4-20-2021

Integrating Common Data Analytics Tools into Non-Technical Undergraduate Curricula

Kurt Kirstein

Integrating Common Data Analytics Tools into Non-Technical Undergraduate Curricula

Kurt Kirstein, EdD

Central Washington University, USA, kurt.kirstein@cwu.edu, https://www.cwu.edu

Abstract: Aside from statistics courses, accessible data analytics skills are often excluded from traditional non-technical university programs. These are topics that are typically the domain of programs that focus on math, statistics and computer science. Yet the need for these skills in non-technical disciplines is changing. A rapid expansion of data-related processes in organizations of many types requires individuals who have at least a working knowledge of common analytic tools. This article briefly describes three categories of data analytics tools that can be useful for graduates in any discipline. The first category covers descriptive tools that allow students to learn what is in a data set and what meaning can be made of it. The second category of tools teaches students how to predict likely outcomes based on relationships in past data. The final category introduces students to tools that allow them to segment data into useful clusters and classes and to build meaningful associations within the data.

Keywords: Data analytics, Non-technical disciplines, University curriculum, Common data tools

Introduction

In recent years, data has had an ever-increasing impact on many aspects of organizations. Mayer-Schonberger and Cukier (2013) point out how companies are often regarding their data as a valuable commodity that can drive strategic and economic decisions. Data provides companies with opportunities to personalize their customer interactions, forecast product demand, and predict market behavior (Rosidi, 2019). And it is not just large companies with "big data" that are taking advantage of this trend. Great insights are possible even in organizations where the data sets are limited in size.

As with many industry shifts, especially those of a technical nature, the skills of the labor force can lag behind the demand for them. This has been the case with data skills (Evans, 2019; Rosidi, 2019). Much of the expertise for handling data has been in the disciplines of statistics and computer science, which are contributing to new domains such as data science or machine learning (Taddy, 2019). However, centralized expertise limits organizations who are dependent on a small number of individuals to fully recognize the value that can be derived from data. It is especially unfortunate when the data knowledge required is readily accessible to

Kirstein, K. (2021). Integrating Common Data Analytics Tools into Non-Technical Undergraduate Curricula. In R. Thripp & I. Sahin (Eds.), *Proceedings of ILSET 2021-- International Conference on Life Sciences, Engineering and Technology* (pp. 1-10), New York, USA. ISTES Organization.

organizational leaders who do not need to be skilled in statistics and computer science to derive value from their data. Too often, they just lack the training.

As the expansion of data continues, the need for data skills in non-technical disciplines is changing. A rapid expansion of data-related processes in organizations of many types requires individuals who have at least a working knowledge of common analytic tools. A collection of such tools is beginning to find its way into the curricular plans of certain disciplines. These tools are practical and useful to help leaders solve common organizational problems or capitalize on strategic opportunities and they don’t require advanced technical or statistical degrees.

This remainder of this article briefly describes three categories of data analytics tools that can be useful for graduates in any discipline. The first category covers descriptive tools that allow students to learn what is in a data set and what meaning can be made of it. The second category of tools teaches students how to predict likely outcomes based on relationships in past data. The final category introduces students to tools that allow them to segment data into useful clusters and classes and to build meaningful associations within the data. Together these make up a analytics skill set that can be infused into the curriculum of university programs that traditionally have not included any focus on data.

The De-mystification of Data Analysis

For many companies, the approach to data analysis and similar skills has been to focus them largely within technical units. They have been seen as skills that reside outside of traditional non-technical disciplines. To make use of data requires the assistance of “data people” who can bring the expertise necessary to turn raw data into something meaningful. This traditional approach has been known to create the false impression that even foundational data skills are inaccessible and are best left to the technical “data people”.

