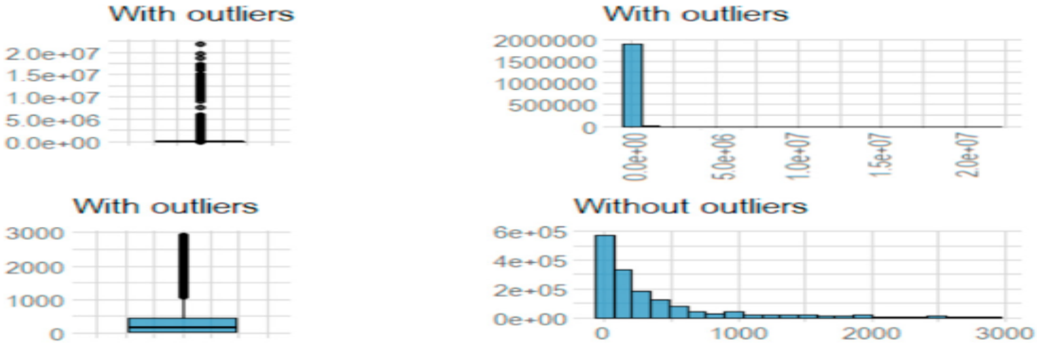


Fig. 2. Outliers in the Global Food Price dataset.

Fig. 4.2. An extract taken from the research article describing the ‘strategy’ (adopted from (Ridzuan & Zainon, 2022))

Gen requirements 1.9: The graphical representations that are required when presenting the diagnostic results must be defined.

*Although this data analytics project is defined as a diagnostics project there are also aspects of prescriptive analytics given that the authors provide suggestions as to what should be done when an “outlier” is found within the data set.

Pro requirement 1.10: The results of the analytics projects must define the factors that are responsible for the “outlier” explicitly as follows “currency, year, prices, location and type of food”.

Gen requirements 1.10: The format in which the results of a diagnostic project are textually presented must be defined

Note: The results of the remaining four case study analyses have been included in Appendix 2.

4.2.2. Generic requirements that relate to diagnostic analytics projects

A list of the defined generic requirements that relate to each phase of the diagnostic analytics project is provided below, these will now be referred to as Gen(dia) requirements since a distinction between generic and project requirements is not necessary but specifying which style of analytics project this generic requirement refers to is needed. ‘dia’ is used to denote diagnostic. The list of the generic requirements that relate to a diagnostic project is as follows:

Initiation:

1. Gen(dia) requirement 1.1: The analytics project must have a clearly defined goal.
2. Gen(dia) requirement 1.2: The analyst must be aware of the level of expertise of the client and define the key terminology within the context of the data analytics project accordingly.
3. Gen(dia) requirement 1.3: The analytics project must have a clearly defined system/object on which the diagnostic analysis is carried out.
4. Gen(dia) requirement 1.4: The analytics project must have a quantitative metric(s) that is used to evaluate the system.
5. Gen(dia) requirement 4.1: The analytics project must define how the results of the data analytics will be validated or verified.

Acquisition:

6. Gen(dia) requirement 1.5: The analytics project must have defined which data sets will be used and where these data sets will be acquired.
7. Gen(dia) requirement 1.6: The properties of the dataset that is used within the analytics project must be defined.
8. Gen(dia) requirement 2.1: The model(s) used within the analytics project along with what said models are used for must be defined for a diagnostic analytics project.
9. Gen(dia) requirement 4.2: The derived data used in the analytics project must be defined.

Analysis:

10. Gen(dia) requirement 1.7: The ‘strategy’ that will be employed to carry out the diagnostic analysis must be defined.

11. Gen(dia) requirement 1.8: An analytics project must have an in depth definition of the strategy that will be used to carry out the analytics. Which includes specific equations that will be used and what the variables within said equation are.
12. Gen(dia) requirements 2.2: The analytics project must have the clearly defined tools that are going to be used and as well as what those tools will be used for.

Presentation:

13. Gen(dia) requirements 1.9: The graphical representations that are required when presenting the diagnostic results must be defined.
14. Gen(dia) requirements 1.10: The format in which the results of a diagnostic project are textually presented must be defined.
15. Gen(dia) requirement 3.1: The format by which the different causes of the issue must be categorized, must clearly be stated when presenting the results of the analytics project.

Table 3.3 shows the requirements that have been elicited from each case study analysis. A total of 15 requirements have been elicited, with the first case study providing the highest number of generic requirements (10 requirements), and case study five being just used to validate the 15 requirements. The average number of requirements elicited within the first four case studies is 3.75, and the standard deviation is 4.19 whereas if you consider all five case studies then the mean and standard deviation becomes 3.00 and 4.00 respectively. The standard deviation being larger than the average meaning that there is a great variability in the number of new generic requirements that had to be defined in order to carry out the analytics project. The conclusions from these findings are incomplete without looking at the validation of the generic requirements.

It must also be noted that the complexity of requirements for diagnostic projects supersedes the complexity of those needed for descriptive analytics projects, shown by the mere amount of text required in order to define said requirements. This is a factor that needs to be taken into consideration when selecting requirement elicitation methods that are suitable for diagnostic analytics projects.

Table 4.3

Generic Requirements That Have Been Elicited From

Each Case Study Analysis

Case study 1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.10
Case study 2	2.1	2.2								
Case study 3	3.1									
Case study 4	4.1	4.2								
Case study 5	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Table 4.4 shows the results of the validation of generic requirements through the definition of project requirements.

Table 4.4
Validation of Generic Requirements That Have Been Elicited From Each Case Study Analysis

Gen (dia)	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.10	2.1	2.2	3.1	4.1	4.2
CS 1	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y					
CS 2	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y			
CS 3	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y		
CS 4	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y
CS 5	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N	Y

Legend: 🟨 Initiation Phase 🟩 Acquisition Phase

🟦 Analysis Phase 🟥 Presentation Phase

Y : The requirement was validated successfully

N : The requirement was not validated

The validation of Gen(dia) requirements shows that Gen(dia) requirements 1.1, 1.2, 1.3, 1.4, 1.5, 1.7, 1.8, 1.9, 1.10, 2.1, and 4.2 can be defined for most (80% of all case studies considered) if not all the analytics cases used for validation. Therefor it can be concluded that those generic requirements apply to all diagnostic analytics projects.

Gen(dia) requirement 1.6 is one that relates to the properties of the datasets used. The reason as to why this requirement could not be defined in some cases might be since the authors did not find it necessary to mention this information within the research article, but it is impossible to use a dataset without knowledge of the properties of the data set.

Gen(dia) requirement 2.1 is the only requirement that relates to the usage of models within diagnostic analytics projects. The models can be considered as a mix between a tool and a dataset in terms of its function within an analytics project but consideration of the need for acquiring the model from a specific source can be considered as part of the acquisition phase of the analytics project. The model also applies to the analysis phase since it influences the method selection process when carrying out the analysis.

Gen(dia) requirement 3.1 is one that relates to the categorical aspect of the solution that the diagnostic analytics project finds. Given that some analytics will result in the discovery that only one factor is affecting the performance whereas others will find multiple factors, Gen(dia) requirement 3.1 only applies in the latter case, but at the start of the project it is not possible to know the number of factors that the diagnostics will reveal. Therefore the recommendation can be made to consider this discussed with a stakeholder in advance to ensure that the opinions of the stakeholder and the analyst are aligned.

Gen(dia) requirement 4.1 is one that relates to the validation of the solution produced by the diagnostic analytics project. The need to validate the solution is something that stakeholders will decide based on the nature of the analytics project, therefore validating a solution does not need to be done unless explicitly mentioned by the clients since this can incur an additional cost of resources to the analytics project.

5. DEVELOPMENT OF GENERIC REQUIREMENTS FRAMEWORK

This phase of the analytics project consists of an analysis of the generic requirements that are define in the previous sub section based on what analytics project variables they can define. The result of this thesis is a generic requirements framework.

5.1. Analysis of Generic Requirements Based on Analytics Project Analytics Project Attributes.

An evaluation of whether a specific project requirement that satisfies a generic requirement can be used to specify the value of an analytics project attribute will be done. This will be done by accessing what project attribute values can be defined from the information provided in the project requirement that initially defines the generic requirement (the first 'Y' within the Tables 3.1, and 3.3).

This will be demonstrated in two tables, with Table 5.1 showing the evaluation of generic requirements that relate to descriptive analytics projects and Table 5.2 showing the evaluation of generic requirements that relate to diagnostic analytics projects.

Table 5.1

Evaluation of Generic requirements that relate to descriptive analytics projects

Gen requirement No.	Case Study No.	Project Attribute	Project Attribute Value
1.1	1	Project Context	Health Science Education
1.2	1	Analytics Type	Descriptive
1.3	1	Object/System of interest	Students
		Datasets used	New data produced by survey
		Data sources	Students
		Dataset cardinality	Single
		Data Extraction Technique	Surveys
		Data integration	Not needed
1.4	1	Analysis techniques	Descriptive Statistics
1.5	1	Analysis Software	IBMs SPSS”
1.6	1	Graphical Results	Bar chart
2.1	2	Dataset used	Pre-existing
3.1	3	Report Specification	Language that must be accessible to all readers.
3.2	3	Data Extraction Technique	Web scraping
3.3	3	Project Documentation	Incomplete data related protocol
4.1	4	Dataset Used	Income Data
		Data types used	Numeric
4.2	4	Data Extraction	ETL process
4.3	4	Data Integration	Data warehouse

Table 5.2

Evaluation of Generic requirements that relate to diagnostic analytics projects

Gen requirement No.	Case Study No.	Project Attribute	Project Attribute Value
1.1	1	Project Context	Data Quality
1.2	1	Report Specification	Basic Concepts within the problem domain must be identified.
1.3	1	System/Object of Interest	Outliers
			Dataset
1.4	1	System/Object of Interest	Evaluation Metric of the Object
1.5	1	Dataset(s) Used	Global Food Prices Dataset
		Data source	Humanitarian Data Exchange
		Dataset Cardinality	Single Source
		Data Integration	Not needed
1.6	1	Dataset Used	The contents of the dataset are defined.
		Data Types Used	Numeric
1.7	1	Analysis Techniques Used	Visual Analysis (Histogram Based)
		Graphical Results	Histograms
1.8	1	Analysis Techniques Used	More detailed description of Histograms
1.9	1	Graphical Results	Histograms, and Box and Wicker Plot
1.10	1	Report Specification	Categorisation of Factors
2.1	2	Analysis Tools	Weather Research and Forecasting (WRF) model
2.2	2	Analysis Tools	IPR module

3.1	3	Report Specification	Categorisation of Factors
4.1	4	Analysis Technique	Model Based Validation
4.2	4	Data Sources	Derived Data

5.2. Creation of Analytics Project Model for Generic Requirements

The creation of the data model is done based on the project attributes that are defined in Table 3.1 and consist of three levels. The first level consists concept of an Analytics Project, the second level consists of analytics project phases and the final level consists of the analytics project attributes that relate each of the project phases as shown in Table 3.1. The final level also contains what project variables can be defined using a specific generic requirement (refer Table 5.1 and Table 5.2). The resulting model can be seen in Figure 5.1.

