Teaching an Introductory Data Analytics Course Using Microsoft Access® and Excel®

Faisal Aqlan
Industrial Engineering
Penn State Behrend
Erie, PA, USA
aqlan@psu.edu

Joshua C. Nwokeji
Computer & Information Science
Gannon University
Erie, PA, USA
Nwokeji001@gannon.edu

Abdulrahman Shamsan
Industrial & Systems Engineering
Binghamton University
Binghamton, NY, USA
ashamsa1@binghamton.edu

***Abstract*— Data analytics has been recently adopted by many researchers and professionals working with data in both academic and industry. With the increase in demand for data analysts, there has been a parallel growth in data analytics training programs within companies and educational institutions. In this paper, we introduce the concepts of data analytics and present practical examples using Microsoft Access and Excel. The four types of data analytics (i.e., descriptive, diagnostic, predictive, and prescriptive) are discussed and practical examples are provided. For descriptive analytics, we discuss the data properties and models and present examples of database design and implementation in Microsoft Access. The example for diagnostic analytics involves an ergonomic assessment application in Microsoft Excel to identify the sources of ergonomic risks in work environments. Predictive analytics examples include regression and clustering models implementation in Microsoft Excel. Finally, the prescriptive analytics example involves optimizing the snow removal process in a local city by developing an optimization model and its implementation in Excel. These examples will help students understand data analytics and be able to implement the different data analytics models in Microsoft Access and Excel.**

Keywords—data analytics, data modeling, teaching and learning, MS Access, MS Excel

I. INTRODUCTION

Data analytics has emerged as an important subject and attracted a growing interest of professionals from both industry and academia. In general terms, data analytics refers to the science and statistical practice of collecting, classifying and analyzing data [1]. Data analytics principles can be used to analyze, visualize and report all categories of data, including big, small, qualitative and quantitative data. The essence of data analytics is primarily to derive information, intelligence and knowledge from raw data, and use these to support decision making in organizations [2]. In addition, data analytics provides insights into the past, present and future trends or events in a particular area of interest [1-3]. Currently, the amount of data generated by industries has grown exponentially in volume, variety, velocity and complexity [4-5]. This growth has given rise to a new branch or specialty of data analytics, called Big Data Analytics or simply Big Data [6]. In this context, big data refers to voluminous, heterogeneous and complex data that cannot be analyzed with traditional data analytics tools and techniques and usually requires advanced techniques [5-6]. Data analytics has its root in decision support systems and decision sciences, and closely relates to various other subjects of study such as business analytics, data science, and business intelligence [6-7].

Data analytics plays vital roles across industries and has been applied to address pressing social issues such as elections and pandemic management. A more recent application of data analytics in social issues is seen during current break-out of the COVID-19 pandemic. Data analytics is contributing immensely to the management and fight against the spread Coronavirus (i.e., the virus that primarily caused COVID-19 Disease). For instance, data analytics principles are being used to provide real time information on the spread of the disease, track recovery and mortality rate, and perform contact tracings. Predictive analytics, which is a type of data analytics, is currently helping health organizations and governments to decide and provide adequate resources needed to defeat this deadly virus. Data analytics principles are also applicable in other disciplines including healthcare (healthcare-analytics), banking (financial analytics), education (learning analytics), sports (sport-analytics), etc.

Employers place a high premium on graduates or professionals who possess data analytics competences or skills [8]. A quick search in leading job portals such as *www.indeed.com* and *www.monster.com* show tens of thousands of vacancies and job opportunities relating to data analytics. For instance, in April 10, 2020, we carried out a quick job search in *www.monster.com* using "data analytics" as a keyword. This search returned over 112,000 data analytics related jobs in the United States alone. Using *www.indeed.com*, we also performed a similar search with the same keyword, this returned over 78, 000 data analytics related jobs. These job openings usually require a broad range of data analytics competencies, ranging from basic skills such as data collections, visualizations and reporting to advanced skills such as big data, predictive analytics and algorithm design. But existing studies have reported that there are insufficient professionals to fill various data analytics related job openings [8-10]. More so, industry demand for data analytics by far exceeds the supply.

In recent times, there have been some concerted efforts by professionals in academia and industry to fill the current open job vacancies. Examples of current initiatives include various efforts to revise or develop new curricula to deliver data analytics skills, as seen in various studies such as [3-4,11-12]. Some other researchers, e.g. [13-15], have developed new academic programs and/or courses to equip students with data analytics skills. Furthermore, attempts have been made by professional organizations to develop and provide open source academic resources to facilitate training of data analytics professionals. For instance, Teradata University Network provides re-usable learning materials and tools to support data analytics education. A similar effort has also been made by SAP Alliance. Despite these efforts, there exist setbacks in teaching and learning of data analytics. One of the major setbacks reported in literature is the difficulty in obtaining the resources, especially software platforms, required for teaching and learning data analytics courses [9].

Available software platforms such as SAS Analytics and Tableau can be difficult and costly to obtain, maintain and teach. Even the available free versions are usually lite editions that usually do not provide the full functionality and features required to cover key analytics topics and deliver the required skill sets. Moreover, software platforms like SAS and Tableau normally have very high learning curves. While bigger and research-intensive universities may be able to afford these software, smaller universities usually lack the resources and hardly can afford them. Hence, it is usually difficult for such smaller universities to teach data analytics to their students. In order to address this challenge, educators, especially those in smaller institutions should be aware of alternative software packages (e.g., Microsoft Excel) that are affordable and easy to use, but provides various features that can be used to teach and learn data analytics. In this way, teaching and learning of data analytics can be inclusive and reach a broad diversity of students; thereby producing sufficient professionals to fill open data analytics positions.

Accordingly, this paper provides a critical perspective on advanced features of Microsoft (MS) Excel and Access and how these can be used to teach core analytics skills to students. We chose MS Excel and Access because they are affordable, and have a considerable very low learning curve, which make them very easy to teach and learn. More so, MS Excel and Access are used by the vast majority of smaller colleges and universities. Although, this research does not aim to develop a new software package for teaching analytics, we contribute by providing an exposition and stepwise guideline for using existing software packages i.e., MS Excel and Access to deliver core analytics competences to students. While we do not undermine the importance and usefulness of other software packages, our practical teaching experiences show that MS Excel and Access have advance features that covers both basic and advanced analytics skills required in the job market. We also found that MS Access and Excel better serve introductory level analytics skills relative to other software packages.

To facilitate readability, we organize the rest of this paper as follows. In Section 2, we provide overview of data analytics and review the popular software packages for teaching and learning analytics. This is followed by Section 3 where we discuss key analytics concepts and show how key features in MS Excel and Access can be used to teach these to students.

II. Relevant Literature

Teaching and learning data analytics have been the focus of many researches in recent times. This is because of the important role data analytics play in both academia and industry. Despite the advances in data analytics research and education, there is currently no generally agreed definition of data analytics. However, there is a consensus among scholars, that analytics refers to the science and statistical practice of collecting, classifying and analyzing all categories of data, including big, small, qualitative and quantitative data [2]. The essence of analytics is primarily to derive intelligence from data, and use this to facilitate business operations and decision making [1-3].

