Journal of Management for Global Sustainability Volume 6, Issue 2, 2018

USING THE BUSINESS CLASSROOM TO HELP FE Y ALEGRÍA-BOLIVIA SCHOOLS WITH ANALYTICS AND PATTERN VISUALIZATION

KATHLEEN CAMPBELL GARWOOD
(corresponding author)
Department of Decision & System Sciences
St. Joseph's University
Philadelphia, Pennsylvania, U.S.A.
kcampbel@sju.edu

JOAO NEIVA DE FIGUEIREDO
Department of Decision & System Sciences
Haub School of Business
St. Joseph's University
Philadelphia, Pennsylvania, U.S.A.
jneiva@sju.edu

HAYLEY F. MILES
Master of Science in Business Intelligence and Analytics
St. Joseph's University
hm671519@sju.edu

MIGUEL ANGEL MARCA BARRIENTOS
Fe y Alegría in Bolivia
arearegular@feyalegria.edu.bo

CITATION OF THIS ARTICLE IN PRESS: Garwood, K. C., de Figueiredo, J. N., Miles, H. F., & Barrientos, M. A. M. 2018. Using the Business Classroom to Help Fe y Alegría-Bolivia Schools with Analytics and Pattern Visualization. ***Journal of Management for Global Sustainability 6(2).***

ABSTRACT. This article describes the analytical support a Saint Joseph's University (SJU) data mining class provided over the past three academic years to Fe y Alegría in Bolivia (FyAB), a Jesuit-sponsored institution dedicated to the education of the poor and looking for a feasible model that could help them identify which students and schools have the most need. SJU undergraduates, working without viable socio-economic household income information for each student in the database, had to be creative in assisting FyAB using only survey data provided by Bolivian school-age pupils. Working in consultation with FyAB school representatives, their goal for each iteration was twofold: 1) create a model that provides evidence, given current sample data, of the students most in need and 2) expand it for application across the larger population of FyAB schools. Such work exemplifies, as noted by Pope Francis in his encyclical *Laudato Si'* (2015), the importance of equality and justice in education as instruments toward sustainability. This article thus provides context for, and a historical background of, this ongoing initiative, and describes its specific characteristics. It reviews sequential cohorts of students by semester, how the requests, focus, and models evolved with new and changing issues, and concludes by sharing a system SJU students created in the fall of 2017—an innovative web-based and easily updated visualization tool that allows for very efficient examination of survey answers—to help make initial analyses easier for those looking to implement immediate student outreach initiatives in Bolivia.

KEYWORDS: social sustainability; education for the marginalized; data analysis for sustainability; data mining in education; data visualization for education; school education equality

INTRODUCTION

Fe y Alegría (FyA)[1] is a Jesuit-sponsored institution dedicated to educating the poorest of the poor in over twenty countries, mostly in Latin America. As Pope Francis states in his encyclical *Laudato Si'* (2015):

> We have to realize that a true ecological approach always becomes a social approach; it must integrate questions of justice in debates on the environment, so as to hear both the cry of the earth and the cry of the poor. (Francis, 2015: #49)

---

[1] Because Fe y Alegría's mission and work is present in over 20 countries worldwide (mostly in Latin America), we use FyA when referring to the broader Fe y Alegría federation and FyAB when referring to Fe y Alegría in the country of Bolivia.

> We are faced not with two separate crises, one environmental and the other social, but rather with one complex crisis which is both social and environmental. Strategies for a solution demand an integrated approach to combating poverty, restoring dignity to the excluded, and at the same time protecting nature. (Francis, 2015: #139)

FyA is also highly consistent with two of the United Nations' Sustainable Development Goals, namely Nos. 4 (Quality Education) and 10 (Reduced Inequality), and indirectly related to Nos. 1 (No Poverty) and 8 (Decent Work and Economic Growth).

This article describes the analytical support a data mining class from Saint Joseph's University (SJU) Haub School provided over the past three academic years to Fe y Alegría in Bolivia (FyAB). Indeed, despite the two institutions having had an ongoing partnership for over fifteen years, the question of how to identify, from survey data alone, which early high-school students[2] are most impoverished was first brought into the SJU data mining classroom only in the fall of 2015. It was essential to identify and prioritize which students were in most economic need so that targeted, effective support could be provided given the limited availability of education support resources for helping students and teachers at FyAB. Undergraduate students, on the other hand, could also benefit from using data analytics in a real-world context and learn more about vastly different realities in the process. The initiative thus resulted in a discovery-based learning opportunity for undergraduate business students to address a real-world challenge with many potential benefits for the underserved.

The Bolivian Context

Bolivia is a landlocked developing country in South America with an estimated population of 11.1 million inhabitants as of 2017, a GDP of $37.78 billion, and an area of 1,098,000 square kilometers. It has, broadly speaking, three geographic regions with very different climates and where different flora, fauna, and human adaptation factors have led to the development of very distinct autochthonous cultures over the centuries. Bolivia thus has the highest percentage of inhabitants with indigenous ethnicity (over 60% percent) among Latin American nations, and cultural plurality permeates Bolivian society to this day. With over three dozen native-American tribal nations represented, the most numerous being the Aymaras, the Quéchuas, and the Guaranis,

---

[2] Specifically the third year of *la secundaria*, i.e., corresponding to high-school freshmen students in the United States.

the country's official name is *Estado Plurinacional de Bolivia* (de Mesa, Gisbert, & Gisbert, 2008: 17, 43, 49). Cultural heritage is thus highly respected, with a bilingual education that includes Spanish now a reality for many Quéchuan and Aymaran children. Pre-university level education includes two cycles, with the primary one encompassing six years of elementary school (ages 6- 11) and the secondary cycle covering another six (ages 12–17). Yet while 92.5% of the mostly young Bolivian population is literate, and despite recent public initiatives having done much to improve education, the country still lags behind other South American nations according to most pedagogical metrics. Bolivia also has, despite recent improvement, very unequal income and wealth distribution, and while raw materials are plentiful given that the country exports mineral commodities including natural gas, crude oil, and tin, Bolivia's original inhabitants have seen the richness of their land being used to benefit outsiders ever since the discovery of the New World. Indeed, with roughly forty percent of the population below the poverty line and despite the country's Gini index for distribution of family income having fallen from 0.60 to 0.47, Bolivia still exhibits extreme poverty (World Bank, 2018).

Fe y Alegría in Bolivia

Founded in 1955 in Caracas, Venezuela, Fe y Alegría ("Faith and Joy") is a Jesuit-sponsored, not-for-profit organization focused on the education and development of the "poorest of the poor" in over twenty mostly Latin American countries but also including Chad, Madagascar, and Spain. Now headquartered in Bogotá, Colombia, FyA acts in each country through a small staff which leverages capabilities and resources across schools within the network to train and develop faculty members, work with individual school personnel in establishing and reaching aggressive goals, identify and develop best practices, and ensure that these are disseminated. In 2018, FyA schools numbered over 1,000 worldwide and reached over 500,000 students. The organization started in Bolivia in 1966 and is present in every Bolivian province (*departamento*), operating in a decentralized structure with departmental (provincial) directors who provide local leadership and a national office that coordinates nationwide support activities. Offering a wide range of educational services and now an integral part of the country's educational system, FyAB counts over 400 schools with more than 10,000 teachers and over 180,000 students in its care. The largest area, "formal education," is where the organization oversees a network of elementary and secondary schools that also offer classes in the widely spoken Quéchua and Aymara indigenous languages. The local impact of FyA is thus apparent because network schools, working in very harsh conditions, not only help individuals become fully integrated members of society with a deep appreciation and respect for their own culture and heritage but also foster within the communities served a very strong sense of self-worth and local identity.

