Applying Product Manufacturing Techniques to Teach Data Analytics in Industrial Engineering: A Project Based Learning Experience

Faisal Aqlan
Industrial Engineering Department
Penn State Behrend
Erie, PA, USA
FUA11@psu.edu

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

***Abstract*— The amount of data generated from industrial processes has dramatically increased in recent times. As a result, data analytics skill has become an essential requirement for industrial engineering jobs. To meet this requirement, universities and colleges are beginning to integrate data analytics into industrial engineering curriculum. However, teaching and learning data analytics to industrial engineering students is by no means an easy task, since both programs have diverging focus. Industrial engineering focuses on process and systems optimization while data analytics focus on the application of information technology and mathematical models to visualize and extract useful information from raw data. To support teaching and learning of data analytics to industrial engineering students, this innovative practice full paper reports a pedagogical method that extrapolates product manufacturing processes to teaching and learning data analytics. We selected product manufacturing because it is a core course in the industrial engineering curriculum. The proposed pedagogical method is developed by first analyzing and comparing product manufacturing processes and data analytics techniques. Afterwards, we used the result of this analogy to develop a teaching and learning method for data analytics. For implementation and validation purposes, we adopt a project-based learning approach where students used our methodology to complete real-world data analytics projects. Data from students' grades shows that this approach improved their performance.**

Keywords—data analytics, data manufacturing, product design and manufacturing, process improvement, system modeling

I. INTRODUCTION

Most universities are beginning to integrate data analytics as a core course in their industrial engineering (IE) curriculum. This is because the amount of data generated from industrial and manufacturing processes has increased dramatically in recent times. In 2013, IBM reported that 2.5 quintillion bytes of data is created every day and 90% of the data in the world today has been created in the last two years alone. Hence, data analytics skills are increasingly becoming essential requirements for a successful career in industrial engineering. However, teaching and learning data analytics, especially in IE can be very challenging; since IE and data analytics have diverging focus. IE focuses on process and systems optimization while data analytics focuses on the application of information technology and mathematical models to visualize and extract useful information from raw data.

Additionally, data analytics is a new course relative to most other courses in the IE curriculum. Therefore, effective pedagogical methods for teaching data analytics are still emerging and being experimented on. These create the need for more effective techniques and methods for teaching data analytics to industrial engineering students. To meet this need, this paper provides the foundation for developing a data analytics teaching technique by exploring the many similarities between data analytics and product manufacturing. We note that both processes start with raw, unrefined inputs. In other words, data analytics starts with raw data while product manufacturing starts with raw materials. These inputs are worked on, processed and manipulated until a final complete finished product is made.

This paper discusses the analogy between "data analysis and modeling" and "product design and manufacturing". Raw data can be considered as raw materials used in the production process. Likewise, the production processes that are applied to convert the raw materials into finish product can be mapped to the analytics techniques used to extract useful information from the raw data as well as visualize and communicate this information to the decision makers. However, it is important to note that unlike raw material, data are generally not depleted through production.

The term "analytics" has been recently adopted by many researchers and professionals working with data in both academic and industry. According to [1], analytics can be defined as the use of information technology, statistical and mathematical models, and data visualization techniques to extract useful information from raw data (see Figure 1). The useful information can then be used to improve processes and decisions. Hence the term "analytics" includes a) data manipulation: extracting the data from the sources b) data

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

analysis: inspecting, cleaning, transforming, and modeling c) communication: data visualization and recommendations. For part a, a knowledge of database and information technology is needed. Part b requires the knowledge of statistics, data mining, and operations research. Part c requires the knowledge of data visualization. Furthermore, the analyst needs to have enough knowledge about the system or process under study. Figure 2 shows the main building blocks for analytics.

Fig. 1. Converting raw data into useful information using analytics

Analytics is used to convert extensive data into powerful insights that can drive effective decisions. It is widely used today by many businesses such as banking, insurance, retail, healthcare, and manufacturing. Many companies today are considering analytics as a key business strategy for making better decisions. The rapid growth in the size, speed, and diversity of data streams requires advanced analytics techniques. Almost every business needs to use analytics to understand its behavior, measure and track its performance, and discover improvement opportunities. Many companies are using analytics to create competitive advantage. With the use of analytics, companies are able to better understand their customers and make better decisions.

In the recent few years, data analytics topics have been given considerable attention by engineering community, especially industrial engineering. The focus includes statistical and data analysis, operations research, decision making, risk analysis, commuting and computational methods, and business analytics.