This is now changing and that change is largely being driven by an expansion in the availability of data. Some disciplines (e.g. business programs) are beginning to recognize the advantage of including data skills in their curriculum but the need is outpacing practice. More university programs need to adopt the inclusion of data skills. In a practical sense, there are a number of data analysis tools, not traditionally built into university-level programs, that could be helpful in addressing common scenarios that future graduates are likely to face. Additionally, it may be best if these skills are not only an elective part of a non-technical program but are tightly infused into several of the core courses that make up those programs.

Any data analysis instruction, built into university programs, should de-mystify these tools and underscore their applicability to common operational problems. When students begin to understand these tools, they can build their own data analysis skill sets leading them to higher levels of productivity and efficacy in their careers. The remainder of this article will largely reference business students and their professional concerns but this is just one example of a non-technical discipline where data skills can have a significant impact on the employability of graduates.

A Data Analysis Skill Set

To successfully integrate data analysis into non-technical education requires a re-framing of students' understanding of data analysis. Sometimes, students approach data instruction with trepidation because they believe themselves to be "not good at math or statistics" and they are hesitant to try. In other cases, graduates have had little experience with data because their programs did not include this type of content into the core instruction of their discipline. However, non-technical students can be successful in the use of these tools when they are explained in a real-world context and used in authentic instructional scenarios. Students can learn these skills through simulated, real-world practice as they begin to build their own data analysis skill sets.

A student's analytic skill set can consist of three parts: data discovery, prediction, and segmentation. In data discovery, students learn to visualize their data, clean up any problems such as missing or flawed data, and reduce its dimensionality to narrow their focus. In prediction, students learn to use past data sets to make predictions about the future. These tools include forecasting, simple and multiple regression, and logistic regression. In segmentation, students learn to understand the relationships and associations of different parts of their data, helping them to focus their analysis on specific groups or scenarios. Such tools include clustering, classification, and association rules. In the sections that follow, each of these tools and their applicability to business professions will be discussed in more detail.

Discovering the Data

Data can come from a number of sources. Regardless of how the data are accessed, the first step in data analysis is to discover what is in the data set. This can help inform later decisions about the most appropriate direction for analysis. In some cases, part of data discovery can be duplicitous because the contents of the data file are already known, or they match the fields that were asked for when the output was requested. In many cases, however, the data analyst must take time to learn what they have. Even when the fields that make up the data file are known, there three important steps to the data discovery process: visualization, data cleansing and data reduction.

Visualization

Visualization can be the first stage of the data analysis process. It serves many purposes, but its main use is as an introduction to the scope and nature of the data. Visualization can be conducted using numeric summarizations or graphical tools. The goal is the same which is to explore data and provide an effective way to present results.

Visualization techniques are primarily used in the preprocessing portion of the data analysis process (Shmueli et al., 2018). They can help us identify clear errors in the data (e.g. customers whose age is 999), replace missing values, remove duplicate rows, and deal with other formatting or content errors that must be corrected or accounted for. Beyond ensuring a clean data set, data visualization techniques support freeform exploration for the purpose of understanding the data structure, identifying interesting patterns, and generating novel questions (Shmuli et al., 2018).

Cleaning the Data

Closely aligned with data visualization is the process of data cleansing. Two key problems with many datasets are missing data and outliers, although data duplication can be an issue as well. At this stage, the questions to be asked are: How should missing data to be handled? Do the ranges of data points make sense or are there obvious outliers? If outliers exist, what strategy should be employed to handle them? Did all data fields convert correctly (e.g. dates and units of measure)? Are there any sections of the dataset that have been duplicated?

Missing data can be a frequent problem in datasets. Should a dataset have empty fields, different applications will handle them in different ways and some data analysis tools will simply fail to run until the missing data problem is resolved. Strategies for handling missing data include row removal or value replacement (with column mean). Future data analysis needs can often drive decisions regarding the handling of missing data.