The Fe y Alegría Bolivia-Saint Joseph's University Partnership

The partnership between FyAB and SJU began over fifteen years ago with the facilitation of an agreement between the Jesuit Provinces of Maryland and Bolivia to collaborate and share resources. In 2001, SJU staff conducted two exploratory visits to Bolivia and tangible steps were taken to initiate a joint collaboration that resulted in three initiatives: FyAB staff attending the English Services Center near SJU; periodic ten-day SJU faculty and staff immersion trips to Bolivia which were scheduled annually at first, subject to availability of funds; and SJU faculty-led workshops for FyAB. Indeed, while there had already been several faculty collaborations for the benefit of FyA, the first long-term community-engaged research project began after the 2008 faculty immersion trip and had examining school efficiencies as its objective (de Figueiredo & Marca Barrientos, 2012). The initiative described in this article took hold in 2015, and was followed by several other service-related in-class initiatives with Fe y Alegría.

The partnership has been able to grow over the years because a solid foundation of trust was gradually built. One factor contributing to this growth was a strengths-based approach where both parties endeavored to identify and recognize each other's abilities. A second factor was mutual respect for cultural characteristics, including the Bolivian culture's gift of focusing on the whole individual and therefore moving beyond the task at hand. A third factor was the concerted mutual deference and curiosity which led to active listening and which in turn led to effective communication.

CONCEPT DEVELOPMENT

Identifying the Fe y Alegría Students Most in Need

Consistent with FyA's mission to provide education support to those who need it most, FyAB focuses very much on obtaining reliable data as exemplified by surveys that students periodically fill out. They were thus searching in 2015 for a way to identify, through survey data alone, students who were most in need of outreach efforts, i.e., those whose families, unbeknownst to their teachers, might be in a difficult socio-economic condition and who were most at risk therefore of not finishing their studies. The outreach efforts occur on an individualized basis and involve several fronts, e.g., making specific efforts to help prevent dropping out, offering special courses on practical skills such as waitressing and basic mechanics, and organizing special initiatives such as introductions to entrepreneurship. FyAB also hopes to identify students with difficult personal situations as early as possible because older students' families tend to be less involved in education.

The Benefits of Using Real-Case Data

Moore and Roberts (1989) initiated work on the merits of teaching "data-driven" courses where the main goal, in the hopes of stimulating discussion and development of statistical techniques, is to lead the class in identifying questions and analyses arising from a set of data that is intrinsically interesting or relevant to the students. In this case, therefore, it is important to identify those most in need. Using household surveys, moreover, had become an increasingly important way to measure poverty and well-being around the world (Deaton, 2003), with student surveys in particular providing important information in which qualitative answer categories are frequently reverted to a Likert scale, thereby allowing for quantitative data analysis. Likert-type questions are often used when the goal is to identify extremes (extreme poverty in this case) while cluster analysis initiates segregation of the objects being studied (Allen & Seaman, 2007).

In 2015, a business data mining course was created. 70% of the classroom time followed a traditional learning model while the remainder was devoted to raw data sets which students were expected to draw conclusions from and provide clear reports on in a short period of time (about 1 to 48 hours, depending on the case). Student teams in each class were in effect acting as pro-bono consultants for FyAB.

The data shared in the case of the FyAB project provided SJU students with a window into a vastly different reality: one of scarce resources and urgent needs. Initiation into a culture different than their own thus became a mind-broadening experience that increased their engagement and emotional attachment, leading them to build on the cohorts' analyses of prior terms and creatively add value in innovative ways. This has resulted in a virtuous cycle over the semesters, with every term culminating in the presentation of new approaches for identifying the most impoverished students in the Bolivian schools surveyed and allowing FyAB schools to provide targeted support. Thus, what began as an analysis of two FyAB schools in the city of Potosí evolved into an examination of institutions across different regions of the country and later incorporated schools outside the FyAB.

The objectives and key results over the course of this analysis have been a moving target, however, much like in the real world where data might not exist at a fixed moment in time (Baumer, 2015). Like a practicing statistician, the lion's share of the time spent on this project was devoted to data cleaning and manipulation (or data wrangling, as it is often called [Kandel et al., 2011]) with less given for conclusions and next steps. Measurement in terms of feedback was necessary, however, to recognize whether what had been done in the past should be used going forward. On this point, then, there were three instances when we stopped to make sure we were moving in the correct direction. After the first data set was analyzed, a list of identified students was shared back with the school where the data was collected and faculty confirmed that those on it were indeed very needy. Many of them, however, had already dropped out during the 18-month lag time between data collection and analysis. In the second data set, a large portion of students came from a similar school and so a shop was immediately implemented to help both the school and the students identified hence. Finally, in the third data set, students identified which among the schools that were requesting for inclusion into the FyA system would benefit the most from it.

The business undergraduate students' contributions, in the meantime, grew in areas that are not as easy to measure, including survey question suggestion (something these students would not usually research and learn) and advanced pattern visualization techniques (which "pushes" them beyond simple bar graphs and into data story-telling). SJU students also suggested, as described below, the inclusion of additional questions to help increase the response accuracy and reliability of the surveys which originally included 22 multiple-choice questions. The entire experience thus led to deeper student involvement and a broadening of impact.

METHODOLOGY

The Process: Data Analysis Consulting

One major (and often overlooked) component of data analysis that is vital when working with real data is the consultative process itself. While the goal is for students to be comfortable diving into any data set, the reality is that those providing the data (e.g., the end-users, the "clients") usually have more familiarity and a better base knowledge of the material. Their proximity to the raw data can be a double-edged sword, however, as they might be used to looking at it in a given way and therefore miss innovative angles or perspectives with which to proceed. In other words, they don't know what they don't know. The ability to ask the right question, then, and even to explore others becomes an important skill that all good analysts need especially in situations like these. Phases important to the process as such can be seen in Figure 1.

Coming in contact with the data. To develop the ability to "think outside the box" and stimulate creativity in problem-solving, students examining raw data for the first time are encouraged to dive in and, within an hour's time, share with the class what they may have found, i.e., to discover anything interesting that the data might be able to tell them. Students will ideally take a moment or two to familiarize themselves with the survey (see Appendix A1) as well as any goal that the stakeholder puts forth before taking this "dive," although they usually have some interesting insights and invariably many questions by the end of the hour. It is at this point that students become intrigued by the FyAB context and spend time learning more about Bolivia and FyA's work, answering some of the overarching questions in the process. Remaining questions are posed to Miguel Angel Marca Barrientos, national advisor for general education for FyAB, often via live Skype chat in class.

Understanding the stakeholder and engaging with the issue. Students gradually deepen their understanding of the data over the succeeding weeks and begin to perceive the bigger picture of what FyAB is trying to accomplish with this effort. They begin to appreciate and understand the FyAB students whom they are analyzing and become familiar with the regions in Bolivia where the schools are located through articles, the Internet, discussion boards, and communication with Miguel and others who may have also been in the country.

Using data analysis techniques. This corresponds to the period when students are applying statistical techniques as enumerated above, tools that are integral to the data mining course (such as ANOVA, PCA, cluster analysis, etc.), and bridging the conceptual models with a real case which they are not only able to envision but also emotionally committed to. Prior to the final analysis, the students work together as a class to clean the data set using class consensus where they debate whether to impute a value or delete a question or student response based on missing data.

