Using JMP® as a Catalyst for Teaching Data-Driven Decision Making to High School Students

Jason Brinkley, East Carolina University, Greenville, NC

ABSTRACT

The presenter was recently asked to develop a 10-day short course in data-driven decision making to present during a program funded by the North Carolina General Assembly. The Summer Ventures program is self-defined as "a cost-free enrichment program for academically motivated high school students potentially interested in a career in science or mathematics." The goal of the course was to expose high school sophomores and juniors to the world of data in a unique and exciting way. It was decided early on that the course would not be an academic overview of statistics, and discussions on analytics would be limited. Instead students spend time describing, visualizing and critiquing real-world data sets from health care, education and business settings. Discussions about the size of data would be key; with students working at first with data sets with hundreds, then thousands, and finally millions of observations. This presentation will highlight key parts of the program and lessons learned while developing a way to teach young people with no statistics background to think critically about data and to discover appropriate ways to visualize and glean important information from different amounts and types of data. JMP Pro 11 was the only data software used.

INTRODUCTION

Working with data is an increasingly vital skill for 21st century employment. The need to gather and examine data is becoming more common in virtually all disciplines. There is a growing desire to explore ways to introduce these skills into younger groups of individuals in order to build an intuition and capacity for understanding data at an early age. While AP Statistics has become a common course in many high schools, there is a desire to provide more hands on training so that individuals understand the nuances of actual data which may overlap with how these students will encounter data in practice.

The outline of this paper is to describe the Summer Ventures program and then go through day by day and discuss the material and examples presented to the students in the 10 day series. Then to provide some discussion of what seemed to work well, possible changes that could be made for subsequent years, and what aspects of this training may work well in a general setting.

SUMMER VENTURES

The Summer Ventures program is self-defined as "a cost-free enrichment program for academically motivated high school students potentially interested in a career in science or mathematics." Funded by the North Carolina General Assembly and governed by the same group that oversees the North Carolina School of Science and Math, Summer Ventures provides unique opportunities for exceptional students to learn advanced scientific and mathematic techniques from University Faculty within the North Carolina University System. The interested reader may see www.summerventures.org for more information about the program.

While Summer Ventures is normally a month long program that starts with a 10 day short course followed by another 2-weeks of small group or individual work with University Faculty culminating in a final research project presentation. The author was approached in 2013 to develop a 10 day short course on data analytics for the summer of 2014. The goal was to expose attendees to the world of data without being (completely) tied to the world of mathematics. That would include discussing how to structure data for analysis and creating summaries and visualizations that could be easily interpreted for making decisions. In discussion with the organizers, it was decided that the amount of traditional statistical analysis would be limited. The idea was to expose students regularly to real data from various areas so that learning could be initiated in a more direct fashion than what is covered in standard mathematics courses.

JMP SOFTWARE

The decision was made early on that JMP Software would be a central focal point for the workshop series. Since the idea was to center learning on hands on data examples, the students would need an appropriate tool for working with data. Many of the topics were geared toward looking at what would work well with JMP Software and some of the advanced topics were chosen in a way to provide learning opportunities for conceptually difficult methods that were relatively easy to implement in software.

DAY 1

Day 1 instruction consisted of introducing structured data for analysis. Students were mostly familiar with using spreadsheet programs such as Microsoft Excel ® but were unclear as to the appropriate way to structure data for use in statistical software. A discussion around gathering data was facilitated by allowing the students to play 'Rock, Paper, Scissors' in pairs. This is a classic teaching exercise in statistics and allows students to play each other a number of times (say 20); discussions can begin on sources of bias, design issues, and appropriate ways to collect and store the observed data.

The examples given the first day include the Air Traffic control data available in the JMP's sample data archive whose Bubble Plot feature allowed students to see a high quality visualization early on that could be used to glean information and make decisions about what they were seeing within the given data. For hands on use, the classic pulse rate dataset from the DASL archive (http://www.statsci.org/data/oz/ms212.html) was used to explore the JMP feature Graphbuilder.