In general, there are two types of outliers. The first is obvious errors in the data. The second are legitimate but extreme values that fall well outside the expected ranges for variables. How each type is handled is up to the analyst. Removal of outliers that are clearly errors can be accomplished by deleting the impacted rows entirely or by replacing erroneous outliers with the column mean that excludes the outlier values. Extreme but legitimate values require more consideration of both their importance and their impact on the analysis. It may be best to remove extreme values that do not impact later analysis. In other cases, the inclusion of extreme values is necessary and the analyst will need to anticipate the impact on results.

Two other challenges that may be encountered at the data cleansing stage are field formatting and data duplication. It is not uncommon that numeric fields, such as currency or dates, lose their proper formats when transferred from one platform to another. This is generally easily fixed by specifying the format of the fields in a separate step. Data duplication can be found using sorting, visualization or other tools that show the number of times that values of a key variable appear in a data set. Erroneously duplicated data rows, once identified, can simply be removed.

Data Reduction

Once the data are visualized and cleaned, and the analyst is familiar with the contents of the data set, more informed decisions can be made about the how the data set will be analyzed. This may lead to data reduction or the removal of fields that will not be used in later analysis. Unneeded columns can simply be deleted from the data set (with proper backups made) but this process should be done judiciously so that further reviews, suggested by the initial analysis, are not hampered or rendered impossible because of missing data.

In some cases, analysis of the data is simplified by removing rows of unneeded records. Regardless of method, the goal of data reduction is to simplify the dataset and the analysis. It is largely driven by the questions that the analysis of the dataset is intended to answer. It allows the analyst to focus energy and effort only on those parts of the dataset that will contribute to the analysis without requiring extra time, storage, or effort.

Learn from the Past

There is a great deal that companies can learn from old data. Information about products that customers bought, or did not buy, is often captured in datasets and this useful in determining product design, feature enhancement, and marketing strategies. Past data can tell analytsts how much customers are willing to spend, which customer segments they should be focused on, information about customer retention or attrition, the likelihood that customers will be interested in our products and how many of those products will need to be produced to precisely meet customer demand (Siegel, 2016).

Included in an analyst's data tool kit should be methods of using past data to predict future customer behavior. At a minimum, analysts should be familiar with forecasting, linear regression, and logistic regression. Each of these are described further in the sections that follow.

Forecasting and Time Series Analysis

Forecasting future customer behavior can be a key part of an organization's success. It can be one of the most important methods of ensuring that the right number of products are available when customers want them. Underproduction can lose customers who go to competitors when products are not available, and overproduction leads to waste. Product forecasts must be as accurate as possible and a key method for forecasting future sales is to include information about past sales.

Time series analysis uses past data to predict future performance by tracking key variables over intervals (months, quarters) to determine how those values have changed over time. When using a time series for forecasting, both the trend and the seasonality of past data must be considered. Line charts and other visualization tools can help the analyst see such patterns. Through this process, it is easy to determine if there an up or downward trend or if the data is flat and consistent with limited variability. It is also possible to detect patterns in the data that suggest seasonality where there are consistent and predictable peaks and valleys in the values. These questions become an important part of the strategy that can be used to determine future forecasts.

Linear Regression

Another tool that can be used for prediction and forecasting is linear regression. With regression, an analyst may be attempting to predict numeric values of an output variable by considering relationships in past data. Regression uses two sets of variables: predictor and output. In more complex models, there can be multiple predictor variables used to predict the output.

The process of building a regression model is to look at how the predictors were related to the output variable in past data. From that analysis, the model provides coefficients that define the relationship between the predictor and output variables. These coefficients can then be applied to scenarios were predictor variables are known but the output values are not. By doing so, predicted output values can be produced. The accuracy of the regression model is highly dependent on the past relationship between the predictor and output variables and on the volume of data that is used to "train" or build the model.