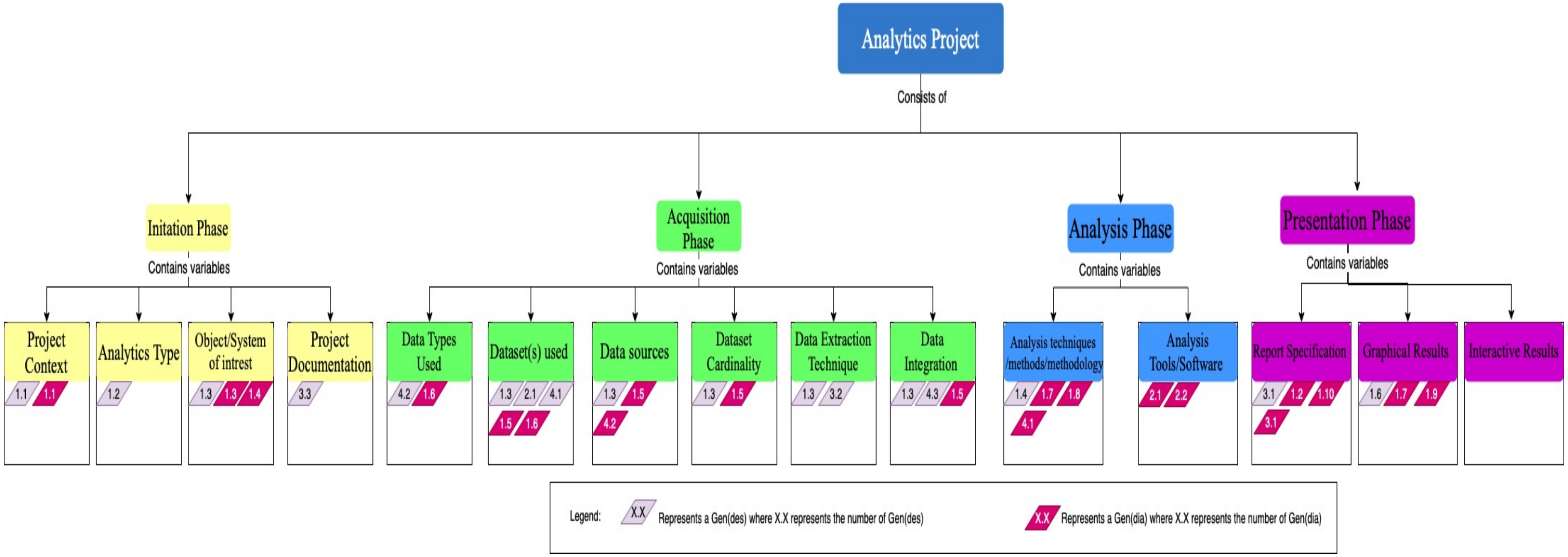


Fig. 5.1: Analytics Project Model for Generic Requirements

5.3. Combination of Gen(des) and Gen(dia) into a single framework based on the results of the analysis

The results of analysis the generic requirements for each type of project revealed that there are some requirements that do not change based on the type of analytics being carried out. Whereas some analytics projects do not need some requirements because they do not have certain processes, e.g., not needing Gen(dia) requirement 4.1 because stakeholders do not want to spend resources on validating results.

The need for said process specific requirements can be defined based on something called ‘secondary requirements’. Secondary (denoted by Sec) requirement is a term used to define a requirement that explicitly specifies a project variable, or a project need which will then indicate if a specific generic requirement needs to be defined for the analytics project or not. Secondary requirements take the form of: if X then define requirement Y. The X denotes a specific project variable or need; the value is selected from a list by the elicitor based either on a previous requirement or is elicited from the stakeholder (client). An example for the use of secondary requirements in case of Gen(dia) requirement 4.1 is as follows:

Sec requirement: If the results of analytics project (<u>must</u>/must not) be validated, then define Gen(dia) requirement 4.1.

Gen(dia) requirement 4.1: The analytics project must define how the results of the data analytics will be validated or verified.

Using above discussed approach the finalized list of generic requirements was ogtained. The generic requirements that seek to define the same project attribute were integrated into a single generic (Gen) requirement (the specific Gen(des) and Gen(dai) are mentioned in comments denoted by ‘*’ symbol). This list of generic requirements also contains newly defined secondary requirements that state whether a specific generic requirement needs to be defined for the analytics project being undertaken, serving as an IF-THEN condition. The list of finalized generic requirements is as follows:

Initiation:

Gen requirement 1: An analytics project must have a clearly defined goal.

Gen Requirement 1 is based on: Gen(dia) 1.1 and Gen(dis) 1.1

Gen requirement 2: The analyst must be aware of the level of expertise of the client and define the key terminology within the context of the data analytics project accordingly.
Gen Requirement 2 is based on: Gen(dia) 1.2 and Gen(dis) 3.1

Gen requirement 3: The analytics project must have a clearly defined system/object on which the analysis is being carried out.
*Gen requirement 3 is based on: Gen(dia) 1.3

Sec requirement 1: If the project is carrying out diagnostic analytics, then define requirement 4.
Gen requirement 4: The analytics project must have a quantitative metric(s) that is used to evaluate the system.
*Requirement 4 is based on: Gen(dia) 1.4

Sec requirement 2: If the results of analytics project must be validated, then define requirement 5.
Gen requirement 5: The analytics project must have defined how the results of the data analytics will be validated or verified.
*Gen requirement 5 is based on: Gen(dia) 4.1

Acquisition:
Gen requirement 6: The analytics project must have the defined dataset(s) that will be used to carry out the analysis, as well as what data is contained within said dataset.
*Gen requirement 6 is based on: Gen(dia) 1.5, Gen(dia) 1.6

Gen requirement 7: The source(s) from which the data (including derived data) is acquired must be specified, alongside how data will be acquired from said source(s).
*Gen requirement 7 is based on: Gen(des) 1.3, Gen(des) 4.2, Gen(des) 4.1, Gen(dia) 4.2

Gen requirement 8: Procedures regarding the use of incomplete data sets must be defined.
*Gen requirement 8 is based on: Gen(des) 2.1

Sec requirement 3: This analytics project will use models then define gen requirement 9.

Gen requirement 9: The model(s) used within the analytics project must be defined alongside the source of said model and its utility within the analytics project.

*Gen requirement 9 is based on: Gen(dia) 2.1

Analysis:

Gen requirement 10: The approach which will be used to carry out the analysis must be defined.

*Gen requirement 10 is based on: Gen(des) 1.4, Gen(dia) 1.7

Gen requirement 11: The specifics that relate to the approach such as mathematical equations, algorithms, analytics techniques must be defined.

*Gen requirement 11 is based on: Gen(dia) 1.8

Gen requirement 12: The tools/software used to in order to carry out the analysis must be defined for the analytics project.

*Gen requirement 12 is based on: Gen(dia) 2.2, Gen(dis) 1.5

Gen requirement 13: Practices that relate to the reliability of analysis results must be defined for the analytics project.

*Gen requirement 13 is based on: Gen(des) 3.3

Presentation:

Gen requirement 14: Specifications regarding the graphical represtations that must be produce by the project.

*Gen requirement 14 is based on: Gen (dia) 1.9 and Gen(des) 1.6

Gen requirement 15: The specifications relating to formatting preferences of the textual report showing the results of the analytics project must be defined for the analytics project.

*Gen requirements 15 is based on: Gen(dis) 1.10, Gen(dia) 3.1 and Gen(des)

The aforementioned framework informs the analyst exactly what requirements need to be defined for their analytics project in order to make the requirements elicitation process much simpler and more straight forward and can be used as guidelines in requirements identification.

6. REQUIREMENTS IDENTIFICATION APPROACHES FOR ANALYTICS PROJECTS

This section of the thesis relates to the identification and elicitation of requirements that relate to analytics project. This chapter consists of:

1. The classification of the generic requirements based on the results of whether they can be met through the use of requirements identification or elicitation.
2. The definition of requirements elicitation techniques that will be considered within the thesis context.
3. The definition what contextual attributes will be used to evaluate the requirements elicitation techniques.
4. The evaluation of requirements elicitation techniques based on the selected contextual project attributes.

6.1. Analysis of generic requirements in relation to the requirement identification and elicitation

Table 6.1 shows a categorization of the generic requirements presented in Section 5.2 based on whether they can be defined using resources that relate to the project instead relying on the user to provide the information. Such as a database system can inform the data analyst what types of data is being held within the database.

Table 6.1

Categorization of Requirements Based on Their Relation to either Data or Users

Requirement	Resource
Gen requirement 1	None
Gen requirement 2	None
Gen requirement 3	None
Sec requirement 1	None
Gen requirement 4	None
Sec requirement 2	None
Gen requirement 5	None

Gen requirement 6	Datafile(s), Database(s)
Gen requirement 7	Datafile(s), Database(s)
Gen requirement 8	Defined Dataset
Sec requirement 3	Defined Dataset
Gen requirement 9	Defined Dataset
Gen requirement 10	Defined Dataset
Gen requirement 11	Defined Dataset
Gen requirement 12	Defined Dataset
Gen requirement 13	None
Gen requirement 14	None
Gen requirement 15	None

If the data analyst has access to the resources that are mentioned in the table, they can be used to define the requirements that satisfy the generic requirements. Project documentation is omitted from this list due to an inability to predict what project documentation will be included and what documents will be presented and what information will be included in the documents.

6.2. Selection of requirements elicitation methods

The elicitation methods were defined based literature except for the case data flow modelling which authors application of data flow diagrams as an means to elicit requirments.

6.2.1. Interview

This can be considered to be one of the oldest requirements elicitation techniques given the fact that it is simply an open conversation (which could be structured or semi-structured) between two individuals within a project development-related context; therefore it is one of the most uncomplicated requirements elicitation techniques. This is proven further by the fact that multiple literature sources ((Quintanilla & Carrizo, 2018),(Carrizo, Dieste et el., 2014)) consider interviews to be a baseline when comparing different elicitation techniques.

(Carrizo, Dieste et el., 2014) points out the various flaws and drawbacks in using this technique such as difficulty in capturing more complex system requirements as well as contextual issues that relate to the stakeholders and problem domain. This is one of the major reasons that lead them to the development of framework for systematizing the selection of requirements elicitation techniques in order to combat the various drawbacks through the use of variety (Carrizo, Dieste et el., 2014). They also go on to say that requirements elicitors tend to overselect or exclusively select this methodology.

Quintanilla & Carrizo, 2018, conclude that requirements elicitation techniques with only one elicitor and one informant (stakeholder) have a cost that is dependent on factors such as stakeholder selection.

When considering all these factors it can be concluded that most if not all projects will use interviews in order to elicit requirements, but should only be relied upon during the earlier stages of a data analytics project or when stakeholder availability is high. Furthermore, this method can be used to get an initial understanding of how constrained a data analytics project is and how many of the attributes within the data analytics project will be defined by stakeholders.

6.2.2. Task observation

This requirements elicitation technique, especially if done anonymously, helps the elicitor to gather requirements independent of direct interaction with the informant. As is mentioned in (Fernandes & Machado, 2016) this methodology requires a significant amount of effort from the elicitor as well as the ability to generalize the result to the entire group using a set of observations. One can see a higher potential utility of this technique within the scope of a diagnostic data analytics project due to the high level of familiarity required between the data analyst and the system that they are diagnosing.

6.2.3. Concept Ranking

This method can be applied within data analytics specifically in descriptive and diagnostic analytics projects because it enables the data analyst to evaluate what aspects of the project context are more important for the client and therefore prioritize them.

It can be recommended that, if concept ranking is selected as a requirements elicitation technique, to alter it from the traditional format and simply incorporate it into another requirements elicitation technique such as questionnaires. Doing so would help to reduce the drawback of being able to elicit requirements from an individual informant or a small group of informants as is highlighted bt Carrizo, Dieste et el., 2014.

This use of the major concepts or components of one elicitation technique within another , i.e., “nested” elicitation techniques can prove to be very useful not only within data analytics but also within other domains that need requirements elicitation techniques.

6.2.4. Questionnaires

Within the context of data analytics, questionnaires can be seen as a better alternative to the interview given the detail-oriented nature of using specific structured questions; but data analytics tends to be very domain specific thereby, if the elicitors do not have the appropriate level of domain knowledge, then their ability to formulate the “right” questions or to interpret the answers given is less reliable. This is especially true for instances where a data analyst's only expertise is in the field of analytical and statistical methodologies and has little to no knowledge of the problem domain.

This is where a deviation occurs with the framework devised in (Carrizo, Dieste et al., 2014) because they state that surveys/questionnaires are suggestable when the level of available information is “Lower”, but it is possible to conclude that this genetically applies assuming that the elicitor has some level of domain knowledge that relates to the project they are doing. This will not be the case within the domain of data analytics if the data analyst is foreign to the problem domain.

6.2.5. Domain Analysis

This method is one that can be especially useful for a data analysts who are new to a specific problem domain or if they are constantly switching to different problem domains. The ontology-based domain requirements elicitation that is formulated in (Lee & Zhao, 2006) is a good approach to domain analysis, were the stakeholders are considered as abstract (because the elicitor gains information from a stakeholder without interacting with them directly but through documents) instead of practical therefor the scope of their methodology is more extensive. This might mean a lot of effort from the data analyst depending on how extensive the domain analysis done is. A counter argument to this is the fact that a domain analysis was done within this thesis when generating the generic requirements for data analytics, and a majority of the generic requirements elicited were done within the first case study analysis, but this might not be true for all problem domains.