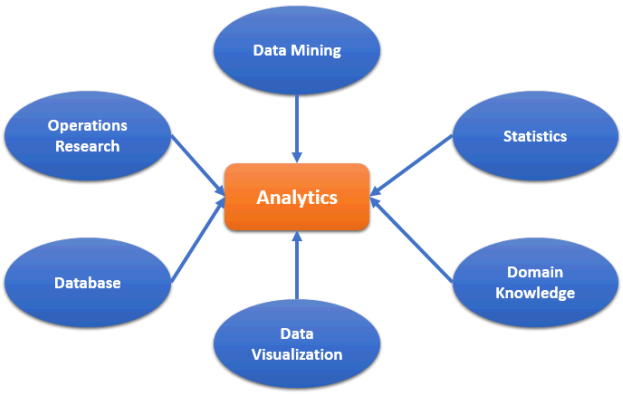

Fig. 2. Building blocks for analytics

Process improvement tools and techniques used in manufacturing such as Lean, and Six Sigma can also be applied to data analytics to improve the analysis process. Furthermore, concepts used in modeling of production systems can also be utilized for data modeling. By studying the analogy between product manufacturing and data analytics, the performance measures and improvement strategies that have been developed and used in manufacturing long time ago can be utilized in data analytics. By comparing information systems to production systems, raw data can be considered as raw material and the information or knowledge can be considered as the finished products. Other processes and concepts that have been used in manufacturing such as product assembly, product quality, and raw material inventory can also be applied in information systems.

To apply the proposed framework, a learning module is developed and used in an IE undergraduate course environment in which data analytics is taught. The learning module consists of two parts. The first part includes only data analytics and the second module considers the analogy between data analytics and product manufacturing. Student's performance in both modules is assessed based on several criteria and is used to compare the two modules.

Data analytics is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information and visualize and communicate it to the decision makers. With today's modern data processing abilities, finding correlations, grouping data by clustering and making decisions with regards to the future can be done much faster than of yesterday's hand calculations and slide rulers. Data and data manufacturing are huge aspects of today's industries. Everyone from hospitals to manufacturing plants use data analytics in some way to improve processes. If the data being used is poor, any results from data manufacturing are next to useless because the data is not a true representation of what it was obtained from. Data checking processes are needed to ensure that the data is both complete and accurate. Many of the sources of errors in data recording can be solved rather cheap and easy ways.

Since high quantity data manufacturing is still relatively new, there are no set standards for measuring the quality of the data. Four measures that are used to determine whether the data quality is poor or not are completeness, accuracy, timeliness, and uniqueness. Completeness is the amount of data that is 100% complete without any missing values. Depending on what is being analyzed can determine whether or not the missing values will have any impact. In some cases, data that is not complete will not matter, but in others it does and can cause the validity as well as the accuracy of the data to decrease. The completeness can be calculated by taking the number of fully complete data sets and dividing it by the total amount of data sets. Accuracy is a binary value, it is either accurate or not. The accuracy of the data measures the values of the data against a real-world example of that data. To calculate the accuracy of all of the data simply take the number of accurate data sets and divide it by the total number of data sets. Accuracy and consistency are two similar qualities of data. If a data set it consistent it means that the data record for one data set matches the record for the same data that was used elsewhere.

Timeliness measures the length of time since a record has been updated. If a data record is not updated or verified then it loses its validity over time. Any incomplete time or date must be flagged since it is an. A data entry is valid if it meets the constraints for each field in the data set. Uniqueness of the data means that there is no more than one copy of each set of data in the database.

Data analytics can be considered similar to the process of producing material goods because both take raw inputs and end with finished or processed outputs (see Table 1). This research investigates the similarities and differences between data analytics and product manufacturing.

By studying the analogy between product manufacturing and data analytics, the performance measures and the improvement strategies that have been developed and used in manufacturing a long time ago can be utilized in data analytics. By comparing information systems to production systems, raw data can be considered as raw material and the information or knowledge can be considered as the finished products. Process improvement approaches that have been used in manufacturing environment such as Lean and Six Sigma can be applied to data analytics.

