INTRODUCING ADVANCED EXPLORATORY DATA ANALYSIS TOOLS IN AN MBA PROGRAM

Joseph McCollum, Siena College
Shahryar Gheibi, Siena College
Necip Doganaksoy, Siena College

ABSTRACT

This paper proposes an approach which business analytics instructors can utilize in order to incorporate advanced exploratory data analysis (EDA) tools into an MBA analytics course. With the recent rise of analytics in business and industry, a new set of advanced analytics tools have proved significantly more effective than simple statistical techniques traditionally used in practice. We address the challenge for academia to augment the students' data analysis skills and to equip them with the analytics tools that prepare them for their future careers.

We, specifically, demonstrate applications of partitioning analysis and association rules mining techniques. Partitioning enables students to incorporate multiple variables into the analysis and obtain results that are easy to interpret. Association rules mining helps enhance the EDA and obtain deeper insights into the relationships among several categorical variables using fundamental concepts of probability theory. We motivate the students' learning by analyzing a practical and relatable dataset which pertains to the railroad safety.

We believe this approach will help promote the pedagogy of business analytics and data science in academic institutions.

***Key words**: Exploratory Data Analysis; Association Rules Mining; Partitioning Analysis; Railroad Safety*

INTRODUCTION

In this age of rapid technological advancements, organizations have become able to collect increasing amounts of data. They eagerly are looking for ways to capitalize on their databases to promote data-driven decision making and enhance their business. It is not surprising that demand for skillful data analysts, who are well trained in utilizing advanced analytics tools, is rising.

The analytics and data science academics have already realized the call for integrating quantitative analysis into their courses (McClure and Sircar, 2008). To practitioners, using data to support and improve the decisions is not new. Traditionally, practitioners have successfully used a number of elementary statistical tools. For example, histogram, control chart, Pareto chart, scatter plot, multi-vari charts and star plot, among others, have been used to seek the root causes of operational problems. However, in recent years the effectiveness of simple tools has been increasingly challenged by growing volumes of operational data containing numerous variables. With the recent rise of analytics in business and industry, a new set of advanced tools for data analysis have proved effective and become readily accessible to practitioners. The challenge, therefore, for data analytics instructors is to not only help students learn fundamentals of statistics and data analysis, but also to enable them to utilize advanced analytics tools (Aasheim et al, 2015; Chaojiang et al, 2015).

The need for introducing data analysis courses in the MBA programs, in particular, has been well recognized. Warner (2013) suggests further that the topics may be covered in one course or spread through several courses within the program. Soule et. al (2018), for example, describe how they created a data analytics course at Nicholls State University in Louisiana. We have taken a similar approach in developing business analytics courses in the Business Analytics concentration of our MBA program. In this paper, we demonstrate how we have utilized a real-world dataset to incorporate advanced exploratory data analysis (EDA) tools into an MBA statistics/analytics course.

The urgent need to revise the first course in statistics has been a long-standing debate within the statistics and quantitative education community. The traditional approach favors mathematical foundations of statistical inferencing and confirmatory analysis (p-values, significance testing, and cookbook formulas for analysis) as opposed to skillful analysis of real life datasets through graphical data visualization.

Horton (2015) advocates revising the first course in order to (1) broaden the role of multivariate thinking and the basics of confounding, (2) develop data-related skills early, and (3) expand the role of simulation and computation. Hartenian and Horton (2015) describe a case study to illustrate how to incorporate these ideas into an introductory level course. They use the Rail Trail and Property Values dataset which includes information on a set of n = 104 homes which were sold in Northampton, Massachusetts in 2007. The dataset contains house information (square footage, acreage, number of bedrooms, etc.), price estimates (from zillow.com) at four time points, location, distance from a rail trail in the community, biking score, and walking score. The dataset is amenable to use with EDA in courses with a focus on visualization, multivariate analysis and sophisticated graphics.

In this paper, using a real-world data set with similar features to those of the dataset used by Hartenian and Horton (2015), we show how graphical data analysis, recursive partitioning and association rules mining can help explore and understand a set of complex data, and shed light on the main drivers of a defective outcome. Specifically, we illustrate the use of partitioning analysis to explore the cost drivers of railroad accidents in the U.S. Partitioning enables us to incorporate multiple variables into our analysis, and obtain results that are easy to interpret by students. Furthermore, we show how association rules mining can be utilized as an efficient tool for discovering and analyzing relationships among numerous categorical variables using fundamental concepts of probability theory. The association rules analysis allows us to augment our EDA and obtain deeper insights into the railroad incidents.

It is worth mentioning that incorporating partitioning analysis and association rules mining into an MBA statistics/analytics course provides a great opportunity for us to enhance students' knowledge of advanced analytics software packages as well. We, specifically, have used R and JMP for the association rules mining and partitioning analysis, respectively. R can be downloaded for free at https://cran.r-project.org/ and it is one of the most-widely used statistics and analytics software packages. Using R effectively, however, requires competency in its scripting language. JMP is a commercial software package by SAS Institute (https://www.jmp.com/). It offers powerful functionality from data visualization through high-end predictive modeling. In general, students are able to gain faster competency with JMP owing to its user-friendly features. Using both R and JMP provides a nice balance between software packages' availability and ease-of-use, which have extensively been discussed in the literature (Liu, 2016; Warner, 2013; Meyer, 2016). We note, however, that both association rules mining and partitioning analysis can be conducted in either R or JMP effectively. Both features are available in other software packages as well (e.g., SAS, SPSS, and Python).

