





Making Sense of Unstructured Data: An Experiential Learning Approach

Sunet Eybers^(✉)  and Marie J. Hattingh 

Department of Informatics, University of Pretoria, Pretoria, South Africa
{sunet.eybers,marie.hattingh}@up.ac.za

Abstract. The need for competent data scientists is recognised by industry practitioners worldwide. Currently tertiary education institutions focus on the teaching of concepts related to structured data (fixed format), for example in database management. However, the hidden value contained in unstructured data (no fixed format) motivated the need to introduce students to methods for working with these data sets. Therefore, an experiential learning approach was adopted to expose students to real-life unstructured data. Third year students were given an assignment whereby they could use any publicly available un-structured data set or an unstructured dataset supplied to them following a set methodology (CRISP-DM) to discover and describe the hidden meaning of the data. As part of the assignments students had to reflect on the process. Twenty student assignments were analysed in an attempt to identify the effectiveness of the experiential learning approach in the acquisition of skills pertaining to unstructured data. Our findings indicate that the experiential learning approach is successful in the teaching of the basic skills needed to work with unstructured data. We discuss the appropriateness of the prescribed methodology, the students' performance, and lessons learnt. On the basis of these lessons we conclude with some recommendations for educating future data scientists.

Keywords: Experiential learning · Big data · Unstructured data · CRISP-DM methodology · Data scientists

1 Introduction

Data has always been a key asset to organisations. Nowadays this asset has become even more important due to the potential value contained in big datasets [11, 33]. *Big Data* refers to data with unique characteristics such as *the three V*: volume, velocity and variety [12, 25, 33]. Some scholars even include an additional characteristic, namely value [11, 33, 36]. Volume refers to the size and subsequent quantity of data sets which are often measured in terabytes or even petabytes [16, 33]. Velocity refers to the continuous generation of data by applications such as social media whilst variety refers to different kinds of data such as operational data from various business systems, xml files and text messages [16, 33].

These different kinds of data are further classified as structured (fixed format), semi-structured (consisting of both fixed format and free text or no fixed format data), and unstructured (no fixed format) [33]. Value refers to the untapped potential worth of the meaning hidden in large data sets [12], which might be economically or financially significant [11, 36]. Unfortunately, unleashing the value contained in these data sets can be challenging due to reasons pertaining to technologies, processes and human aspects [12]. For example, working with technologies such as advanced data mining tools requires specialised statistics tools; processes to combine data sources from various locations might be unclear; data scientists working with big datasets require a *combination* of business-, technical and analytical skills which is rarely found in today's human resources [34].

Despite the current scarcity of data scientists the position is said to be a "*most exciting career opportunity of the 21st century*" with above-average remuneration packages [2]. The demand for data scientists and data engineers is projected to grow with 39% [27]. The challenge from an educational perspective is to ensure that students studying in the areas of informatics, information science and computer science (ICT) are ready to meet the demands of industry practitioners upon graduation [3, 32]. Although the majority of educational institutions currently focus on skills to work with data, the curriculum has a strong focus on working with structured data such as databases, data marts and data warehouses; for comparison see [20, 21].

A current challenge in the curriculum, in particular the institution under study, is to introduce students to ways and methods of working with unstructured data obtained from social media platforms. Also, students often, as part of their post-graduate research projects (or fourth-year projects), are faced with challenges to work with unstructured data—a skill they have often not been exposed to during undergraduate studies. Therefore, our third-year semester module aimed at introducing students to working with unstructured data from social media platforms. A set methodology was prescribed to guide students through the assignment (namely the CRISP-DM Cross-Industry Standard Process for Data Mining explained later in this paper).

We followed an experiential learning approach where students could select their own set of unstructured data from any social media source and subsequently any tool or technique to extract meaning from those data. Our aim was to evaluate how effective the learning process was. The research question was: *how effective is an experiential learning approach in the teaching of basic skills to work with unstructured data?*

This paper continues with a brief introduction to the experiential learning approach, followed by related work focused on the topic of data science education. The CRISP-DM methodology is explained followed by a description of the case study and proposed research method. The analysis and discussion section describes the findings after evaluating the student assignments in relation to the phases of the experiential learning approach and the six steps of CRISP-DM. The discussion also includes the lecturers' reflection.