Logistic Regression

While we may be looking for predicted numeric values in linear regression, logistic regression allows the prediction of likelihood or probability. Similar to linear regression, the relationship between predictor and output variables is determined. Also, similar to linear regression, there can be single or multiple predictors. But with logistic regression the output is converted to a probability, ranging from 0 to 1. Logistic regression helps determine the likelihood of customer behavior by looking at the relationship between predictor variables and customer behavior in past data. A common use of logistic regression is in loan approval decisions. Predictors such as credit score and income can be used to determine the likelihood of loan repayment which drives the loan approval decision.

Data Segmentation

For a variety of purposes, many companies find it useful to segment their customers into groups based on patterns of behavior (Linhof & Berry, 2011). Customers are separated based on their purchasing patterns, income level, frequency, or a combination of similar fields. Each customer is assigned to a group according to which categories or combinations of categories they match best. Then organizations can mount targeted campaigns leading to improved returns and expanded customer relationships. The expectation is that by focusing on groups of similar customers, these efforts will be more effective than a one-size-fits-all approach.

Applying segmentation to large customer transaction databases helps companies to understand the behavior of their different customer groups including what they are buying and how much they are spending. Each segmented group of transactions becomes its own cluster. Once the transactions are segmented into clusters, association rules can be built on each separate cluster further personalizing the marketing approach that a company can take.

The following sections will cover three tools that can be used for segmenting or associating data: clustering, classification and association rules. These are useful tools for a number of scenarios where there is benefit to understand grouping and associations that are found in a data set. All three should be part of the data analyst's skill set.

Cluster Analysis

In simple terms, clustering allows a large dataset to be broken into smaller sets based on the values of selected variables. These variables could be customer demographics, purchase frequency, or product selection. The goal is to identify subsets of customers who can be associated with each other because they have similar patterns across selected variables. Then companies can reach out to those subsets using customized approaches.

There are two common methods for clustering a large dataset. The first divides the data into a pre-determined number of subsets based on a geometric determination of distance. K-means clustering is one common method of dividing data into groups. The newly clustered records can be reviewed in both graphical and tabular formats making it easy to see which cluster each record has been assigned to. One downside of the k-means clustering process is that it forces all records to be in a limited number of k clusters and that can be complicated by outliers.

The second form of clustering is agglomerative or hierarchical clusters. Under the agglomerative clustering process, all records start out as their own cluster. Next, the two records that are most similar to each other, across a selected set of variables, are joined. The process of joining similar records continues until every record is part of a single cluster. The value of this strategy is that the entire process of joining records is captured and represented in a graphical tree structure called a dendrogram (Linhof & Berry, 2011). Similar records are on the same branches and the lengths of the branches indicate the level of similarity.

Both clustering tools help to separate customers or transaction records into smaller sections, each of which can be treated separately which helps to build stronger customer relationships. It is important to note that k-means and agglomerative clustering are only two of the many different types of clustering available to data analysts. There are many more accessible clustering methods, the use of which depends on the analyst's needs.

Classifiers

Very similar to clustering is the process of classification and, like clustering, there are many methods of classifying. One method that is easy to understand and set up is the k nearest neighbor (kNN) classification process. As was the case with k-means clustering, similarity is most often determined by distance which is usually a geometric measure of distance between the new record and the closest k records. Each existing record has already been assigned to a class. The new records are compared to their closest neighbors and the class represented by highest number of close neighbors becomes the class of the new record.

Classification can be used to learn more information about new data. From a customer relationship standpoint, it is useful to determine how to approach new customers who do not have established buying patterns. The assumption is that customers who are similar, in many aspects, to already known customers are likely to have similar purchasing patterns and this can help define relationship strategies for unknown customers.

Association Rules

The creation of association rules, also often known as market basket analysis, allows companies to determine item sets or collections of products that have a higher likelihood of being purchased together. Through careful analysis of purchase records, confidence measures can help companies determine product sales as a function of their associations with other products. Counting the number of items of each type that are purchased and then counting associations of items purchased together allows for the determination of the likelihood that items will be bought together.