In what follows, we first discuss the dataset we have used and the goals of the analysis. Next, we demonstrate some graphical data analysis. Then, we present the partitioning and association rules analysis in details. Finally, we conclude by summarizing the value each of these advanced tools can add to any MBA curriculum.

DESCRIPTION OF THE DATA

In competitive business environments, companies strive to find ways to differentiate from competitors. This is true for railroad companies as they face stiff competition from within the railroad industry and other transportation industries such as trucking, air, and marine. Safety is one of the key performance indicators of a railroad company, and receives a great deal of attention from the public and the shareholders as well (Basch, 2018). In our EDA, we examine the recent railroad incidents in the U.S.

We obtained our dataset from the Federal Railroad Association (FRA) website (https://www.fra.dot.gov/Page/P0037). The dataset –which we occasionally refer to as “railroad incidents dataset” hereafter—is developed based on the safety reports that every railroad company has to submit to the FRA through the 6180.54 form on a regular basis. The 6180.54 form includes more than thirty variables of study ranging from incident number, company name, type of accident, and type of rail to weather condition and visibility at the time of incident (accident). The variables which we refer to in our analysis are described in Table 1.

It is worth noting that railroad incidents dataset offers features of typical datasets encountered in business applications: It contains more than a dozen variables with mixed types, i.e., qualitative and quantitative, and provides a good opportunity for students to apply their analytics skills to practical business situations. Moreover, the context of the application is readily relatable by students irrespective of their past training and experience. The learnings from this example can be generalized to other situations as well.

Table 1
Main variables of interest in the railroad incidents dataset

Variable	Description
TypeOfAccident	Has 13 options including derailment, head-on collision, read-end collision, side collision, highway-rail crossing, obstruction, explosion, and others
TypeOfEquip	Has 14 options including freight train, passenger train pulling, single car, yard/switching, and others
TypeOfTrack	Either main, yard or siding
TrackClass	Dictates the maximum speed allowed
HIGHSPD	The maximum speed allowed by the Federal Track Safety Standards for freight trains on the track where the accident occurred (this speed depends on the track class as described in Table 2)
ACCDMG	The total reportable damage in dollars on all reports submitted for the accident/incident
ACCAUSE	One of the 390 causes that can be used to characterize the main cause of the accident/incident
CauseCode	Either E (electrical), M (miscellaneous), T (track), S (signal), or H (human)
AMPM	Time of the accident/incident: either AM or PM
Visibility	Either Dawn, Day, Dusk, or Dark
Weather	Either Clear, Cloudy, Rain, Fog, Sleet, or Snow
Temp	Temperature in Fahrenheit at the accident site at the time of accident
Season	Spring, Summer, Fall, or Winter
TOTINJ	Total number of people injured in the accident
TOTKLD	Total number of people killed in the accident
SourceName	The railroad company which were involved in the accident and reported it

Table 2
Maximum speed allowed for freight trains based on the track class

Track Class	x	1	2	3	4	5	6	7	8	9
Max Speed	10	10	25	40	60	80	110	125	160	200

Having carried out a significant amount of data pre-processing and cleaning, we prepared a dataset containing 12,847 records (rows) where each record (row) of data represents an incident occurred in the 2008-2017 period. We focused on the incidents which have occurred on commercial tracks (i.e., main, yard, siding) and where the Class 1 railroad companies were involved. We focused on these companies because over 90% of the freight revenue is generated by few Class 1 companies which include: BNSF Railway Company (BNSF), Canadian Pacific Railway Company (CP), CSX Transportation (CSX), Kansas City Southern Railway Company (KCS), Norfolk Southern Railway Company (NS), Union Pacific Railroad Company (UP). A Class 1 railroad is defined as of 2016 as "having operating revenues of, or, exceeding $453 million annually" (american-rails.com).

GOAL OF THE ANALYSIS

In this section, we provide a brief overview of what EDA is and how incorporating partitioning and association rules mining can create a synergy between multiple topics in an MBA program.

Exploratory Data Analysis (EDA) is a data analysis approach which uses a variety of techniques, mostly graphical, to:

- Gain insight into the dataset,
- Uncover underlying structures (e.g., association between variables),
- Identify important variables,
- Detect outliers and anomalies,
- Test potential underlying assumptions.

By its very nature, the main objective of EDA is to openly explore the data in order to gain new and often unexpected insights into the underlying operations. This is why EDA has traditionally been taught using popular graphing techniques like bar charts, boxplots, histograms, scatterplots, pivot tables and basic trend analysis. We suggest that adding advanced yet easy-to-use tools such as partitioning and association rules mining to the list of EDA techniques taught in an MBA course will significantly promote the students' EDA skills.

Relevance to practice has proved effective in facilitating the learning process and motivating the students. To this end, we believe using data similar to the railroad incidents dataset is highly beneficial in teaching the analytics methods like EDA while reinforcing the concepts such as business performance measure. For example, considering the railroad incidents dataset, an MBA instructor can motivate the analysis by discussing the fact that safety is a crucial performance metric in the railroad industry, and thus, examining the railroad incidents can provide some important insights into the industry's (or a particular company's) performance. This helps students develop an applied understanding of the links between business concepts and methods of analysis.

The next three sections of the paper are organized as follows: First, we present some graphical data analysis using conventional EDA tools. In the two sections following graphical analysis, we describe our partitioning and association rules analysis, respectively, and discuss how they can be useful for the analytics pedagogy.