In conclusion it can be said that domain analysis has some utility within the realm of data analytics, and it is advisable in most cases, but all requirements that are gained through domain analysis must ideally be verified by a stakeholder.

6.2.6. Brainstorming

Given these techniques group-based approach, a lot more information than from a single stakeholder can be gathered in general. But given the specificity required for the data analytics related requirements, these techniques might be overly chaotic specially if the data analyst has no prior experience in requirements engineering. Newertheless, this technique can be recommended if the project has a lot of stakeholders but their availability in very low. The experience and the confidence the data analyst will have when it comes handling a room full of stakeholders is also a big consideration for techniques consisting of actively handling multiple stakeholders simultaneously, because failure to do so can result in a loss in stakeholders trust in the data analyst abilities.

6.2.7. Data Flow Modelling

Data flow modelling through the creation/analysis of Data Flow (DFD(s)) can be considered as a very effective methodology within the context of data related fields specially if there is a movement (flow) of said data within the system. When undertaking an analytics project, one of two things can occur: a preexisting data flow diagram can be provided by the client(s); the data flow model must be created by the data analyst.

Within the first case the modelling has already been done and the task of the data analyst is to analyze said model and elicit requirements that mostly likely relates to the extraction the data.

An example of elicitation through this method is as follows:

Figure 6.1 shows an example of a DFD which will be used to elicit a requirement.

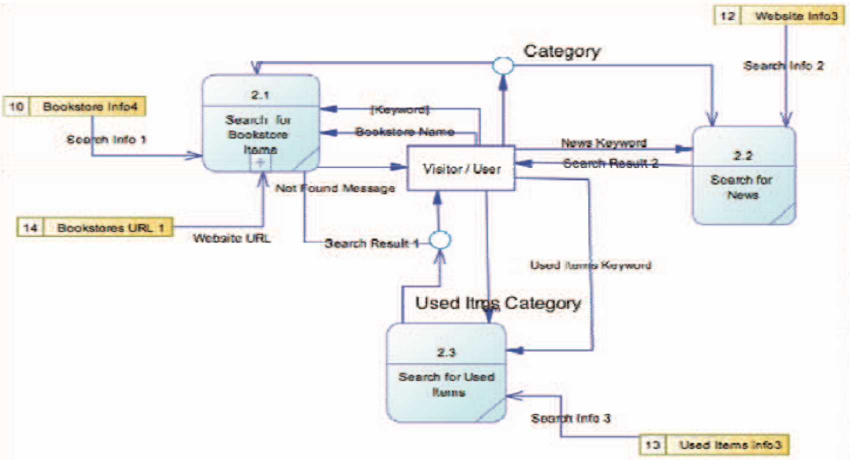


Fig. 6.1. Level-1 DFD for the process "Search" (adopted from (Olayan, Patu et al., 2013))

Form this diagram a data analyst can elicit requirements that relate to the acquisition of data regarding the "News keyword". The requirement can be defined as follows:

"This data analytics project will gather data regarding "News keywords" during the search for new process of this system".

In the second case the data analyst must analyze a preexisting system and create the data flow model, the only reason this can be justified will be when: the client requests the creation of the DFD(s) as a result of the data analytics; the complexity of the system that will modeled allows data flow modelling within the time constraint of the project; the complexity of the system is large and there is a lower time constraint that justifies the creation of a DFD.

6.3. Contextual project attributes that relate to the selection of requirements elicitation techniques.

Describe here what is contextual attribute, why you consider them and what is

6.3.1. Selection of Contextual Attributes that relate to the selection of requirements elicitation techniques

The definition of context attributes are discussed in detail in (Carrizo, Dieste et al, 2014), (Carrizo, 2016) and (Carrizo, Dieste et al, 2017). These three papers, for which D. Carrizo has contributed, follow a sequence were (Carrizo, Dieste et al, 2014) defines a framework for the selection of the requirements elicitation techniques based on "influential contextual attributes" and (Carrizo, Dieste et al, 2017) (which is developed by the same authors) further exploring the contextual attributes influence on requirements elicitation techniques. Table 6.1 shows an interpretation of the results of Table 7 of Carrizo, Dieste et al, 2017, where the list of project attributes (that have been confirmed to have an influence on the effectiveness of a requirements elicitation technique) is discussed regarding it fit within an analytics project context (Comment column in the table).

Table 6.2

Contextual Attributes that will be Used for the Selection of Requirements Elicitation Techniques

Aspects	Context Attributes	Comment	Will attribute be considered in creation of evaluation metric
Elicitor	Domain Familiarity	This attribute applies to how familiar an analyst is with the problem domain. This attribute is especially useful within the context of an analytics projects which can relate to a multitude of domains.	Yes
Stakeholder	Geographical Aspects (Temporal Co-Location)	This attribute relates to whether the changes brough on by a stakeholder being living in a different time zone can affect a requirements elicitation technique.	Yes
	Personal Aspects	The term “personal aspect” relates to a lot of different things therefore this attribute will be defined as anything noteworthy regarding the stakeholder that will influence his/her ability to provide viable requirements. Methods used for the definition of attribute values lies outside the scope of this thesis therefor it will not be considered.	No
	Information Source (Level of expertise)	This attribute looks at the stakeholder as a	Yes

		source of information and any factor that influences the stakeholder's validity as a source of information i.e., "stage of development of an expert" (Carrizo, Dieste et al, 2017)..	
Problem Domain	Information type	The information type within the project domain does not apply to analytics project because the primary focus of these projects is data; therefore this attribute only takes on the value "data".	No
	Task types	The tasks that relate to the analytics is already being addressed by the definition of generic requirements. Therefore task types are irrelevant for the selection of attributes.	No
	Complexity	The complexity of the project is a very subjective metric within analytics projects; therefore this attribute is not the best way to define the effectiveness of a requirements elicitation technique for analytics projects.	No
Solution Domain	Product Types	This attribute remains consistent within the scope of an analytics project because the result of a diagnostic and descriptive analytics projects will always be the insight regarding some object or system.	No
Elicitation Process	Communication type	This relates to the definition of medium by which the elicitation process can occur. This can help define whether	Yes

		the elicitation process can be done remotely using a specific technique.	

Table 6.1 shows contextual project attributes that can be used to evaluate how suitable a requirements elicitation technique is based on context attributes. These attributes can be divided into four different aspects (Elicitor, Stakeholder, Problem domain, solution domain, Elicitation process) namely. According to the relevance of the aspect attributes to data analytics and business intelligence projects, discussed in Comments column and depicted in with “Yes”and “No” in the last column of Table 6.1, Table 6.3 is created that can be used as a guideline for choosing the requirements elicitation technique in analytics projects. The R/N (relevant/not relevant) values are inherited from the related works of Carrizo et al.

The Table 6.3 condenses the results of the evaluation of the requirements elicitation techniques based on the four contextual attributes that were defined in section 6.1. In the table, “R” is used to indicate that a particular technique is recommended when an attribute takes a certain value, whereas “N” stands for not recommended. “R+” means that the effectiveness of that technique benefits from a particular attribute value and “– “means that these contextual attributes do not affect the requirements elicitation technique.

Table 6.3

Evaluation of Requirements Elicitation Techniques Based on Context Attributes

Contextual Attribute	Values	Requirements Elicitation Technique						
		Interview	Task Observation	Questionnaires	Concept Ranking	Domain Analysis	Brainstorming	Data Flow Modelling
Domain Familiarity of Elicitor	High	R	N	R	R	N	-	R
	Low	R	R	R	N	R	R	R
	None	R	R	N	N	R	N	N
Temporal Co-location	Same Time	R	R	R	R	-	R	R
	Different Time	N	N	R	R	-	N	N
Level of expertise of stakeholder	Novice	N	N	R	R	N	R	R
	Knowledgeable	R	R	R	R	R	R	R
	Expert	R	R	R	R+	R+	R	R
Communication Type	Remote	R	N	R+	R+	-	R	R+
	Physical	R	R	R	R	-	R+	R

Within the context of an analytics project these contextual attributes must be defined based on the specifications presented in sections 6.3.2 to 6.3.5. Based on this information the user can redact the columns that include values that relate to contextual attribute values that don't relate to the project. Followed by a counting of all the number of Ns in each column, this number will be negative ranking the requirements elicitation techniques based on the specific project context.

6.3.2. Domain Familiarity of Elicitor

Values that the attribute can take (Carrizo, Dieste et al, 2017):

High: The elicitor has 1 or more years of experience within a specific domain.

Low: The elicitor has 6 months to 1 year of experience within a specific domain.

None: The elicitor is new to the domain.

Evaluation of requirements elicitation methods based on attribute "Domain Familiarity of Elicitor:

Interview: Can be recommended regardless of the domain familiarity due the luxury of being able to clarify any and everything that the stakeholder says.

Task Observation: best suited when elicitor has low or no domain knowledge because task observation can provide insight into the processes that occur within the domain.

Concept ranking: the reliable use of this techniques requires familiarity and ability to define domain concepts, therefore it can only be recommended when there is high domain familiarity.

Questioners: given that the questions contained within the questioner will relate to the definition of analytics project requirements, which will need only a minimal amount of domain knowledge, this method can be recommended if domain familiarity is Low or High.

Domain Analysis: Domain analysis is required for the acquisition of domain knowledge and the benefit of carrying out a domain analysis decreases if domain familiarity of the elicitor is high; therefore the use of domain analysis can only be recommended if domain familiarity is low or none.

Brainstorming: this method consists of the creation of ideas, therefore, there can be a certain level of technicality involved. This can help bridge the gap between an analyst who has little domain knowledge and lot of analytics knowledge and stakeholders that have a lot of domain knowledge and little analytics knowledge. Therefor it can be recommended at any level of domain familiarity except None.

Data Flow Modelling: this method has a similar nature to brainstorming where the analyst can be an expert in data flows and the client can provide the domain knowledge therefor it can be recommended at any level of domain familiarity.

6.3.3. Temporal Co-Location

Values that the attribute can take (Carrizo, Dieste et al, 2017):

Same time: Elicitor and stakeholder are in the same time zone.

Different time: Elicitor and stakeholder are in different time zones.

Evaluation of requirements elicitation methods based on attribute “Temporal Co-Location”:

Interview: an interview becomes increasingly difficult to conduct (online) if the the time difference between the location of the elicitor and the informant increases because the number of working hours that coincide for the two parties decreases. Therefore interview can be recommended only if the time difference is small or none.

Task observation: requires the elicitor to be present with the informant ideally in the same room; therefore task observation works the best when it is in same time.

Questionnaires: given that questionaries can be filled out by the client at any time of their choosing, they can be recommended regardless of time.

Concept Ranking: Assuming that concept ranking is done electronically this can be recommended regardless of the temporal attributes.

Domain analysis: can be done through research that does not involve human systems therefor it is not affected by this contextual attribute.

Brainstorming: due to this methods similarities interviewing means that it reacts in a similar way to this attribute. Therefor it can also be recommended only if there is a small-time difference.

Data flow modelling: unless done without the help of the client this method has the same reaction as interviews and brainstorming to this attribute.

6.3.4. Level of Expertise of Stakeholder

Values that the attribute can take (Carrizo, Dieste et al, 2014):

Novice: More than five years in domain or role.

Knowledgeable: 2 to 5 years in domain or role.

Expert: Less than two years in domain or role.

Evaluation of requirements elicitation methods based on attribute”Level of Expertise of Stakeholder”:

Interview: the higher the level of expertise of the stakeholder the more knowledgeable the stakeholders are and the more viable the requirements which they provide are. there for interviews may be ineffective if the stakeholder is a novice therefor interviews cannot be recommended.

Task Observation: the level of expertise of the stakeholder being observed can affect how accurate the information elicited is so doing task observation of a novice can result in an inaccurate representation of a process. So, task observation can be recommended only if the stakeholder is not a novice.

Questionnaires: the questionnaires can be considered to be independent of the expertise of stakeholder assuming that the elicitor is aware of the level of expertise and interprets the results accordingly. If this awareness of expertise is maintained, then questionnaires can be recommended regardless of this attribute.