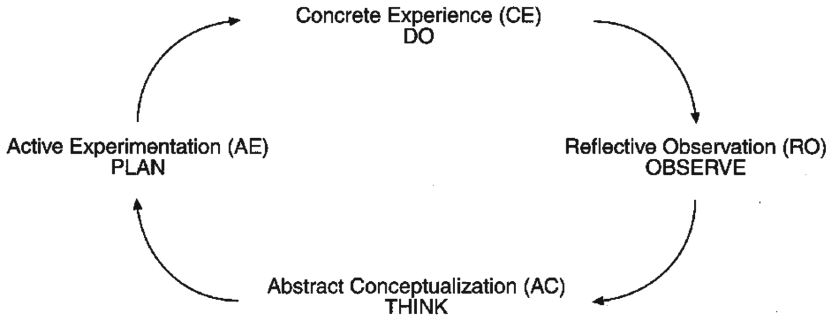


Fig. 1. Kolb's experiential learning cycle [15]

2 Experiential Learning

Experiential learning theory was introduced by Kolb in 1984 [18] and widely adopted in various educational environments [19] across fields such as medical and health [14], information systems [10], and marketing [4] (to name a few). The theory postulated that learners acquire new knowledge through practically completing tasks, i.e. their experience of interaction with the construct under discussion [18]. Figure 1 illustrates how learning is perceived as a continuous process that consists of four cycles, namely: experiencing (i.e. interaction with the construct), reflecting (review and evaluate the experience), thinking (drawing conclusions after reflecting on the experience), and acting (apply what has been learned from the process).

Learning can start at any point in the cycle. The benefits associated with experiential learning are: (1) increased opportunities for 'analytical' reflection on tasks completed — in particular 'short term experiential learning' where, similar to our study, the task do not have a long duration [28,38]; (2) 'substantive' benefits which refer to the ability of students to relate theoretical constructs to the practical exposure on a deeper level than just theoretical exposure [38]; (3) 'methodological', which refers to practically applying concepts in a structured way [38]; (4) 'pedagogical', which refers to active participation of the learners in their own and peer learning [38]; (5) 'transition', which refers to bridging the gap between applying concepts during theoretical studies and the practical requirements of industry practitioners. Lee identified lower-level benefits after a comparison between in-classroom learning and field-based experiential learning activities [23], which included an increase in 'soft' skills (e.g. the ability to adopt to change), 'leadership' and 'financial management' skills. Accordingly, learners could establish their own networks of practitioner contacts [4].

3 Education in Data Science

Davenport and Patil describe a data scientist as a "hybrid" of "data hacker, analyst, communicator and trusted adviser" [8] with the abilities to: write

program code, understand contextually the environment in which they function, communicate well in order to convey the message contained in the data to various audiences [8, 25]. These skills can only be acquired through the exposure to real-life scenarios.

Goh and Zhang acknowledged the challenge of exposing students to real-life scenarios when working with data [13]. They referred to current educational efforts to teach students about data analytics as artificial and simplified due to the utilisation of ‘canned’ data (a term used to refer to clean, structured data). They, too, adopted an experiential learning approach whereby they offered students an opportunity to work on a data analytics project in partnership with a large ‘Fortune 500’ company (as a live case study). The objectives of their study were: to investigate the influence of the adoption of an experiential learning approach on the teaching of data analytics; to evaluate students’ perceptions and attitudes towards the experiential learning approach; and to identify challenges associated with their experience when working with big data. Their findings suggested that, although learning outcomes were met and student motivation increased, the students were overwhelmed by the task of statistical analysis of big datasets. Their project entailed additional time challenges as students required more communication with the instructors as well within their groups. Their groups also experienced many failures which—although part of the normal learning process—had to be explained to the (disappointed) students.

Serrano (et al.) adopted experiential learning methods as part of an ongoing data science teaching project focusing on deep learning [9]. An ‘incremental’ teaching approach was used to allow students to adequately reflect their experiences when engaging with the content presented. From the ‘lessons learned’ they proposed the development of a platform for experiential learning that will act as a repository for capturing and storing students’ experiences, to be used by both other students and instructors. They furthermore provided a detailed list of functionalities that such a repository should offer, such as a rating system to gauge the ‘difficulty’ of students’ experiences as well as an anonymous peer review functionality to facilitate students’ reflections.

As part of their investigation of the utilisation of predictive analytics in a supply chain management environment, Schönherr and Speier-Pero evaluated the curriculum of data scientists [29]. They found that in one particular instance, where an experiential learning approach was adopted (in which students had to complete a ‘corporate analytics project’ within an organisation), students were immediately employable by organisations at above-average remuneration (in line with industry requirements).

4 CRISP-DM: The Cross-Industry Standard Process for Data Mining

The CRISP-DM methodology was introduced by a consortium of manufacturing companies, as well as software and hardware organisations in an attempt to standardise and formalise a method for data mining [26]. Their result was

an independent conceptual model which proposed data as central to a six-step process to be followed during a data mining cycle. These six steps started with a clear understanding of the business under investigation, the data to be analysed, the preparation of those data (e.g. cleansing), the application of specific models to analyse the data (e.g. linear regression models), the evaluation of the results as a consequence of modelling, and finally the deployment or distribution of the results and/or the model to the stakeholders.

The CRISP-DM methodology is similar to other data mining approaches such as the Knowledge Discovery Databases (KDD) process model or the Sample, Explore, Modify, Model, Assess (SEMMA) model [30]. Although the numbers of data mining steps differ amongst those (9 in KDD, 6 in CRISP-DM, 5 in SEMMA), the meanings of the steps are similar. For example, the ‘data understanding’ step (2) of the CRISP-DM methodology corresponds to the ‘selection of a data sample set’ and the ‘exploration of the data sample set’ in SEMMA as well as to the ‘selection and preprocessing of data’ in KDD [1].