GRAPHICAL DATA ANALYSIS

Graphical presentation of the data helps students (and professionals) develop an initial understanding of the underlying activities visually. Figure 1 displays the number of railroad accidents and the resulting damage (in millions of dollars) by month for the period 2009 through 2017.

Even though both variables exhibit considerable variation, they generally remain stable over the observation period without an apparent trend or a shift. The monthly number of accidents (total damage by month) typically varies between 90 ($7 million) and 150 ($40 million) with an average of 120 ($20 million).

As mentioned before, the dataset contains information on 12,847 incidents. Figure 2 displays the distribution of damages at the accident level. Individual accident damages typically range from about $1,000 to amounts well in excess of $10 million.

Considering accidents as safety failures, in order to improve safety as one of the most important performance measures, business executives would naturally wish to investigate the factors involved in the costly damages.

In order to identify the main drivers of the damage associated with railroad accidents, we consider a set of potential factors (or explanatory variables) displayed by Table 3 (see Table 2 for the explanations of these variables.)

In EDA, a great deal of insight can be obtained through appropriately constructed data graphs and their interpretation. Figure 3 displays the graphs constructed to examine the associations between a subset of the variables in Table 3 and accident damages.

Figure 1
Number of railroad accidents and total damage by month (2009-2017)

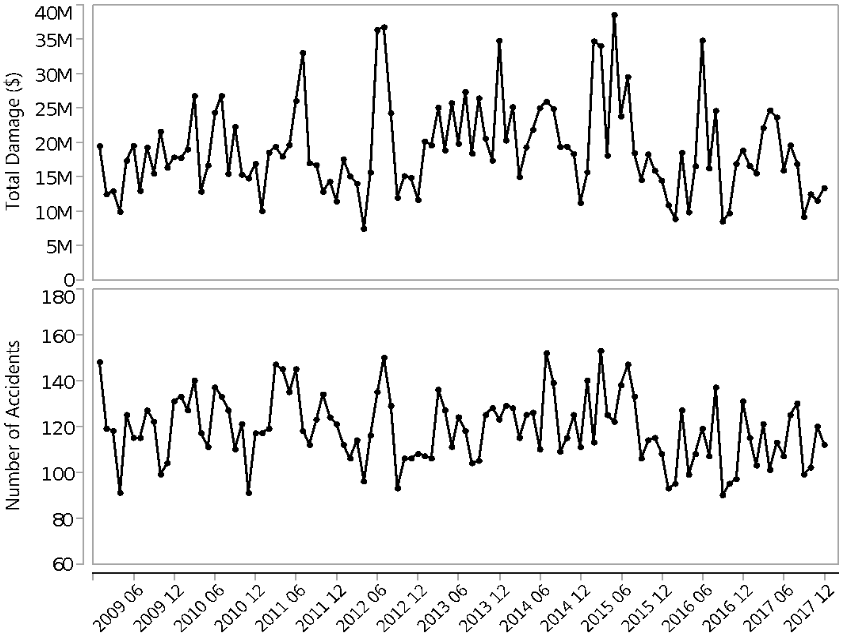


Figure 2
Distribution of individual accident damages ($1000)

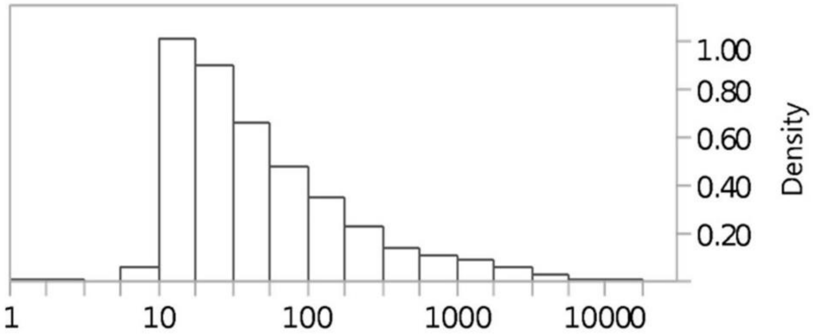


Table 3 **Potential factors that affect cost of damage**	
CauseCode	AMPM
TrackClass	HIGHSPD
TypeOfAccident	Temp
TypeOfEquip	Season
TypeOfTrack	TOTINJ
Visibility	TOTKLD
Weather	