Concept Ranking: a similar approach as was considered in the case of questionnaires applies here were as well the ranking done by experts, and knowledgeable individuals will be considered more important than those done by the novices.

Domain Analysis: The domain analysis is done through the use of various online sources , then the credibility of the creator of said resource must is important. Therefor if the author is a novice, their information domain knowledge is not credible.

Brainstorming: this method will consist of multiple stakeholders therefore whatever is being said can be peer reviewed therefor the information that will be acquired will be reliable regardless of the level of expertise.

Data Flow Modelling: given that data flow modeling will be done primarily by the analyst and the stakeholder provides only domain specific validation or insight this process can be effectively done cooperatively with someone how has a limited amount of domain knowledge.

6.3.5. Communication Type

Values that the attribute can take (Carrizo, Dieste et al, 2017):

Remote: Using remote conferencing technology

Physical: Where the elicitor and stakeholder are in the same room.

Evaluation of requirements elicitation methods based on attribute "Communication Type":

Interview: interviews can be done in person or remotely and but face to face communication although "face to face has the highest level of satisfaction, comfort, and perceived engagement, during the negotiation and elicitation stages" (Rodina, Amjed et al., 2012). This finding is pre COVID pandemic therefore might be seen as situation dependent.

Task Observation: Doing task observation virtually is not recommended since facilitating that process remotely sounds unnecessarily complicated. Doing this process in person is the only plausible way of eliciting requirements.

Questionnaires: there added functionality of creating virtual questionaries since additional requirements elicitation methods like concept ranking can be incorporated within questionaries. Taking this factor into account it can be recommended to use virtual questionnaires as opposed to physical ones.

Concept ranking: If a concept map is done virtually it takes the form of questionnaire therefore there cannot be virtual concept ranking, but concept ranking remains its own elicitation technique if it is done in person and can be recommended.

Domain Analysis: if virtual resources are used then effect that this attribute has is negligible.

Brainstorming: holding a brainstorming session in person is better given that there can be a higher level of interaction between the participants and the larger(more than one) number in of stakeholder participating in each session is higher (when compared to an interview) the likelihood that interferences will happen is higher. Therefore doing a brainstorming secession in person is recommended over a virtual one.

Data Flow Modelling: the creation of DFD diagrams have been made much more convenient with software tools and they also enable a higher level of cooperation by allowing multiple users to edit the same diagram. If a software tool is being used this means that collaborators can be more effective if they are working on their own computers; therefore remote meetings are superior to physical ones in this context.

7. APPLICATION OF THE PROPOSED GUIDELINES WITH REAL ESTATE ANALYTICS PROJECT

This section of the thesis relates to the application of the proposed guidelines within a real estate analytics project. This chapter contains:

1. Description of the project case.
2. How the generic requirements were satisfied using requirements elicitation techniques.
3. How the evaluation of the defined reequipments.

7.1. Definition of the Project Case

The project that will be used to validate the guidelines that are defined is a muti contextual real estate analytics project. The goal of this project is to undertake research that results in the development of an analytics solution. The progress of the project so far has consisted of a single brainstorming session where the researchers and stakeholders from a real estate management company have discussed the project. Therefore it is possible to conclude that project is still in incubation.

A single stakeholder (hereinafter referred to as John (not the stakeholders' real name)) who acts as a researcher for this project was corresponded with. This stakeholder participated in a single brainstorming session that relates to the project and this was the only thing done so far that relates to the project.

The goal was to apply the findings of this thesis which were guidelines for eliciting generic requirements, and the guidelines for selecting requirements elicitation methods will be applied within the context of this real estate analytics project to define requirements that would help stakeholders of this project to undertake it.

7.2. Application of guidelines for selecting requirements elicitation techniques for the project case

The author will play the role of elicitor and John will act as the informant. The lack of data files, data bases, defined website the sole reliance must be placed in requirements elicitation (as opposed to requirements identification) within the context of this project.

To apply the guidelines for the selection of requirements elicitation technique the following contextual project attribute values where defined:

- Domaine Familiarity of the Elicitor: No prior experience within the field of real estate analytics (defined through asking the stakeholder the question of how long he haswork in the field of real estate industry).
- Temporal Colocation: Same Time Zone (Eastern European Summer Time)
- Level of Expertise of the Stakeholder: Less than two years in the analytics field therefor a Novice.
- Communication Type: Remote

Based on identified values of contextual factors the guidelines in the form of Table 6.3 were applied as shown in Table 7.1.

Table 7.1
Evaluation of Requirements Elicitation Techniques Based on Context Attributes

Contextual Attribute	Values	Requirements Elicitation Technique						
		Interview	Task Observation	Questionnaires	Concept Ranking	Domain Analysis	Brainstorming	Data Flow Modelling
Domain Familiarity of elicitor	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	None	R	R	N	N	R	N	N
Temporal Co-location	Same Time	R	R	R	R	-	R	R
Level of expertise of stakeholder	Novice	N	N	R	R	N	R	R
	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Communication Type	Remote	R	N	R+	R+	-	R	R+
	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
No of Ns per column		1	2	2	2	1	1	1

Based on the results of table 7.1 the author decided to utilize an interview followed by a questionnaire.

7.3. Application of generic requirements identification guidelines

7.3.1. Application of the requirements elicitation technique to identify the generic requirements

The discussion of the case prior to the intercview provided that the project requires descriptive analytics. Based on this, once the introductions within the interview session was done and then the following questions that comply with the guidelines proposed in Section 5.2 were asked in the questionnaire:

1. Could you please describe the goal of this project?
2. What is the expected outcome of this project and how will it be used?
3. What system or object is this analysis being aimed at?
4. What other systems or objects affect or are affected by the system or object in question?
5. What are the quantitative metrics that are being used to evaluate the performance of the system or object that is being considered in this project?
6. Does the solution produced by this project need to be validated, and if so, do you have a specific strategy for the validation?
7. How would you like the final report for this project to be formatted, and what information needs to be placed in this report?

The questionnaire was presented to presented to Mr. John as a Google form which can be found clicking on the following link blow:
https://forms.gle/q7enbwR34CwY4qg29

7.3.2. Results of the requirements elicitation process.

The interview that lasted for less than 40 minutes and a questionnaire that took Mr. John less than 30 minutes to fill out produced the following results. The requirements that were defined are as follows:

- Initiation Phase

Requirement 1: This analytics project must develop an analytics solution in the form of a dashboard that allows the user to evaluate the value of a real estate property based on available data.

Requirement 2: The analytics solution must have a user interface that can be used by individuals with little to no experience with data visualisation and BI (business intelligence) dashboards.

Requirement 3: The analytics solution must analyse a selected real estate property based on real estate market economics, latvian energy economics, and the latvian economy (inflation).

Requirement 4: The analytics solution must be validated by applying it to historical data provided by a real estate management company and comparing it with current data to see how well the evaluations made by the analytics solution about the value of the real estate property compare with the actual value.

- Acquisition Phase

Requirement 5:

The analytics solution must utilize the following datasets:

1. Real estate management company (R.E.M.C.) billing records and estimates
2. Energy prices
3. Inflation indexes
4. EURIBOR rates
5. House prices
6. Construction prices

Requirement 6: The analytics system must acquire data from the following sources:

1. R.E.M.C. billing system
2. Energy price index source
3. Inflation index sources
4. EURIBOR rate source.

Requirement 7: This analytics project must utilize data export and data crawling.

Requirement 8: The analytics project must be carried out following protocols when they are defined.

- Analysis Phase

Requirement 9: The analytics project will result in the development of the mathematical model(s) that will be utilized within a business intelligence dashboard.

Requirement 10: The analytics project must carry out the following tasks within the analysis phase of the project:

Requirement 10.1: Statistical analysis of the created data set.

Requirement 10.2: Regression analysis of the data set.

Requirement 10.3: Development of a mathematical model which carries out real-time analytics of the input data.

Requirement 10.4: Integration of mathematical model within a business intelligence dashboard.

Requirement 11: The software/tools that are utilized within the analysis phase must include but is not limited to Python 3, Oracle DB, and Tableau.

Requirement 12: The analytics project must report all factors that impact the reliability of the result to stakeholders.

- Visualisation Phase

Requirement 13: The results of the analytics project will be a business intelligence dashboard that emphasizes pie charts, bar charts, and line graphs in that order of importance.

Table 7.2 shows which of the generic requirements have been satisfied via the requirements which were defined for the project, as well as possible causes for being unable to define a specify a generic requirement.

Table 7.2

Specification of which Generic Requirements are being Satisfied by the Project Requirements

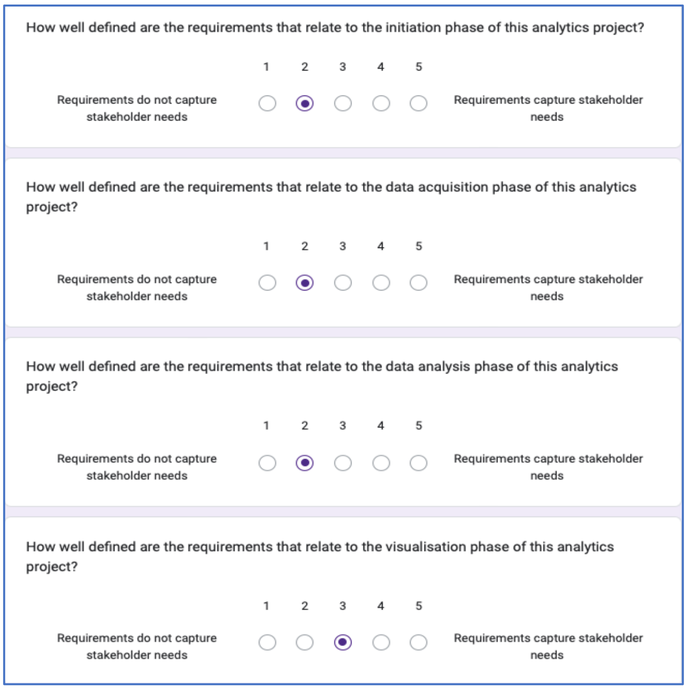
Generic Requirement No.	Requirement defined for project case that satisfies generic requirement	Comments regarding the unsatisfied generic requirements
1	Requirement 1	
2	Requirement 2	
3	Requirement 1	
4	Does not apply because the project is carrying out descriptive analytics	This conclusion is based on Sec requirement 4.0.
5	Requirement 4	
6	Not defined	The progress of the project has not come to a point where datasets can be defined.
7	Requirement 6	
8	Requirement 8	
9	Not defined	Mr. John could not recommend any models for utility within the project.
10	Requirement 10	
11	Not defined	Without a defined dataset it is close to impossible to define specific regarding analytics techniques
12	Requirement 11	
13	Not defined	No reliability related procedures have been defined.
14	Requirement 13	
15	Does not apply because the result of the analytics project is	

	a dashboard and not a report.	

From the results obtained, we can conclude that the guidelines for the selection of requirements elicitation techniques helped to choose the interview and questionnaire as the data elicitation techniques, which in turn, applying gidelines for generic requirements elicitation gave a satisfactoruy result in ability to identify the requirements for data analytics pfoject in the domain of real estate (real estate business intelligence) as prescribed by the guidelines. The stakeholder's opinion about the elicited requirements is provided in Section 7.3.1.

7.3.1. Stakeholder evaluation of the requirements that have been identified.

The resulting requirements were presented to Mr. John in the form of a report after which, he was asked to fill in a question based on his evaluate of the requirements. His responses to the questions were included within a Google form that relate to the requirements that relate to each phase of the analytics project are shown in Figure 7.1.



How well defined are the requirements that relate to the initiation phase of this analytics project?

	1	2	3	4	5	
Requirements do not capture stakeholder needs	○	●	○	○	○	Requirements capture stakeholder needs

How well defined are the requirements that relate to the data acquisition phase of this analytics project?

	1	2	3	4	5	
Requirements do not capture stakeholder needs	○	●	○	○	○	Requirements capture stakeholder needs

How well defined are the requirements that relate to the data analysis phase of this analytics project?

	1	2	3	4	5	
Requirements do not capture stakeholder needs	○	●	○	○	○	Requirements capture stakeholder needs

How well defined are the requirements that relate to the visualisation phase of this analytics project?