TABLE 1. COMPARING DATA ANALYTICS & PRODUCT MANUFACTURING

Item	Product Manufacturing	Data Analytics
Inputs	Raw material	Raw data
Suppliers	Raw material owners	Data sources
Inventory	Collection of parts	Datasets
R&D Infrastructure	Labs and prototyping workshops	Data labs or discovery environment
Production Process	Manufacturing/assembly	Analytics/visualization
Defects	Defective products	Outlier data
Outputs	Products	Useful information

II. RELATED LITERATURE

The quality of data analysis has been given an extensive attention of many researchers including statisticians, mathematicians and computer scientists. However, only a few published studies have attempted to develop effective pedagogical technique for data analytics education, especially in higher education. In [1], Kennedy et al proposes a technique for teaching and learning data analytics using infographics tools. Results from their experiment shows 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 conducted in [2] compares the learning propensities of grades 10 to 12 high school students to data analytics education. However, none of these studies developed effective methods for teaching and learning data analytics to industrial engineering students in higher education. Additionally, these available studies did not consider the application of project-based learning to data analytics education. Project-based learning is a vital learning approach that helps students to acquire knowledge and skills by investigating real-world problems [3]. Our approach focuses on higher education and extrapolates product manufacturing processes to teaching and learning data analytics. We also, combined these with project-based learning to help students develop problem, critical thinking, and other benefits [3] provided by project-based learning approach.

Aside from the work reported in [1] and [2], there are other published worked on data analytics. But these do not focus on educational and project-based learning approach. For instance, a framework for evaluating the data quality was presented in [4]. A Multidimensional Robust Data Quality Analysis technique was proposed to improve the quality of data. The technique ensures the consistency of the database before and after the cleansing process. The model was applied in the field of Italian market domain. The results proved the usefulness of the developed approach in improving the cleansing process. In [5], metrics were proposed for investigating whether the value of a data attribute is still has the same value at the time of assessment. The metrics were developed based on probability theory. The value of the metric can be expressed as expected values so it would be easier to understand for a decision maker. A framework was developed for considering the quality preferences of data user while querying [6]. Three major issues related to quality aware query system were considered: measuring quality of data, modeling user preferences, and providing the query while considering the defined preferences and measures.

Various approaches have been utilized to assess data quality. For example, a comparative method was proposed to evaluate data quality [7]. The approach was used to evaluate a data for the behavior of a primate group. The collected data was compared with past data that was collected from the literature. Specific relevant criteria were taken into account. In [8], it was stated that restricting the participant in the sampling stage improve the data quality. The data was collected from high reputation workers to ensure the quality of Amazon Mechanical Turk. A “fit-for-use” model was proposed for assessing the quality of healthcare data that are collected electronic health record in either single or multiple sites [9]. The model involves prioritization of quality dimension, standardization of assessment approaches within and between sites, linking quality problems, and documentation of the outcomes.

With regards to the analogy between product manufacturing and data manufacturing. The use of total quality management as a tool to consistently define, measure, analyze, improve and control the quality of manufacturing data was discussed in [10]. In [11] and [12], it was stated that the data quality is similar to physical product as they are multidimensional problem. Some researches also viewed data as a product and used statistical process control (SPC) tools to maintain the data quality [13, 14]. A process centered approach was proposed to maintain the data quality [15, 16].

He utilized various tools such as histogram, cause and effects diagram and Pareto chart. After conducting the initial efforts that lead to in-control stage, process monitoring tools should be utilized to obtain the required quality level. Control chart can be used for monitoring and controlling the quality of data [17, 18, 19]. A robust algorithm for assessing the consistency of large volume of data was developed in [20]. The assessment process was conducted at different point of time for the collected data. An efficient computation method was utilized to estimate the expected values and variance of consistency metrics. Three sigma limits are used to reveal the inconsistent data. Also, Cumulative Sum (CUSUM) and Exponentially Weighted Moving Average (EWMA) plans are used to reveal the biased data.

In this paper, the analogy between data analytics and product manufacturing is discussed with the emphasis of enhancing student understanding of analytics. Before and after learning modules are proposed and course projects are used to evaluate students' understating of analytics.