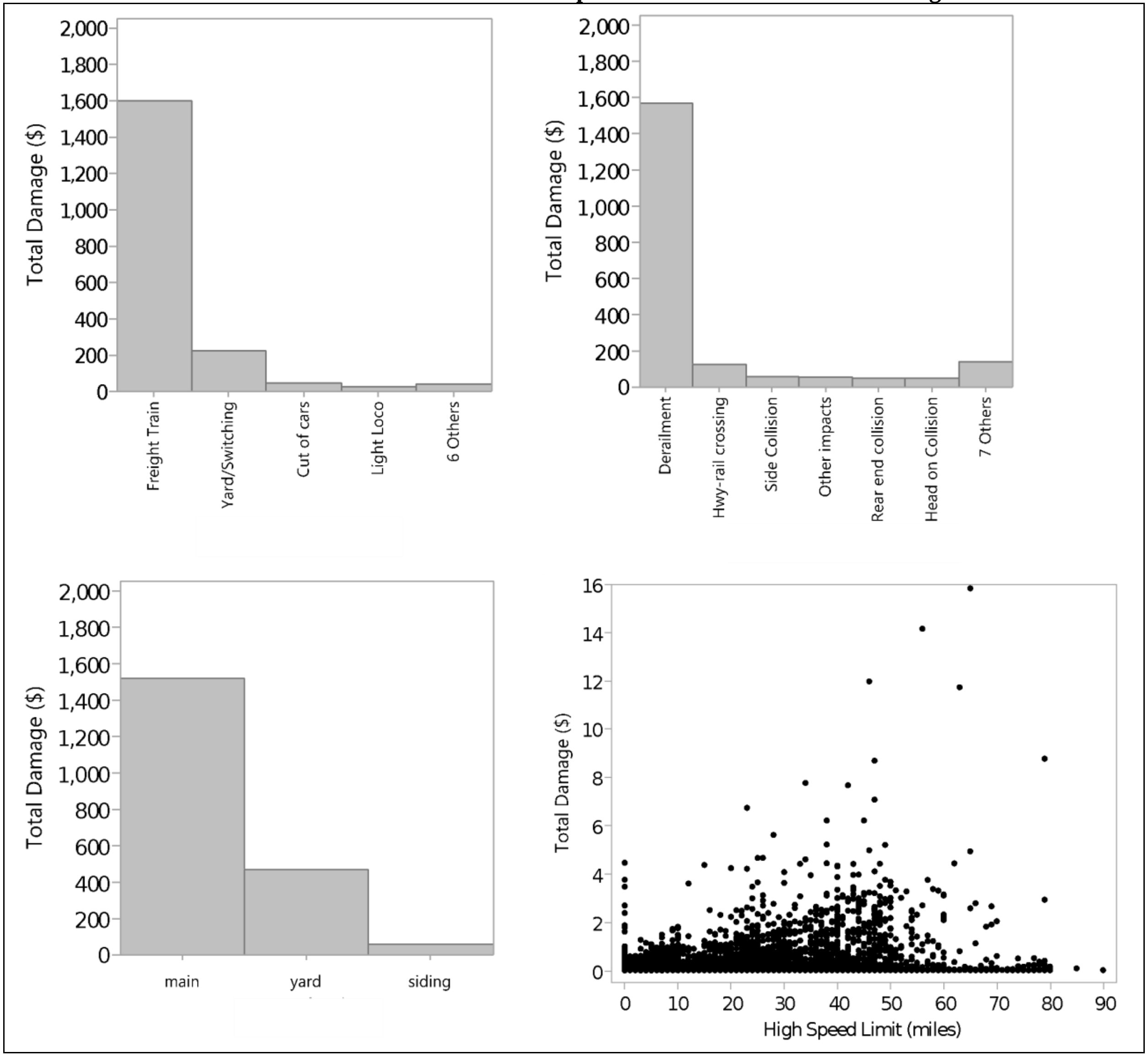
The graphs in Figure 3 show that a major portion of damages could be attributed to accidents that involve

- freight trains (82% of all damages),

- derailments (76%), and
- main track (74%).

The scatter plot between speed and damage shows that the propensity for high damage tend to increase at high speeds. The scatter plot also shows a curious grouping of relatively high cost accidents at near zero speeds, which we discuss in the next section on partitioning analysis.

Figure 3
Associations between a subset of the potential factors and cost of damage



Even though these observations provide useful initial insights, they are far from being definitive. They offer fragmented insights into the relationship of each variable with total damage one variable at a time. Further exploration of the data is called for in order to attain a more coherent and holistic understanding of the variable associations. Among other things, one of the important paths to analyze the data is to consider multiple explanatory variables simultaneously. For example, it would be of interest to explore the association between speed and damage by accident type and equipment type. However, as the number of variables under consideration increases, such analyses could become highly painstaking. We offer partitioning analysis as a useful tool to gain early insights

EDA USING PARTITIONING ANALYSIS

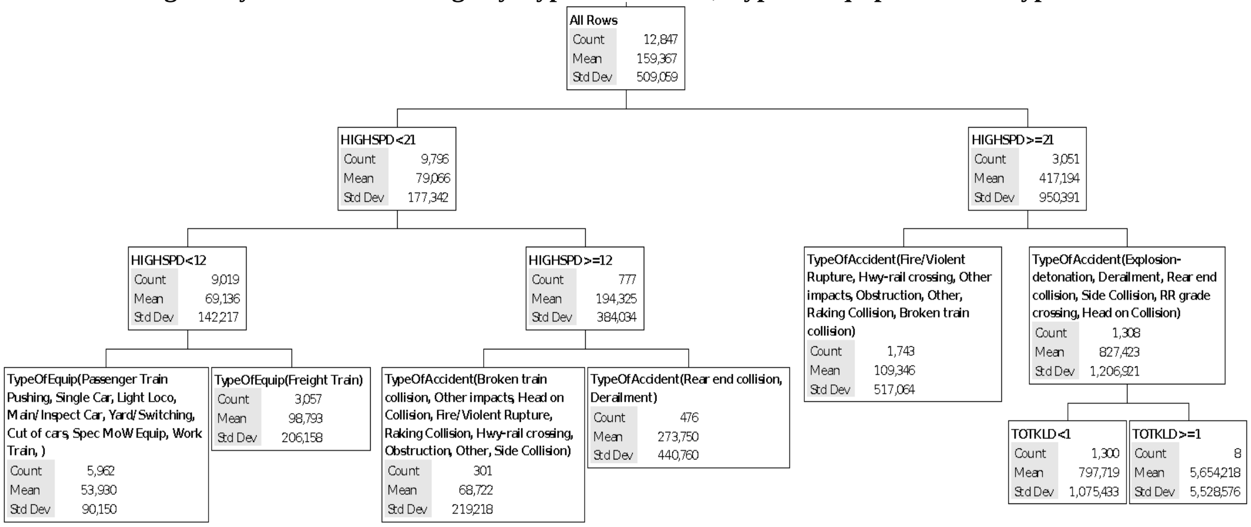
Partitioning analysis is a popular advanced method for data mining and exploring unfamiliar datasets. It is increasingly becoming available in statistical software packages, which provides a great opportunity to teach students this technique using an applied approach. MBA students, in particular, with a solid understanding of the underlying business concepts, can conduct this analysis efficiently, and interpret the results fairly easily.