	1	2	3	4	5	
Requirements do not capture stakeholder needs	○	○	●	○	○	Requirements capture stakeholder needs

Fig. 7.1. Screenshot showing the questions and the answers provided by Mr. John within Google Forms

The results shown in Figure 7.1 indicate that the requirements defined fail to properly capture the stakeholders needs that is shown by the less than 3 rating that is provided by Mr. John proceeding his review of the defined requirements. An exception to this is the requirements that relate to the visualization phase of the analytics project, which has been given a 3 instead of a 2 which in this case represents a 20% ((1 divided by a 5 which is the number of options) multiplied by 100) improvement when compared to the other phases.

A further comment provided by the stakeholder was that 'requirements are too general'. A deduction can be made based on this statement that the satisfaction of generic requirements results in requirements are themselves 'too' generic.

Nevertheless, the groups of requirements were well understood by a stakeholder and the purpose of this work was to see what are the requirements that have to be identified in data analytic projects, which actually was achieved.

From this the following conclusions can be derived:

1. The provided guidelines for identifying generic requirements are useful for identifying necessary requirement types (or groups) for data analytics or business intelligence projects.
2. The provided guidelines for the selection of eleicitaiton techniques could be used in the real project case.
3. It is a matter of further research whether the additional guidelines for acquisition of detailed requirements are needed or the generic types of requirements are sufficient to help to avoid missing requirement types in the data analytics projects.
4. The results obtained in this thesis can be used in the initial phases of data analytics projects to ensure the completeness of discussed issues before the details of the project are decided.

8. CONCLUSIONS

The goal of this thesis was to create guidelines that will enable personnel within data analytics and business intelligence projects to define requirements for their project. The author accomplishes this by identifying requirements that need to be defined for all analytics projects, defining methods that can be used to define these requirements either through requirements identification or elicitation, and finally making recommendations on which requirements elicitation techniques to use based on contextual project attributes, the result being a set of guidelines that inform the user what requirements need to be defined for an analytics project and what requirements identification/elicitation techniques can be used to define these requirements. These guidelines are then applied within a project case to demonstrate its utility and to evaluate resulting requirements based on how they define stakeholder needs. This project had the limitation of still being in a very early stage evidenced by the project progress at the time consisting of a single brainstorming session.

Based on the recommendations made by the guidelines an interview and a questionnaire were utilized that took less than forty and thirty minutes respectively of the project stakeholders' time which enabled the author to define a total of 13 requirements. An evaluation of how well the defined requirements capture a stakeholder's (through a questionnaire) needs, resulted average score of 2.25 out of 5.

These results are not promising within the context of attempting to elicit requirements based on a framework that defines requirements for all projects, but taking into consideration that these guidelines enabled a user with no prior experience in requirements elicitation to achieve these results provides some hope for the future of this endeavor.

9. LIST OF REFERENCES

Arif, M., Cheung, S.C.P., Andrews, J. , *Diagnostic analysis of a single-cell Proton Exchange Membrane unitized regenerative fuel cell using numerical simulation*, International Journal of Hydrogen Energy, Volume 46, Issue 57, 2021, Pages 29488-29500, ISSN 0360-3199, https://doi.org/10.1016/j.ijhydene.2020.11.165.

Borodin, A. Mityushina, I., Streltsova, E., Kulikov, A. et al., Mathematical Modeling for Financial Analysis of an Enterprise: Motivating of Not Open Innovation, Journal of Open Innovation: Technology, Market, and Complexity, Volume 7, Issue 1, 2021, 79, ISSN 2199-8531, https://doi.org/10.3390/joitmc7010079.

Carrizo, D., *Comparison of Research and Practice Regarding What We Mean by The Right Software Requirements Elicitation Technique*," 2016 10th International Conference on the Quality of Information and Communications Technology (QUATIC), Lisbon, Portugal, 2016, pp. 79-82, doi: 10.1109/QUATIC.2016.022.

Carrizo, D., Oscar , D., Juristo, N., *Contextual attributes impacting the effectiveness of requirements elicitation Techniques: Mapping theoretical and empirical research*, Information and Software Technology, Volume 92, 2017, Pages 194-221, ISSN 0950-5849, https://doi.org/10.1016/j.infsof.2017.08.003.

Carrizo, D., Oscar , D., Juristo, N., *Systematizing requirements elicitation technique selection*,Information and Software Technology,Volume 56, Issue 6,2014,Pages 644-669, ISSN 0950-5849, https://doi.org/10.1016/j.infsof.2014.01.009.

Chawda, R. K., Thakur, G., *Big data and advanced analytics tools*, *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, Indore, India, 2016, pp. 1-8, doi: 10.1109/CDAN.2016.7570890.

Cote, C. (2021, October 19). 4 *Types of Data Analytics to Improve Decision-Making*. Business Insights Blog. https://online.hbs.edu/blog/post/types-of-data-analysis .

Davis, A., Dieste, O., Hickey, A., Juristo, N. et al., *Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review*, *14th IEEE International Requirements Engineering Conference (RE'06)*, Minneapolis/St. Paul, MN, USA, 2006, pp. 179-188, doi: 10.1109/RE.2006.17. Davis

Dinmohammadi, M., Mohammadi, , S., Taherkhani, M., Yadegary, M.A., *Factors contributing to coronavirus disease 2019 vaccine hesitancy among healthcare workers in Iran: A descriptive-analytical study*, Clinical Epidemiology and Global

Health, Volume 18, 2022, 101182, ISSN 2213-3984,https://doi.org/10.1016/j.cegh.2022.101182.

Fernandes, J.M., Machado, R.J. (2016). *Requirements Elicitation. In: Requirements in Engineering Projects*. Lecture Notes in Management and Industrial Engineering. Springer, Cham. https://doi-org.resursi.rtu.lv/10.1007/978-3-319-18597-2_5.

Ghahramani, A., Prokofieva, M.,*Visualisation for social media analytics: landscape of R packages*, *2021 25th International Conference Information Visualisation (IV)*, Sydney, Australia, 2021, pp. 218-222, doi: 10.1109/IV53921.2021.00042.

Goguen, J. A. Linde, C. *Techniques for requirements elicitation*, *[1993] Proceedings of the IEEE International Symposium on Requirements Engineering*, San Diego, CA, USA, 1993, pp. 152-164, doi: 10.1109/ISRE.1993.324822.

Gradwell, D. J. L., *Data dictionary standardisation*, *IEE Colloquium on Data Dictionary Systems: Present and Future*, London, UK, 1988, pp. 4/1-425.

Gururajan, R., Clark, K., Moller, S. et al., *Reliability of Qualitative Data Using Text Analysis - A Queensland Health Case Study*, 2014 3rd International Conference on Eco-friendly Computing and Communication Systems, Mangalore, India, 2014, pp. 303-308, doi: 10.1109/Eco-friendly.2014.68.

Hakami, F., Pramanik, A., & Basak, A. K. (2022). *Statistical Analysis*. SpringerBriefs in Applied Sciences and Technology, 121–133. https://doi.org/10.1007/978-981-19-2908-3_8

Hardin-Fanning, F.,Hartson, K.R.,Galloway, L., Kern, N. et al. *Students' perceptions of a community health advocacy skills building activity: A descriptive analysis*, Nurse Education Today, Volume 120, 2023, 105627, ISSN 0260-6917, https://doi.org/10.1016/j.nedt.2022.105627

Hillier, W. (2022, December 19). *A Step-by-Step Guide to the Data Analysis Process*. CareerFoundry. https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/.

Irzavika, N.,Supangkat, S.H., "*Descriptive Analytics Using Visualization for Local Government Income in Indonesia*," *2018 International Conference on ICT for Smart Society (ICISS)*, Semarang, Indonesia, 2018, pp. 1-4, doi: 10.1109/ICTSS.2018.8550006 .

ISO/IEC/IEEE International Standard - Systems and software engineering -- Life cycle processes -- Requirements engineering, in ISO/IEC/IEEE 29148:2018(E) , vol., no., pp.1-104, 30 Nov. 2018, doi: 10.1109/IEEESTD.2018.8559686.

Kadadi, A., Agrawal, V., Nyamful, C., Atiq, R.,*Challenges of data integration and interoperability in big data*, *2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2014, pp. 38-40, doi: 10.1109/BigData.2014.7004486.

Kumar, A.,Ali, A.S.,Jamnadas, H., Sharma, V., *Big Data Visualisation - An Update until Today*, *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Melbourne, VIC, Australia, 2019, pp. 1-8, doi: 10.1109/CSDE48274.2019.9162356.

Lee, Y., Zhao, W. (2006). Domain Requirements Elicitation and Analysis - An Ontology-Based Approach. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds) Computational Science – ICCS 2006. ICCS 2006. Lecture Notes in Computer Science, vol 3994. Springer, Berlin, Heidelberg.https://doi-org.resursi.rtu.lv/10.1007/11758549_108

Li, L., Xie, F., Li, J. , Gong, K. et al., *Diagnostic analysis of regional ozone pollution in Yangtze River Delta, China: A case study in summer 2020*, Science of The Total Environment, Volume 812,2022,151511, ISSN 0048-9697,https://doi.org/10.1016/j.scitotenv.2021.151511.

Lim, S., Henriksson, A., Zdravkovic, J. *Data-Driven Requirements Elicitation: A Systematic Literature Review*. SN COMPUT. SCI. 2, 16 (2021). https://doi-org.resursi.rtu.lv/10.1007/s42979-020-00416-4.

Liu, S., Li, Y., Fan, W.D., *Mixed logit model based diagnostic analysis of bicycle-vehicle crashes at daytime and nighttime*, International Journal of Transportation Science and Technology, Volume 11, Issue 4, 2022, Pages 738-751, ISSN 2046-0430,https://doi.org/10.1016/j.ijtst.2021.10.001.

Lopes, A.S.D.A, Santos, L.D.N., Razé, M.D.C.,Lazzarini, R.,*Alopecia areata: descriptive analysis in a Brazilian sample*,Anais Brasileiros de Dermatologia,Volume 97, Issue 5,2022,Pages 654-656,ISSN 0365-0596,https://doi.org/10.1016/j.abd.2021.04.016.

Milella, P., Bisantino, T., Gentile, F., Iacobellis, V. et al., *Diagnostic analysis of distributed input and parameter datasets in Mediterranean basin streamflow modeling*, Journal of Hydrology, Volumes 472–473, 2012, Pages 262-276, ISSN 0022-1694,https://doi.org/10.1016/j.jhydrol.2012.09.039.

O'Regan, G. (2017). *Requirements Engineering. In: Concise Guide to Software Engineering. Undergraduate Topics in Computer Science.* Springer, Cham. https://doi-org.resursi.rtu.lv/10.1007/978-3-319-57750-0_3.

Olayan, N.,Patu, V.,Matsuno, Y., Yamamoto, S., *A Dependability Assurance Method Based on Data Flow Diagram (DFD)*, 2013 European Modelling Symposium, Manchester, UK, 2013, pp. 113-118, doi: 10.1109/EMS.2013.20.

Ozimek, A., *The future of remote work*., 2020 Available at SSRN 3638597.

Praveena, M. D. A. , Bharathi, B., *A survey paper on big data analytics*, *2017 International Conference on Information Communication and Embedded Systems (ICICES)*, Chennai, India, 2017, pp. 1-9, doi: 10.1109/ICICES.2017.8070723.

Quintanilla, I., Carrizo, D. (2018). *Formalizing a Cost Construct Model related to the Software Requirements Elicitation Techniques*. In: Mejia, J., Muñoz, M., Rocha, Á., Quiñonez, Y., Calvo-Manzano, J. (eds) Trends and Applications in Software Engineering. CIMPS 2017. Advances in Intelligent Systems and Computing, vol 688. Springer, Cham. https://doi-org.resursi.rtu.lv/10.1007/978-3-319-69341-5_3

Ridzuan, F. , Zainon, W.M.N.W., *Diagnostic analysis for outlier detection in big data analytics, Procedia Computer Science*,Volume 197,2022,Pages 685-692,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2021.12.189.

Rodina, A., Amjed, T. , Zarinah, *M. K. An empirical assessment of the use of different communication modes for requirement elicitation and negotiation using students as a subject*, 2012 IEEE Symposium on Computers & Informatics (ISCI), Penang, Malaysia, 2012, pp. 70-74, doi: 10.1109/ISCI.2012.6222669.