The pulse rate dataset was key early on for understanding different features in JMP. Since not routinely available in JMP, a description of the data is given below:

"*Students in an introductory statistics class, taught by Professor John Eccleston and Dr Richard Wilson at The University of Queensland, participated in a simple experiment. The students took their own pulse rate. They were then asked to flip a coin. If the coin came up heads, they were to run in place for one minute. Otherwise they sat for one minute. Then everyone took their pulse again. The pulse rates and other physiological and lifestyle data are given in the data. Five class groups between 1993 and 1998 participated in the experiment.*"

DAY 2

Day 2 instruction included some early lecture based discussion on summarizing and visualizing data. Topics included histograms, bar charts, pie charts, time plots, and scatter plots. Brief overviews with examples of interpreting the mean, standard deviation, median, and correlation were discussed.

The students returned to the pulse rate dataset for an in-depth discussion regarding a multitude of basic JMP features which included Distribution, Tabulate, and Graphbuilder. In addition, time was devoted to highlighting subsets of data, excluding data from analysis, saving scripts to data tables for easy recall, moving output from JMP into Microsoft Word ®, and other assorted features that the students would have to rely on again and again throughout the workshop series.

DAY 3

Day 3 instruction centered on data collection; time was spent identifying the differences between experiments and observational studies. Issues such as confounding were discussed and the differences between association and causation. A small portion of the day was spent giving an overview of clinical trial methods and explaining to students how a drug goes from a chemical compound to the market. The idea was to impress on the students how the data play a pivotal role in the decision to allow a drug to be marketed as a potential therapy in the United States and the benefits allotted by experimentation.

Data example for that day was a small dataset that the author had used in previous consulting experiences with clients locally. The data were not shared here due to potential privacy issues. However, an adequate description can be given; the data was from an Autism intervention study where patients were randomized to two groups: the first who cared for and fed real animals while the other had 'pretend play' to care for and feed stuffed animals. The students were allowed to see crossover effects in that the students in the 'real' animal group saw declines in measures that quantified 'Aberrant Behaviors' while the 'pretend play' group saw increases in social reciprocity metrics. Since the data did not have one definitive answer, clinical thought was needed to interpret results with the prevailing idea that when autistic children work with animals we see declines in 'acting out behaviors' but when similar children spend time with a clinician guiding them through 'pretend play' we see gains in the ability to socially interact. The purpose of this dataset was to illustrate that the end result may not always be something along the lines of 'treatment is better than control'.

DAY 4

Day 4 instruction consisted of data cleaning, merging and manipulating tasks. The goal was to illustrate that the data are rarely in a 'nice' format for additional analysis. Students were exposed to the stack, split, concatenate, and join commands in JMP along with the recode feature and the formula option for new variables.

The data example for the day consisted of a 10,000 song subset available via the 1 million song dataset (http://labrosa.ee.columbia.edu/millionsong/). The dataset had song titles and artists along with measures like length

and popularity of artist. Each song had a series of 'tags' that classified the song or artist into several groupings. Students discovered the difficulty in working with non-independent data as some artists had multiple songs in the dataset and that essentially the same artist was coded in the dataset multiple times if they appeared in duos or had another artist featured in a small aspect of the song (which was very common among songs from the R&B and Hip Hop genres). Students joined two datasets, one with the song information and the other with some of the tag information; then subsetted out those identified as American in the tags to compare with those that were identified as British to see if American songs were more popular or longer than the British songs in the registry. Given the multiple tags identifying the same thing: "US", "USA", "American", etc; this became a difficult challenge.

The second half of this session focused on gathering our own observational data, the students scoured online menus to provide nutritional information for cheeseburgers from over 30 restaurants. That data was collected and brought together to examine calorie, fat, and carbohydrate information between so called 'fast food' restaurants versus 'sit down' restaurants. The end result was that in general 'sit down' restaurants had higher calories and fat and especially high sodium levels, however students did not have access to serving size information and the prevailing thought was that 'fast food' cheeseburgers may be smaller in size.

DAY 5

Day 5 instruction consisted of a discussion on maps. The mapping feature in JMP's Graphbuilder was the central focus along with pre-coded examples of the Bubbleplot feature. Student's examined JMP sample datasets to examine Napoleon's March, as well as the Air Traffic and Crime Data. After the students had good exposure to the mapping elements of Graphbuilder, it was decided to gather and examine our own data.