For the analysis described here, we used the JMP software. The R package rpart can be used to conduct a similar analysis.

A recursive partitioning analysis was performed on the data. In this analysis, the variables displayed in Table 3 were used as independent variables and the total damage was the dependent variable. The results are shown in Figure 4.

Partitioning analysis systematically examines each of the independent variables to identify the variables and the associated cut-off value that splits the data set into the two most dissimilar partitions with regard to the total damage. The process gets repeated on each partition (see, for example, James et al. 2017).

Figure 4
Partitioning Analysis of Total Damage by Type of Accident, Type of Equipment and Type of Track


All Rows
Count 12,847
Mean 159,367
Std Dev 509,059
HIGHSPD<21
Count 9,796
Mean 79,066
Std Dev 177,342
HIGHSPD>=21
Count 3,051
Mean 417,194
Std Dev 950,391
HIGHSPD<12
Count 9,019
Mean 69,136
Std Dev 142,217
HIGHSPD>=12
Count 777
Mean 194,325
Std Dev 384,034
TypeOfAccident(Fire/Violent Rupture, Hwy-rail crossing, Other impacts, Obstruction, Other, Raking Collision, Broken train collision)
Count 1,743
Mean 109,346
Std Dev 517,064
TypeOfAccident(Explosion-detonation, Derailment, Rear end collision, Side Collision, RR grade crossing, Head on Collision)
Count 1,308
Mean 827,423
Std Dev 1,206,921
TypeOfEquip(Passenger Train Pushing, Single Car, Light Loco, Main/Inspect Car, Yard/Switching, Cut of cars, Spec MoW Equip, Work Train,)
Count 5,962
Mean 53,930
Std Dev 90,150
TypeOfEquip(Freight Train)
Count 3,057
Mean 98,793
Std Dev 206,158
TypeOfAccident(Broken train collision, Other impacts, Head on Collision, Fire/Violent Rupture, Raking Collision, Hwy-rail crossing, Obstruction, Other, Side Collision)
Count 301
Mean 68,722
Std Dev 219,218
TypeOfAccident(Rear end collision, Derailment)
Count 476
Mean 273,750
Std Dev 440,760
TOTKLD<1
Count 1,300
Mean 797,719
Std Dev 1,075,433
TOTKLD>=1
Count 8
Mean 5,654,218
Std Dev 5,528,576

The top box in Figure 4 provides the summary statistics for the entire sample (n=12,847 accidents with an average damage of $159,367). We consider the accident damage (cost in dollar amount) as our response variable. Partitioning analysis hierarchically divides the observations into two non-overlapping groups (nodes or partitions) based on a rule defined by a predictor variable identified by the method. The predictor variable picked to split the observations is the one that maximizes the difference between the average values of the response variable (here, damage) of the observations assigned to the two nodes. Continuous predictors (e.g., speed) are split according to a threshold (or cut) value. Qualitative predictors (e.g. accident type) are split based on their levels. In general terms, the partition tree displayed in Figure 4 is referred to as a regression tree since the response variable of this application is continuous. Partitioning analysis can also be used with a categorical response variable which results in a classification tree.

The partition tree is fairly self-explanatory:

- The top variable identified by partitioning is the speed limit. Specifically, the average damage of accidents that occurred on tracks with speed limits less than 21 miles was $79,066 (n=9,796). For speeds exceeding 21 miles, the average cost was $417,194 (n=3,051).
- At speed limits less than 12 miles, the most important factor with regard to accident damage is the type of equipment. The more costly accidents on low speed tracks involved freight trains with an average damage of $98,793 vs. the average damage of $53,930 of accidents involving other types of equipment.
- At medium speed limit (12 to 21 miles) tracks, the type of accident was the key factor. In this case, the higher cost accidents involved derailments and rear-end collisions. The average damage of such accidents was $273,750 vs. $68,722 of other types of accidents.
- Likewise, for accidents that occurred on high speed tracks, accidents involving explosions, collisions and grade crossings incurred higher damages (average damage of $827,423 vs $109,346 of other types of accidents.)
- On average, the highest damages are associated with accidents on high speed tracks that resulted in fatalities (average damage $5,654,218 vs. $797,719 for accidents with no fatality.)