Runkler, T. A. (2020). Data Analytics. *Springer EBooks*. https://doi.org/10.1007/978-3-658-29779-4

Salah, M., Okush, B. A., Rifaee, M. A., *A Comparison of Web Data Extraction Techniques*, 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 785-789, doi: 10.1109/JEEIT.2019.8717519.

Sedov, D.,*Restaurant closures during the COVID-19 pandemic: A descriptive analysis*, Economics Letters, Volume 213, 2022, 110380, ISSN 0165-1765,https://doi.org/10.1016/j.econlet.2022.110380.

U.S. Bureau of Labor Statistics, *Data Scientists: Occupational Outlook Handbook,* (2023). https://www.bls.gov/ooh/math/data-scientists.htm

Unpingco, J. (2021). *Python Programming for Data Analysis*. Springer EBooks. https://doi.org/10.1007/978-3-030-68952-0

W. Maalej, M. Nayebi and G. Ruhe, *Data-Driven Requirements Engineering - An Update*, *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, Montreal, QC, Canada, 2019, pp. 289-290, doi: 10.1109/ICSE-SEIP.2019.00041.

Wirth, R., Hipp, J., *CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000, (Vol. 1, pp. 29-39).

Wu, J., Niu, Z., Li, X., Huang, L. et al., *Understanding multi-scale spatiotemporal energy consumption data: A visual analysis approach*, Energy, Volume 263, Part D, 2023, 125939, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2022.125939.

Yousuf, F., Zaman, Z. and Ikram, N., *Requirements validation techniques in GSD: A survey*, 2008 IEEE International Multitopic Conference, Karachi, Pakistan, 2008, pp. 553-557, doi: 10.1109/INMIC.2008.4777800.

APPENDIXES

Appendix 1
Descriptive analytics case studies analyses

Case Study 2:Alopecia areata: descriptive analysis in a Brazilian sample

(Lopes, Santos et al., 2022)

Initiation:

Pro requirement 1.1: This project must carry out the "assessment of cases followed at the dermatology outpatient clinic in a quaternary hospital between 2000 and 2017".

Pro requirement 1.2: The project will consist of "Data were collected retrospectively and submitted to the statistical program R, version 3.4.2. (R Core Team, 2016), descriptively analyzed and compared".

Acquisition:

Pro requirement 1.3: The project will get data "collected retrospectively" from the "assessment of cases followed at the dermatology outpatient clinic in a quaternary hospital between 2000 and 2017".

Pro requirement 2.1: The data collected can include the "159 cases (34.1%) with no information".

Gen requirement 2.1: The permissions regarding the use of incomplete data sets must be defined within the context of the analytics project.

Analysis:

Pro requirement 1.4: The data collected will be "descriptively analyzed and compared using Pearson's chi- square test".

Pro requirement 1.5: The collected data will be analyzed using "statistical program R, version 3.4.2."

Presentation:

Pro requirement 1.6: The findings must be presented in a table showing the "Distribution of 466 patients".

Case study 3: Restaurant closures during the COVID-19 pandemic: A descriptive analysis (Sedov, 2022)

Initiation:

Pro requirement 1.1: The goal of this analytics project is to answer the question "Which restaurants were more likely to exit the industry in this challenging time?"

Pro requirement 1.2: The goal of this analytics project will be achieved through "descriptive evidence on this question in the context of major US urban areas using data from the review platform Yelp and the location data company SafeGraph."

Pro requirement 3.1: The author will complete the following steps within the context of this project "I provide descriptive evidence on this question in the context of major US urban areas using data from the review platform Yelp and the location data company SafeGraph. Specifically, I explore location- and restaurant-specific characteristics that explain variation in restaurant closure decisions. First, I document the across-cities differences in observed restaurant exit rates, which range from 9.6% in El Paso to 21.5% in Honolulu. Next, I estimate binary response econometric models and summarize the association between restaurant characteris- tics and exit. I find that higher rating scores and review counts are robustly associated with lower restaurant exit probabilities."

Gen requirement 3.1: The level of detail and technicality required when describing the methodology used within the project must be defined based on the knowledge level of the client(s).

Acquisition:

Pro requirement 1.3: The sources of data for this analytics project are "Three data sources are used for the analysis discussed in this paper. The data from the Yelp restaurant review platform provides information on restaurant characteristics and exit deci- sions. I also use data from the location data company SafeGraph, which collects information on US points-of-interest (defined as places outside of home where people spend time and/or money), and the U.S. Census to construct additional covariates related to restaurant location characteristics."

Pro requirement 2.1: The collected data creates a "combined dataset covers 128,285 restaurants in 42 major US cities"

Pro requirement 3.2: The data used in the analytics project "using a scraping routine that systematically parsed Yelp Fusion API".

Gen requirement 3.2: How data was collected from the data source must be defined for a data analytics project if said data source does not have ready to use data.

Analysis:
Pro requirement 1.4: The data collected will be used to "estimate binary response models (LPM, logit or probit linking closures and restaurant characteristics."
Pro requirement 1.5: No information regarding the tools used were mentioned within the research paper.

The author mentions the limitations of the results provided due to an inadequacy of the data that is collected not because the data is incomplete but because the indicator used to derive the data is not fully reliable. These limitations must be highlighted in order to prevent providing clients with misinformation.

Pro requirement 3.3: The project must convey that the result of the analysis is affected by the "imperfection of Yelp's exit data".
Gen requirement 3.3: Practices that relate to the reliability of analysis results must be defined for the analytics project.
Presentation:
Pro requirement 1.6: The results of the project will be graphically presented using 1 bar chart showing which "depicts the exit rates across sample cities" , 1 line graph which "displays the relationship between market size (measured as restaurant count on the city level) and restaurant closure rates" and 4 tables: "Restaurants dataset summary statistics" , "Coefficient estimates for the binary response models", "Coefficient estimates for LPMs with an extended set of location controls", "Average Partial Differences for the binary response models".

Case study 4: Descriptive Analytics using Visualization for Local Government Income in Indonesia(Irzavika & Supangkat, 2018)
Initiation:
Pro requirement 1.1: The goal of this analytics project is "to give exposure to decision makers about phenomena that occur on PAD in order to become new information for decision makers and as a material consideration in the decision-making process to increase PAD in a city."

Pro requirement 1.2: The methodology used with this analytics project consists of “descriptive analytics is done using data visualization to conduct historical analysis, and to know and understand the current condition”.

Pro requirement 3.1: This project must provide the client with an introductory level of information regarding concepts such as “Local Government Income (PAD)”, “Business Intelligence”, “Data Analytics”, “Descriptive Analytics”.

Acquisition:

Pro requirement 1.3: The data used with this analytics project consists of “the real data of the Expenditure Budget Report from one of the cities in Indonesia.”

Pro requirement 4.1: The dataset used within this analytics project “consists of income data, expenditure data, and financing data. The data used in this study is only income data.”

Gen requirement 4.1: The data contained within the data source must be defined as well as which of the data will be used for the data analytics.

Pro requirement 4.2: This analytics project will use ETL process consisting of “Pentaho Data Integration (PDI) or kettle is software provided by Pentaho that can perform the ETL process”

Gen requirement 4.2: Specification regarding the ETL (Extract Transform Load) must be defined for analytics projects.

Pro requirement 4.3: The analytics will use “Dimensional tables and fact tables are integrated with the data warehouse scheme as in Figure 4”

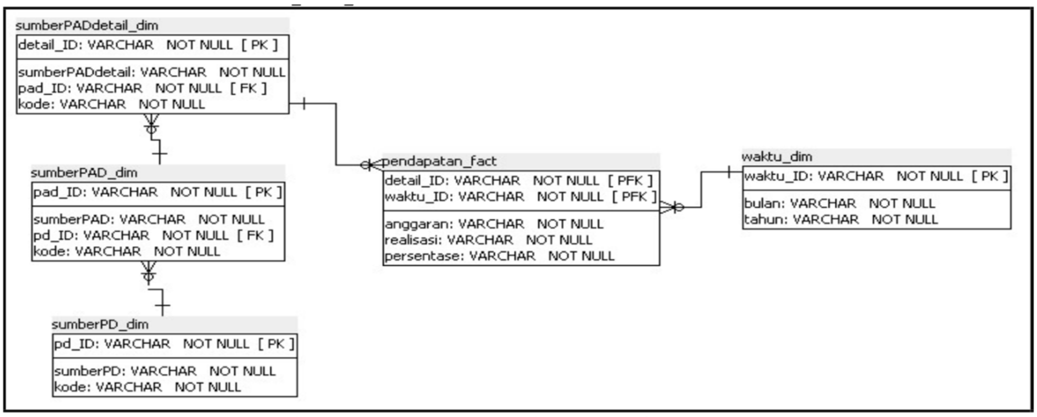

sumberPADdetail_dim
detail_ID: VARCHAR NOT NULL [PK]
sumberPADdetail: VARCHAR NOT NULL
pad_ID: VARCHAR NOT NULL [FK]
kode: VARCHAR NOT NULL
sumberPAD_dim
pad_ID: VARCHAR NOT NULL [PK]
sumberPAD: VARCHAR NOT NULL
pd_ID: VARCHAR NOT NULL [FK]
kode: VARCHAR NOT NULL
sumberPD_dim
pd_ID: VARCHAR NOT NULL [PK]
sumberPD: VARCHAR NOT NULL
kode: VARCHAR NOT NULL
pendapatan_fact
detail_ID: VARCHAR NOT NULL [PFK]
waktu_ID: VARCHAR NOT NULL [PFK]
anggaran: VARCHAR NOT NULL
realisasi: VARCHAR NOT NULL
persentase: VARCHAR NOT NULL
waktu_dim
waktu_ID: VARCHAR NOT NULL [PK]
bulan: VARCHAR NOT NULL
tahun: VARCHAR NOT NULL

Figure 4. Data warehouse scheme

Fig. A1.1. An extract adopted from (Irzavika & Supangkat, 2018)

Gen requirement 4.3: How data warehousing is carried out in the analytics project, then the specifications regarding the data warehouse must be specified.

Pro requirement 2.1: The data will be transformed using "The ETL process involves" "filling the missing data, delete unnecessary data and repairing inconsistent data.".

Pro requirement 3.2: Data for this analytics project will be collected through "One of the integration processes can be seen in Figure 3. This process consists of data input, changing the name of month and year, checking for redundant data, and performing a lookup database to retrieve the id from the dimension table."

Analysis:

Pro requirement 1.4: The analysis for this project will be carried out using "data visualization that is making a dashboard and displaying meaningful information."

Pro requirement 1.5: The analysis will be carried out "~~Data visualization in this research was designed~~ using Tableau"

Pro requirement 3.3: No factors regarding reliability were mentioned within the research paper.

Presentation:

Pro requirement 1.6: "

KnowingconditionandtrendofPADeveryyear

KnowingthecontributionofeachsourcePAD

KnowingthebiggestcontributionofsourcePAD

KnowingtheamountofPADeverymonth

KnowingwhenthelargestPADforthelast5years

Knowing the ratio of PAD realization to PAD budget every year

Based on the visualization goals, the dashboard can be seen in Figure 5. "

Pro requirement 4.4: The result of the analysis must be used to make "predictions for the next five years using a linear regression algorithm"

Gen requirement 4.4: If and how the findings of the analysis must be used in order to do further types of analysis must be defined.

Case study 5:Factors contributing to coronavirus disease 2019 vaccine hesitancy among healthcare workers in Iran: A descriptive-analytical study(Dinmohammadi, Mohammadi et al., 2022)

Initiation:

Pro requirement 1.1: The goal of this "This cross-sectional descriptive-analytical" was to "assess the factors contributing to COVID-19 vaccine hesitancy (VH) among HCWs in Iran".

Pro requirement 1.2: The data analytics project will consist of using "the SPSS software (v. 20) and through the independent-sample *t*-test, the one-way analysis of variance, and the multiple linear regression analysis".

Pro requirement 3.1: Nothing indicating the author's awareness of readers knowledge level has been provided within the research article.

Acquisition:

Pro requirement 1.3: The data used within the analytics project consists of "Study population consisted of all 8000 HCWs with or without the history of COVID-19 vaccination in four leading hospitals affiliated to Zanjan University of Medical Sciences, Zanjan, Iran."