III. Learning Modules

The goal of teaching data analytics to industrial engineering students is to introduce them to the fundamental concepts, methods, and tools of data analytics and visualization. The Industrial Engineering course discussed in this study is Engineering Analytics which is offered during the sixth semester. The course discusses the use of different quantitative techniques to identify patterns and trends in data to obtain insights from large data sets and communicate findings in useful terms. The course focuses on developing an in-depth knowledge of data storage, analysis, and visualization related to manufacturing and service domains. In particular, the areas of descriptive analytics, predictive analytics, and the use of Big Data in enterprise systems are covered from an engineering perspective. The three main objectives of the course are to: (1) provide both a theoretical and practical understanding of database and data analytics, (2) provide engineering students with simple yet effective ways to implement the database and data analytics concepts and techniques using Microsoft Office Access and Microsoft Office Excel, (3) illustrate the concepts and techniques with real case studies and course projects.

The main topics covered in the course are: Data properties and models, Database design and implementation, Structured Query Language (SQL), Data visualization, Ordinary Least Square (OLS) regression techniques, Logistic regression, Cluster analysis, Principal component analysis, unstructured data concepts, and Hadoop and MapReduce.

Two learning modules (before and after) that were designed to deliver the analytics knowledge to IE undergraduate students. The first module does not include the analogy between product manufacturing and data analytics (see Figure 3). Students were assign a course project on database implementation and data analytics. The students worked on the projects in groups of three or four. Figure shows the main steps of the course project.

In the second learning module (see Figure 4), the students were exposed to the analogy between data analytics and product manufacturing. The Industrial Engineering students take several courses on product design and manufacturing and they were already familiar with these topics when the analytics course was offered. The course project steps were linked to the product design and manufacturing steps.

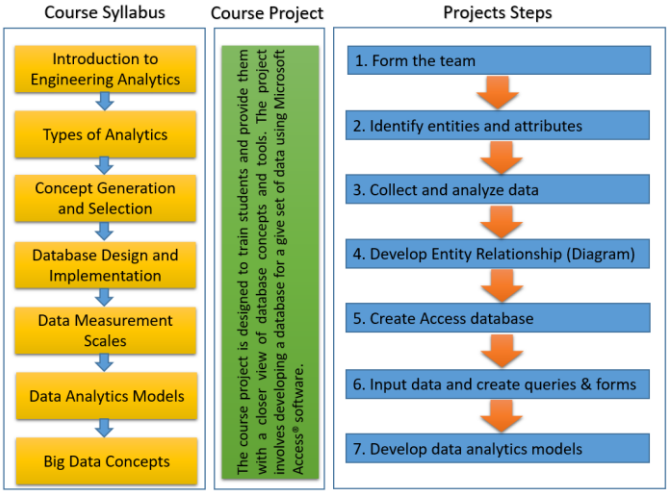

Fig. 3. First learning module

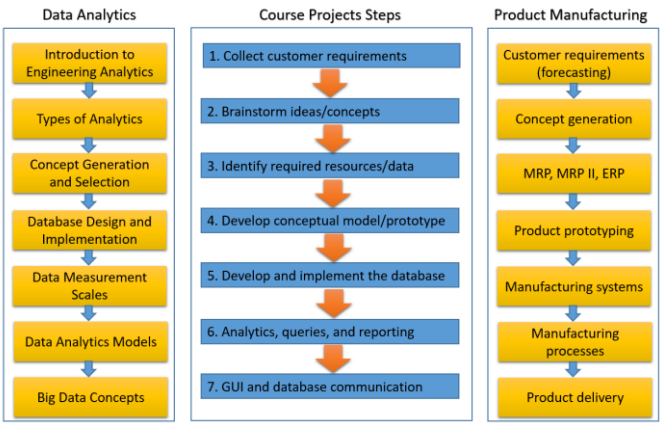

Fig. 4. Second learning module

IV. Course Project Description

The course project is designed to train students and provide them with a closer view of database concepts and tools. The project involves developing a database for a given set of data using Microsoft Access® software. The students started by collecting customer requirements about the problem and brainstorm some ideas for developing the database. Then, the required data and resources were identified. To create a

database, the data has to be analyzed to determine how the data can be organized most efficiently [21]. This includes understanding the entities and thier relationships. Students will also create queries, forms, and reports. Students were assigned two data sets; the first data set is for a product manufacturing company and the second data set in on a real estate company. Each data set consists of seven or eight tales that will be used to create a database and develop analytics models. An Entity Relationship Diagram (ERD) is created to visually show the entities, attributes, and relationships of the datasets. Figure 5 shows an example of ERD for the first data set. The ERD represents the conceptual model (or prototype) for the database.