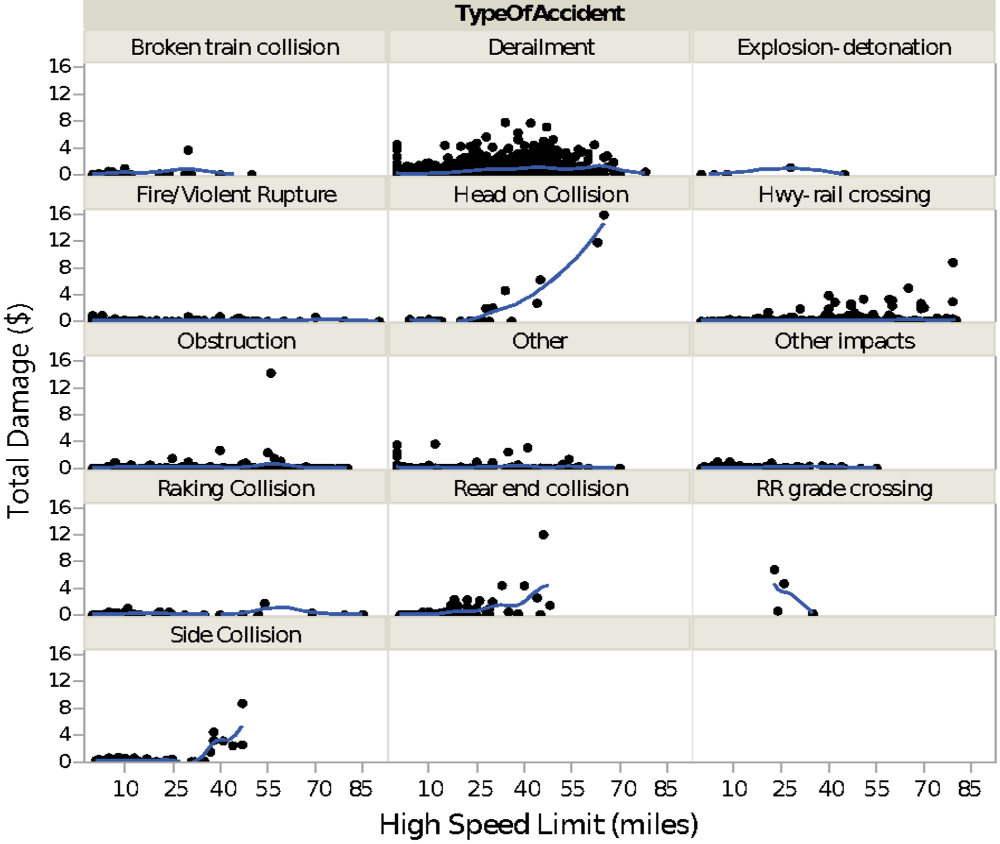
As the example shows, the partitioning analysis is easy to interpret. It provides an insightful description of the way in which multiple variables jointly act on the response variable (damage). For example, the association between accident type and damage depends on the speed limit. At speed limits lower than 12 miles, type of accident does not appear to have a marked effect on damage. Damage depends on type of accident at speed limits exceeding 12 miles per hour. Partitioning often leads to further follow up with graphical analysis. For example, consider the association between speed and damage. Our earlier examination of Figure 3 did not lead to a straightforward interpretation. Partitioning suggests that the effect of speed depends on the type of the accident, as depicted by Figure 5.

The results in Figure 5 lend themselves to an easier interpretation than Figure 3. It is now evident that the nature of the association between speed and damage depends on the type of accident. For most accident types, there is no discernible association of speed with cost of damage. The most notable effect is with head-on collision in which case the cost of damage increases proportional to speed; an association which also makes sense on intuitive grounds. This perspective on the data also shows that the high-cost accidents at near-zero speeds resulted from derailments.

It is worth noting that it often is desired to further explore the insights resulted from partitioning analysis. Regarding the analysis conducted on the railroad incidents data, we note that HIGHSPD, TypeOfAccident, TypeOfTrack, and TypfOfEquipment are picked as main factors driving the cost of damages (Figure 4). It may be beneficial to examine the association between these variables and other variables of study in the dataset to gain further insights. For example, one may seek to examine whether there is any strong relationship between type of accident with other factors such as type of track, weather, and visibility or whether there are certain causes responsible for particular types of accidents. This may help executives make more informed decisions as they try to prevent or reduce the likelihood of accidents.

In the next section, we describe how association rules mining can help discover the relationships among several qualitative variables.

Figure 5
Cost of damage vs. speed limit by accident type



EDA USING ASSOCIATION RULES MINING

In this section, we present our analysis of associations among a number of important variables in the railroad incidents dataset. Note that our dataset contains a large number of categorical variables, which hampers use of some common methods developed to explore relationships among numeric variables such as pairwise correlation analysis. Association rules mining, however, enables us to analyze the associations among qualitative variables. In our analysis of the accidents data, it also helps provide further insights into the partitioning analysis.

Association rules are probabilistic if-then statements intended to uncover the associations among seemingly independent variables within large transactional datasets (Camm et al. 2017). That is, association rules indicate if a certain transaction item occurs, then how probable it is for another item to occur as well.

In the marketing discipline, this method is known as market basket analysis by which marketers examine the customer purchase transactions and study the likelihood of particular products being purchased together. For example, consider this rule: If a customer buys orange juice, then there is an 80% probability that he buys ice cream. The “if” part is called “antecedent”, and the “then” part is called the “consequent” of the rule. The 80% conditional probability is called “confidence” in association rules mining. Clearly, the association rules that apply to more frequent transactions have a larger scope of impact. The frequency of transactions is known as “support.” For example, a support of 20% for the above-mentioned rule indicates that in 20% of purchase transactions orange juice and ice cream are bought together. In other words, the joint probability of a customer buying orange juice and ice cream is 20%. Now suppose the frequency of orange juice and ice cream purchases are 25% and 40%, respectively. If the two items were assumed independent, one would expect that only 40% of the customers who bought orange juice would buy ice cream as well, (or alternatively said, they would be purchased together only 10% of the time). But, according to the association rule between orange juice and ice cream, the probability is 80%, rather than 40%. Therefore, the analysis of association rules has doubled the confidence in this case. This added value, i.e., 80%/40% = 2, is called “lift ratio.” Clearly, higher lift ratios indicate stronger associations.