Pro requirement 4.1: The source of data being health care worker there does not need to be the requirement of defining what data the data store contains and what data will be used

Pro requirement 4.2: The data source being information collected from people does not require a ETL.

Pro requirement 4.3: Data warehousing is not required in order to complete this analytics project.

Pro requirement 2.1: The "sample size was determined to be 500 and was increased to 551 due to a potential attrition rate of 10%."

Pro requirement 3.2: The data for this analytics project used "Data collection instruments were a demographic questionnaire and a COVID-19 VH questionnaire"

Analysis:

Pro requirement 1.4: Data analytics for this project was carried out using the following methods: "Data description was done through the measures of descriptive statistics, namely frequency, mean, and standard deviation", "Kolmogorov-Smirnov test indicated the normality of the data" , "independent-sample *t*- test, the one-way analysis of variance, and the multiple linear regression analysis with the Enter method"

Pro requirement 1.5: Data analysis for this project was done by "using the SPSS software (v. 20)"

Pro requirement 3.3: The limitations of this data analytics project are as follows "This study was conducted on HCWs with an age mean of 34.40 ± 7.77 years and hence, its findings may not be generalizable to adoles- cents and elderly people."

Presentation:

Pro requirement 1.6: The graphical representation of the data will be done using tables.

Pro requirement 4.4: Explanations for why the results were present were provided such as "people usually showed limited adherence to COVID-19 prevention protocols and refused vaccination because they believed that a new wave would never happen" , "Another explanation for the higher VH prevalence in the pre- sent study compared with previous studies is that most of those studies assessed individuals' attitudes during the period of COVID-19 vaccine production, testing, and approval, while our participants had free access to COVID-19 vaccination services".

Case study 2: Diagnostic analysis of regional ozone pollution in Yangtze River Delta, China: A case study in summer 2020 (Li, Xie, et al., 2022)

Initiation:

Pro requirement 1.1: The goal of this analytics project is to conduct “a comprehensive diagnostic analysis of O3 formation during a 1-week regional O3 pollution event in August 2020 in the YRD region”.

Pro requirement 1.2: “Emission based model (WRF-CMAQ)”, and “OBM” were defined within the context of the data analytics project.

Pro requirement 1.3: This analytics project “aims to understand the causes of O3 pollution during ” “A regional ozone (O3) pollution event occurred in the Yangtze River Delta region during August 17–23, 2020 (except on August 21)”.

Pro requirement 1.4: This project will use “O3 pollution” as well as “O3 sensitivity to its precursors during the O3 pollution” both of which will be measured using “O3 concentrations”.

* This information is not clearly defined and within the introduction of the paper and this results in a level of ambiguity when defining the requirement.

Acquisition:

*This research paper is very unique because the authors predominantly rely on models as opposed to a dataset inorder to carry out the diagnostic analysis, although inputs required for one of the models was defined using a dataset. Therefore it is advisable to define both the datasets and models used within a diagnostic analytics project given that models and datasets are not interchangeable.

Pro requirement 2.1: The models used within this project are “Weather Research and Forecasting (WRF) model version 4.2.1 was used to provide the meteorological fields for the chemical transport model”,

“emission-based model (EBM) (i.e., a 3-D chemical transport model) was used to simulate the air quality during the episode”,

“The Community Multiscale Air Quality version 5.2 (CMAQv5.2), developed by the United States Environmental Protection Agency (US EPA), was employed in this study to simulate the air quality and explore the causes of O3 pollution during the summer 2020 in the YRD region" ,

"a source-oriented CMAQ model was utilized in this study to assess the contributions of different emissions sources and emitting regions to O3, which was based on an improved sensitivity regime classification (i.e., VOC-limited, NOx-limited, and transition regimes) approach for O3 formation",

"OBM developed by Cardelino and Chameides (1995), incorporating the Master Chemical Mechanism version 3.3.1 (MCMv3.3.1, available at http://mcm.leeds.ac.uk/MCM/) in this study, was used to simulate the O3 photochemistry and further identify the sensitivity of O3 formation to precursor concentrations at a certain monitoring site".

Gen requirement 2.1: The model(s) used within the analytics project along with what said models are used for must be defined for a diagnostic analytics project.

Pro requirement 1.5: The datasets used within this data analytics project are "1° × 1° FNL reanalysis dataset with a temporal resolution of 6 h from the National Centers for Environmental Prediction" , the "hourly observation data of trace gases (e.g., O3, NO2, and CO) for major cities in the YRD region were obtained from the China National Environmental Monitoring Center (CNEMC, http://106.37.208.233: 20035/) from August 17 to 23, 2020.", "Continuous field measurements of VOCs were also carried out at a typical urban monitoring site (32.057°N, 118.749°E, Fig. 1(b)) in Nanjing that was surrounded by commercial and residential districts. Hourly data of 57 VOC species, consisting of 29 alkanes, 16 aromatics, 11 alkenes, and acetylene, were collected. The observed data of NO2, CO and VOCs were as input in OBM. The real-time hourly data of the meteorological parameters (i.e., temperature, wind speed, wind direction, relative humidity, and precipitation) in Nanjing were obtained from the weather website (http://q-weather.info/weather/)."

Pro requirement 1.6

Analysis:

Pro requirement 1.7: The analytics will be carried out using "an emission- based model" and "an emission- based model".

Pro requirement 1.8: Although the strategy used within this study was defined in an in-depth manner it is very extensive and hard to grasp, owing to the fact that the authors had a reader that is more well versed in the subject matter in mind when developing the research article. Although it can be said that the methodology could have been more concisely defined.

Pro requirement 2.2: Within this data analytics project “integrated process rate (IPR) module, a process analysis tool based on solving the mass continuity equation, was available in CMAQv5.2 and applied in this study”.

Gen requirements 2.2: The analytics project must have clearly defined the tools that are going to be used and as well as what those tools will be used to do.

Presentation:

Gen requirement 1.9: The results of this analytics project must be formatted such that “~~The OBM analysis determined that~~ the O3 formation was in the ~~VOC-limited regime on August 19,~~ and in the transition regime on all the other polluted days. ~~Although neither aromatics nor alkenes were~~ the most abundant groups, ~~they were~~ the top two contributors to O3 formation in terms of the shares in OFP among all the VOCs measured”, and “The process analysis indicated that ~~the photochemical process~~ was the predominant factor in the formation and accumulation of O3 during the daytime.”

Pro requirement 1.10: This data analytics project must visually represent:

“Modeling domain. (a) The three nested domains with different horizontal resolutions (d01: 36 km, d02: 12 km, and d03: 4 km) for WRF simulation. The blue rectangle indicates the CMAQ simulated domain. (b) The various colored and patterned areas represent 15 cities tagged in the YRD region. The black dot identifies the location of the VOCs field measurement.”

“Time series of observed concentrations (black dotted line) of (a) NO2 and (b) total VOCs; (c) comparison between simulated (red dotted line) and observed O3 concentrations (black dotted line); the daily average concentrations of (d) NO2, (e) total VOCs, and (f) MDA8 O3 (the red dash line for the limit exceeding 160 μg/m3) in Nanjing during study period.”

“Time series of O3 change rate caused by individual atmospheric processes in Nanjing in the PBL. DEPO, HTRA, VTRA, and CHEM mean deposition (dry deposition and cloud process), horizontal transport, vertical transport, and chemical process, respectively. Total O3 variation is the sum of these processes.”

“The RIR values for O3 precursors (i.e., AVOCs, NOx, BVOCs (isoprene), and CO) at Nanjing urban site during O3 pollution episode.”

"(a) RIR of top 10 AVOCs for O3 formation at Nanjing urban site, and (b) concentrations and (c) OFP proportions of different VOCs groups (ALKA: alkanes; ALKE: alkenes; AROM: aromatics; and ACET: acetylene) to total observed VOCs during O3 pollution episode."

"The percentage of source contributions to average MDA8 O3 attributed to (a) power, (b) industry, (c) residential, (d) transportation, (e) biogenic source, (f) IC/BC, and (g) background during O3 pollution episode in the YRD region. Contributions of IC/BC to O3 are attributed to NOx and VOCs entering the domain through the initial and boundary conditions. Background O3 is regarded as that directly entering the domain through the initial and boundary conditions. The area scope of Nanjing city is marked in bold on the map."

"The percentage contributions of different sources to hourly O3 in Nanjing from August 17–23 (excluding August 21), 2020. Predicted O3 concentrations from different sources are represented by the corresponding colored areas."

"Source contributions of transport from individual cities to hourly O3 in Nanjing from August 17 to 23 (excluding August 21), 2020. "BG" means background. The percent in the pie chart is the average MDA8 O3 during O3 pollution episode. "Local" refers to the contribution of Nanjing city itself. "Non-Local" refers to the contribution from cities tagged other than Nanjing. "Other" refers to contributions from those cities not tagged in the target area."

Note: This case study provided much insight into the complexity of diagnostic analytics when compared to descriptive analytics where although one cannot elicit requirements that capture the complexity of the methodology adequately, having the foundational requirements set as to what methods will be used will provide the data analysts a much-needed foundation to build upon doing his or her projects.

Case study 3: Mixed logit model based diagnostic analysis of bicycle-vehicle crashed at daytime and nighttime (Liu, Li et al.,2022)

Initiation:

Pro requirement 1.1: The goal of this analytics "is to explore the underlying factors to injury severity in crashes involving cyclists in the daytime and nighttime separately".

Pro requirement 1.2: "Mixed logit (ML) model", and "Marginal effect analysis" must be explained within the context of the data analytics project.

Pro requirement 1.3: This data analytics project will look at "crashes involving cyclists in the daytime and nighttime separately"

Pro requirement 1.4: This project will use “Five injury severity levels are identified, which are no injury (NI), possible injury (PI), suspected minor injury (SMI), suspected severe injury (SSI), and fatal injury (FI)” in order to evaluate the severity of “crashes involving cyclists”.

Acquisition:

Pro requirement 2.1: This analytics project will utilize “mixed logit model to analyze the underlying factors towards injury severities in crashes involving cyclists”

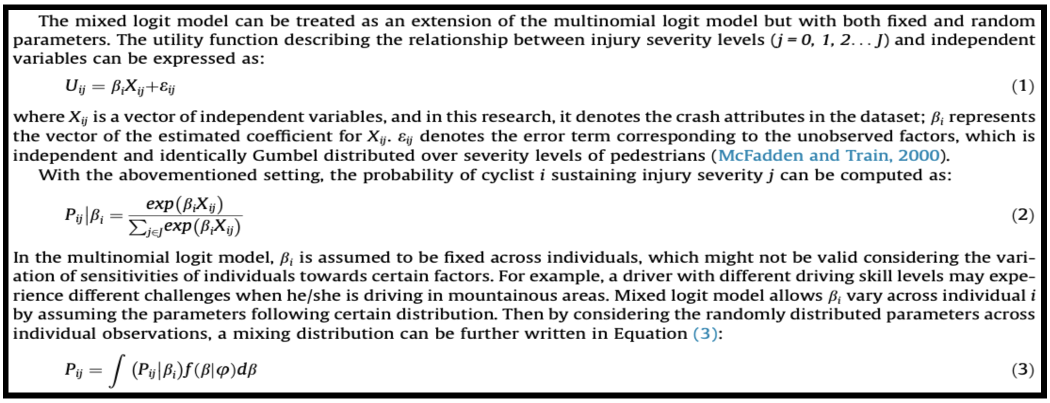
Pro requirement 1.5: Inorder to carry out the analysis for this project “data used to estimate mixed logit models are retrieved from the police report data of North Carolina Department of Transportation (NCDOT) between 2007 and 2018”.

Pro requirement 1.6: The properties of the data “include many categorical explanatory variables, which are cyclist, driver, vehicle, road, environment, and crash characteristics. 8049 out of 11,196 are filtered for model estimation via the data cleaning process. Essentially, the data without the necessary information were filtered out in the cleaning process. The cyclist’s characteristics contain the gender and age of the cyclists, as well as alcohol usage. The characteristics of drivers include the same variables as those of cyclists. Vehicle characteristics mainly refer to vehicle type. Variables in road characteristics are traffic control, speed limits, road configuration and road condition, rural and urban. Environmental char- acteristics include weather, light condition, region, and development type. Crash characteristics contain variables of crash types, crash time, and crash location. Five injury severity levels are identified, which are no injury (NI), possible injury (PI), suspected minor injury (SMI), suspected severe injury (SSI), and fatal injury (FI). In this study, no injury is selected as the base injury severity level in the mixed logit model. Details of the data utilized in this study are summarized in Table 2 by category and injury severity level.”