Data analytics helps organizations to gain insights into the past, current and future events; as well as provide avenues to answer questions critical to organizational transformations. Examples of such questions are: what happened or what is happening in an organization; why did it happen or why is it happening; what is likely to happen in the future; and what do I need to do? Respectively, the answers to these questions inform the various types of analytics, taught in educational institutions, namely descriptive, diagnostic, predictive and prescriptive analytics [3,13,16].

In recent times, the amount of data generated from organizations has grown exponentially in volume and these data are largely heterogeneous [17-19]. Such data are too complex to be processed by traditional statistical tools and techniques [4]. Consequently, a new area of analytics, called big data and analytics (BD&A), has recently emerged [6]. Big data are generated at exorbitant amount that tends towards exabytes (volume); and produced at excessive rates (velocity) from heterogenous sources (variety). Other characteristics of big data include the ability to yield actionable intelligence (value) and the quality of being truthful, honest and accurate (veracity) [5]. Big Data and analytics are interdisciplinary in nature and thus, applicable in a variety of disciplines including healthcare banking, education, sports, and others [12].

A. Software and Resources for Teaching Data Analytics

Educators and practitioners are currently leading laudable initiatives to develop and provide software packages and resources that can be used to train data analytics professionals, thereby contributing to efforts to close the current skill gap in data analytics. Results from these initiatives have produced useful learning consortiums such as Teradata University Network, SAS academic Programs; SAP University Alliances, IBM Big Data University, and Tableau Academic Programs. These consortiums provide collaborative platforms such as free

data sets, learning materials, tools and computing environment to facilitate teaching and learning data analytics. Furthermore, universities and colleges have also been in the forefront of various initiatives aimed to address the shortage of skills in big data and data analytics. Furthermore, many commercial or open source software packages have been developed by leading software vendors such as SAS, Microsoft and Tableau. However, most of these software packages are costly to buy and maintain, making it difficult for smaller institutions to use them as teaching tools. Although, there are free and open source data analytics platforms that can be used to teach data analytics, these often have relatively high learning curves and installation challenges that can dissuade their usability. In Table 1, we compared eight popular software packages that can be used to teach data analytics. These tools were selected because the authors have used them to teach data analytics courses and are commonly used. We used four factors as the bases for this comparison. These include the purchasing and maintenance cost, learning curve, ease of use and installation process.

Table 1. Comparison of Popular Software Used for Teaching DA

Factors / DA Software	Cost	Learning Curve	Ease of Use	Installation Process
MS Excel	Low	Low	High	Easy
SAS	High	Very High	Low	Challenging
Tableau	High	Very High	Low	Moderate
JASP	Free	High	Low	Easy
Python	Free	Medium	Medium	Moderate
R	Free	Medium	Medium	Moderate
JAMOVI	Free	High	Low	Easy
NLTK	Free	High	Low	Moderate

The comparison shows that commercial data analytics software packages have a relatively high purchasing and maintenance cost, except MS Excel, which has a lower cost when compared with others. MS Excel relatively has a very low learning curve. Recall that most undergraduate students would have taken one or more MS related courses in their high school or freshman years. Thus, they would have some MS skills by the time they are ready to take data analytics courses. MS Excel is also relatively easy to use and install. Despite myriad benefits MS Excel offers, its capacity is currently underutilized as a data analytics teaching software. Hence, this research aims to demonstrate the various capacities of MS Excel and they can be useful in teaching data analytics.

B. Relevant Studies

Several studies in the literature have discussed the learning and teaching of data analytics. For example, a study proposed a technique for teaching and learning data analytics using infographics tools [20]. The study found that the use of infographics to teach data analytics helped students to acquire data analytics skills. Nonetheless, the study focused on students between grades 9 to 12 in a high school. A similar study was conducted to compare the learning propensities of grades 10 to 12 high school students to data analytics education [21].

Table 2 provides a summary of current researches in teaching and learning of data analytics. As shown in the table, we summarize the available studies into four focus areas. These include studies that focus on evaluating software packages and tools used for teaching and learning analytics. For instance, the authors in [10] performed a comparative analysis of various platforms used for teaching data analytics. We also found studies [13]–[15] that focus on the design and development of data analytics academic courses and programs. These studies contribute a variety of courses that can be useful in delivering key data analytics skills to students at both graduate and undergraduate levels. Studies such as [22]–[23] analyze various skills required by employers in the data analytics profession. These analyses are based on job descriptions found in leading job portals. The last category of studies we found focus on the design and development curriculum for graduate and undergraduate data analytics programs, for instance see [3-4]. Indeed, these studies made useful contributions to teaching and learning data analytics skills.

Table 2. Related Studies in DA Teaching and Learning

Study	Focus	Summary
[10]	Data analytics Tool Evaluation	Study evaluates a number of tools that are currently used for data analytics and business intelligence education
[13-15]	Design and development of data analytics courses	These studies contribute to the design and development of academic programs and courses to support the propagation of data analytics skills
[22-23]	Data analytics Competences and Skills	Studies investigate competences required in data analytics and proposes a theoretical framework for training data analytics professionals
[3-4]	Design and development of data analytics curricula	These studies present sample curricular for teaching analytics related courses at both graduate and undergraduate levels. They also discuss approaches that can be used to develop curricula

In this paper, we build on existing research and contribute by providing an insight and exposition into how affordable and easy to learn and use software packages such as MS Excel and Access for data analytics. We anticipate that the insight we provide in this paper will be useful to instructors and academic institutions that are new or not currently teaching data analytics.

III. Pedagogical Context

This paper supports the teaching and learning of data analytics by providing insights into concepts, tools and techniques that are useful in teaching data analytics. These can be adapted and used by instructors to teach introductory data analytics courses to undergraduate students at the junior level. The target audience for this course are students who are new entrants to data analytics and do not have any previous exposure or knowledge in data analytics related courses such as Information Systems, Computer Science, Mathematics, and Statistics. Instructors can choose to structure this course into a 2-hour lecture and 2-hour lab each week.

Figure 1 shows examples of course learning objectives that can be covered using the tools and techniques discussed in this paper. During the lectures, core concepts and principles of data analytics can be covered, including the theories behind the four data analytics categories: descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics. Essentially covering the first three learning objectives i.e., a to c. The labs cover the application of the four data analytics categories and utilizes MS Access and Excel. The students can be given real-world datasets and asked to apply the different analytics techniques they learned during the lectures to extract useful information from these datasets. The course assessment includes exams, lab reports, quizzes, homework, and project.