The CRISP-DM methodology was furthermore chosen for the purpose of this experiential learning exercise for the following reasons: (1) it was conceptual and therefore applicable to any scenario in working towards understanding data; (2) it is a complete and workable methodology [30,37]; (3) the methodology was already prescribed in fourth-year postgraduate studies and therefore seemed appropriate for preparatory purposes also at undergraduate level; (4) the SEMMA method is tightly linked to the SAS enterprise software suite and is thus too software-tool-specific [30].

5 Case Study Description

123 students were enrolled in our third-year second-semester course ‘Trends in Information Systems’. The course offered an introduction to a variety of novel concepts such as IS security and bitcoin technologies (to name a few). As part of the course, students were also exposed to the concept of Big Data and the subsequent challenge of working with unstructured data. The *learning outcomes* specifically for these sessions were:

- *Understand* the Data Lifecycle as part of the Software Development Lifecycle;
- *Describe* the characteristics of unstructured data;
- *Overcome* challenges associated with unstructured data;
- *Understand and implement* the six steps of the CRISP-DM methodology.

At the end of the sessions students were given a practical assignment to complete. The main objective of the practical assignment was to use the current concepts explained during the sessions and apply it in order to work with unstructured data. The task instruction was to follow the CRISP-DM methodology to identify, clean and interpret data from any publicly available, unstructured social media platform (for example Twitter or Facebook) or to select one of the unstructured datasets supplied by the instructors. Students who chose their own datasets were free to use any publicly available source, whereby no further information

was supplied to them (for example about how to use Twitter or what topics or ‘threads’ to use). Students were allowed to use any free tool to assist them in the process of data acquisition, data cleansing, analysis and presentation. To help them on the way, a practical example of what the intended outcome should look like was presented in class.

Table 1. Practical assignment rubrics

CRISP-DM STEP	
1. Business understanding:	What type of business will benefit from this analysis? What are the goals of the business, i.e. what do they want to achieve as a business? What question would you like to answer with the exercise? How will you go about to answer the business question (high level plan)?
2. Data understanding:	Collect unstructured data (from any source). Tip: Twitter might be the easiest. Describe, explore, verify data
3. Data preparation:	ETL: extract, transform (i.e. clean), load data into another structure (flat file, table, etc.). Include copies of your screen where you prepared your data
4. Data modelling:	Decide what you are going to do with the data: apply complex statistical algorithms, or do basic modelling, for example categorisation. Include copies of your screen where you modelled your data
5. Evaluation:	What does the results mean? Do I need to repeat the analysis? Include copies of your screen where you show your results
6. Deployment:	If you were to share your results with the other students: how would you do that? What was your experience working with the datasets (good or bad)? Was the dataset appropriate for what you wanted to achieve? What challenges did you face?
7. Conclusion:	Did you enjoy the assignment? What did you like? What did you dislike? What did you learn?

As part of the assignments the students had to write a report using the six steps of the CRISP-DM methodology. Table 1 outlines the details of each step. Students were also instructed to include copies of the screen(s) where the actual process of data preparation, modelling and evaluation was followed. There was no need to actually deploy or implement their proposed solutions, though the students had to make plausible suggestions about how an organisation could possibly use the results of their analyses.

A conclusion section (7: not originally in CRISP-DM) was added for our students to summarise and reflect what they have learned. Marks were allocated to each of these sections. Table 1 also shows a summary of the rubric used for evaluating the assignments.

6 Method

We followed an *interpretive approach* to analyse a sample of 20 (out of 123) assignment submissions. Saturation was reached with 15 assignments (i.e., no new concepts emerged), but another 5 assignments were analysed to confirm saturation. All data sets used by these students were publicly available and not password-protected. Although the individual data records used by the students did not disclose any identifying attributes, the organisations associated with these data sets were in some instances revealed. We followed the ethical procedure recommended by Langer and Beckman who suggested that if public data, that is not password-protected, is used, researchers do not need to obtain any usage permission [22]. Anyway, the anonymity and privacy of the users were guaranteed as we anonymised particular organisation names by grouping them into ‘industries’.¹ All assignments followed the structure of the rubrics of Table 1.

Thematic content analysis was used to analyse the students’ project reports.² Thereby we followed the six steps proposed by Braun and Clarke [6]. Initial *codes* were captured as they emerged and were recorded under every step of the CRISP-DM methodology. As the analysis continued and themes emerged (and were reviewed and named), it was easy to see how the *six steps of the CRISP-DM methodology mapped to the four stages of the experiential learning approach*. The following section presents the analysis and discussion of these findings.