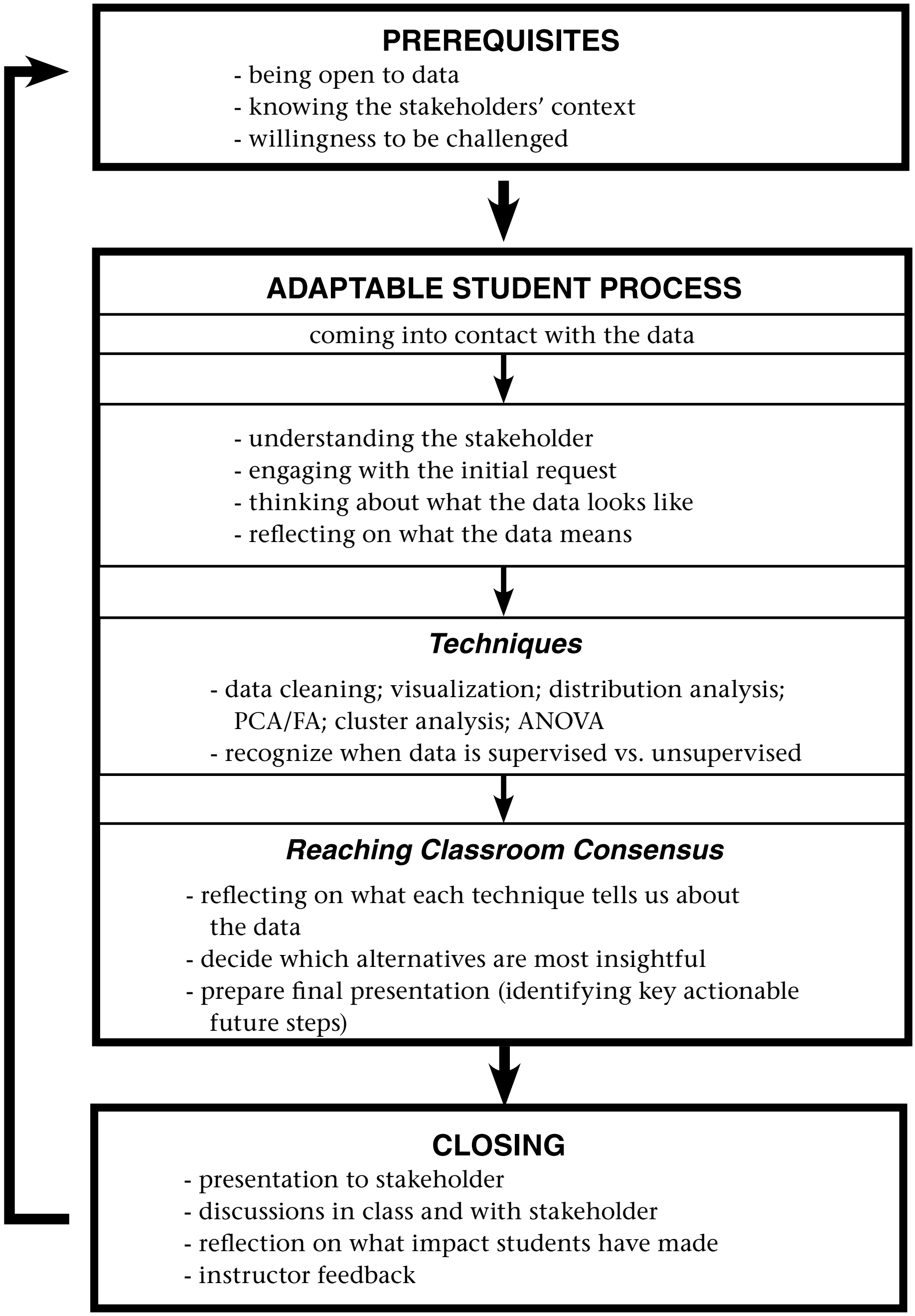

PREREQUISITES
- being open to data
- knowing the stakeholders' context
- willingness to be challenged
ADAPTABLE STUDENT PROCESS
coming into contact with the data
- understanding the stakeholder
- engaging with the initial request
- thinking about what the data looks like
- reflecting on what the data means
Techniques
- data cleaning; visualization; distribution analysis; PCA/FA; cluster analysis; ANOVA
- recognize when data is supervised vs. unsupervised
Reaching Classroom Consensus
- reflecting on what each technique tells us about the data
- decide which alternatives are most insightful
- prepare final presentation (identifying key actionable future steps)
CLOSING
- presentation to stakeholder
- discussions in class and with stakeholder
- reflection on what impact students have made
- instructor feedback

Figure 1: Consulting process applied in this data analysis effort

Reaching classroom consensus. The objective at this point is for the class to reach agreement on which one among the data sets is to be analyzed. This is because while each team started out with the same raw data, each cleaned it in a different way and therefore performed analyses that produced slightly different results. Once students agree on a final data set, they are given two days to come up with a final analysis and presentation which they share live, translated in real-time, with FyAB on Skype. It is a consultative process that has evolved since its first iteration in the fall of 2015 and yet has been applied in each succeeding iteration, for while FyAB's specific questions and goals have changed slightly over time, their main objective of identifying students with the highest socio-economic need from survey data alone has remained the same.

Examples of Impact

In this subsection, we briefly consider impact related directly to activities done within the schools in Bolivia and look at some reflections from the students who ran the analysis at SJU.

Impact on FyAB. As mentioned earlier, the application of this FyAB-SJU initiative has focused mostly on the analysis of surveys filled out by students in the third and fourth years of *la secundaria*. This faces some challenges: First, families in Bolivia tend to not be directly involved with their children's school work at this level (as opposed to lower school, for instance, where family involvement is intense and indeed a prerequisite for FyA participation). This lack of involvement means school officials may not have relevant information about the socio-economic condition of individual students. Second, students in adolescence may be reluctant to provide fully accurate information in a survey for various personal reasons (including, perhaps, feelings of inadequacy), making the importance of data analysis (cross-checking different responses in the same survey, for example) even more desirable. Third, the main objective of the effort is to provide outreach and individual support within enough time to make a difference. Thus, because most FyAB high-school students in unfavorable socio-economic conditions will likely find themselves unable to go to college upon graduation, it is important to identify those in need who still have one or two more years of high-school instruction ahead so their learning can be tailored appropriately.

Depending on the school, FyAB has several programs in place that provide support for students with the most socio-economic need. In the schools of Sucre and Potosi, for example, selected students were prioritized for training and entrepreneurship initiatives under the assumption that these would be helpful for them should they be unable to go to college. These included both basic training for school and job training in areas such as bakery, cooking, basic electrical installations, and various forms of retail selling. One school, for instance, set up a small taco-vending operation to serve as a training ground for their students. Such initiatives are indeed helpful for students not only because of the specific skill training provided but also because they become involved, usually for the first time, with entrepreneurial opportunities in the process.

Impact on SJU Students. Methods for assessing the direct impact on SJU students are still being developed for future cohorts. We do, however, collect reflections from students throughout the duration of the course and, in this case, at the end of each phase of the project.

ANALYSES AND FINDINGS

To give a sequential view of the ongoing process and share, by semester, the analyses considered, relationship flow, successes, and next steps, the in-class analyses conducted by SJU undergraduate business students in response to the original FyAB inquiry is presented in chronological order below.

Data Set 1 (*Fall Semester, 2015;* four schools in Potosí, with 272 student responses before data cleaning and 261 after). A data mining class of SJU students was asked to help assess a simple two-part question posed by Miguel Marca Barrientos:

1. What was the meaning of the principal component analysis/factor analysis (PCA/FA) output which an outside source had generated for him?

2. Could the coefficients from said analysis be applied as weights to produce a meaningful measure that might be indicative of poverty for each student (poverty score)?

The class, working through the PCA and FA topics, collectively decided to focus on the first question as a real-world example of survey analysis. They met with FyAB via Skype to identify the most important variables and discussed how these might be formed into factors—the very goal of PCA/FA. Nevertheless, while this provided insight into the analysis that Miguel had in hand, it did not help answer the more important question of how to create a predictive model for identifying poverty or helping in any way to partition the FyAB students.

The SJU students then continued working through the analysis in an attempt to help with the second question, leading them to the use of data mining to address social issues. They created a poverty model with insights and results that were presented to Miguel at the end of the semester. PCA/FA analysis was performed after the data was first cleaned (which resulted in 11 students being dropped), thereby allowing for data reduction and refined independent variables. Cluster analysis was then applied to partition the 261 remaining students into subgroups of similar attributes to help identify those most in need; the class eventually settled on using eight clusters along with a preset goal of identifying the bottom 25% of the most impoverished individuals. The SJU students expected at first that creating more clusters would result in more equal partitioning and a clearer way to observe the requested threshold of 25%; instead, two very large clusters remained unchanged even as more clusters were made. The smaller clusters, in the meantime, continued to break down into smaller and smaller partitions, thereby clarifying the groups so identified (e.g., one cluster lacking water or another lacking electricity); such resulting uneven partitions thus made the original request to identify the bottom 25% more elusive.

