Data Cleansing:
An Omission from Data Analytics Coursework

Johnny Snyder
josnyder@coloradomesa.edu
Business Department
Colorado Mesa University
Grand Junction, CO 81501

Abstract

Quantitative decision making (management science, business statistics) textbooks rarely address data cleansing issues, rather, these textbooks come with neat, clean, well-formatted data sets for the student to perform analysis on. However, with a majority of the data analyst's time spent on gathering, cleaning, and pre-conditioning data, students need to be trained on what to look for when generating or receiving data. A critical scan of the data needs to be performed (at a minimum) to look for errors in the data set before data analysis can be performed.

Keywords: Data cleansing, data pre-conditioning, data analysis, data formatting, Pareto Principle

1. INTRODUCTION

Data gathering and cleansing is the first task an analyst must perform before analytical tools can be applied to the data. Data issues such as non-printing characters, misspellings, text embedded in the quantitative data, interpretation, imputation, or unit conversions all must be accomplished before the data is ready for analysis.

In this, the information age, new positions in corporate structures call for positions such as "data steward" and "data analyst" who, among other things are responsible for:

- extracting existing data
- performing data validation
- confirming data correctness
- confirming data upload
- identifying data quality issues
- identifying and analyze defects in data sources and processes
- receiving, inspecting, validating, transforming, cleaning, and loading data received in a variety of formats (Monster, 2018).

These responsibilities illustrate that the corporate data person spends most of their time managing data rather than analyzing data. Ruiz (2017), while dubbing data science as "the sexiest job of the 21st century," also noted that "most data scientists spend only 20% of their time on actual data analysis" (para 1). These data cleansing items that a data analyst must be (primarily) responsible for consume a large part of their day, so merit inclusion in the quantitative methods classroom.

The 80/20 Rule (aka the Pareto Principle) appears in many situations in business and other human activities (Koch, 1998). There are many examples of the 80/20 rule online, in the academic literature, and in books such as Koch (1998). The definition of the Pareto Principle is simple, "a prediction that 80% of the effects come from 20% of the causes" (Mar, 2013, para. 4).

Many people have used the Pareto Principle in business, in computer coding, in describing computer trouble shooting activities, in product management, and in organizing one's personal life activities! One recent application of the 80/20 rule can be useful to new job titles such as: data steward, data analyst, business analyst, data scientist…of the information age, or the age of big data. This rule is stated as: 80% of a data scientist's time is spent collecting, organizing, and cleansing the data, while only 20% of the time is spent analyzing the data.

However, this rule of thumb is not being taught in many quantitative methods textbooks. The data sets a student sees in these classes are neat, clean, organized, and ready for analysis – not quite the way data generally comes to an analyst in its native form.

This case illustrates that data is messy, full of human errors or misinterpretations, incorrect, misspelled, illegible, or incorrectly formatted; thus, in need of pre-conditioning (cleansing) before analysis can begin.

2. LITERATURE REVIEW

Data cleaning has traditionally been a "lower status" of data quality activities, bordering on data manipulation (Van den Broeck, Cunninghan, Eeckels, & Herbst, 2005). Part of this reputation could be due to the prevalence of how data errors can be "fixed." For example, missing data values can be addressed by:

- deletion – exclude the instance
- hot deck – replace using values from the same data set
- imputation – assign a representative value (mean, median) to a missing one (Corrales, Corrales, & Ledezma, 2018)

How one does data cleansing is still a topic open to debate, and might have factors such as type of data, application for data, source of data, and discipline specific conventions to consider when making data cleansing decisions. However, it has been observed that students are not well trained in the methods of data preparation (for analysis) but seem to be able to come up to speed rather rapidly (Yue, 2012).

In the 20th century, the 80/20 Rule was shown to describe library usage patterns – 20% of the patrons use 80% of the resources (Trueswell, 1969), posting to electronic bulletin boards – 20% of the participants post 80% of the content (Echavarria, Mitchell, Newsome, Peters, & Wentz, 1995), consumer spending patterns – 20% of the customers account for 80% of the revenue (Fitzsimmons, 1985), and of course, Pareto's original assertion that 20% of the population of a country owns 80% of the land (Pareto, 1971).

More recently, in the 21st century, the 80/20 rule has been observed in computer code – 20% of the code contains 80% of the errors (Pressman, 2010), healthcare – 20% of the patients use 80% of healthcare services (Weinberg, 2009), and 80% of the defects can be explained by 20% of the causes in a quality control environment (the famous Pareto Chart) (Larson, 2018).

In the case of business analytics, or the study of data and what information can be gained from the data, the 80/20 rule becomes: 80% of the time spent by a data scientist is on gathering, cleansing, and storing the data, while 20% of the time is spent on analyzing the data. However, this concept is not discussed in most quantitative methods textbooks, thus, students enter the workforce with unrealistic expectations of how data will be coming to them. For example, Render, Stair, Hanna, and Hale (2018) state: "…collecting accurate data can be one of the most difficult steps in performing quantitative analysis" (p. 4). This is a true statement, and methods for collecting data are then presented, but there is no mention of cleansing data, or examples of data needing cleansing presented. Groebner, Shannon and Fry (2014) discuss how to collect data (surveys, observation, personal interviews), collection issues (bias, accuracy, error), and sampling techniques, but no mention of data cleansing or examples or problems/exercises are presented.