7 Analysis and Discussion

The analysis and discussion outline section followed the phases of the experiential learning approach, namely: abstract conceptualisation, concrete experience, reflective observation, and the lecturers’ reflection. Each of the six CRISP-DM steps could be related to the four phases of the experiential learning approach.

7.1 Abstract Conceptualisation: Business Understanding

The abstract conceptualisation stage is concerned with learners trying to make sense of a problem at hand. For our particular assignment the students could use any publicly available data set from social media or a variety of unstructured datasets supplied to them, (see above). The objective was to understand the message(s) the data can communicate to various audiences, and questions that can be formulated which can be answered with the help of the gathered data. Most students chose data from Twitter from diverse industries namely: Gaming, Finance, Music, Government/Activism, Government/Treasury, App Store, Fast-moving consumer goods (Beverage), Marketing/Communications.

¹ For this paper we obtained the consent from our participating students as well as a permission from our faculty’s ethics committee.

² The 1st author (S.E.) was the examiner of the assignment and was thus familiar with the content of the assignment. The 2nd author (M.J.H.) familiarised herself with the content by reading two assignments before the analysis began.

It turned out that our students were able to contextualise those data and could identify the parties that might be interested in answers to questions that relate to the data. For example, Participant #5: *“These businesses want to know their market/customers better while situating themselves to a favourable position in their operating markets”*. Contextualising results is a very important skill of any data scientist [8, 17, 25]. Also Kennan stated that in a business environment it is important for a data scientist to know what the organization does, who its customers are, and what the operating environment is [17]. She further stated that the context is different for different countries due to different governments’ regulations. Hence it is not required for graduates to know all the contextual details; nonetheless they must be aware that *“the contexts in which data and information are used are highly varied”*; they must also *“understand examples, and where to look for specific contexts, and be prepared to continue learning on the job”* [17].

In order to contextualise their potential results, the students were required (as part of the CRISP-DM methodology) to find out more about the companies, understand their ‘missions’, ‘visions’ and goals. This aspect of the assignment was very important, as the students were exposed to real-life companies and had to make sense of how a data set supports an organisation’s purposes. A few students did this very well by studying a business, understanding its goals and how the social media data relate to those goals. This exposure to real-life scenarios is important in delivering industry-ready graduates [13, 32]. For example, participant #3 noted: *“it is very important that the academy has a strong social media presence to attract potential donors, spread the word about the music programs on offer, and to promote any upcoming events”*. Accordingly, Kennan states that in order to understand data within its context one has to understand the intended audience of the information [17].

7.2 Active Experimentation: Data Understanding and Preparation

The active experimentation stage is concerned with planning the *“forthcoming experience”* [15]. In this instance the students had to plan, in accordance with the CRISP-DM methodology, how to approach the collecting, extracting, cleaning, loading and storing of their chosen data sets. Part of this Extract-Transform-Load (ETL) process was the verification of the sources.

The students used a number of techniques to clean the data sets, such as splitting datasets into smaller parts according to the original date into day, month and year, then combining those attributes into a new column. The columns were furthermore labelled with meaningful names, classified according to types of Tweets (for example RT for retweet, OR for original tweet), and classified according to keywords. The ‘noise’ of hyperlinks was removed. All of these activities are essential for their ability to work with data [8, 25].

Furthermore, students understood the importance of recognising missing data and the potential implications it might have on the analysis results—or that it might not affect the result depending on the way the data is analysed. Through the ETL process the students were also able to identify and discard redundant