Students downloaded data from two sources: first was the foreign assistance data made available from the United States government (http://www.foreignassistance.gov/web/dataview.aspx) while the second was world-wide GDP data for countries from the World Bank (http://data.worldbank.org/data-catalog/GDP-ranking-table). The goal of the exercise was to look at the distribution of U.S. Aid to foreign countries and compare that to the GDP of those nations.

Students were not surprised at the amount of aid being given to countries like Iraq and Afghanistan; however, there was some surprise at how much aid was being given to Russia and China. Mapping the data allowed students to see great disparities in aid to African countries. When Africa became the central focus of the in-class discussion, it was found that similar aid levels were being given to smaller GDP countries like Nigeria as well as larger GDP countries such as South Africa.

DAY 6

Day 6 instruction consisted solely of statistical theory. No hands-on analysis was performed that day. A lecture covering statistical inference was given. The prevailing ideas that govern confidence intervals and statistical hypothesis testing were discussed. This was the only day in which lecture based material exceeded 30 minutes and the goal was to provide a general overview of some of the tools required for clinical decision making by some individuals.

DAY 7

Day 7 instruction contrasted day 6 by starting the students with a new dataset to work on without guided instruction. The data come locally from the Heart Healthy Lenoir project (http://www.hearthealthylenoir.com/high-blood-pressure-study) and was made available to the author from work with local colleagues. The students were given some background on high blood pressure (http://www.nhlbi.nih.gov/hbp/hbp/whathbp.htm) and tasked with exploring the data on their own. They had been given several patient level predictors, side effects, comorbidities, as well as beginning and end of study blood pressure values (systolic and diastolic) and a variable labeled 'Treatment' with levels 'high' and 'low'. The students analyzed the data using the standard features discussed regularly in class (Distribution, Tabulate, and Graphbuilder) and the conclusion that was reached from the analysis was that 'high' treatment levels were associated with higher changes of blood pressure levels across the study as well as higher rates of many different side effects (fatigue, sleeplessness, headaches, and so on).

The students were asked to make a decision and they suggested that the high treatment level must be the source of the negative impact, that indeed there was a causal relationship between taking high treatment and poor outcomes. The students were informed that the 'Treatment' was actually medication adherence. What they just told me was that high medication adherence caused poor outcomes. I then informed them that clinical insight illustrated that for these patients the causality was the other way around; that individuals (in this study) only adhered to their medication when they "felt bad" and that was the reason for the higher rates and values. Indeed, testing showed significant differences but the 'high adherence' group had deeper family histories of diabetes and high blood pressure. Clinical thought was these individuals take their medicine regularly because they feel the true impact of their disease.

After finishing this dataset more discussion was had around causality and a deeper discussion of association versus causation was given. Examples included local anecdotes and the website (http://www.tylervigen.com/) which looks at ridiculous spurious correlations between variables that should logically have no meaningful connection.

The second half of Day 7 saw the students being guided through a dataset of over 60,000 documented UFO sightings with text descriptions (http://www.infochimps.com/datasets/60000-documented-ufo-sightings-with-text-descriptions-and-metada). The original data comes from the National UFO Reporting Center. The goal in data exploration was simple text based analysis. The data was first brought into Microsoft Excel ® so that the data could be deliminated by spaces, commas, and periods so that each word was given its own cell in a spreadsheet. The data were then imported into JMP and stacked so that the new dataset (with well over 2 million rows) indicated each word that had been used in the reported descriptions. Tabulate was then used to coalesce and organize the key words from which an output dataset was created with frequency of occurrence. The students could see that this method still left important delimiters in the mix (e.g. semicolons and dashes) but that they had been successful in getting most of the important key words. Early analysis found that California was a high frequency reporting state and given the size of the data it was decided to focus later efforts on that state. Key word searches found some interesting phenomena, for example words for the colors white, yellow, red and orange occurred much more frequently than blue, green, purple, and black. Likewise, the words two and three were more common than one (and their numeral counterparts). Discussion ensued on how to interpret this in the context of text reports. Students pointed out the high frequency of the word 'Santa' which led to several classroom jokes. However, later thought realized that the subsetted data came from California with populated cities such as Santa Cruz and Santa Monica had several reports. Student discussion was vibrant and the purpose was to convey an appreciation of how difficult it is to work with text based data with later discussion on the difficulties in creating quality search engines online such as Google ®.