The class ranked the clusters in order using key "poverty" features as defined by the United Nations based on some key questions concerning food, shelter, electricity, water, and parents' work. Using this logic, the analysis identified the 23 most impoverished students (bottom 9%), i.e., those who lacked the basic necessities of food, electricity, or water. The cluster next most in need identified 36 students (moderately low 13.8%) who had 2.28 meals a day on average, no meal before school, and whose fathers were less educated and less likely to work a full five days in each given week.

At this point, the class was able to provide initial identification of students most in need according to school and name. It is important to note, however, that the data was collected two years before (in 2013). As such, while teachers in the respective schools confirmed that a majority of the students identified in the cluster analysis were indeed the most impoverished, many had already dropped out or graduated.

The successes of Data Set 1 included data cleaning, integrity, PCA and FA, cluster analysis, achievement of initial concepts, and proper student identification. More work needed to be done, namely, the use of simpler cluster techniques and creation of a usable predictive model.

Data Set 2 (*Spring Semester, 2016 to Spring Semester, 2017;* six schools in Potosí and Sucre, with 838 student responses before data cleaning and 731 after). This data set, with its techniques and conclusions advancing due to continuous analysis improvement and tool fine-tuning, was used from Spring 2016 through Spring 2017. Until it arrived, however, the class focused on the previous 261-student Data Set 1 for the first half of the semester.

The new data set required intense cleaning, eventually dropping 107 responses due to missing or invalid answers. The class also read up on the two Bolivian regions as they cleaned the data and provided suggestions on how to adapt questions to get more useful insights; these were considered and implemented in subsequent surveys.

Correlations were checked once the data was cleaned, revealing what seemed to be mild correlation (.54) between two variables (mothers' and fathers' education levels). As with previous data, PCA/FA was also run, generating results that were similar to those of the first data set, and used once again for reducing variables and identifying which survey questions would be considered in running the cluster analysis. The class tried several iterations of cluster analysis; they eventually chose eight clusters from which 38 of the most impoverished Bolivian students were identified and immediately made known to FyAB.

School	# of students from school	% of students from the school
Luis Espinal Camps	17	11
Gualberto Paredes	6	7
Sagrada Familia	5	4
Loyola De Fe Y Alegría B	4	2
Jose Maria Valez	4	4
Fray Vicente Bernedo B	2	2

Table 1: Counts and percentage of students in-need by school.

A new application was then considered with this second data set. The 38 students so identified were assigned a binary dependent value of "1" which indicated poverty while the remaining students were assigned a value of "0" which indicated a less impoverished status. The class then attempted to generate a logistic regression using this newly created binary dependent variable. Such a model would help identify weighted averages that might be applicable for future survey-based data sets, although with fewer than 5% of students coded as "1" for impoverished, it was less accurate due to the small sample bias in the logistic model's maximum likelihood estimation (Allison, 2012). The attempt to obtain a logistic regression needed further investigation, therefore, to ensure model accuracy.

The successes during this term included data cleaning and integrity, PCA and FA, cluster analysis, and creation of an initial dependent variable. More work needed to be done on the number of most impoverished students so identified (38 was too small) and the predictive model relied heavily on two survey questions, namely, about electricity and water.

Fall 2016. In the Fall semester of 2016, the class embraced a major change that was prompted by one group who decided to pre-partition a subset of the data. This group classified any Bolivian student as impoverished if they responded that they were without electricity or water. Once classified as such, these respondents were removed from the data set, thereby allowing for the clustering technique to be applied to the remaining 693 students.

After this revised data set was partitioned, the number of impoverished students so identified increased to a total of 69 (9%), nearly doubling the previous number of students identified as such. This change thus helped to create a slightly better logistic model that could feasibly be applied to survey collection and data analysis in the future. Indeed, while the new dependent variable produced a feasible model, logistic models do not have easily identifiable weights; some care about the magnitude of the effect, for example, while others about the magnitude of the odds ratios, yet both are easy to misinterpret (Norton & Dowd, 2018). However, despite the challenges of interpreting the outputs in layman's terms, this was the first time a reasonable model was created. The results were thus shared with Miguel with the goal of producing coefficients for future survey applications.

The successes during this term included data cleaning and integrity, PCA and FA, cluster analysis, creation of an initial dependent variable, identification of a larger sample, and creation of a decent logistic model. More work needed to be done on the number of most impoverished students so identified (69 was still small) and the creation of a continuous dependent variable.

Spring 2017. Two major changes occurred during this term. First, a class of graduate students, in addition to two undergraduate classes, was invited to analyze the data. Second, a resampling method called bootstrapping was considered in this iteration in an attempt to create a continuous dependent variable, i.e., once all the initial steps of cleaning the data and reducing the independent variables (survey questions) through PCA/FA had been successful, cluster analysis could be applied using simulated samples from within the survey to try and make estimates about the student population in FyAB. The technique is often useful for analyzing small, expensive-to-collect data sets for which prior information is sparse, distributional assumptions are unclear, and further data may be difficult to acquire (Henderson, 2005). In this case, no prior data was available, the data was unsupervised, and we had already established a viable method for model-creation that needed verification.

After SJU students performed the initial data cleaning and reduction, smaller subsets (about 25%) of the larger data set were chosen at random. These subsets were run through cluster analysis by hand (using four clusters typically) and each cluster was ranked according to need based on the averages of responses to questions being considered. Here the two undergraduate classes noticed, on their part, that the rankings naturally followed an ordered pattern that most often depended on number of meals the students consumed on a daily basis, i.e., having fewer meals was typically consistent with the group considered most impoverished. Working in groups (with each group running an iteration), the SJU students also assigned a value for the FyAB students within each cluster once the rankings were put in order. Thus, due to random generation and 20 separate groups running the analysis, each Bolivian student was chosen multiple times and assigned a ranked cluster value between 1 (most in need) and 4 (least in need) each time they were chosen for an iteration. The classes averaged these recorded values to see if a continuous variable might be created according to which each student might be attributed a continuous value (between 1 to 4 inclusive).

The graduate students, working separately in the meantime, came to a similar conclusion. Having written a program in R that was simulated 10,000 times, this group recognized that true bootstrapping required resampling to be done thousands of times over to assure the most appropriate results. Each time a sample was taken, a cluster value was recorded for each FyAB student (resulting in approximately 2,000 values assigned to each); these values were then averaged out, creating a feasible continuous dependent variable ranging from 1 to 4 inclusive. In this way, a multiple linear regression, by possibly identifying weights applied to questions, became a feasible option for getting a poverty value. A distribution of this continuous dependent variable revealed clear partitions of students, with the lowest section (y-value ≤ 2.3) assigned a value of one for impoverished and the remaining students given a value of zero indicating less impoverished (see Figure 2). Thus, at this point in time, both logistic and multiple regression models became available for SJU students to use as feasible methodologies for evaluating the data in view of the ideal scenario: generating a model with viable weights that FyAB can apply to future student responses to identify those who might be the most in need.