Although extensively used in marketing, association rules mining has applications in several other disciplines. For example, it can help medical researchers identify how likely it is for an illness to occur if certain symptoms are observed.

In order to shed light on the relationships among important categorical variables within the railroad incidents dataset we utilized association rules mining considering each accident as a transaction, similar to purchase transactions in market basket analysis. The categorical variables that we included in the analysis were: SourceName, TypeOfAccident, TypeOfTrack, Visibility, Weather, TypeOfEquipment, TrackClass, and ACCAUSE. Note that this set of variables include four main predictors in the partitioning analysis, i.e., TypeOfAccident, TypeOfTrack, TypeOf Equipment, and HIGHSPD (there is a direct relationship between TrackClass and the highest speed allowed on a track, as described in Table 1).

We used R package arules to conduct the association rules analysis. Table 4 demonstrates the top 20 rules that resulted from the analysis.

It is worth noting that the number of possible associations among the categorical variables included in the analysis can be so high that it becomes almost impossible for anyone to evaluate the associations manually. However, the Association Rules mining tools enable us to overcome this computational hurdle and conduct the analysis efficiently.

Consider rule 7 in Table 4. It implies that if an incident occurs on a class-4 track, it is 98.3% likely that the type of track is main track. This confidence is boosted by the lift ratio of 2.63. It, therefore, implies that the probability of an incident occurring on a main track is 98.3% / 2.63 = 37.37%. But, if we know that it has happened on a Class-4 track, then we are 98.3% confident that it has taken place on a main track. The support of 0.14 indicates that this rule can be applied to 14% of transactions, i.e., the portion of transactions on which the incident has occurred on a track which is both a Class-4 and a main track.

Not all association rules provide new insights. Marketing professionals know this well. For example, suppose this rule has resulted from a market basket analysis: If a customer buys salsa, there is a 90% chance that he buys chips. This rule is fairly obvious and unable to provide novel information. However, it suggests that the analysis has been conducted properly.

Rule 20 in Table 4, for example, seems like an obvious association: If the equipment involved in the incident is yard/switching, then the incident must happen in the yard. However, even this obvious rule may be useful. A natural question regarding Rule 20 may be why the confidence is 96.4%, rather than 100% then? A follow-up examination shows that the reason is the missing data points on the type of track in the dataset. Specifically, for a few incidents where type of equipment is yard/switching, the type of track, which must be yard, is not recorded. This can eventually result in a root-cause analysis of missing data and potentially improving the data entry/gathering process.

Table 4
Association rules for the railroad incidents data

	Antecedent		**Consequent**	**Support**	**Confidence**	**Lift**
1	{TypeOfAccident=Hwy-rail crossing}	=>	{CauseCode=M}	0.10	0.999	4.00
2	{TypeOfAccident=Hwy-rail crossing,TypeofTrack=main}	=>	{CauseCode=M}	0.10	0.998	4.00
3	{TypeOfAccident=Hwy-rail crossing}	=>	{TypeofTrack=main}	0.10	0.985	2.64
4	{TypeOfAccident=Hwy-rail crossing,CauseCode=M}	=>	{TypeofTrack=main}	0.10	0.985	2.64
5	{Weather=clear,TrackClass=4}	=>	{TypeofTrack=main}	0.12	0.985	2.64
6	{TypeOfEquip=Freight Train,TrackClass=4}	=>	{TypeofTrack=main}	0.14	0.984	2.63
7	{TrackClass=4}	=>	{TypeofTrack=main}	0.18	0.983	2.63
8	{TypeOfEquip=Freight Train,CauseCode=M}	=>	{TypeofTrack=main}	0.11	0.878	2.35
9	{TypeOfAccident=Derailment, TypeofTrack=main}	=>	{TypeOfEquip=Freight Train}	0.16	0.869	1.83
10	{Visibility=dark,TypeofTrack= main}	=>	{TypeOfEquip=Freight Train}	0.10	0.837	1.76
11	{Source.Name=UP,TypeofTrack= main}	=>	{TypeOfEquip=Freight Train}	0.11	0.817	1.72
12	{TrackClass=4,TypeofTrack=main}	=>	{TypeOfEquip=Freight Train}	0.14	0.782	1.65
13	{TrackClass=4}	=>	{TypeOfEquip=Freight Train}	0.14	0.782	1.65
14	{TypeOfEquip=Yard/Switching, TrackClass=1}	=>	{TypeofTrack=yard}	0.28	0.982	1.64
15	{Source.Name=CSX,TypeofTrack= yard}	=>	{TrackClass=1}	0.10	0.964	1.64
16	{TypeofTrack=main}	=>	{TypeOfEquip=Freight Train}	0.29	0.775	1.63
17	{Source.Name=UP,TypeOfEquip= Yard/ Switching}	=>	{TypeofTrack=yard}	0.11	0.971	1.63
18	{Visibility=dark,TypeOfEquip= Yard/ Switching}	=>	{TypeofTrack=yard}	0.13	0.968	1.62
19	{Weather=clear,TypeOfEquip= Yard/ Switching}	=>	{TypeofTrack=yard}	0.20	0.965	1.62
20	{TypeOfEquip=Yard/Switching}	=>	{TypeofTrack=yard}	0.30	0.964	1.61

