Case Study Analyses

This document contains the results of a case study analysis conducted as part research regarding the creation of a requirements template for the data analytics projects.

Descriptive analytics case studies 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"

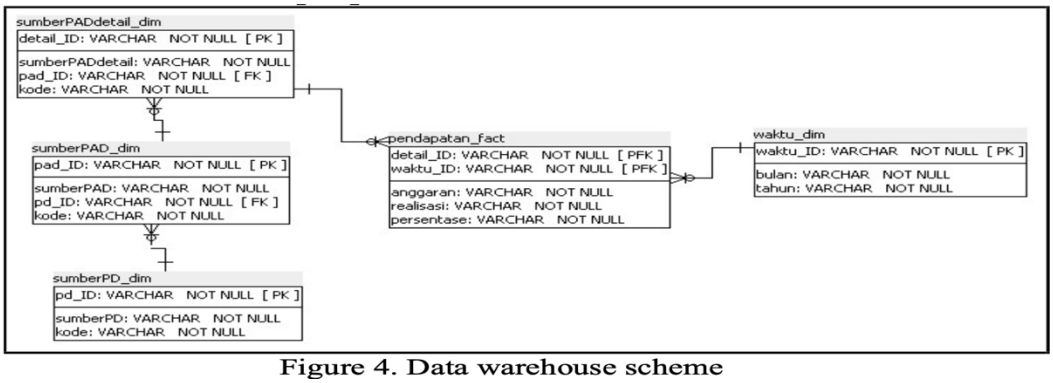

Figure 4. Data warehouse scheme

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".

Diagnostic analytics case studies analyses

Case Study 1: Diagnostic Analysis for outlier detection in big data analysis[6]

Initiation:

Just as with descriptive analytics, diagnostics analytics must have a goal, as well as an understanding of the level of expertise of the users of the results of the analysis.

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: Data 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 data 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.

- The nature of diagnostics means that a system/object must be looked at, therefore when doing diagnostic analytics the system/object/event that will be analyzed/diagnosed must be defined alongside a quantitative metric(s) that will be used to evaluate the system.

Pro requirement 1.3: This data 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 data 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 data analytics project must have a quantitative metric(s) that is used to evaluate the system.

Data acquisition:

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

Gen requirement 1.5: The data analytics project must have a source(s) of data.

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 data analytics project must be defined.

Data analysis:

- Diagnostics analytics has more complex methodologies when compared to descriptive analytics therefore the requirements that need to be specified are as expected larger.

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:

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 $\theta_{tf} \in (0,1)$.

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

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.

An extract taken from research paper [14]

Gen requirement 1.8: A data analytics project must have an in depth definition of the strategy that will be used to carry out the analytics.

Pro requirement 1.9: "Fig. 2 shows the visualization of the price in the dataset with the outlier and without the outlier. It can be seen that the presence of the outlier provides a significant difference in the graph."

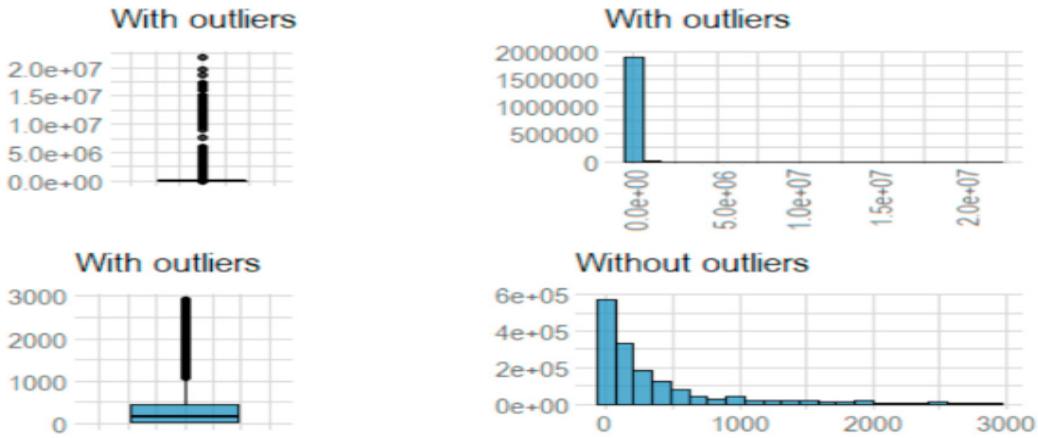

Fig. 2. Outliers in the Global Food Price dataset.

An extract taken from research paper [14]

- 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.

Presenting findings:

- Unlike descriptive analytics which can culminate in a more visualisable end result, diagnostic analytics has a more textual result. Where things are more definitive and exact.

Pro requirement 1.9: "From the analysis, it can be seen that three main factors contribute to the outlier, which is; currency, year, prices, location and type of food".

Gen requirements 1.9: The diagnostic data analytics project must clearly state what are the causes of issues within the system.

Case study 2: Diagnostic analysis of regional ozone pollution in Yangtze River Delta, China: A case study in summer 2020 [7]

Gen requirement 1.1: The goal of this data 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".

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

Gen requirement 1.3: This data 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)".

Gen 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".
*Note: 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.

Data 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 with the data analytics project as well as how said model(s) were used must be clearly defined.

Gen 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/)."

Gen requirement 1.6

Data analytics:

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

Gen 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. Although it can be said that the methodology could have been more concisely defined.

Gen 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 data analytics project must have clearly defined tools that are used and must be defined alongside the application of said tool.

Presentation of results:

Gen requirement 1.9: Within the context of this data analytics project "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, accounting for 39.2% and 34.6%, respectively", 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 2.3: 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."

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 [8]

Initiation:

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

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

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

Gen 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".

Data acquisition

Gen requirement 2.1: This data analytics project will utilize "mixed logit model to analyze the underlying factors towards injury severities in crashes involving cyclists"

Gen requirement 1.5: Inorder to carry out the data analytics 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".

Gen 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."

Data analysis:

Gen requirement 1.7: The data analytics will be carried out using a "mixed logit model" as well as "Marginal effect analysis".

Gen 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 1** and **Figure 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:

$$U_{ij} = \beta_i X_{ij} + \varepsilon_{ij} \quad (1)$$

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:

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

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):

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

Figure 1 :An extract taken from research paper regarding the mixed logit methodology used within the study [14]

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

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:

$$E_{X_{ijk}}^{P_{ij}} = P_{ij}(X_{ijk} = 1) - P_{ij}(X_{ijk} = 0) \quad (4)$$

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.

Figure 2 :An extract taken from research paper [14]

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

Presentation of results:

Gen requirement 1.9: The data analytics project results in the conclusion that "32 variables are identified with significant impacts on at least one of the cyclist injury severity levels, among which 5 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 it should be noted that, due to the fact that no random parameter has been found in the nighttime model, the nighttime model finally collapses into a multinomial logit model (MNL) with 25 statistically significant contributing factors being identified."

Gen 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 3.2: 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.

List of research articles referenced

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*.

6. Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon,Diagnostic analysis for outlier detection in big data analytics,Procedia Computer Science,Volume 197,2022,Pages 685-692,ISSN 1877-0509,

7. Lin Li, Fangjian Xie, Jingyi Li, Kangjia Gong, Xiaodong Xie, Yang Qin, Momei Qin, Jianlin Hu,Diagnostic analysis of regional ozone pollution in Yangtze River Delta, China: A case study in summer 2020,Science of The Total

8.