Upon satisfactory completion of this course, students should be able to:

a) Understand data types and models and design simple relational databases

b) Understand the meaning and basic concepts of analytics including descriptive and predictive analytics, and big data

c) Understand the importance of segmentation and classification and the impact of big data on decision making

d) Prepare and present real-life case studies on database systems and data analytics

e) Use Microsoft-based tool such as Excel and Access, and other software tools for data modeling and analysis

Figure 1. Sample Learning Objectives

IV. Data and Data Analytics

Data is a set of values for qualitative or quantitative variables. Throughout this paper, we will use the term "data" to represent the raw data or the preprocessed data. This includes any facts, numbers, or text that can be processed using the analytics techniques. The raw data is processed to extract useful information. The information is then converted to knowledge that will be used for making decisions. Figure 2 shows the relationship between data, information, knowledge, and decisions.

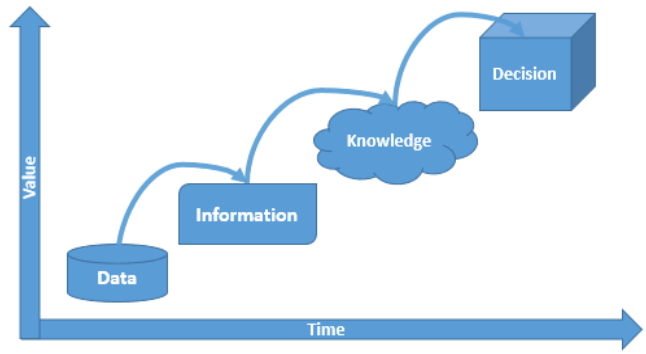

Figure 2. Relationship between data, information, knowledge, and decisions

There are several **classifications of data**. We present five classifications in Table 3. For example, in terms of accessibility and security, data can be classified into: public, sensitive, confidential, and regulatory.

Table 3. Classifications of data

Criteria	Classification
Accessibility	Public, sensitive, confidential, regulatory
Structure	Structured, unstructured
Measurement scale	Qualitative: Nominal, ordinal Quantitative: interval, ratio
Domain	String, number, date, time
Number of values	Discrete, continuous

Data analytics differs from data since in term of focus. Data analytics focuses on using machine learning and visualization to derive insights from the data whereas data science encompasses data analysis, computer science, and business domain experience to solve business problems. Figure 3 shows the difference between data analytics and data science. It also shows the evolution of the fields from artificial intelligence to machine learning and then deep learning.

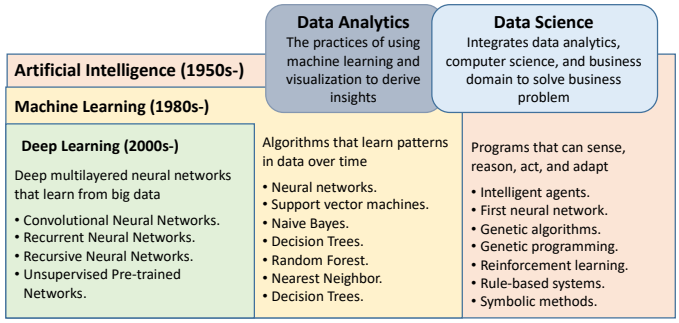

Figure 3. Data analytics vs. data science

Data analytics is classified into descriptive, diagnostic, predictive and prescriptive. This classification is based on the phase of workflow and the kind of analysis required. Figure 4 shows the four types of data analytics. Before performing the analytics, data should be collected and cleaned.

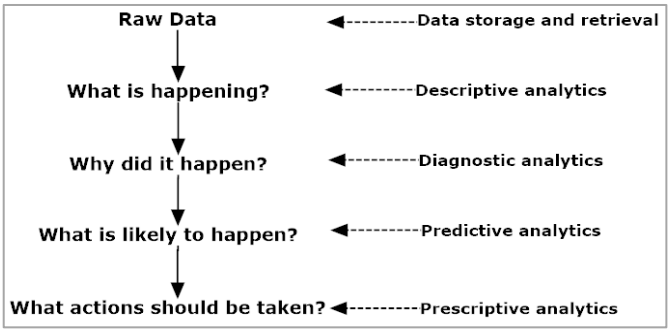

Figure 4. The four types of data analytics

Descriptive analytics focuses on using descriptive analysis or descriptive statistics to summarize the data. This includes identifying patterns in the data or modeling the data and relationships in the form of a database. In this paper, we discuss the use of MS Access to perform descriptive analytics. The process requires four main steps: (1) conceptual modeling, (2) data collection, (3) database implementation, and (4) reporting

and analysis. For step 1, we usually used Entity Relationship Diagram (ERD). An example of an ERD for engineering lab database is shown in Figure 5. The database consists of six main tables: safety table, ergonomic table, lab information table, lab equipment table, lab units table, and research activity table. The safety table includes all the safety related data such as safety issues and resolutions. Ergonomic table includes ergonomic related data and information such as ergonomic risks, assessment date, and ergonomic status (e.g., Green, Yellow, Red). The lab equipment table stores the equipment data including maintenance and repair data and ownership and location information. The other tables include data about labs, lab users, and research activities.

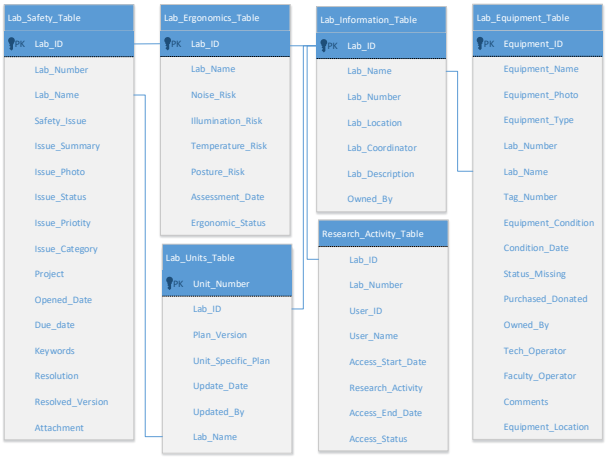

Figure 5. A sample ERD for a database

In step 2, data is collected from different resources to build the database. The ERD serves as a guide for the data collection. The list of engineering labs is shown in Table 4. For each lab, the collected data includes equipment, research activities, lab users, lab coordinators, safety information, etc.

