DESCRIPTIVE ANALYTICS CASE STUDIES ANALYSES

This document consists of two sections that contain the following information:

1. Analyses of descriptive analytics case studies.
2. Analysis of the results produced by the case study analysis.

1. Case Study Analyses

Case Study 1:Students' perceptions of a community health advocacy skills building activity: A descriptive analysis [1]

The goal of this project is to understand the effectiveness of a community health advocacy building activity. When looking at the overall project one can infer that there were four stages within the project they are as follows;initiation, data acquisition, statistical analysis, and finally presenting the results(data visualization). The requirements that required to carry out each of these phases is as follows:

Initiation:

- Before the start of the project as with all research the authors had to clearly define what the objective of this project is.

Pro requirement 1.1: This 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: Data analytics project must have a clearly defined goal.

- Once the goal is clearly defined the authors have to select what type of analytics (method) is required in order to achieve the specified 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: Data analytics project must clearly define the 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.

Data acquisition:

- Data acquisition is necessary regardless of what type of data analytics is being done, and defining form where and how data will be acquired is vital.

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.

Data analysis:

- Once data has been acquired the type of descriptive analysis that will be carried out must be decided.

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

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

- The algorithm selection is also accompanied by the select of which tool will be used to execute that algorithm.

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

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

Presenting the results:

- Selection of how to present the insights provided by the data is the final stage of the data analytics projects. Both format and content are very important given that any given format such as a graph can include any metrics therefore what information should be confided using a given format is very important.

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 predefined.

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

This case study is very similar in terms of having the generic requirements inexplicitly defined but there are additional requirements that can be derived by analyzing the literature.

Initiation:

Gen 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".

Gen 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".

Data acquisition:

Gen 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".

- The completeness of the data must be considered when data acquisition is being carried out, and not just when considering the data set but also individual data entries.

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 entities must be mentioned.

Data analysis:

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

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

Presenting results:

Gen requirement 1.6: The findings must be presented in a table showing the "Distribution of 466 patients".
Overall this research provided only one new generic requirement regarding the quality (completeness) of the data that can be used to do the analytics.

Case study 3: Restaurant closures during the COVID-19 pandemic: A descriptive analysis [3]

Initiation:

Gen requirement 1.1: This project seeks to answer the question "Which restaurants were more likely to exit the industry in this challenging time?"

Gen requirement 1.2: The author provides "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."

- This author provides an in depth description of the steps that are taken in order to complete the project in the project introduction.

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).

Data acquisition:

Gen requirement 1.3: The project will use data form "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."

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

- The source from which the data is collected is very important, but some data sources do not have ready to use data and this will require the use of data acquisition methods.

Pro requirement 3.2: Data was collected "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.

Data analysis:

Gen requirement 1.4: The data collected will be used to "estimate binary response models (LPM, logit or probit linking closures and restaurant characteristics."

Gen 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: The data analytics project must uncover and convey if any factor(s) may be affecting the reliability of the result of the data analysis.

Presenting results:

Gen 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[4]

Initiation:

Gen requirement 1.1: The goal of this data 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."

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

Gen 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".

Data acquisition:

Gen requirement 1.3: The data is "the real data of the Expenditure Budget Report from one of the cities in Indonesia."

- This project places a comparatively greater emphasis on data acquisition. With the use of data integration and data warehousing and a subtle but more specific description of the data that is being used.

Pro requirement 4.1: "The data obtained 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: "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 projects where it is applicable.

Pro requirement 4.3: "Dimensional tables and fact tables are integrated with the data warehouse scheme as in Figure 4"

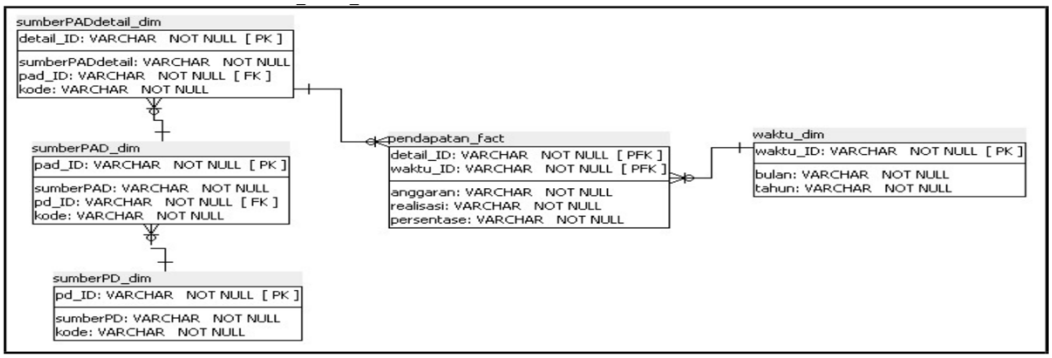

Figure 4. Data warehouse scheme

Gen requirement 2.1: "The ETL process involves" "filling the missing data, delete unnecessary data and repairing inconsistent data.".

Gen requirement 3.2: "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."

Data analysis:

Gen requirement 1.4: "data visualization that is making a dashboard and displaying meaningful information."

Gen requirement 1.5: "Data visualization in this research was designed using Tableau"

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

Presenting results:

Gen requirement 1.6: “

1. KnowingconditionandtrendofPADeveryyear
2. KnowingthecontributionofeachsourcePAD
3. KnowingthebiggestcontributionofsourcePAD
4. KnowingtheamountofPADeverymonth
5. KnowingwhenthelargestPADforthelast5years
6. Knowing the ratio of PAD realization to PAD budget every year

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

- Even though this project mainly focuses on doing descriptive analysis using data visualization the authors go a step further and use a linear regression algorithm in order to predict PAD for the next five years.

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 data 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[5]

Initiation:

Gen 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”.

Gen 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”.

Gen requirement 3.1: An in depth description of the context(COVID-19) to which this study relates is provided

Data acquisition:

Gen requirement 1.3: “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.”

Gen 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

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

Gen requirement 4.3: Data warehousing is not required.

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

Gen requirement 3.2: "Data collection instruments were a demographic questionnaire and a COVID-19 VH questionnaire"

Data analysis:

Gen 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"

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

Gen 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."

Presenting result:
Gen requirement 1.6: The graphical representation of the data will be done using tables.

Gen 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".

2. Analysis of Case Study Analyses

Table 2.1 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 2.1

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 2.2 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 2.2

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.

References

1. Frances Hardin-Fanning, Kimberly R. Hartson, Lynette Galloway, Nancy Kern, Rebecca Gesler,Students' perceptions of a community health advocacy skills building activity: A descriptive analysis, Nurse Education Today, Volume 120, 2023, 105627, ISSN 0260-6917,

2. Andressa Sato de Aquino Lopes, Leopoldo Duailibe Nogueira Santos, Mariana de Campos Razé, Rosana Lazzarini,Alopecia areata: descriptive analysis in a Brazilian sample,Anais Brasileiros de Dermatologia,Volume 97, Issue 5,2022,Pages 654-656,ISSN 0365-0596,

3. Dmitry Sedov,Restaurant closures during the COVID-19 pandemic: A descriptive analysis, Economics Letters, Volume 213, 2022, 110380, ISSN 0165-1765,

4. N. Irzavika and S. H. Supangkat, "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

5. *M. S. Albahly and M. E. Seliaman, "Evaluation of the impact of Clinical Decision Support Systems: Descriptive Analytics," 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2020, pp. 1-5, doi: 10.1109/ICCIS49240.2020.9257686*.