Although, association rules mining may result in several obvious associations, the goal is to find a few rules that can provide beneficial insights. We found rule 1, for example, interesting and insightful. It suggests that if the type of accident is rail-highway crossing, it is almost certain that the cause code is M. An interesting note is that the M code refers to "miscellaneous." This rule by itself does not shed much light on the rail-highway crossing accidents. Nevertheless, if examined in more depth, it leads to an insightful result. Given the high lift ratio and confidence, further investigation of the rule seemed to be worth pursuing. We conducted a focused analysis on the incidents with miscellaneous causes and found out that, in fact, in most highway-rail crossings the car driver was at fault. This suggests that defining a separate cause code for car driver's fault, rather than putting that under the miscellaneous category, may be more illuminating.

It is noteworthy that highway-rail crossing is one of the influential factors in determining the accident damage according to our partitioning analysis. Association rules mining essentially provides deeper insights into this particular factor by enabling us to analyze and discover the associations among several categorical variables efficiently.

As a pedagogical learning point, we note that if a data analyst (or student) sought to analyze the associations between TypeOfAccident and ACCAUSE using conventional cross-tabulation techniques such as Pivot Table, they would end up having to analyze a 13 × 389 matrix, since type of accident and accident cause can take 13 and 389 different (qualitative) values, respectively. It would, therefore, be burdensome to gain any insights from such a large table, whereas association rules analysis can be carried out in seconds using software packages like R. This is a great opportunity to discuss the advantages of advanced EDA tools in our classrooms. Association rules mining, in particular, also provides a nice applied framework for reviewing basics of probability theory such as joint probability, conditional probability, Bayes' Theorem among others, in an MBA class where students are already familiar with the theoretical concepts.

CONCLUDING REMARKS

The recent growth in businesses' abilities to collect increasing amounts of data, and their strive for promoting data-driven decision making have led to a rising demand for competent data analysts. This, in turn, has called for enhancing academic curricula in order to equip students with the advanced tools that have proved effective in practice.

We propose an approach for incorporating advanced EDA tools in an MBA statistics/analytics course. We, specifically, use a real-world dataset to teach partitioning analysis and association rules mining techniques. Partitioning analysis is a popular, advanced method for data mining and exploring unfamiliar datasets. It enables students to incorporate multiple variables into the analysis and obtain results that are easy to interpret. Association rules mining helps analyze and discover relationships among several qualitative variables, which can enhance the EDA and provide richer insights into the data using fundamental concepts of probability theory. The availability of both methods in multiple software packages provides a valuable opportunity to help students become familiar with different packages. We have used JMP and R for partitioning analysis and association rules mining, respectively.

Using a practical and relatable dataset, like the railroad incidents dataset that we have used for our analysis, motivates students' learning significantly. It also provides the opportunity to help students develop an applied understanding of the benefits of advanced analytics methods in examining business performance metrics like railroad safety.

We believe that our suggested approach can help bridge the gap between academia and practice and hope that it will be utilized as an effective approach for enhancing the students' data analysis skills in other academic institutions.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Vikkie McCarthy, Associate Editor, and the two anonymous referees for their review of an earlier version of our manuscript.

REFERENCES

Aasheim, C., S. Williams, P. Rutner, and A. Gardiner (2015). Data analytics vs. data science: A study of similarities and differences in undergraduate programs based on course descriptions. *Journal of Information Systems Education*, 26(2), 103-115.

Basch, M. 2018. Jacksonville-Based CSX Says It Will Put More Resources Into Safety At Annual Shareholders Meeting. Retrieved from https://news.wjct.org/post/jacksonville-based-csx-says-it-will-put-more-resources-safety-annual-shareholders-meeting.

Camm, J. D., J. J. Cochran, M. J. Fry, J. W. Ohlmann, D. R. Anderson, D. J. Sweeny, and T. A. Williams (2019) Business Analytics, 3rd Edition, Cengage Learning.

Chaojiang, W., M. Feng, and Y. Yan (2015). Teaching data mining to business undergraduate students using R. *Business Education Innovation Journal*, 7(2), 64-73.

Hartenian, E. and N. J. Horton (2015). Rail Trails and Property Values: Is there an Association? *Journal of Statistics Education*. 23(2), 1-24.

Horton, N. J. (2015). Challenges and Opportunities for Statistics and Statistical Education: Looking Back, Looking Forward. *The American Statistician*. 69(2), 138-145.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2017). An Introduction to Statistical Learning: with Application in R, 7th Edition, Springer.

Liu, K. (2016). What your industrial analytics course needs. *Industrial Engineer*, 48(4), 43-46.

McClure, R. and S. Sircar (2008). Quantitative literacy for undergraduate business students in the 21st century. *Journal of Education for Business*, 83(6), 369-374.

Meyer, C. (2015). 8 tips for teaching big data. Retrieved from https://www.aicpa.org/interestareas/accountingeducation/newsandpublications/pages/how-to-teach-big-data.aspx

Soule, L., R. Fanguy, B. Kleen, R. Giguette, and S. Rodrigue (2018). Evolution of a First Course in Data Analytics for Business Students. *Journal of Research in Business Information Systems*, 10(10), 55-73

Warner, J. (2013). Business analytics in the MBA curriculum. *Proceedings of the Northeast Business & Economics Association*, Bretton Woods, New Hampshire, 251-254.