Table 4. List of engineering labs

Lab Name	Lab #	Lab Name	Lab #
Plastics Labs		**Electrical Labs**	
Plastics Processing Lab	L - 127	Embedded Systems & Microprocessor Lab	L - 140
Plastics Materials Lab	L - 129	Electrical Technician's Lab	L - 141
Plastics Characterization Lab	L - 130	Programmable Logic Controller Lab	L - 143
Plastics Secondary Processing Lab	L - 132	Motors Lab	L - 144
Mechanical Labs		Circuit Fabrication Lab	L - 145
Thermo/Fluids Lab	L - 112	Circuits & Devices Lab	L - 146
Wind Tunnel	L - 113	Software Engineering Lab	L - 147
Heat Treat Lab	L - 116	Signal Processing Lab	L - 148
Energy Center / Supermileage Lab	L - 117	Measurements & Instrumentation Lab	L - 151
Rapid Prototyping Lab	L - 119	Game Development Lab	
Materials Characterization Lab	L - 121	**General Engineering Labs**	
Metrology Lab	L - 122	Innovation Commons	L - 108
Manufacturing Lab	L - 123	Innovation Commons	L - 109
Advanced Manufacturing Lab	L - 124	Senior Design Lab	L - 111
Technicians' Machine Shop	L - 125	Engineering Research	
Materials Testing Lab - MTS Room	L - 118a	**Outreach Lab**	
Materials Testing Lab - Hardness Testing	L - 118b	K-12 Outreach	L - 209

Figure 6 shows the main window of the database in MS Access. The tables were linked using primary keys such as lab IDs, equipment IDs, etc. The button "Ergonomic Evaluation" is linked to the ergonomic assessment table that includes the ergonomic data for the all the labs. Similarly, the "Safety Information" button is linked to the safety assessment table that includes the safety data for all the labs. The other buttons shown in the main menu include information related to the labs and the research activities and equipment in the labs.

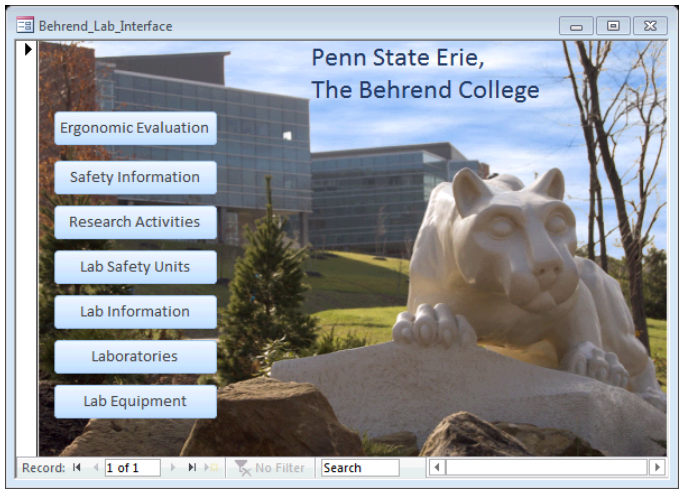

Figure 6. Main window of the engineering lab database

Figure 7 shows a sample lab equipment form, which is used to manipulate equipment data. The database is updated using the forms created for equipment, safety, and ergonomic tables.

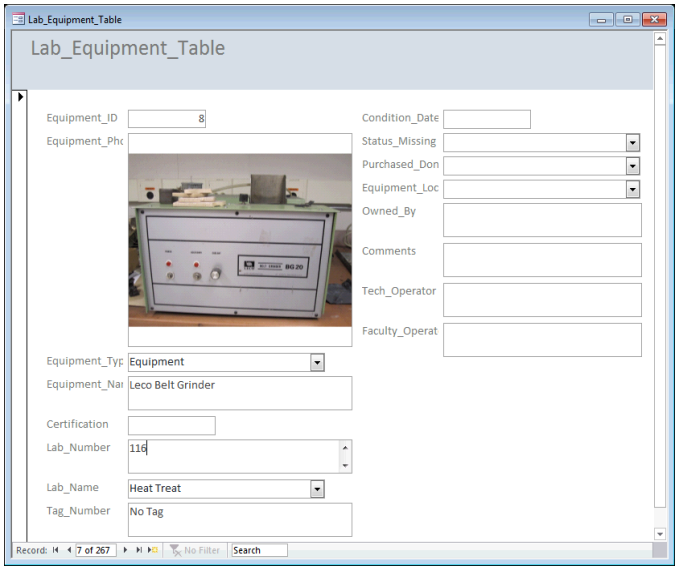

Figure 7. Sample form from the database

Figure 8 presents a sample descriptive analytics for the engineering lab safety assessment. The scores are calculated based on 58 questions used in the assessment and a score of 100 means that the lab does not have safety issues. Scores below 100 indicate the existence of some safety issues in the lab. The lower score means higher number of safety issues. The range of the scores is between 97 and 86 and the lowest score is for the electrical circuit labs (e.g., L-145). The lower score is because laboratory documentation and training were not complete when the assessment was performed. Immediate corrective actions are undertaken to resolve safety issues and a comprehensive assessment will to be performed annually.

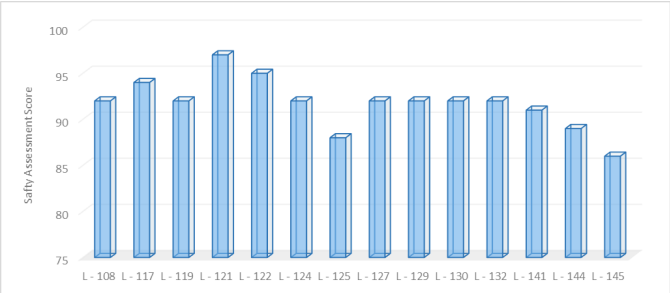

Figure 8. Descriptive analytics for lab safety

Diagnostic analytics focuses on examining the data to identify the root causes of an event. Some of the techniques used in diagnostic analytics are data discovery and correlations. In this paper, we present two examples of diagnostic analytics linked the database presented earlier. The first example represents an Ergonomic Heat Map (EHM) application that was developed in VBA (see Figure 9). EHM is a way to visualize ergonomic risks on the different parts of the human body. As indicated in Figure 9, the human body is divided into 24 parts in addition to the eyes, mouth, and ears. Noise is illustrated by the ears, illumination is illustrated by the eyes, and the posture and force risks are represented by the other body parts. The mouth color is an indicator of worker's feedback regarding the workplace ergonomic conditions and is also used to validate the ergonomic evaluation of the workplace. Four color codes are used: green for low or no risk, yellow for medium risk, brown for high risk, and red for very high risk. Each part of the body is evaluated to identify the ergonomic risk in the workplace.