Analysis:

Pro requirement 1.7: The analytics will be carried out using a “mixed logit model” as well as “Marginal effect analysis”.

Pro requirement 1.8: The in-depth description for the methods used within this data analytics project for the mixed logit model and the marginal effect analysis are shown in **Figure A2.1** and **Figure A2.2** respectively.



The mixed logit model can be treated as an extension of the multinomial logit model but with both fixed and random parameters. The utility function describing the relationship between injury severity levels ($j = 0, 1, 2 \ldots J$) and independent variables can be expressed as:

$$U_{ij} = \beta_i X_{ij} + \varepsilon_{ij} \tag{1}$$

where X_{ij} is a vector of independent variables, and in this research, it denotes the crash attributes in the dataset; β_i represents the vector of the estimated coefficient for X_{ij}. ε_{ij} denotes the error term corresponding to the unobserved factors, which is independent and identically Gumbel distributed over severity levels of pedestrians (McFadden and Train, 2000).

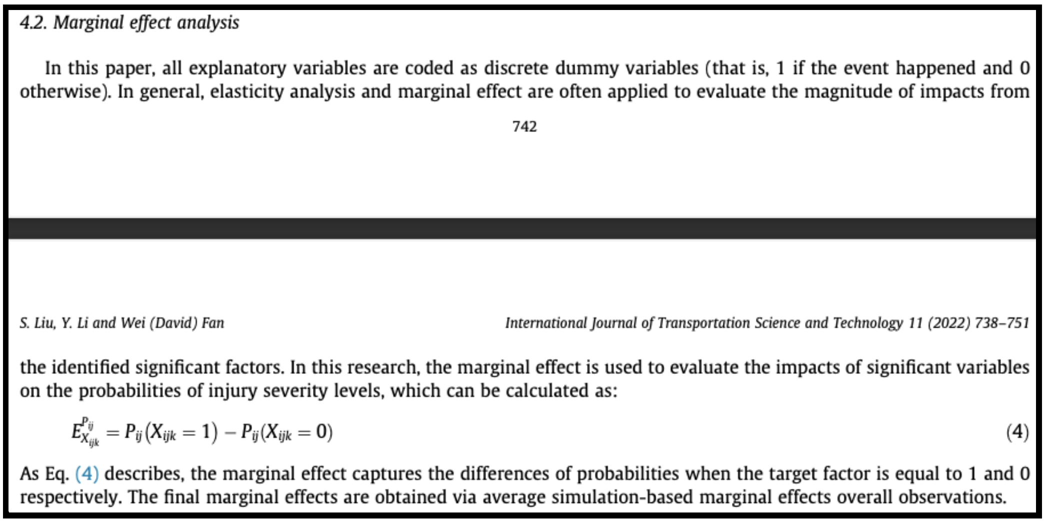
With the abovementioned setting, the probability of cyclist i sustaining injury severity j can be computed as:

$$P_{ij}|\beta_i = \frac{exp(\beta_i X_{ij})}{\sum_{j \in J} exp(\beta_i X_{ij})} \tag{2}$$

In the multinomial logit model, β_i is assumed to be fixed across individuals, which might not be valid considering the variation of sensitivities of individuals towards certain factors. For example, a driver with different driving skill levels may experience different challenges when he/she is driving in mountainous areas. Mixed logit model allows β_i vary across individual i by assuming the parameters following certain distribution. Then by considering the randomly distributed parameters across individual observations, a mixing distribution can be further written in Equation (3):

$$P_{ij} = \int (P_{ij}|\beta_i) f(\beta|\varphi) d\beta \tag{3}$$

Fig. A2.1. An extract taken from research paper regarding the mixed logit methodology used within the study (adopted from (Liu, Li et al.,2022))



4.2. Marginal effect analysis

In this paper, all explanatory variables are coded as discrete dummy variables (that is, 1 if the event happened and 0 otherwise). In general, elasticity analysis and marginal effect are often applied to evaluate the magnitude of impacts from

742

S. Liu, Y. Li and Wei (David) Fan *International Journal of Transportation Science and Technology 11 (2022) 738–751*

the identified significant factors. In this research, the marginal effect is used to evaluate the impacts of significant variables on the probabilities of injury severity levels, which can be calculated as:

$$E^{P_{ij}}_{X_{ijk}} = P_{ij}(X_{ijk} = 1) - P_{ij}(X_{ijk} = 0) \tag{4}$$

As Eq. (4) describes, the marginal effect captures the differences of probabilities when the target factor is equal to 1 and 0 respectively. The final marginal effects are obtained via average simulation-based marginal effects overall observations.

Fig. A2.2. An extract taken from research paper regarding the marginal effect analysis methodology used within the study (adopted from (Liu, Li et al.,2022))

Pro requirement 2.2: Specifics regarding the software that are used within this study are not explicitly mentioned.

Presentation:

Pro requirement 1.9: The data analytics project must result in the “variables ~~are~~ identified-with significant impacts on at least one of the cyclist injury severity levels”, “variables ~~(that is, male cyclist, cyclist on crosswalk, rural area, adverse road condition, and no traffic control) are~~ found to have random effects across all observations under different severity levels”, and “random parameter ~~has been~~ found in the nighttime model”.

Pro requirement 3.1: The causes that relate to the “Human characteristics”, “Vehicle characteristics”, “Environmental characteristics”, and “Crash characteristics” must be clearly defined for the “crashes involving cyclists in the daytime and nighttime separately”.

Gen requirement 3.1: The format by which the different categorical causes of the issue must be clearly stated when presenting the results of the analytics project.

Pro requirement 1.10: The data analytics project must “show the marginal effects for each significant contributing factor to the fatal injury of the cyclist in both models” using a histogram.

Case study 4: Diagnostic analysis of distributed input and parameter datasets in Mediterranean basin streamflow modeling (Milella, Bisantino et al. , 2012)

Initiation:

Pro requirement 1.1: The goal of this analytics project is to “analyze the impact of different data sources in the input and parameterization phase of a water balance hydrological modeling application”.

Pro requirement 1.2: A table containing the nomenclature that relates to this data analytics project must be defined.

Pro requirement 1.3: This analytics project will analyze “different data sources in the input and parameterization phase of a water balance hydrological modeling application” as well “the comparison of different configurations of input and parameter datasets” .

Pro requirement 1.4: The data analytics project will use “LAI, reference evapo-transpiration, crop coefficients and volumetric soil moisture contents at wilting point, field capacity and saturation”.

Pro requirement 4.1: The results of the data analytics must be validated by applying the “model” “at daily scale in a semi-arid basin of Southern Italy (Carapelle river, basin area: 506 km2)”.

Gen requirement 4.1: The analytics project must define how the results of the data analytics will be validated or verified.

Acquisition:

Pro requirement 2.1: For the evaluation of data sources “semi-distributed model was used, based on a discrete grid for the representation of vertical water fluxes (rainfall, evapotrans- piration, infiltration and groundwater recharge) and a lumped representation of sub-horizontal fluxes (overland runoff, lateral flow and groundwater flow) ”

Pro requirement 1.5: This analytics project will use: “Detailed data of watershed physical information, land uses and climate” taken from “Time series of rainfall, temperature and wind speed, recorded by the Hydrometric Office of Regione Puglia”; “Continuous streamflow data are provided by the gauging station at the Ordona Castelluccio dei Sauri bridge ”; “Land use and vegetal coverage were obtained by the Corine land Cover (scale: 1:100,000)”.

Project requirement 4.2: The derived data used within this analytics project is: “other climatic quantities required by the FAO Penman–Monteith equation were derived by temperature and wind speed”; “topographic features were defined using the Digital elevation map (90 m X 90 m) of the Carapelle watershed”; “Soil parameters such as the textural classes, saturated hydraulic conductivity, soil depths and porosity were extracted from the ACLA2 project ”; “percentage of organic matter was derived from the Octop Project of the Euro- pean Soil Data Centre ”.
Gen requirement 4.2: The derived data used in the analytics project must be defined.

Pro requirement 1.6: The “Main characteristics of the Carapelle watershed at Ordona Bridge closing section. ” is defined but the properties of the other data sets are not mentioned.

Analysis:
Pro requirement 1.7: The diagnostic analytics will be carried using “semi-distributed hydrologic model”.
Pro requirement 1.8: The methodology has been defined but its complexity is beyond the scope of this thesis.
Gen requirement 2.2: The software or tool that were used within the analytics project are not explicitly mentioned.

Presentation:

Pro requirement 1.9: The data analytics project must result in the "evaluation of reference evapotranspiration ~~the FAO Pen- man Monteith formulation~~ provided the best model performance", the best of "two different pedotransfer function sets provided" for "soil hydraulic properties", which of the "~~Among all~~ the metrics ~~the KGE~~ provided more sensitivity and convincing consistency with the recognized scientific value of the information and/or the methodology used to evaluate distributed model input and parameters. "

Pro requirement 3.1: There is no well-defined formatting of the solution therefor the results are very hard to comprehend.

Pro requirement 1.10: The data analytics project must result in: graphs showing the "Duration curves using different values of the" of "subsurface flow coefficient c ", "position parameter in the gamma distribution h", and "scale parameter kb in the gamma distribution "; as well as a "Comparison between observed and simulated discharges with the optimal dataset: hydrographs (a) and duration curves (b). ".

Note: This case study was unique one given that it wasn't truly a diagnostic analytics project although the title claimed to be the diagnostic analysis of distributed input and parameter datasets, where as the result of this project was a diagnostic tool. Although for the purposes of this thesis this case study can be considered it does exceed the scope of this data analytics project. The scope of this case study shows the potential use of the generic requirements that are produced as a result of thesis to define what is required of a data analytics project with anything exceeding that being a beyond the scope of what a data analytics project is.

Case study 5: Diagnostic analysis of a single-cell Proton Exchange Membrane unitized regenerative fuel cell using numerical simulation (Arif, Cheung et al., 2021)

Initiation:

Pro requirement 1.1: The goal of this project is to "to identify key performance limiting factors in fuel cell mode of a PEM Unitised Regenerative Fuel cell (URFC) fabricated at RMIT".

Pro requirement 1.2: A table containing the nomenclature that relates to this data analytics project must be defined alongside the "source terms of governing equations".

Pro requirement 1.3: This data analytics project will carry out diagnostic analysis on the "single-cell PEM URFC designed and made at RMIT University".

Pro requirement 1.4: The performance of the fuel cell is measured using the "maximum power of the RMIT cell " in "W/cm2".

Pro requirement 4.1: No validation methodology is mentioned.

Acquisition:

Pro requirement 2.1: The diagnostic analytics will be done using "computer simulatione namely the ANSYS PEM Fuel Cell Module"

Pro requirement 1.5: The data inputs for the model used within this project can only be acquired as derived data.

Pro requirement 4.2: The model used in this analytics project uses "estimated values of the input parameters obtained from" "the simulation polarization curve"

Pro requirement 1.6: The estimated data used in this diagnostic analytics project must contain data that relates to "model input parameters".

Analysis:

Pro requirement 1.7: The diagnostic analytics project is carried out using the "diagnosis by simulation".

Pro requirement 1.8: The steps for this project have been defined by Arif, Cheung et al. but the complexity of said steps are beyond the scope of this thesis.

Pro requirement 2.2: This diagnostic analytics project "ANSYS" to work with the "Fuel Cell Module"

Presentation:

Pro requirement 1.9: The results of the data analytics project include "the performance limiting factors of RMIT cell ", the effects of "Hwang's operating conditions ", as well as varying " the values of input parameters related to selected cell properties until the polarization curves of both URFCs matched in this region".

Pro requirement 3.1: The results of the diagnostic project are not categorically defined.

Pro requirement 1.10: The diagnostic analytics project must line graphs that demonstrate the "Matching of simulated polarization curve RMIT single-cell PEM URFC with Hwang's URFC experimental".