Culling through a transaction database makes it possible to create meaningful association rules leading to different approaches to product placement and marketing. Products with strong associations can be placed together in a store on a website encouraging customers to purchase both (or all) products at once. Beyond just a retail application, understanding associations between events can have far wider applications.

Accessibility of Analysis Tools

The previous sections described a number of analysis tools and a key argument of this article is that these tools are and should be accessible to a wider group of individuals who will benefit from their use in an increasingly data-driven environment. While advanced and deep-level analysis of large data sets may remain primarily the domain of a company’s technical professionals, there are a number of widely accessible packages that provide access to many of the technical tools described in this article. These tools can be also included as a key part of non-technical university programs.

Microsoft Excel is, perhaps, the most widely accessible tool and its Data Analysis add-on package extends its capabilities in important ways. Certainly data discovery, visualization, cleansing and reduction can all be done using the standard version of Excel but the newer versions and the Data Analysis add-on package provide tools such as forecasting and linear regression. Rudimentary segmentation, that resembles clustering, is also available in Excel but more advanced tools that can be used for logistic regression, clustering, classification and association rules can be accessed using open source scripting languages such as R or Python. Should analysts have access to an advanced statistical package, such as SPSS, the use of these more advanced tools is made even easier.

Data analysis tools are useful at many levels in a company and the tools needed to conduct effective analysis are easily accessible. What may stop non-technical majors from using these tools upon graduation may have more to do with the extent to which they were, or were not, effectively covered in their university curriculum. Thus, there is an increasing need to infuse the use of these tools throughout the curriculum and to cover their use in real-world, authentic learning scenarios.

Conclusion

This article contains a brief description of analysis tools that could make up a non-technical analyst's data skill set. It is intended to describe common tools that are accessible and that should be included in a non-technical university education where they have the potential to be useful for graduates seeking employment in a data-driven economy. They are just a subset of analysis skills, yet ones that are useful and relevant to many of today's current data challenges.

But tools are just tools. The primary driver in any effort to utilize data and data analysis tools should be an analytic mindset, one in which questions are asked or opportunities are identified that drive data exploration (Rosidi, 2019). This is similar to the generation of hypotheses that drive research. Directive questions and goals should be identified at the outset of any data project as the selection of data sources and analysis methods are dependent on them. A data project without a purpose is analogous to a journey without a destination.

When planning university curricula, instructors need to consider the ongoing digital disruption that is currently impacting so many fields (Arthur, 2013). It is not going away and employees in today's companies need to adapt to the technological demands that are being placed on them. This includes data and the ability to analyze it. No longer can the responsibility for data analysis be relegated to technical departments who are likely overloaded with their own demands. Data analysis skills should and must become a standard part of the education that prepares future graduates, in many disciplines, for a data-driven marketplace.

References

Arthur, L. (2013). *Big data marketing: Engage your customers more effectively and drive value*. Wiley.

Evans, J. R. (2019). *Business analytics: Methods, models and decisions*. (3rd ed.). Pearson.

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data*. Houghton Mifflin Harcourt Publishing.

Linhof, G. S., & Berry, M. J. A. (2011). *Data analysis techniques for marketing, sales, and relationship management* (3rd ed.). Wiley.

Patel, N. (2019). *10 ways data analysis can help you get a competitive edge.* Retrieved from https://neilpatel.com/blog/data-analysis/

Rosidi, N. (2019). *How I teach analytics to non-technical students*. Retrived from https://towardsdatascience.com/how-i-teach-analytics-to-non-technical-students-2db4a900f0cf

Siegel, E. (2016). *Predictive analytics: The power to predict who will click, buy, lie or die*. Wiley.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2018). *Data analysis for business analytics: Concepts, Techniques, and applications in R.* Wiley.

Taddy, M. (2019). *Business data science: Combining machine learning and economics to optimize, automate and accelereate business decisions*. McGraw Hill.