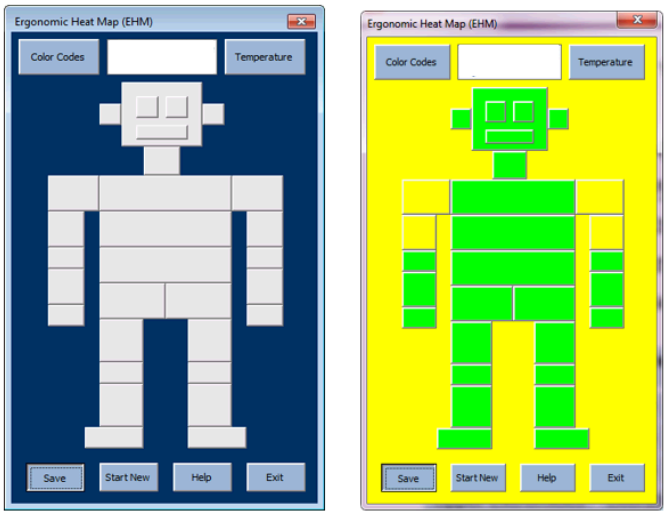

Figure 9. Ergonomic Heat Map (EHM)

EHM pulls the data for the ergonomic assessment table the database and displays the result in on the heat map to identify the ergonomic risks. For example, for noise level, EHM uses the following equation for evaluation:

$$D = 100 \sum_{i=1}^{n} \frac{C_i}{T_i}$$

where D is the noise dose, C_i is the total time of exposure at a specific noise level measured in hours, T_i is the reference duration recommended by The Occupational Safety and Health Administration (OSHA), in consideration to the sound level. Figure 10 shows a radar diagram for the scoring of noise levels. Th figure identifies which labs have high noise risks to the lab users. In the figure, all the labs have noise scores below the green line which means the noise levels are acceptable.

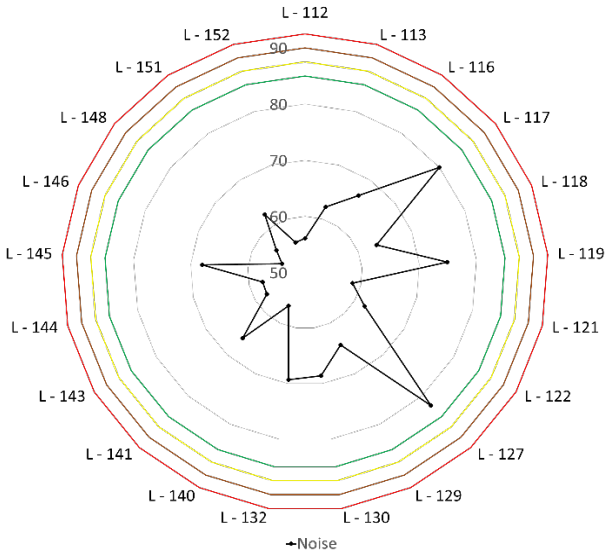

Figure 10. A radar diagram for lab safety scores of noise levels

Predictive analytics focuses on making predictions about future outcomes based on historical data using statistical modeling and machine learning techniques. In predictive analytics, we use historical data to models that can capture important trends in the data and predict future events. Common predictive analytics techniques include ordinary least square regression, logistic regression, clustering, decision trees, and neural networks. Below, we present an example of predictive analytics in Excel using logistic regression.

Figure 11 shows sample diabetes data for female patients who are at least 21 years old. The data was obtained from the data repository of the University of California Irvine. We want to build a logistic regression model for predicting the diabetes class (0 or 1) based on the other input variables. Below is the description of the variables:

A. Number of times pregnant
B. Plasma glucose concentration 2 hours in an oral glucose tolerance test
C. Diastolic blood pressure (mm Hg)
D. Triceps skin fold thickness (mm)
E. 2-Hour serum insulin (mu U/ml)
F. Body mass index (weight in kg/(height in m)^2)
G. Diabetes pedigree function
H. Age (years)
I. Class variable (0 or 1)

Logistic regression model, which is constructed by an iterative maximum likelihood procedure, was developed. First,

we define the logistic regression parameters (logit, odds, and likelihood) as shown in Figure 12. The "logit" in column "K" is calculated as:

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where P is the probability, β is the constant, and X's are the input variables (i.e., A-H). The β constants are shown in Figure 13. In the Excel model, Cell K2 is calculated as:

K2 = R2 + S2*A2 + T2*B2 + U2*C2 + V2*D2 + W2*E2 + X2*F2 + Y2*G2 + Z2*H2

The odds are calculated as:

$$Odds = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}$$

In the Excel model, Cell L2 is calculated as:

L2 = EXP(K2)

Columns M and N are used to calculate the corrected probability which is then used to calculate the log likelihood in column O. The Cells M2, N2, and O2, are caulked as:

M2 = =L2/(1+L2); N2 = =IF(I2=1, M2,1-M2); O2 =LN(N2)

In order to find the optimal constants for the logistic regression model, we optimize the sum of the log likelihoods. To do so, we will first calculate the sum of the log likelihood as shown Figure 13. The Excel function is Cell O 771 is:

O771 = SUM(O2:O769)

Then, we use Excel solver to minimize the sum of log likelihood by changing the values of the logistic regression constants as shown in Figure 15.

	K	L	M	N	O
1	**Logit**	**Odds**	**P(Y=1)**	**Corrected P(Y=1)**	**Log Likelihood**
2	0.952962	2.59338088	0.721710547	0.721710547	-0.326131125
3	-2.97304	0.05114778	0.048658984	0.951341016	-0.049882694
4	1.36558	3.91799381	0.796665055	0.796665055	-0.227320946
5	-3.1362	0.04344772	0.04163862	0.95836138	-0.042530348
6	2.221958	9.2253761	0.902204086	0.902204086	-0.102914525
7	-1.76112	0.17185155	0.146649591	0.853350409	-0.15858502
8	-2.63996	0.07136413	0.066610523	0.066610523	-2.708892706
9	0.595853	1.81457819	0.644706974	0.355293026	-1.034812405
10	0.891496	2.438776	0.709198854	0.709198854	-0.34361932
11	-3.27811	0.03769949	0.03632987	0.03632987	-3.315115011
12	-1.26767	0.28148705	0.219656567	0.780343433	-0.248021158
13	2.172396	8.77929665	0.897743157	0.897743157	-0.107871268
14	1.291669	3.63885349	0.784429493	0.215570507	-1.534467243
15	0.525365	1.69107544	0.62840135	0.62840135	-0.464576224
16	0.521405	1.68439192	0.627476155	0.627476155	-0.466049609
17	-0.40109	0.66959037	0.401050691	0.401050691	-0.913667448
18	-0.52713	0.59029412	0.3711855	0.3711855	-0.991053342
19	-1.40736	0.24478987	0.196651563	0.196651563	-1.626321831
20	-0.58581	0.55665754	0.357597945	0.642402055	-0.442540918
21	-1.18461	0.30586392	0.234223423	0.234223423	-1.45147982
22	-0.43707	0.64592577	0.392439185	0.607560815	-0.498303001
23	-0.76905	0.46345328	0.316684713	0.683315287	-0.380798904
24	2.758597	15.7776925	0.940397048	0.940397048	-0.061453101
25	-0.87223	0.41801963	0.294791145	0.294791145	-1.221488157
26	0.856681	2.35533064	0.701966779	0.701966779	-0.3538692
27	-0.23584	0.78990554	0.4413113	0.4413113	-0.818004756

Figure 12. Calculating the logistic regression parameters

	R	S	T	U	V	W	X	Y	Z
1	**B0**	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	**B8**
2	-8.4034	0.1232	0.0352	-0.0133	0.0006	-0.0012	0.0897	0.9453	0.0149