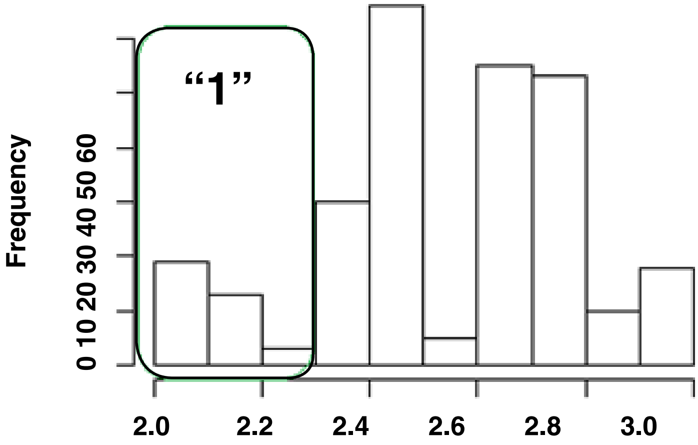
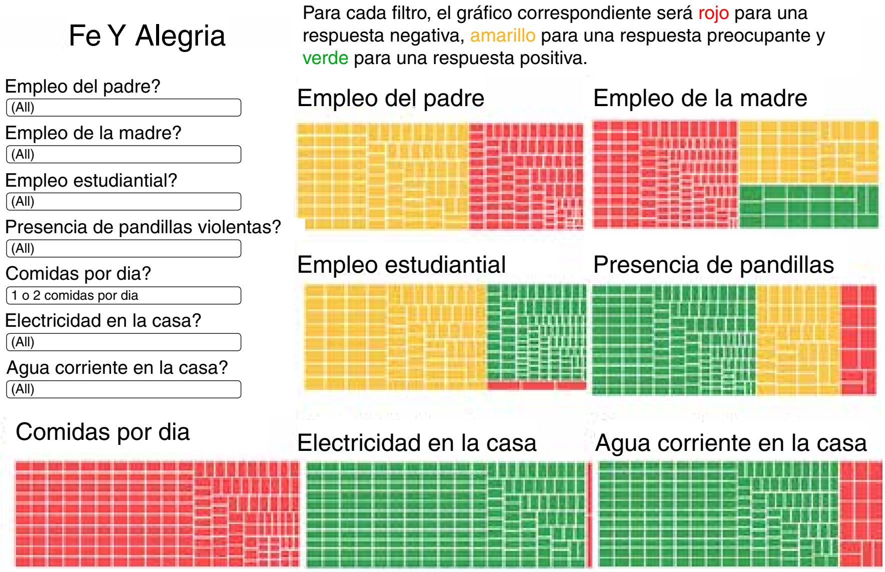


Figure 2: Distribution of dependent averages after bootstrapping is applied

Two reasonable and useful models, then, were created during this term, yet more work on the use of other data samples needed to be done to verify the process. This was carried out over the summer to assess the most viable cluster size and best model. A more advanced R program was thus created that could be applied to this data set and to future survey results, thereby providing an open source code and model which could be used by FyAB and others who might give similar surveys (Garwood & Dhobale, 2018).

Data Set 3 (18 to 32 students from each of eight schools in various regions of Bolivia, totaling 204 responses before data cleaning). This final data set was dramatically different from previous ones—student names were not provided, the schools included were not part of the FyAB network at that time, and many students left blank answers. The underlying question/goal here was noticeably different, namely, to identify schools with students similar to those within FyAB, assuming perhaps that these schools might be interested in being adopted by the FyA administration.

Fall 2017. While waiting for access to Data Set 3, the students considered Data Set 2 for the first iteration of the project. They were asked to clean and organize the data as well as analyze it for insights and anomalies. One group proposed a dashboard that could filter through all the responses based on eight questions that provided an overall visual of the data. A filter applied to as little as one or two questions, for instance, could produce a list of students who could be high in need, with the resulting visual dashboard narrowing down to those students falling within a need category based on the filter(s) chosen. Possible responses within each question were color-coded in red (likely impoverished), yellow (possibly impoverished), and green (less likely impoverished) to indicate a different need level for those filtered in each. Indeed, while such a method does not undergo the statistical processes required to make a model, it is useful for immediate visualization and was shared with FyAB soon after its creation. This tool can be updated with new data easily and efficiently and provides immediate poverty insights to FyAB. A grayscale snapshot of this dashboard is shown in Figure 3; immediate access to the live version is available at https://public.tableau.com/profile/hayley.miles#!/vizhome/FeYAlegria/Dashboard1.


Fe Y Alegria

Para cada filtro, el gráfico correspondiente será rojo para una respuesta negativa, amarillo para una respuesta preocupante y verde para una respuesta positiva.

Empleo del padre? (All)
Empleo de la madre? (All)
Empleo estudiantial? (All)
Presencia de pandillas violentas? (All)
Comidas por dia? 1 o 2 comidas por dia
Electricidad en la casa? (All)
Agua corriente en la casa? (All)

Figure 3: Visualization of students by need, with drop down choices on student employment, parents' work, food, number of rooms, and electricity. Choosing different responses presents how each student answered.

Once students had familiarized themselves with FyAB, they were asked to compare Data Set 2 with the new and very different Data Set 3, with one objective proposed to the class being to advise FyAB as to how the two data sets compared to one another and see if any patterns existed. The students thus ran analyses of variance in an effort to identify any schools in the 2017 data that aligned more closely with those in the 2016 set, i.e., that seem to have similar students. All the schools were also analyzed as one whole in an effort to identify the 25% most impoverished students overall. Similar to the analysis conducted in the previous semester, the students ran PCA/FA to identify the characteristics that provided the most variability and isolate clusters of impoverished students. Finally, efforts were made to identify the schools the students came from (school) and who the students were (according to student names for the 2016 data and student numbers for the 2017 set).

As an example of the ANOVA used to identify students in need, the average number of meals per day by data set (see Figures 4a and 4b) was analyzed. This produced evidence of students in two schools within each data set who receive less than the primary average of three meals a day.

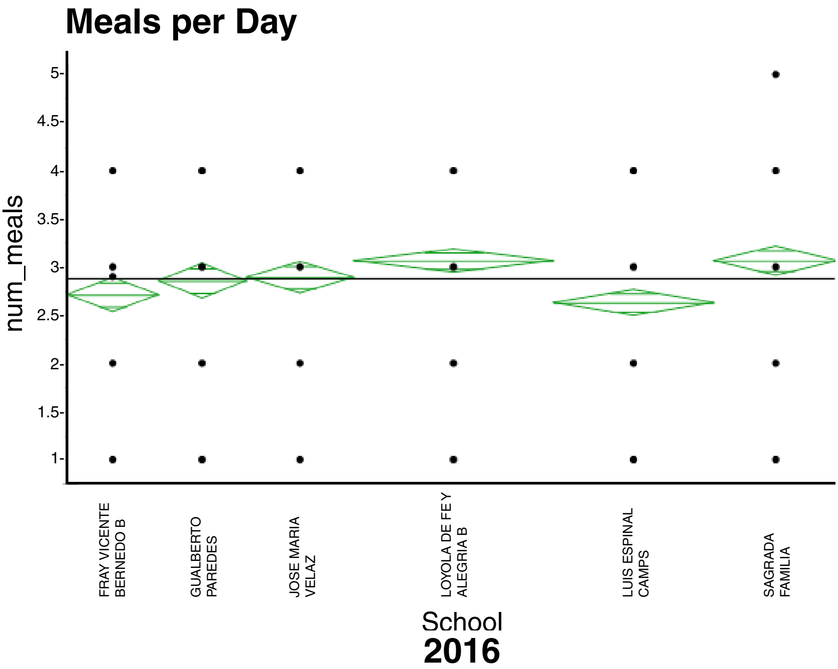


Figure 4a: ANOVA of the number of meals per day by school, showing 2016 data. The majority of students for both years did answer that they ate three meals per day. Note that Fray Vicente Bernedo B and Luis Espinal Camps had a larger portion of students with fewer than three meals and are the focus of the 2016 data.

The overall analysis thus provided early insights into patterns as well as a recommendation to FyAB regarding which schools in Data Set 3 seemed to have the most need and how these schools compared with currently known FyAB schools.

The successes during this term included adapting to new and different data and identifying similar schools across data sets. More work needs to be done, however, in creating a poverty index that can help find schools most in need given only small samples and a lack of names.