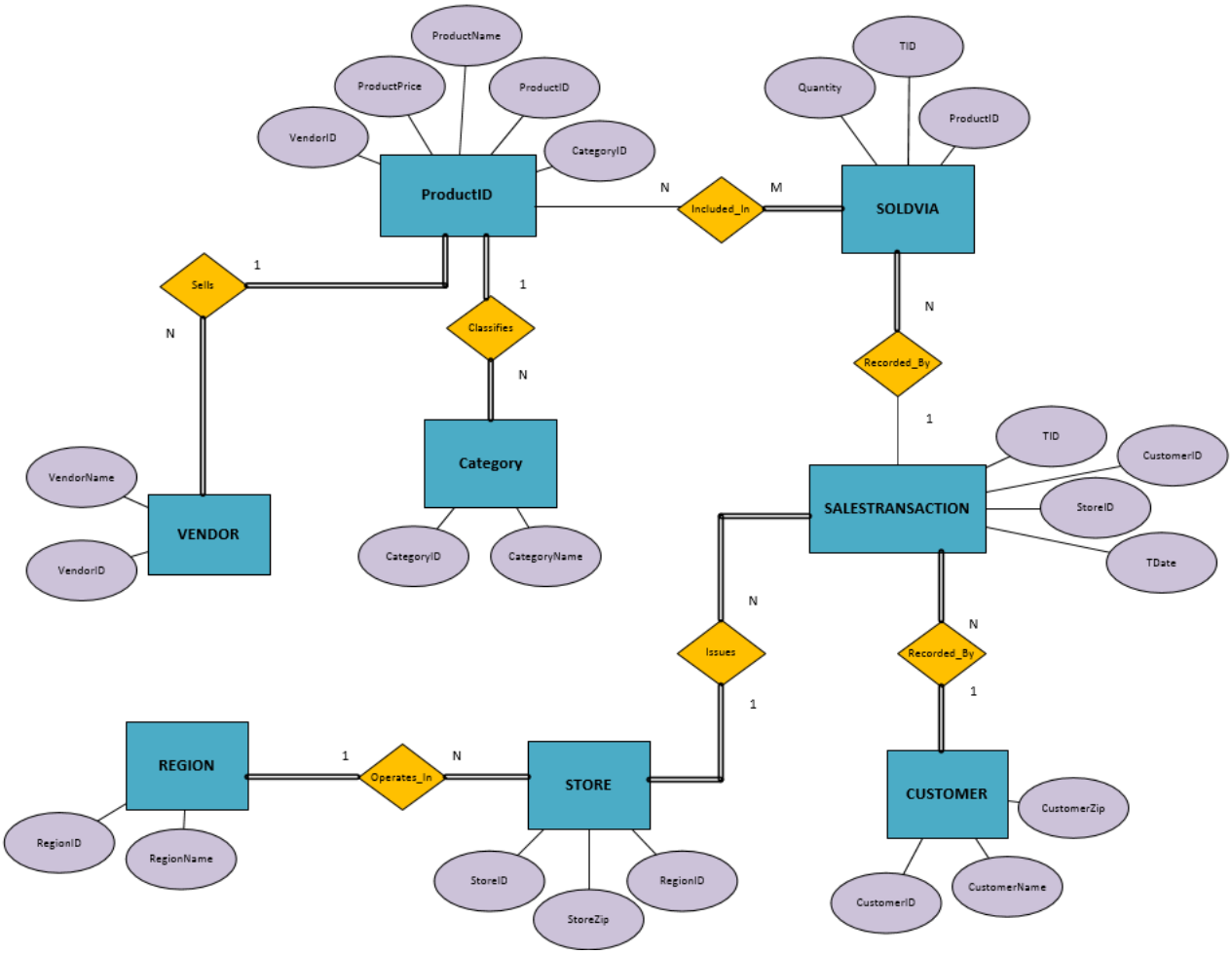

Fig. 5. ERD for the course project database

An example of a sample report generated by the one query is shown in Figure 6. The graphical user interface (GUI) for the database is shown in Figure 7. The main GUI is linked to the sub-forms and queires as wlll as analytics results. An example of an SQL query to retreive the number if items sold in each region is shown below.

```
SELECT region.regionname, Count(soldvia.noofitems) AS
CountOfNoOfItems
FROM region INNER JOIN ((store INNER JOIN salestransaction ON
store.StoreID = salestransaction.StoreID) INNER JOIN soldvia ON
salestransaction.TID = soldvia.TID) ON region.RegionID = store.RegionID
WHERE (((region.regionid)=store.regionid) And
((store.storeid)=salestransaction.storeid) And
((salestransaction.tid)=soldvia.tid))
GROUP BY region.regionname;
```