Figure 13. Logistic regression model constants

	K	L	M	N	O
768	-0.91991	0.39855607	0.284976825	0.284976825	-1.255347419
769	-2.55581	0.07762936	0.072037162	0.927962838	-0.074763592
770					
771				Sum	-361.7226933

Figure 14. Calculating the sum of log likelihood

	A	B	C	D	E	F	G	H	I
1	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	9	119	80	35	0	29	0.263	29	1
26	11	143	94	33	146	36.6	0.254	51	1
27	10	125	70	26	115	31.1	0.205	41	1
28	7	147	76	0	0	39.4	0.257	43	1

Figure 11. Sample data foe the logistic regression model

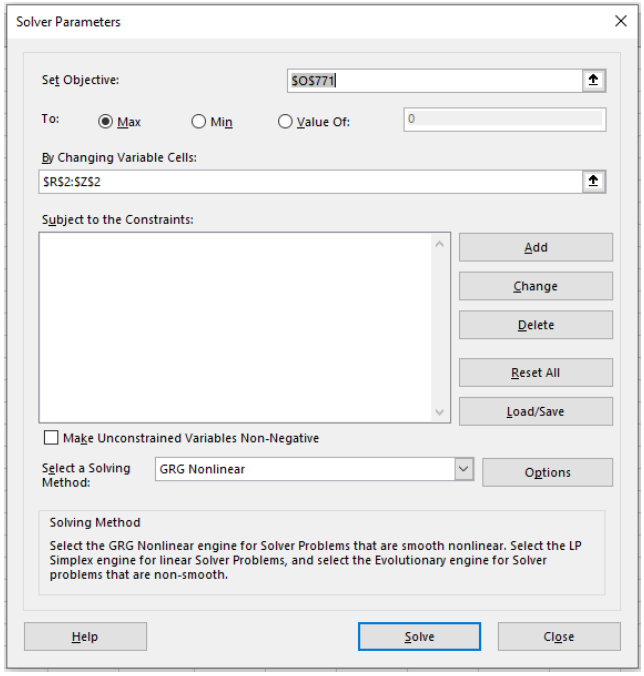

Figure 15. Minimizing the sum of log likelihood using Solver

Another example of predictive analytics using K-means clustering in MS Excel in presented below. K-means clustering algorithm is one of the simplest and most popular clustering techniques. The algorithm makes two assumptions: (1) each input data point can belong to a single cluster only, and (2) the number of clusters in predetermined by the user [21]. In k-means, k is the predefined number of clusters and "means" is the average location of all the members of a single cluster. An illustration of the k-means algorithm is shown in Figure 16. k: cluster index, c_k: k^{th} cluster set, μ_k: the mean of cluster c_k, K: total number of clusters, and x_i: data set. The objective of K-means is to minimize sum of squared error over all K clusters.

Pseudo-Code for K-means Algorithm

Determine the required number of clusters, K

Initialize cluster centroids $\mu_1, \mu_2, \ldots \mu_k \in \mathbb{R}^n$, randomly

While *termination criterion is not met* **do**

  for ($i = 1, i \le N, i = i + 1$) **do**

    assign set of points x_i to the nearest cluster;

    $y(i) = \underset{k\in\{1,2,\ldots,K\}}{\operatorname{argmin}} \|x_i - \mu_k\|^2$;

  end

  Recalculate the cluster centroids;

  for ($k = 1, k \le K, k = k + 1$) **do**

    assign x_i to cluster c_k based on nearest distance to μ_k ;

    $c_k = \{x_i \mid y(i) = k\}$;

    calculate new centroid μ_k ;

    $\mu_k = \frac{1}{|c_k|} \sum_{x_i \in c_k} x_i$;

  end

end

Figure 16. An illustration of the K-means algorithm

The model clusters universities based on acceptance rate and graduation rate. First, cluster distances are calculated as shown in Figure 17.

	A	B	C	D	E	F	G	H
1	Univ	Accept	GradRate	Distance to C 1	Distance to C2	Distance to C3	Min Distance	Class
2	Brown	22	94	0.390475837	52.32577564	20.62134003	0.390475837	Cluster 1
3	CalTech	25	81	12.95207696	45.12192238	18.53492486	12.95207696	Cluster 1
4	CMU	59	72	42.79695579	10.04973988	23.8332194	10.04973988	Cluster 2
5	Columbia	24	88	5.939488502	48.10392357	17.91323091	5.939488502	Cluster 1
6	Cornell	33	90	11.51943845	40.70614029	9.011697538	9.011697538	Cluster 3
7	Dartmouth	23	95	1.672946414	51.88436225	19.95555671	1.672946414	Cluster 1
8	Duke	30	95	8.057675326	45.79289182	13.51228837	8.057675326	Cluster 1
9	Georgetown	24	92	2.522156983	49.658709	18.22213851	2.522156983	Cluster 1
10	Harvard	14	97	8.744287466	60.83571786	29.13863214	8.744287466	Cluster 1
11	JohnsHopkins	44	87	22.91317069	29.68152453	2.640519455	2.640519455	Cluster 3
12	MIT	30	91	8.357175328	43.82908879	12.14164101	8.357175328	Cluster 1
13	Northwestern	39	89	17.55501541	34.98558993	2.929813503	2.929813503	Cluster 3
14	NotreDame	42	94	19.94139606	35.4681859	5.396182794	5.396182794	Cluster 3
15	PennState	54	80	34.71844218	17.49273425	14.84511862	14.84511862	Cluster 3
16	Princeton	14	95	8.180506537	60.00820452	28.6265373	8.180506537	Cluster 1
17	Purdue	90	69	72.2592704	21.09515886	51.93903274	21.09515886	Cluster 2
18	Stanford	20	93	2.151941629	53.71206825	22.33967613	2.151941629	Cluster 1
19	TexasA&M	67	67	52.22765031	4.472084722	33.11526663	4.472084722	Cluster 2
20	UCBerkeley	40	78	23.78179109	29.83273374	10.77408534	10.77408534	Cluster 3
21	UChicago	50	87	28.71002321	24.83937444	8.254458203	8.254458203	Cluster 3
22	UMichigan	68	85	46.73841762	14.03564882	26.34475368	14.03564882	Cluster 2
23	UPenn	36	90	14.398732	38.07874302	6.065687194	6.065687194	Cluster 3
24	UVA	44	92	21.99700165	32.64954508	3.990671118	3.990671118	Cluster 3
25	UWisconsin	69	71	52.10145765	0.000135775	32.31363933	0.000135775	Cluster 2
26	Yale	19	96	3.881794254	55.9015737	24.06738543	3.881794254	Cluster 1
27								
28						Sum of Min D	176.4115271	

Figure 17. Calculating the distance to the clusters

Then the centroids of the clusters are calculated based on initialization of the clusters' centroids (shown in Figure 18 as x1 and x2). The error is then calculated and optimized using Excel Solver to find the optimal clusters. This is shown in Figures 18 and 19.