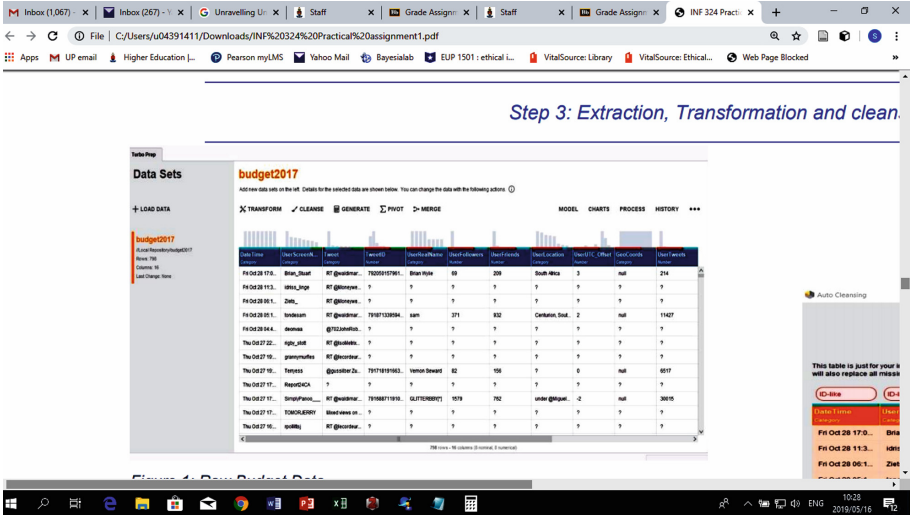


Fig. 2. Example of data set before cleansing

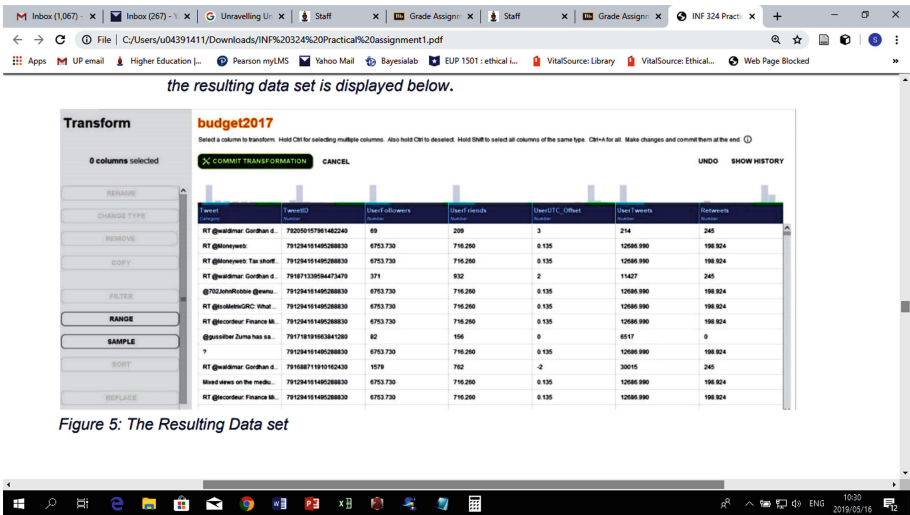


Figure 5: The Resulting Data set

Fig. 3. Data set from Fig. 2 after cleansing

data as they recognised that it would have no purpose in the analysis. Figures 2 and 3 illustrate examples of a data set before and after cleaning.

As mentioned above, one of the challenges for a data scientist is the requirement to work with various technologies [12]. In this assignment our students were exposed to a variety of tools to complete the ETL tasks, for example: RapidMiner, ParallelDots AI in MS Excel (sentiment analysis), the Twitter API in RapidMiner,

Twitter Analytics, Zoho Reports (now Zoho Analytics), Tableau, MS Excel Azure Machine Learning add-in, and Jupyter Notebook in Python. The variety of tools available to students indicated the evolving nature of data science. Accordingly, exposure to a variety of such tools will increase the students' abilities.

7.3 Concrete Experience: Data Modelling, Evaluation and Deployment

The concrete experience stage is concerned with the actual completion of the activity. During this assignment, this stage refers to the modelling, evaluation and deployment of the results.

During this stage we observed that the students used a variety of *visualisation* methods to communicate their results, for example: bar charts, pie charts, scatter plot, bubble chart, ring graph, line chart, and location map. One student used a histogram to indicate how brand sentiment changed over a period of time. Some students used more than one visualisation method to communicate different messages. One student used a more advance modelling technique, namely the 'predict' and 'simulate' functions of RapidMiner, to build a Deep Learning Simulator based on the data set. Accordingly, Kennan stated that there is a great need for graduates to have such visualisation skills [17], which allow them to present the data in such a way that it enables managers to make better and quicker decisions and to communicate messages more clearly to stakeholders outside an organisation.

A second observation at this stage was the students' ability to interpret the visualised results. The majority of the students were able to correctly interpret the meaning behind the visualisation. The power of big data lies indeed in the *interpretation* of analysis results. McAfee and Brynjolfsson stated accordingly: "*Big data's power does not erase the need for vision or human insight*" [25]. Our assignment was not too complex, and some of the analyses were quite basic, but nonetheless contained powerful messages. For example, participant #3 found that "*the company does not have to change what it is tweeting; rather when it is tweeting. In addition, gaining more followers should increase impressions and engagement rate. If they do these two things, they should see an improved Twitter performance*". Students were thus able to derive *meaning* from the data, such as: "*the rebranding campaign was not well received*" (based on the tweets analysed) "*as the audiences did not understand the campaign concept when it came to the brand messaging and intent. Some found it offensive and insensitive whilst others either engaged positively or were indifferent*" to what the company had communicated.

An important component of the methodology was to evaluate the results to verify if they were plausible. This requires students to critically double-check their initial findings. Whilst most students reported that their results were in line with their expectations, participant #12 found that a sentiment analysis done by Azure were not correct, as sometimes there was a colloquial misunderstanding which skewed the results. Participant #7 evaluated the data post modelling and concluded that the results were not accurate because re-tweets

skewed the analysis. This illustrates a level of awareness about the nature of the data set which is very important for any data scientist. Accordingly, Costa and Santos describe data scientists as having an ‘inquisitive mind’ with which they ‘interrogate’ the data to understand their deeper meaning [7].