Product Report — Monday, March 26, 2018 4:01:02 AM

ProductID	ProductName	ProductPrice	VendorID	CatergoryID
1x1	Zzz Bag	$100.00	PG	CP
1x2	Comfy Harness	$150.00	WL	CL
1x3	Sunny Charger	$125.00	OA	EL
1x4	Safe T-Helmet	$40.00	PG	CY
2x1	Mmm Stove	$80.00	WL	CP
2x2	Easy Boot	$70.00	MK	FW
2x3	Reflect-o Jacket	$35.00	PG	CY
2x4	Strongster Carribe	$20.00	MK	CL
3x1	Sleepy Pad	$25.00	WL	CP
3x2	Bucky Knife	$60.00	WL	CP
3x3	Cosy Sock	$15.00	MK	FW
3x4	Treado Tire	$30.00	OA	CY
4x1	Slicky Tire	$25.00	OA	CY
4x2	Electra Compass	$45.00	MK	EL

Fig. 6. A sample generated report

Fig. 7. GUI for the developed database

V. DISCUSSION

In this Section, we present statistical analysis of the student grades before and after implementing the proposed approach. Student grades for two semesters was collected and analyzed. Table 3 shows the summary of the analysis where both the mean and standard deviation of the course project grades were compared. The results are also plotted in Figure 8. It can be concluded from the results that both the mean and variability of the students' grades have improved after implementing the second learning module.

TABLE 3. COMPARISON FOR THE COURSE PROJECT

Semester	n	Mean	Std. Dev.
Before	24	87.26	11.26
After	17	92.38	2.07
P-value		0.039	0.000

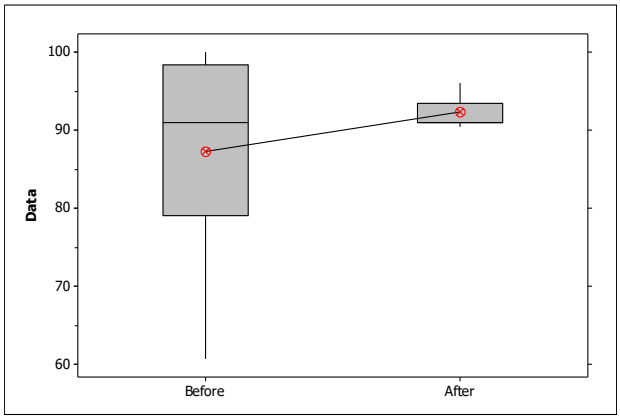

Fig. 8. Comparing the project grades using box plot

We also compared the exam scores for both cases, "Before" and "After" (see Table 4 and Figure 9). It can be noted that the means of the exams ("Before" and "After") grades are not statically different whereas the variability in the

exam grades is statistically different and it was reduced after introducing the analogy between data analytics and product manufacturing. This indicates that the analogy between analytics and manufacturing can reduce the differences in understanding the topics among the students. However, more data should be collected in the future to further support this hypothesis.

TABLE 4. COMPARISON FOR EXAM SCORES

Semester	n	Mean	Std. Dev.
Before	24	83.81	11.06
After	17	86.88	6.92
P-value		0.275	0.028

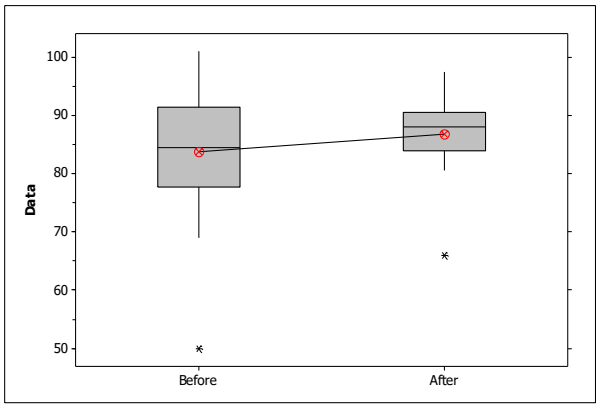

Fig. 9. Comparing exam grades using box plot

VI. CONCLUSIONS AND RECOMMENDATIONS

In this paper, we discussed the analogy between data analytics and product manufacturing and how this analogy can improve students' understanding of analytics. Students were assigned course projects on database and analytics and their performance was evaluated before and after the analogy of analytics and manufacturing was introduced. The results showed that the analogy can help students to understand the topics and perform better on solving problems.