Cluster 1		Cluster 2		Cluster 3	
x1	x2	x1	x2	x1	x2
22.06232986	93.61453	68.99986	71.00001041	41.9030219	88.6046887
				Error	2.97145324

Figure 18. Calculating the final clusters' centroids and error

	J	K	L	M	N	O
1	Cluster 1 Centroid		Cluster 2 Centroid		Cluster 3 Centroid	
2	22	94	0	0	0	0
3	25	81	0	0	0	0
4	0	0	59	72	0	0
5	24	88	0	0	0	0
6	0	0	0	0	33	90
7	23	95	0	0	0	0
8	30	95	0	0	0	0
9	24	92	0	0	0	0
10	14	97	0	0	0	0
11	0	0	0	0	44	87
12	30	91	0	0	0	0
13	0	0	0	0	39	89
14	0	0	0	0	42	94
15	0	0	0	0	54	80
16	14	95	0	0	0	0
17	0	0	90	69	0	0
18	20	93	0	0	0	0
19	0	0	67	67	0	0
20	0	0	0	0	40	78
21	0	0	0	0	50	87
22	0	0	68	85	0	0
23	0	0	0	0	36	90
24	0	0	0	0	44	92
25	0	0	69	71	0	0
26	19	96	0	0	0	0
27	22.27273	92.45455	70.6	72.8	42.44444	87.44444

Figure 19. Calculating the cluster centroids

Prescriptive analytics focuses on identifying the best course of action in a scenario given the available data. It is the final phase of data analytics which can also incorporate descriptive and predictive analytic. Below, we present an example of prescriptive analytics implementation in MS Excel®. The example focuses on developing a decision support system for snow removal considering the City of Binghamton, NY as a case study. The problem is formulated as an integer programming model and an Excel application was developed. The City snows on an average of 64.5 days, 83.4 inches in depth, also as an average [21]. The snow removal and disposal is an important and expensive activity that is performed every winter in the City. Even though small amounts of snow can be removed from streets by plowing and shoveling, huge amounts of snow and the freezing temperatures require snow trucks that remove the snow in a timely manner. To facilitate the snow removal in a timely manner, Binghamton city is divided into 14 sectors based on its geographical nature, and the city has eight disposal sites with different capacities, separated to five surface site

disposal areas and three river disposal areas where dumping of snow also is used, “Susquehanna River” and “Chenango River” (see Figure 20).

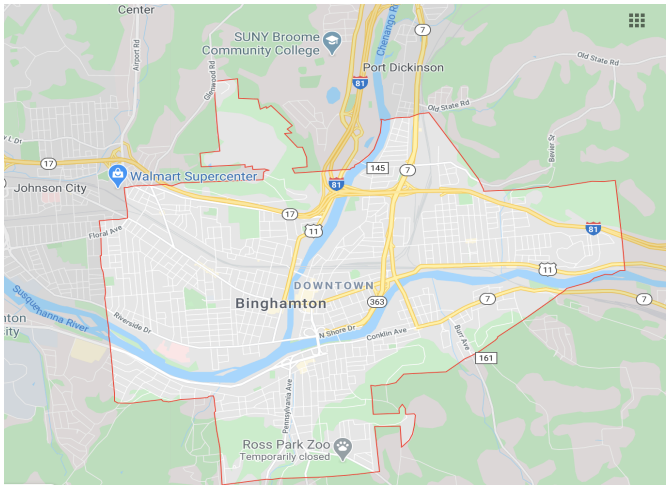

Figure 20. Google Map for Binghamton NY

To optimize the snow removal process, a prescriptive analytics model is developed. The snow removal problem was formulated as an integer programming model and solved using Excel Solver. Let: $x_{ij} \geq 0$ be the amount of snow moved from sector i to site j. If $x_{ij} = 0$, it means the site j is not assigned to the sector i. The objective function will optimize the transportation cost of moving a cubic feet of snow one mile and is formulated as:

$$Min \sum_{i=1}^{14} \sum_{j=1}^{8} h_{ij} d_{ij} x_{ij}$$

subject to

$\sum_{i=1}^{14} c_i x_{ij} \leq C_j$, for all $j = 1, 2, \ldots, 8$ (annual capacity)
$\sum_{i=1}^{14} r_i x_{ij} \leq R_j$, all $j = 1, 2, \ldots, 8$ (snow removal rate)
$\sum_{j=1}^{8} x_{ij} \geq 0$, all $i = 1, 2, \ldots, 14$ (assigning sector to sites)

where

d_{ij} is the distance from sector i to disposal site j (mi).
c_i is the annual capacity of sector i ($ft^3/year$).
C_j is the annual capacity of site j ($ft^3/year$).
h_{ij} is the corresponding handling cost from sector i to site j ($\$/ft^3$).
r_i is the snow removal rate in sector i (ft^3/hr).
R_i is the maximum snow receiving rate of site j (ft^3/hr).

A software application was designed in Excel using VBA as shown in Figure 20. Figure 21 shows sample optimization results for the snow removal. This includes the optimal assignment of the sectors to the sites as well as the optimal cost and quantity of now that will be removed from each sector and dumped into the assigned sites.

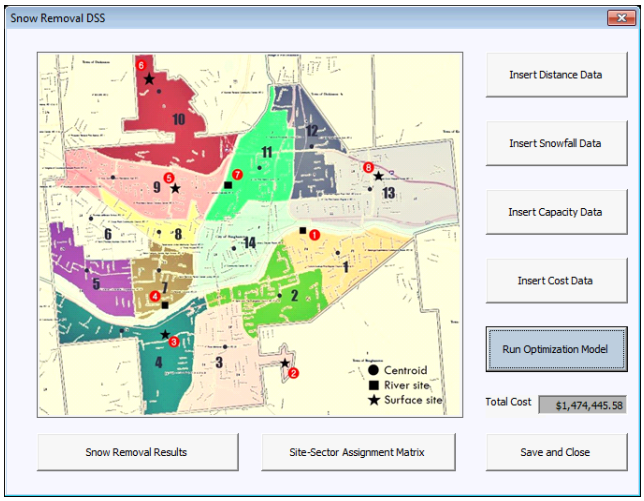

Figure 20. Snow removal optimization software

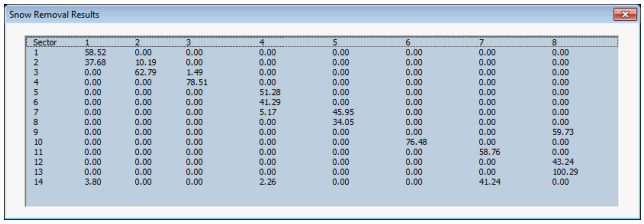

Figure 21. Sample snow removal results

The software application also allows for sensitivity analysis by changing some variables. Sensitivity analysis was performed to calculate the total cost of snow removal based on changing the total snowfall volume (±5%, ±10%, ±20%) and the results are shown in Table 5.