Finally, students recognised that the *dataset can potentially answer different questions depending on the analysis*. For example, participant #7 generated nine different visualisations from the dataset which included a pie chart, five different bar charts, a scatter plot and two line charts. The scatter plot was used to indicate location. Participant #8 developed a generalised linear model between the categories. However, the fact that one data set can communicate different messages is something students struggle with. Our assignment, that prescribed the adopted CRISP-DM methodology, allowed students to do an arbitrary number of modelling iterations whereby a result obtained after the completion of one cycle of the methodology introduced a new question or problem to be investigated.

7.4 Reflective Observation

The reflective observation stage is concerned with the students’ reflection on their completed activity. This stage offered students the opportunity to indicate what aspects of the assignment as well as experiential approach they enjoyed and what they found challenging. From the lecturer’s perspective, the reflective observation stage allowed her to make a judgement about the success of the exercise, (see Subsect. 7.5). Some positive feedback by the students regarding the assignment included the following points:

- Students enjoyed the ‘mining’ aspect of the assignment. This refers to practically using an identified software tool, (see Subsect. 7.2). Accordingly, Kolb explained that students preferring a practical approach to solving problems refer to the ‘converging’ learning style [18].
- The practical component of the assignment also extended to learning new software. Participant #10 stated: “Overall, I enjoyed working on the assignment because I enjoy working with new software, and the learning that comes with it, especially when it allows you to apply your theory, making it interactive”. Learning how new software works whilst completing an assignment is an example of ‘incidental’ learning which is imperative in assisting students to get ‘hands-on’ experience and to prepare them for the information age. Incidental learning is a ‘side effect’ of learning whereby the students were initially not aware of the fact that they would have to learn new software, but then suddenly had to acquire new skills in order to complete the entire CRISP-DM lifecycle [31].
- Students learnt about the potential impact of data. Participant #4 said that he could “*see the potential impact of big data*” whereby “*it was best to see this when it was done practically*”. Participant #2 confirmed this by stating: “*I enjoyed seeing how the steps are done practically and I learnt a great deal about how unstructured data can be used to make better business decisions*”. Participant #1 indicated that “*simple data can give good answers to important questions*”.

- Students learnt about a variety of options to visualise data as well as the power of tools to manipulate data. Participant #7 stated: “I was amazed how these tools could formulate meaningful graphs and charts based on unstructured data. Even if a field was null, the tool was able to identify it without mixing it with the rest of the data”. Accordingly, Wang (et al.) stated that the adoption of good visualisation methods can transform the challenges introduced by big data (such as the vast volumes of seemingly unrelated data) into meaningful ‘pictures’ [35].
- The assignment allowed for Self-Motivated Incremental Learning (SMIL) [5]. SMIL is concerned with the intrinsic motivation by a person which will allow him/her to learn a hierarchy of skills freely by repeating three phases: exploring the environment, identifying interesting situations, and obtaining skills to cope with these situations. Participant #15 reported that *“we were never really taught how to actually do a full data analysis. We were only ever shown the theory behind data analysis and data modelling, and we had to use the information along with all the other knowledge we have from Statistics, Mathematics, IT, etc., and figure out ourselves how to work with data and analyse the raw data. I learned that Excel is a much more powerful tool than I first anticipated, but it cannot compete with how strong Python is and how easily you can achieve the same results”*. The assignment furthermore introduced students to the area of data science, whereby one participant noted: *“It sparked an interest in Python in me and I already enrolled in two short online courses on Python and R in data analytics”*.

Some challenges observed by the students included the following points:

- Data preparation took a long time due to the volume of data they had to work with. The majority of the students indicated that this was their least favourite part of the assignment. Participant #10 indicated that *“having to read through the data, and generating a question or problem statement, took a while”*.
- Some students struggled to understand the data set. Participant #3 said that it was a *“challenge to understand what data I was working with, and the relevance it might have”* to the company in question. *However, after doing plenty of research I was able to overcome this*. This case is another example of SMIL [5], as the student has to analyse the environment in order to solve the data problem.
- One student indicated that he would have preferred more direction in the assignment, and *“often felt lost and unsure of whether I was doing things correctly”*.
- A few students reported that it was difficult to identify the correct technology for the task. Participant #7 stated that *“finding a data analyser tool which would best fit the dataset was challenging”*. Participant #20 stated that *“being able to choose the right model, and making the data fit the model, was also a challenge”*. Participant #16 struggled with the method: *“I did not know which algorithms to use and how to correctly use the datasets”*. Obviously this challenge was two-fold:

- firstly w.r.t. the students' ability to identify the technology with the correct functionality to obtain the envisaged results,
 - secondly w.r.t. the students' ability to use the chosen software (as some software requires deeper knowledge for its proper utilisation).
- Finally, some students also had problems to identify an appropriate visualisation method.

The above-mentioned challenges are linked to key competencies of data scientists [7]. Exposing the students early to these challenges should give them an opportunity to continue with SMIL in preparation for a workplace after graduation.