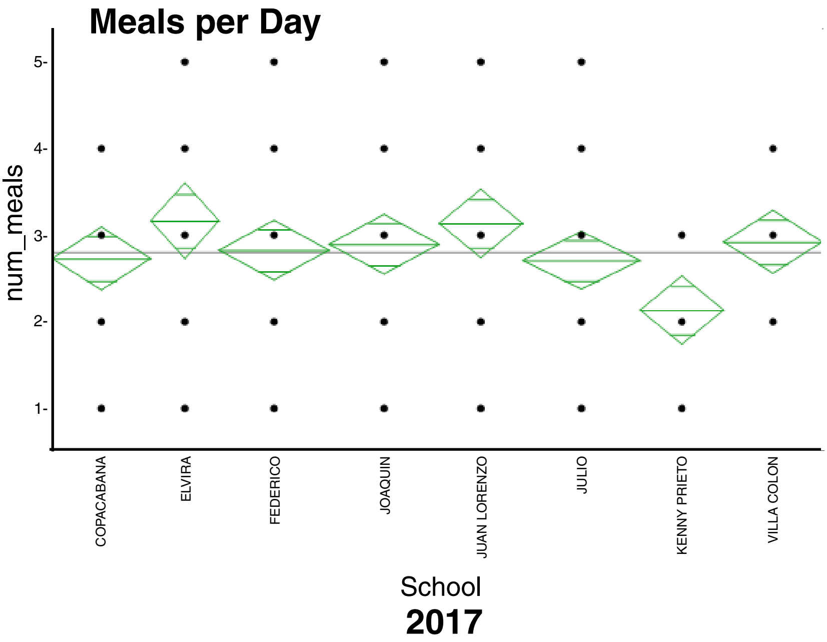


Figure 4b: ANOVA of the number of meals per day by school, showing 2017 data. The three schools that should be the focus for 2017 due to the larger portion of their students having less than three meals per day are Copacabana (Virgin de Copacabana), Julio (De Julio B Senakata), and Kenny Prieto.

Tableau Dashboard: An Expanded Example of Students Moving the Project Forward

When the class created the original Tableau dashboard, it contained tree maps partitioned by eight separate questions. A tree map is a space-constrained diagram (visualization) of a hierarchical structure which uses an enclosure to visualize trees and size and color coding to map sub-trees onto a sequence of nested rectangular areas (Shneiderman & Wattenberg, 2001). Survey respondents were thus easily compared to one another by nesting the data into separate categories (eight in this case) and coloring the level of poverty based on answers to the survey questions. The visual also allows for quick interpretation regarding their level of need—dropdown menus at the top of the dashboard contain filters that partition these tree maps and allow for a user-friendly and interactive experience while the color-coding of responses visually articulates each student's need environment.

The questions used in the original Tableau dashboard can be seen in Table 2. To ensure a more successful final product that could be shared within FyAB, several improvements were suggested, including increased participant confidentiality, translation of materials into Spanish, and categorization of answers based on need.

1	How many meals per day do you eat?
2	Is there running water in your house (do you have access to potable water)?
3	How many people sleep in your bedroom?
4	Is there a shower in your house?
5	Which one of the following groups best fits your mother's current work situation?
6	Which one of the following groups best fits your father's current work situation?
7	Does your house have electricity?
8	Do you work and get paid for it?

Table 2: Questions used in original Tableau dashboard

Confidentiality is critical to a shared dashboard. In the original version of this one, each cell represented a student and displayed that student's name as the user hovered over it. To update the initial dashboard, therefore, the first element that needed to be rectified was the confidentiality of the participants. Second, the original dashboard was in English but it needed to be in Spanish as the main users are Spanish speakers. Third, only one question was partitioned into three categories even though color-coding was a success of the original data compilation. The three-level scale was thus adopted for all questions that had three or more responses to reduce polarization among the survey questions.

The privacy of participants was protected by adapting each tree map to include alternate identifiers for both school and student names. To create the identifier codes for the school names, three letter acronyms for each school were created as shown in Table 3. Each student was also assigned an identifying number, with the first on the list of each school assigned 001, the second 002, and so on in increments of 1. The school and student identifiers were then concatenated to form confidential and unique student ID numbers that retained regional information. This list can thus be supplied to users who need to decode a student ID and identify the original student.

Code	School Name
FVB	FRAY VICENTE BERNEDO B
GBP	GUALBERTO PAREDES
JMV	JOSE MARIA VELAZ
LEC	LUIS ESPINAL CAMPS
LFA	LOYOLA DE FE Y ALEGRÍA B
SAF	SAGRADA FAMILIA

Table 3: Codes given to each school used in the tree diagram.

The survey originally came to the schools in Spanish and was translated into English for the use of the data mining students. For the dashboard to be efficient and user-friendly, however, it had to be translated from English to Spanish to allow for ease of use at FyAB. Once this was achieved, the questions to be used were discussed with Miguel and a final dashboard was created. These questions in the final Tableau dashboard can be seen in Table 4. Although the survey asked about the highest education level achieved by the mother and father, SJU students elected to include their employment status in the dashboard as that more directly reflects how the respondent is being provided for. These two questions, combined with the students' need to work, deliver to the user an idea of the financial situation of the respondent's family that is consistent with the United Nations' definition of poverty (United Nations, 1995: 6–12).

1	How many meals per day do you eat?
2	Is there running water in your house (do you have access to potable water)?
3	Which one of the following groups best fits your mother's current work situation?
4	Which one of the following groups best fits your father's current work situation?
5	Does your house have electricity?
6	Do you work and get paid for it?
7	Are there any violent gangs in your neighborhood or school?

Table 4: Questions used in final Tableau dashboard

The survey given to students also offered multiple choice questions with possible answers that ranged from 1 to 2 for yes/no queries and 1 to 4 or 1 to 5 when multiple individual responses were an option. Most responses to each question considered were grouped into three color categories: green, yellow, and red, indicating a low, medium, and high level of need, respectively. There was no corresponding medium or intermediate need (yellow) category in the case of two-answer (yes/no) questions. Table 5 provides an example of how the questions were color-coded for the tree maps and indicates how answers were categorized by needs.

Are there any violent gangs in your neighborhood or school?		
Provided Answers	**Need Base**	**Color**
There aren't any gangs. Neither at school nor in my neighborhood.	Low Need	Green
There are gangs in my neighborhood, but they do not come close to school.	Low Need	Green
When I come to school or go home I see gangs.	Medium Need	Yellow
There are gangs at my school.	High Need	Red
Some of my friends are part of a gang.	High Need	Red

Table 5: Responses to gang question and how these were color coded.

By translating the dashboard back into Spanish, adjusting the confidentiality of students, and altering how the information is presented, the final product became a user-friendly and informative dashboard. Thus, as more data is collected over time, this tool will be helpful in identifying broader trends (such as per region and per school) and a surface level of socio-economic well-being among FyAB students. It is easy adaptable as new data comes in, and can be changed if new or different questions become significant in poverty identification. Moreover, since the tool is web-based, it can feasibly be implemented on a broader scale and shared with school officials quickly and easily as other schools in Bolivia continue to collect data. The impact over the long term can help with an immediate overview of possible student need.

Impact on SJU Students

The following are student testimonials representing qualitative examples of impact.

> From this experience, we have not only learned to improve upon our skills, but how we may also make use of them to help others. So many groups took a unique approach to dissecting this issue and with such a wide array of perspectives, I really feel as though we have taken the first step to alleviating this problem. It is remarkable how by using just a few tools: data, software, and our minds, we were able to help in making a difference in someone's life. (Student A from Spring 2017, after Phase 1 [data dive])

> When combined with insights based on relationships, insights from this data could be used to alleviate some basic needs of children and families in the area. I believe that education is one pathway to break the cycle of injustice or poverty. By identifying students to target and programs to implement, students may be able to focus more on their education and ultimately have more of a fighting chance in this world. (Student B from Fall 2017, after Phase 1 [data dive])
>
> The final presentation was overall very rewarding because it did show that Miguel had the potential to use our findings to better the lives of some of the children in Bolivia. This was definitely the most fascinating and most rewarding project that I have participated in during my time at St. Joe's. (Student C from Fall 2016, after Phase 2 [Final Presentation])

CONCLUSION

The FyA's mission of serving the poorest of the poor through education is one of the most eloquent examples of social sustainability as idealized by Francis and outlined in the United Nations' Sustainable Development Goals. This article describes a model that shows how this social aspect of sustainability can be embedded in the learning and practice of undergraduate business students, a model that integrates the application of statistical analysis techniques in support of Jesuit mission objectives while engaging undergraduate business students in imparting values of community stewardship, global citizenship, and social responsibility.