Table 5. Sensitivity analysis by changing snowfall volume

Change in Total Snowfall Volume	New Cost ($)	Change in Cost
+20%	1,474,428.56	+24.8%
+10%	1,320,671.83	+11.8%
+5%	1,251,030.78	+5.9%
0%	1,181,389.72	-
-5%	1,111,748.67	-5.9%
-10%	1,045,141.9	-11.5%
- 20%	913,704.07	-22.6%

Student feedback from the course has revealed that the use of the problem-based examples and software applications enhances their understanding of data analytics.

V. Conclusions and Future Work

Data analytics has emerged as an important subject in higher education. Existing studies have repeatedly shown a sustained increased in demand of data analytics professionals across industries. Yet, the number of professionals currently produced by higher education is not sufficient to meet the demands. While there are clear efforts and desires to teach data analytics in higher education, challenges abound. One of major challenge reported in literature is the high cost and learning curve of

software packages like SAS and Tableau used for teaching data analytics. This challenge can deter institutions, especially smaller colleges and universities, from teaching data analytics. Microsoft applications such as Excel and Access are familiar to most students and instructors and are used in most higher education institutes. In addition, Microsoft applications can provide cheaper alternative, with lower learning curve, for teaching and learning data analytics. However, many students and professionals are not yet aware of features and functions in these applications that are similar to those found in SAS and Tableau. The aim of this paper is to provide an exposition into some features and functions in MS Access and Excel that can be used to teach introductory data analytics courses and skills. The paper presented several examples on data analytics that were developed in MS Access and Excel. These examples can be used by instructors to help students understand data analytics and be able to implement data analytics models. Future work will focus on extending this work to include more examples other data analytics approaches that were not presented in the paper. We also plan to develop an online repository to make these examples and associated Access and Excel files available to students and educators. We also plan to evaluate and validate these tools and techniques to ensure that they help students to acquire data analytics skills. To do so, we will collect data from students after each semester, and then analyze the data and identify areas of strengths or weaknesses.

REFERENCES

[1] T. T. Goh and P.-C. Sun, "Teaching social media analytics: An assessment based on natural disaster postings," *J. Inf. Syst. Educ.*, vol. 26, no. 1, pp. 27–36, 2015.

[2] T. Davenport and J. Harris, *Competing on Analytics: Updated, with a New Introduction: The New Science of Winning*. Harvard Business Press, 2017.

[3] C. R. Wilder and C. O. Ozgur, "Business analytics curriculum for undergraduate majors," *INFORMS Trans. Educ.*, vol. 15, no. 2, pp. 180–187, 2015.

[4] B. Gupta, M. Goul, and B. Dinter, "Business Intelligence and Big Data in Higher Education: Status of a Multi-Year Model Curriculum Development Effort for Business School Undergraduates, MS Graduates, and MBAs," 2015.

[5] V. N. Gudivada, R. Baeza-Yates, and V. V Raghavan, "Big data: Promises and problems," *Computer (Long. Beach. Calif).*, no. 3, pp. 20–23, 2015.

[6] B. Wixom, T. Ariyachandra, D. Douglas, M. Goul, and B. Gupta, "The Current State of Business Intelligence in Academia: The Arrival of Big Data," *Commun. Assoc. Inf. Syst.*, vol. 34, 2014.

[7] B. Wixom, T. Ariyachandra, M. Goul, P. Gray, U. R. Kulkarni, and G. E. Phillips-Wren, "The current state of business intelligence in academia.," *CAIS*, vol. 29, p. 16, 2011.

[8] S. Miller and D. Hughes, "The quant crunch: How the demand for data science skills is disrupting the job market," 2017.

[9] A. Y. Yap and S. Drye, "The challenges of teaching business analytics: finding real big data for business students," *Inf. Syst. Educ. J.*, vol. 16, no. 1, p. 41, 2018.

[10] C. Kollwitz, B. Dinter, and R. Krawatzeck, "Tools for Academic Business Intelligence and Analytics Teaching: Results of an Evaluation," in *Analytics and Data Science*, Springer, 2018, pp. 227–250.

[11] C. Wymbs, "Managing the innovation process: Infusing data analytics into the undergraduate business curriculum (lessons learned and next steps)," *J. Inf. Syst. Educ.*, vol. 27, no. 1, p. 61, 2016.

[12] F. Jacobi, S. Jahn, R. Krawatzeck, B. Dinter, and A. Lorenz, "Towards a design model for interdisciplinary information systems curriculum development, as exemplified by big data analytics education," 2014.

[13] D. A. Asamoah, R. Sharda, A. Hassan Zadeh, and P. Kalgotra, "Preparing a data scientist: A pedagogic experience in designing a big data analytics course," *Decis. Sci. J. Innov. Educ.*, vol. 15, no. 2, pp. 161–190, 2017.

[14] Y. Gil, "Teaching big data analytics skills with intelligent workflow systems," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[15] D. Brandon, "Teaching data analytics across the computing curricula," *J. Comput. Sci. Coll.*, vol. 30, no. 5, pp. 6–12, 2015.

[16] R. Sharda, D. A. Asamoah, and N. Ponna, "Research and pedagogy in business analytics: Opportunities and illustrative examples," *J. Comput. Inf. Technol.*, vol. 21, no. 3, pp. 171–183, 2013.

[17] J. Nwokeji, F. Aqlan, A. Anugu, and A. Olagunju, "Big Data ETL Implementation Approaches: A Systematic Literature Review (P)," 2018, pp. 714–721.

[18] J. C. Nwokeji, R. Stachel, T. Holmes, F. Aqlan, E. C. Udenze, and R. Orji, "Panel: Addressing the Shortage of Big Data Skills with Inter-Disciplinary Big Data Curriculum," in *2019 IEEE Frontiers in Education Conference (FIE)*, 2019, pp. 1–4.

[19] F. Aqlan and J. C. Nwokeji, "Applying Product Manufacturing Techniques to Teach Data Analytics in Industrial Engineering: A Project Based Learning Experience," in *2018 IEEE Frontiers in Education Conference (FIE)*, 2018, pp. 1–7.

[20] J. Kennedy, P. Abichandani, and A. Fontecchio, "Using infographies as a tool for introductory data analytics education in 9--12," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, 2014, pp. 1–4.

[21] J. Kennedy, P. Abichandani, and A. Fontecchio, "An initial comparison of the learning propensities of 10 through 12 students for data analytics education," in *2013 IEEE Frontiers in Education Conference (FIE)*, 2013, pp. 916–918.

[22] R. Dubey and A. Gunasekaran, "Education and training for successful career in big data and business analytics," *Ind. Commer. Train.*, vol. 47, no. 4, pp. 174–181, 2015.

[23] M. A. Meyer, "Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings," *J. Am. Med. Informatics Assoc.*, vol. 26, no. 5, pp. 383–391, 2019.