DAY 8

Day 8 instruction centered on a discussion of k-means clustering. K-means clustering algorithms are iterated computer programs that create k distinct subgroups in a dataset. Each subgroup has a centroid point, the clusters are chosen so that the distances from the points to the subgroup centroid are minimized. While typically a topic covered in advanced undergraduate and graduate statistics course, the method is easy to implement in JMP and can be a unique tool in creating meaningful subgroups from a large amount of data. The students in the course were all familiar with basic geometry and simulated datasets provided a good starting point for illustrating how k-means methods work. Many of the already explored dataset were revisited to see the effects of clustering. Could height, weight, and baseline pulse rate clusters from the pulse rate dataset adequately separate males from females within the data? Could the treatment groups in the autism dataset be separated by their behavior measure scores? In addition the Fitness data from the JMP Sample Data archive was employed to created cluster groups of fitness levels based on measures within the data.

DAY 9

Day 9 instruction centered around big data. The students explored two big data datasets to explore the difficulties in working with very large datasets. The first was simulated data whose purpose was to explore problems with visualizations such as scatterplots in datasets with tens of millions of observations. The focus was that it was hard to know 'how much' data was being visualized in different portions of the plot. The data was manufactured specifically to highlight these issues and to illustrate the required time in analyzing such large scale data. The need for caution and planning for what types of analyses would be done was stressed.

The other big dataset came from the American Time Use Survey (http://www.bls.gov/tus/datafiles_0312.htm) which looks at diary data recorded from a cross section of Americans as to how they spend their day. The files were large with many entries per respondent. The data had a guidebook for which students also needed to examine to understand the specific field within the dataset. It was decided that the file was simply too large to completely analyze in one setting and the students were allowed to decide on a specific question of interest. Since family demographic information was available it was decided to explore diaries of individuals with and without young children. It was found that families with young children spent more time on the weekend at the grocery store and slightly less time watching television.

DAY 10

Day 10 instruction allowed the students another opportunity to work independently on a particular dataset. The Titanic dataset from the JMP sample data archive was accessed and students used the tools they had been given to look at factors that were associated with mortality on the Titanic. The students were given the freedom to use anything they had seen in the course to tackle the problem with some interesting results. In the end, the instructor illustrated the use of a decision tree as a method by which one could screen the data for important subgroups of individuals who did and did not survive.

The session closed with a discussion on frameworks for decision making and ideas on next steps. Six sigma was introduced as a conceptual framework for decision making that involved the continuous and systematic use of data. Options for further study and possible college programs to study data analysis were also covered as well as employment opportunities for individuals who found great interest in routinely working with data.

DISCUSSION

Feedback from the course was almost exclusively positive, students responded well to the hands on approach and these students in particular adapted well to the barrage of varying datasets that they were tasked with looking at. There was a growing appreciation of how much subject matter was needed to understand the data for each dataset. Some students were disappointed in the level of mathematics presented in the course but were very happy for the discussions on statistical inference and k-means clustering. By contrast, those were the least popular aspects of study to several other students who enjoyed working hands on with data.

There are several aspects from this kind of training that can be taken into a more general setting. These students in some ways because of the variety of data they were working with. Most of the tools from JMP being utilized were standard (Distribution, Tabulate, Graphbuilder) but seeing how they could be utilized in different ways from problem to problem helped gain an appreciation for the tools as well as a recognition of the commonalities in approaching many different types of data. Spending significant time on data cleaning and organization was time consuming but necessary as many students suggested that was their least favorite portion of the work. When informed that this was very common in practice and many applied analysts spend much of their time in this setting, it was a true wake up call. While changes will be on-going as this program evolves, it seems clear that it was effective. By the final day, these students (mostly 15-17 year olds) had developed a capacity for working with data and an understanding for the need in clinical and expert insight into the results of data in order to make proper decisions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Jason Brinkley, Ph.D.
Enterprise: East Carolina University
Address: Mail Stop 668, 2435 Health Sciences Building
City, State ZIP: Greenville, NC 27834
E-mail: brinkleyj@ecu.edu