7.5 The Lecturer's Reflection

This subsection presents the reflection of the lecturer during and after the assignment. We consider (1) the appropriateness of the prescribed methodology, (2) the students' performance, and (3) 'lessons learnt'. Each of these will be briefly discussed.

Appropriateness of the CRISP-DM Methodology. Prescribing the CRISP-DM methodology was appropriate, as students easily grasped the six steps of the process while working with data. The methodology guided students through the process and provided a structure for approaching an assignment that seemed daunting to some students at first. Due to its conceptual nature, students can hopefully reuse this methodology when working on similar projects in the future.

Students' Performance. 11% of our 123 students obtained an assignment mark between 80% to 89%. About half of all students obtained a mark between 70% to 79%. About a quarter of the students obtained a mark between 60% to 69%. The remaining students obtained a mark between 50% to 59%. The average mark for the assignment was 68%, the highest 88%. The students who obtained a mark between 50% and 59% did either not provide enough detail in their reports, or misinterpreted their found results. For example, two students discovered the most prominent words associated with their data but failed to synthesise the findings to draw a meaningful conclusion. As a consequence, they did not adequately answer the initial question. Overall, given the performance of the students, the assignment was successful in teaching students the basic skills necessary to work with unstructured data.

Lessons Learnt, and Recommendations. Our experience can be summarised as follows.

- *Working with social media data* was enjoyed by our students, as they are already familiar with these platforms. On the basis of this experience we recommended that educators use data sets that students can relate to (such as social media).

- Students found the extracting, cleaning and transforming of data very challenging and time-consuming. This was their biggest and to some extent overwhelming tasks. Therefore we recommended that students be provided with more practical demonstrations of how to extract, clean and transform data, as the ETL process is the biggest task when working with any form of data.
- The lecturer should cater for and accommodate students with different learning styles. Our assignment was best suited for students with the ‘converging’ learning style (who prefer to solve problems and apply their knowledge to practical implementations) as well as to students with the ‘accommodating’ learning style (who prefer to ‘do things’ practically) [18]. Students with the ‘diverging’ style of learning (who prefer to watch rather than do) as well as with the ‘assimilating’ style (strong analysts when given good information) should be accommodated by group-work opportunities.
- The lecturer should offer students more consultation time when students have more questions particularly at the beginning of the assignment; for comparison see Goh and Zhang [13].

8 Conclusion

This paper describes the effectiveness of teaching basic skills to third year undergraduate students to work with unstructured data by following an experiential learning approach (ELA). We found that the ELA enabled ‘novices’ to acquire basic skills in working with unstructured data. The students were exposed to a structured methodology that allowed them to tie the data sets to the goals of a business. They used a variety of tools and technologies to obtain, prepare, model and interpret unstructured data sets. Our assignment enabled the students to experience the ‘nature of data’, the influence of missing or redundant items on the analysis results, and how one data set can provide different answers to a variety of questions. The students reported that they enjoyed the ‘(data) mining’. The students reported that they also enjoyed the acquisition of skills to work with new software and tools. They realised the impact social media data has on an organisation. The challenges experienced by the students in completing this assignment do not outweigh the benefits that students derived from it. As more organisations become data-driven, graduates need to be prepared to assist organisations in their data needs [3]. With our educational efforts we also gave our undergraduate students an initial preparation for possible postgraduate studies [24] in this new field.

References

1. Azevedo, A., Santos, M.F.: KDD, SEMMA and CRISP-DM: a parallel overview. In: Proceedings European Conference on Data Mining, p. 6 (2008)
2. Baškarada, S., Koronios, A.: Unicorn data scientist: the rarest of breeds. Program 51(1), 65–74 (2017)