Future work will focus on collecting more data on the analogy as well as expanding this idea to other topics. Moreover, an online simulation for "data manufacturing" can be developed to simplify the understanding of database and data analytics. Other factors such as the overall GPA of the students and their interests in manufacturing and data analytics topics can also be considered.

REFERENCES

[1] J. C. Nwokeji and P. S. T. Frezza, "Cross-course project-based learning in requirements engineering: An eight-year retrospective," *IEEE Frontiers in Education Conference (FIE)*, Indianapolis, IN, pp. 1-9, 2017.

[2] J. Kennedy, P. Abichandani, and A. Fontecchio, "Using infographics as a tool for introductory data analytics education," *IEE Frontiers in Education Conference (FIE)*, Madrid, Spain, pp. 1-4, 2014.

[3] J. Kennedy, P. Abichandani and A. Fontecchio, "An initial comparison of the learning propensities of 10 through 12 students for data analytics education," *IEE Frontiers in Education Conference*, Oklahoma City, OK, pp. 916-918, 2013.

[4] M. Mezzanzanica, R. Boselli, M. Cesarini, and F. Mercorio, "A model-based evaluation of data quality activities in KDD," *Information Processing & Management*, vol. 51, no. 2, pp. 144-166, 2015.

[5] B. Heinrich, and M. Klier, "Metric-based data quality assessment-Developing and evaluating a probability-based currency metric," *Decision Support Systems*, vol. 72, pp. 82-96, 2015.

[6] N. K. Yeganeh, S. Sadiq, and M. A. Sharaf, "A framework for data quality aware query systems," *Information Systems*, vol. 46, pp. 24-44, 2014.

[7] S. K. Patterson, A. A. Sandel, J. A. Miller, and J. C. Mitani, "Data quality and the comparative method: The case of primate group size," *International Journal of Primatology*, vol. 35, no. 5, pp. 990-1003, 2014.

[8] E. Peer, J. Vosgerau, and A. Acquisti, "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk," *Behavior Research Methods*, vol. 46, no. 4, pp. 1023-1031, 2014.

[9] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner, "A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research," *Medical Care*, vol. 50, pp. 21-29, 2012.

[10] R. Y. Wang, "A product perspective on total data quality management," *Commun. ACM*, vol. 41, no. 2, pp. 58-65, 1998.

[11] D. A. Garvin, "What does "product quality" really mean?" *Sloan Manage. Rev.*, vol. 26, no. 1, pp. 25-43, 1984.

[12] D. A. Garvin, "Competing on the eight dimensions of quality," *Harvard Bus. Rev.*, vol. 65, no. 6, pp. 101-109, 1987.

[13] D. Dey, S. Kumar, "Reassessing data quality for information products," *Manage. Sci.*, vol. 56, no. 12, pp. 2316-2322, 2010.

[14] A. Parssian, S. Sarkar, and V. S. Jacob, "Assessing data quality for information products: impact of selection, projection, and Cartesian product," *Manage. Sci.*, vol.50, no. 7, 967–982, 2004.

[15] T. C. Redman. *Data Quality: Management and Technology*. Bantam Books, New York, 1992.

[16] T. C. Redman. *Data Quality for the Information Age*. Artech House Publishers, Norwood, MA, 1996.

[17] T. C. Redman, "The impact of poor data quality on the typical enterprise," *Commun. ACM*, vol. 41, no. 2, pp. 79-82, 1998.

[18] T. C. Redman. *Data Quality: The Field Guide*. Digital Press, Boston, MA, 2001.

[19] W. H. Woodall, "Controversies and contradictions in statistical process control," *J. Qual. Technol*., vol. 32, no. 4, pp. 341-350, 2001.

[20] R. Sparks, and C. OkuGami, "Data quality: algorithims for automatic detection of unusual measurements," *International Workshop on Intelligent Statistical Process Control*, Seattle, WA, 2010.

[21] J. C. Nwokeji, T. Clark, B. Barn, V. Kulkarni and S. O. Anum "A data-centric approach to change management," *IEEE International Enterprise Distributed Object Computing Conference*, Adelaide, SA, pp. 185-190, 2015