Having provided analytical support over the past three academic years to Fe y Alegría in Bolivia, the efforts of a data mining class' cohorts at the Haub School of Saint Joseph's University have led to a process of identifying, from survey data alone, which early high school (*la secundaria*) students are most impoverished so that support efforts can be targeted sooner rather than later to those most in need. There were changes in approach, analysis, and instruction as more data and new questions arose, and the initiatives led to an innovative, web-based Tableau dashboard visualization tool which allows for very efficient examination of survey answers and updates as new data is gathered. Indeed, such outcomes contribute to FyAB's mission while engaging students and faculty at the Haub School of SJU in this important social-sustainability work.

REFERENCES

Allen, I. E., & Seaman, C. A. 2007. Likert scales and data analyses. ***Quality Progress,*** 40(7): 64– 65.

Allison, P. 2012. ***Logistic regression for rare events.*** Available at https://statisticalhorizons.com/logistic-regression-for-rare-events (accessed May 19, 2018).

Baumer, B. 2015. ***A data science course for undergraduates: Thinking with data.*** ArXiv.org.

de Figueiredo, J. N., & Marca Barrientos, M. 2012. A decision support methodology for increasing school efficiency in Bolivia's low-income communities. ***International Transactions in Operational Research,*** 19(1–2): 99–121.

de Mesa, J., Gisbert, T., & Gisbert, C. D. M. 2008. ***Historia de Bolivia*** (7th ed.). La Paz, Bolivia: Editorial Gisbert.

Deaton, A. 2003. Household surveys, consumption, and the measurement of poverty. ***Economic Systems Research,*** 15(2): 135–159.

Francis. 2015. ***Laudato si': On care for our common home.*** Available at http://w2.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html.

Garwood, K. C., & Dhobale, A. 2018. A comparison of cluster algorithms as applied to unsupervised surveys. ***International Journal of Business Intelligence and Data Mining,*** in press.

Henderson, A. R. 2005. The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. ***Clinica Chimica Acta***, 359(1–2): 1–26. DOI: 10.1016/j.cccn.2005.04.002.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., & Buono, P. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. ***Information Visualization,*** 10(4): 271–288.

Moore, T. L., & Roberts, R. A. 1989. Statistics at liberal arts colleges. ***The American Statistician,*** 43(2): 80–85.

Norton, E. C., & Dowd, B. E. 2018. Log odds and the interpretation of logit models. ***Health Services Research,*** 53(2): 859–878.

Shneiderman, B., & Wattenberg, M. 2001. Ordered treemap layouts. In ***Proceedings of the IEEE symposium on information visualization 2001 (INFOVIS' 01):*** 73. Washington, DC: IEEE Computer Society.

United Nations. 1995. ***World summit for social development: The Copenhagen declaration and programme of action.*** New York: United Nations.

World Bank. 2018. ***The World Bank in Bolivia.*** Washington, DC. Available at http://www.worldbank.org/en/country/bolivia/overview (accessed May 14, 2018).

We would like to acknowledge with sincere gratitude the help and guidance provided by the Managing Editor who took time to engage with our paper and authors.

Kathleen Campbell Garwood has been an Assistant Professor in the Department of Decision & System Sciences at Saint Joseph's University since 2004. Her teaching focuses primarily on data mining and modeling with the goal of introducing students to real data and analytical techniques. Her research interests include data visualization (best practices and techniques), rank order comparisons (with a focus on sustainability rankings), modeling applications with real world settings (including identifying the most impoverished for Fe y Alegria-Bolivia), and data collection and analysis related to science, technology, engineering, and math (STEM).

João Neiva de Figueiredo is Professor of Management at Saint Joseph's University's Haub School of Business. His research and teaching focus on international business and organizational sustainability, and he has published extensively in these areas. He co-authored the second edition of *Organizational Change and Strategy: An Interlevel Dynamics Approach* (Abingdon/Routledge, 2016) and co-edited the books *Green Products: Perspectives on Innovation and Adoption* (Taylor & Francis, 2012) and *Green Power: Perspectives on Sustainable Electricity Generation* (CRC Press/Taylor & Francis, 2014). With over 20 years' experience in international business prior to joining academia, Dr. Neiva holds a Ph.D. in Business Economics from Harvard University.

Hayley F. Miles is a graduate student in the Master of Science in Business Intelligence and Analytics program at SJU. Her interests include data analytics and visualization. Hayley has begun her research career with this project and is excited to continue communicating the value of data through interactive dashboard displays.

Miguel Angel Marca Barrientos is Fe y Alegria's National Coordinator for Education in Bolivia where he has been responsible for quality improvement processes in the network's four hundred schools since 2005. In his thirty-year career in education in the country, he has been a school teacher and principal from 1989 to 2001, a homeless educator in La Paz from 1989 to 1995, and head of research in innovative education at the Bolivian Center for Education Research and Implementation (*Centro Boliviano de Investigación y Acción Educativas*) from 1995 to 2005. He is currently a member of the Board of the Bolivian Campaign for Education Rights (*Campaña Boliviana por el Derecho a la Educación*) and has published research on education as well as educational materials in Bolivia and abroad. His B.A. in Pedagogy and Certificate in Educational Management are from the Universidad Mayor de San Andres, La Paz, Bolivia.

APPENDIX A1: 32-question survey given by FyAB, filled out by students in the third and fourth years of *la secundaria,* and analyzed by SJU students throughout this ongoing project.

Context Analysis. Carefully read the following instructions before answering this questionnaire. *IMPORTANT: DO NOT WRITE ANYTHING HERE, USE THE ANSWER SHEET.*

- Now we will ask you to answer some questions regarding your home and your life.
- This is not a test/exam: if you do not understand a question or if you are not sure about what to answer, you can ask your professor/teacher.
- Please be as honest as possible when answering this questionnaire. None of your answers will be made public.
- Use the answer sheet to indicate the number that corresponds to the answer you selected.

1. How present and involved is your father in daily life in your home?
    1) Deceased
    2) Not present in the house
    3) Little or nothing present / involved
    4) Present / involved sometimes
    5) Always present and very involved

2. What is the highest level of education achieved by your father (or guardian if you do not live with your father)?
    1) Does not know how to write or read
    2) Finished primary school then stopped studying
    3) Some high school education but no diploma
    4) Finished high school education and has a diploma
    5) At least some college education (not necessarily with a degree)

3. Which one of the following groups best fits your father's (or guardian's if you do not live with your father) current work situation?
    1) Does not work (either physically/mentally unable or just unemployed)
    2) Holds an occasional/temporary job in agriculture, construction, cleaning, security, etc., or he performs self-sustaining agriculture; works as a domestic employee; or does partial jobs that may come up

   3) Works a salary job in a store, office, workshop, the army/police, agriculture, or as an independent driver
   4) Owns a store or a workshop with less than ten employees; is an office or a store manager; works independently performing specific jobs; is a teacher; holds a high position in the army/police; or is a farmer who owns the land where he works and sells what he plants
   5) Is the owner or general manager of a business with more than ten employees; is a government official or an army general; or holds a profession as an architect, doctor, lawyer, etc.

4. How present and involved is your mother in daily life in your home?
   1) Deceased
   2) Not present in the house
   3) Little or nothing present / involved
   4) Present / involved sometimes
   5) Always present and very involved

5. What is the highest level of education achieved by your mother (or guardian if you do not live with your mother)?
   1) Does not know how to write or read
   2) Finished primary school then stopped studying
   3) Some high school education but no diploma
   4) Finished high school education and has a diploma
   5) At least some college education (not necessarily with a degree)

6. Which one of the following groups best fits your mother's (or guardian's if you do not live with your mother) current work situation?
   1) Is physically/mentally disabled to work or only performs domestic labor
   2) Works as a domestic employee, or holds a temporary position with no salary
   3) Works independently and/or has a stable job in a store, workshop, or business/office
   4) Works as a teacher or nurse, or is the owner or general manager of a small business or workshop with less than ten employees
   5) Is the owner or general manager of a business with more than ten employees; is a government official or an army general; or holds a profession as an architect, doctor, lawyer, etc.