3. van Belle, J.P., Scholtz, B., Njenga, K., Serenko, A., Palvia, P.: Top IT issues for employers of South African graduates. *CCIS* **963**, 108–123 (2019)
4. Bobbitt, L.M., Inks, S.A., Kemp, K.J., Mayo, D.T.: Integrating marketing courses to enhance team-based experiential learning. *J. Mark. Educ.* **22**(1), 15–24 (2000)
5. Bonarini, A., Lazaric, A., Restelli, M.: Incremental skill acquisition for self-motivated learning animats. In: Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J.C.T., Marocco, D., Meyer, J.-A., Miglino, O., Parisi, D. (eds.) *SAB 2006. LNCS (LNAI)*, vol. 4095, pp. 357–368. Springer, Heidelberg (2006). https://doi.org/10.1007/11840541_30
6. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**(2), 77–101 (2006)
7. Costa, C., Santos, M.Y.: The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *Int. J. Inf. Manage.* **37**(6), 726–734 (2017)
8. Davenport, T.H., Patil, D.J.: The Sexiest Job of the 21st Century. *Harvard Business Rev.* (October), 8 (2012)
9. Emilio, S., Martin, M., Daniel, M., Luis, B.: Experiential learning in data science: from the dataset repository to the platform of experiences. In: *Ambient Intelligence and Smart Environments*, pp. 122–130 (2017)
10. Eybers, S., Hattingh, M.J.: The last straw: teaching project team dynamics to third-year students. *CCIS* **963**, 237–252 (2019)
11. Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., Gnanzou, D.: How big data can make big impact: findings from a systematic review and a longitudinal case study. *Int. J. Prod. Econ.* **165**, 234–246 (2015)
12. Gantz, J., Reinsel, D.: *Extracting Value from Chaos*. IDC, p. 12 (2011)
13. Goh, S., Zhang, X.: Incorporating experiential learning into big data analytic classes. In: *Proceedings 21st Americas Conference on Information System*, p. 10 (2015)
14. Grace, S., Innes, E., Patton, N., Stockhausen, L.: Ethical experiential learning in medical, nursing and allied health education: a narrative review. *Nurse Educ. Today* **51**, 23–33 (2017)
15. Healey, M., Jenkins, A.: Kolb's experiential learning theory and its application in geography in higher education. *J. Geogr.* **99**(5), 185–195 (2000)
16. Janvrin, D.J., Watson, M.W.: Big data: a new twist to accounting. *J. Account. Educ.* **38**, 3–8 (2017)
17. Kennan, M.A.: In the eye of the beholder: knowledge and skills requirements for data professionals. *Inf. Res.* **22**(4), 1–21 (2017)
18. Kolb, D.A.: *Experiential Learning: Experience as the Source of Learning and Development*. Prentice Hall, New Jersey (1984)
19. Kolb, Y.A., Kolb, D.A.: Learning styles and learning spaces: enhancing experiential learning in higher education. *Acad. Manage. Learn. Educ.* **4**(2), 193–212 (2005)
20. Kotzé, E.: A survey of data scientists in South Africa. *CCIS* **730**, 175–191 (2017)
21. Kotzé, E.: Augmenting a data warehousing curriculum with emerging big data technologies. *CCIS* **730**, 128–143 (2017)
22. Langer, R., Beckman, S.C.: *Sensitive Research Topics: Netnography Revisited* (2005)
23. Lee, S.A.: Increasing student learning: a comparison of students' perceptions of learning in the classroom environment and their industry-based experiential learning assignments. *J. Teach. Travel & Tourism* **7**(4), 37–54 (2008)

24. Marshall, L., Eloff, J.H.P.: Towards an interdisciplinary master's degree programme in big data and data science: a South African perspective. *CCIS* **642**, 131–139 (2016)
25. McAfee, A., Brynjolfsson, E.: *Big Data: The Management Revolution*, vol. 9. Harvard Business Review, Brighton (2012)
26. North, M.: *Data Mining for the Masses*. CreateSpace Independent Publication Platform, Scotts Valley (2016)
27. Piatetsky, G.: How many Data Scientists are there and is there a Shortage? Technical Report, (2019). <https://www.kdnuggets.com/2018/09/how-many-data-scientists-are-there.html>
28. Scarce, R.: Field trips as short-term experiential education. *Teach. Sociol.* **25**(3), 219 (1997)
29. Schöherr, T., Speier-Pero, C.: Data science, predictive analytics, and big data in supply chain management: current state and future potential. *J. Bus. Logistics* **36**(1), 120–132 (2015)
30. Shafique, U., Qaiser, H.: A Comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *Int. J. Innov. Sci. Res.* **12**(1), 217–222 (2014)
31. Sleight, D.: *Incidental Learning from Computerized Job Aids*. Technical Report (1994). <https://msu.edu/~sleightd/inclearn.html>
32. Smuts, H., Hattingh, M.J.: Towards a knowledge conversion model enabling programme design in higher education for shaping industry-ready graduates. *CCIS* **963**, 124–139 (2019)
33. Tanwar, M., Duggal, R., Khatri, S.: Unravelling Unstructured Data: A Wealth of Information in Big Data. In: *Proceedings ICRITO 4th International Conference on Reliability, Infocom Technologies and Optimization*, pp. 1–6, IEEE (2015)
34. Tole, A.: Big Data Challenges. *Database Syst. J.* **IV**(3), 31–40 (2013)
35. Wang, L., Wang, G., Alexander, C.A.: Big data and visualization: methods. *Challenges Technol. Progress. Dig. Techn.* **1**(1), 33–38 (2015)
36. Wang, S., Yeoh, W., Richards, G., Wong, S.F., Chang, Y.: Harnessing business analytics value through organizational absorptive capacity. *Inf. Manage.* **56**, 103–152 (2019)
37. Wirth, R., Hipp, J.: CRISP-DM: towards a standard process model for data mining. In: *Proceedings 4th International Conference on the Practice Application of Knowledge Discovery and Data Mining*, pp. 29–39 (2000)
38. Wright, M.C.: getting more out of less: the benefits of short-term experiential learning in undergraduate sociology courses. *Teach. Sociol.* **28**(2), 116 (2000)