7. How many rooms does your household/residence have (including living room, dining room, bathroom, and kitchen)? Do not include back or front yards, or any other exterior properties.
    1) One common room for everything
    2) Two
    3) Three
    4) Four
    5) Five or more

8. How many people live in your household (including you)?
    1) Three or less
    2) Four or five
    3) Six or seven
    4) Eight or nine
    5) Ten or more

9. Is there running water in your house (do you have access to potable water)?
    1) Yes
    2) No

10. Which family members do you live with?
    1) I live with my mom and dad (also including siblings, grandparents, and other relatives).
    2) I live with only one of my parents (also including siblings, grandparents, and other relatives).
    3) I do not live with my parents. Instead, I live with my grandparents or my aunt/uncle (also including siblings and other relatives).
    4) I live with my siblings only or with other relatives who are not my parents, grandparents, aunt, or uncle.
    5) I live with a family different than mine.

11. How many people sleep in your room?
    1) No one else (I sleep by myself)
    2) One or two more people (two to three people in total)
    3) Three or four more people (four to five people in total)
    4) Five or six more people (six to seven people total)
    5) More than seven people total

12. How many meals per day do you eat?
    1) One
    2) Two
    3) Three
    4) Four

13. Do you have breakfast or lunch before going to school?
    1) Yes
    2) No

14. Is there a shower in your house?
    1) Yes
    2) No

15. Does your house have electricity?
    1) Yes
    2) No

16. Is there a table/desk in your house where you can sit down and study?
    1) Yes
    2) No

17. Is the entire floor of your house comprised of concrete?
    1) Yes
    2) No

18. Does your house have a working [telephone] landline?
    1) Yes
    2) No

19. Are there any books in your house?
    1) There are no books
    2) Only school books/textbooks
    3) Less than twenty books, without taking the school books/textbooks into account
    4) More than twenty, without taking the school books/textbooks into account

20. Are there any purchased toys or games in your house?
    1) None
    2) Less than five
    3) More than five purchased toys or games

21. How many people in your family constantly contribute with money to the house?
    1) None
    2) One or two
    3) Three or four
    4) More than four

22. What do you usually do on weekends (Saturdays and Sundays) or during holidays?
    1) I have a job outside of the house.
    2) I help my parents with their jobs outside of the house.
    3) I work at home (for example, doing laundry, cleaning the house, taking care of my siblings …).
    4) I play with my friends or siblings.
    5) I go on roadtrips with my parents (to the city, visit family or friends, etc.).

23. Do you work and get paid for it?
    1) I do not work.
    2) I help my family when I do not have to go to school, but I do not get paid.
    3) I work sometimes when I do not have to go to school in order to make some money.
    4) I work and get paid, but that does not stop me from going to school.
    5) Sometimes I work to earn some money and that prevents me from going to school.

24. Are there any violent gangs in your neighborhood or school?
    1) There aren't any gangs neither at school nor in my neighborhood.
    2) There are gangs in my neighborhood, but they do not come close to school.
    3) When I come to school or go home I see gangs.
    4) There are gangs at my school.
    5) Some of my friends are part of a gang.

25. Do you currently have any disease that requires permanent/special treatment (e.g., asthma, allergies, gastrointestinal problems)?
    1) Yes
    2) No

26. Are you currently affiliated to any health insurance/service?
    1) Yes
    2) No

27. How long does it take you to get to school from home?
    1) Less than 15 minutes
    2) Between 15 to 30 minutes
    3) Between 30 minutes and one hour
    4) Between one and two hours
    5) More than two hours

28. The access and circulation roads (streets, avenues, roadways, etc.) where you live are:
    1) Not paved and unattended; they flood easily.
    2) Not paved but taken care of; they rarely flood.
    3) Paved but unattended; there are potholes on the roads.
    4) Paved and are in good condition.

29. Are the walls in your home made of any of these materials: rough wood, plank, bamboo, any vegetable/plant material, zinc, cloth, or cardboard?
    1) Yes
    2) No

30. Is there access to the public sewage system in the house where you live in?
    1) Yes
    2) No

31. Do you have access to a mobile phone?
    1) No, I do not have access to a telephone.
    2) Yes, I use it one to ten times a week.
    3) Yes, I use it eleven to twenty times a week.
    4) Yes, I use it more than twenty times a week.

32. Are you involved in any sports, cultural, musical, or other activity outside of class?
    1) No activity.
    2) One activity
    3) Two activities
    4) Three activities
    5) Four activities or more

APPENDIX A2. Individual statistical techniques and tools applied throughout the process of data mining the surveys for FyAB.

The Content: Data Analysis Topics Covered

The following statistical and data analysis tools were applied by students through many iterations, with slight adaptations each semester based on current and prior class interactions, changing data sets over time, consultations with FyAB and changing goals of the project. The goal is to provide a succinct rationale and basic definitions without excessive statistical discussion.

Data cleaning. Incorrect survey answers such as blanks and mistakes were identified throughout each phase. Consistent and reasonable techniques ensure that all iterations considered the data through a similar lens, leading to the development of identical data-cleaning rules. Typically respondents with more than two blanks or errors were removed from the data. With two or fewer errors substitutions were made.

Data Visualization. Histograms, boxplots, bar charts, tree diagrams, cluster plots, and pivot tables were used in each analysis iteration: every step provided graphics and visuals that often helped guide both the students doing the analysis and the team at FyAB who relied on the results.

Distribution analysis. In every data iteration, the distributions of survey responses were considered both on a per-school basis and on an aggregate basis to identify underlying data patterns. Histograms provided quick insights into how students responded to each question and helped quickly examine differences among schools regarding pertinent questions.

Principal Component Analysis (PCA)/ Factor Analysis (FA). PCA allows the user to reduce the number of independent variables in a model while capturing as much of the total variability among all the independent variables as possible. The independent variables are regrouped into *factors*, which combine associated variables into orthogonal (uncorrelated) groups using weights identified in the factor analysis. Overall, PCA/FA was used to identify which variables might be eliminated from consideration without substantial loss of explanatory variation.

Analysis of Variance. One of the main goals was often to find schools that were similar and schools that were different from one another focusing on survey questions that might be indicative of poverty. In each data iteration, analyses of variance were run in order to assess the overall average of any specific question answered as well as how each school compared to that average revealing differences among the schools being analyzed and a lack of difference within the schools.

Cluster Analysis. This technique considers regrouping the students (row data) into like groups that have minimal variation within each group while maximizing the variability between any two different groups. The output provides an "average" student answer to each question by cluster. These averages provide insights into students who have fewer meals, minimal access to electricity or water, or parents who have fewer job prospects (i.e., who work fewer hours each week). Once created, these clusters can be used to rank the students in need using the UN definition of poverty[3] or other reasonable indicators of poverty.

Multiple Linear Regression. The data set of survey answers is formed by purely independent variables with no clear dependent variable that could be used to create a predictive regression model. Several possibilities have been considered with the long term goal of establishing a dependent variable representing a poverty score. Such a score could then be used to develop a weighted model to help categorize future survey takers.

Bootstrapping. This advanced simulation technique is an iterative process of sampling from within a larger data set in an attempt to identify key data patterns. Simulations were applied in an effort to create a meaningful dependent variable by which a poverty value might be created for each current student within the collected data.

Logistic Regression. From each cluster iteration, groups of students may be considered to be "more impoverished" than other groups when considering the average answers in each cluster. A binary dependent variable where the more impoverished students receive a value of one while the remaining students receive a value of zero was considered. In this way a logistic regression model can be made to evaluate which variables are most useful in identifying poverty, possibly creating a predictive tool that could be implemented with future students.

---

[3] In 1995, the United Nations defined absolute poverty as "a condition characterized by severe deprivation of basic human needs, including food, safe drinking water, sanitation facilities, health, shelter, education and information. It depends not only on income but also on access to services" (see https://www.un.org/development/desa/dspd/world-summit-for-social-development-1995/wssd-1995-agreements/pawssd-chapter-2.html).