The literature regarding data cleansing includes the ETL (extraction, transformation, and loading) process for a database or data warehouse (Boyno, 2003), data quality in regression models (Corrales, Corrales, & Ledezma, 2018), as well as harvesting, cleaning and analyzing Twitter data (Hill & Scott, 2017).

These papers point to what Hellerstein (2008) infers when he states: "Data collection has become a ubiquitous function of large organizations – not only for record keeping, but to support a variety of data analysis tasks that are critical to the organizational mission" (p. 1). In short, business has become data driven, and as the old acronym tells us: GIGO. Keeping the data accurate, formatted correctly, and timely has become a new job in corporations – that of the data scientist or data steward (Experian, 2018). These "newer" positions in the corporate structure illustrate the importance of obtaining, storing, and utilizing data in business decision-making processes. As Hellerstein (2008) states: "Data errors can creep in at every step of the process from initial data acquisition to archival storage" (p. 1).

Data collection and cleansing is the first step of the analysis process and must be taken seriously (GIGO). Preparing data for analysis is critical if good information is to be extracted from data flows. The first steps are illustrated in this paper, where data is collected, cleansed, and prepared for simple analysis. Many further activities for data cleansing are context dependent as illustrated previously in the literature review. However, the steps illustrated in this case are universal and should be performed on any data set an analyst receives.

3. METHOD

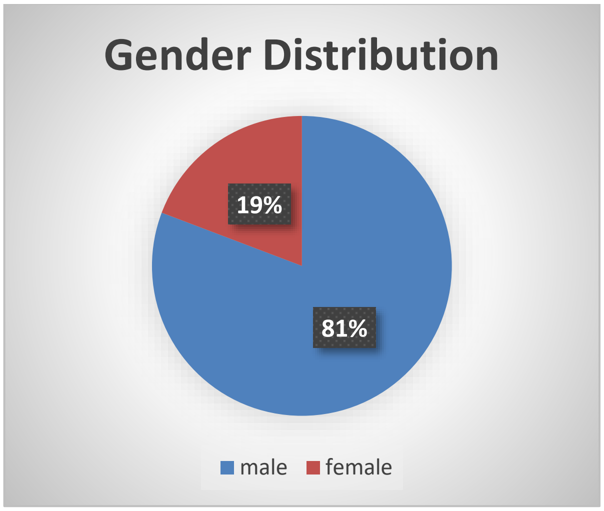
A survey instrument was created to gather student data from an introduction to business analysis class (see Appendix A). This instrument contains questions intended to solicit answers from all data categories (Nominal, Ordinal, Interval, Ratio - NOIR) to aid further classroom discussions about data types and graphical and analytical techniques associated with them. The survey is anonymous and is distributed on the first day of class. The instructor collects the survey instrument and inputs the data into an Excel (Excel, 2016) spreadsheet and distributes the spreadsheet to the class. (See Appendix B) The resulting spreadsheet is used to discuss (throughout the class) data types (NOIR), data errors (units missing), different units from different survey respondents, interpreting what the survey respondent "meant", text characters input into Excel cells (Excel refuses to do analysis on these cells), data conversions (from feet/inches to inches for example), and the dreaded non-printing character (space, for example) which can foul up the simplest Excel operations.

4. RESULTS

One of the first exercises for students could be to graph a nominal variable such as gender which is easy due to the pre-defined selection on the survey instrument. Manipulating the data into a form ready to graph, we obtain Table 1. Note for students: be sure to track n, the sample size, to be sure all data values have been accounted for. It should also be mentioned that this is one of the only "clean" parts of the data set…one of the columns that are ready for analysis!

Gender	Frequency	Percent
male	21	0.807692308
female	5	0.192307692
n =	26	1

Table 1 - Gender



Graph 1
Gender Distribution

The next exercise could be to evaluate the students' favorite color or type of security software they use. Analyzing these, using a pie chart or bar chart (or a Pareto Chart), would be straight forward if the data were clean, but looking at Appendix B, the results from the survey are not ready for analysis.

Attempting to organize the favorite color column results in questions about the data that must be addressed before a frequency distribution can be constructed. Some of these questions include:

- Is maroon brown or red or its own color?
- What color is blue/black? (counting both would artificially increase n)
- Should navy blue be counted as blue?

After a first pass at constructing a frequency distribution, depending on the Excel count function utilized, it could be found that n = 24 instead of 26. This is an interesting result for students, as the difference is due to colors being entered with a space (a non-printing character) at the end (or beginning) of the cell, resulting in Excel not counting these data points. Once these two cells have been identified and cleansed, the resulting frequency distribution can be seen in the third column of Table 2.

Non-printing characters can give the data analyst a lot of grief! Space is the most common non-printing character, but many others exist, such as carriage return (enter), end of record and end of file characters from various software packages that the analyst might have to import into their computing environment.

Upon cleansing the data, the color black was removed from the distribution (blue/black was cleansed to blue, the respondents first color choice). In a practical application, such as scheduling the percentage of cars to paint of each color for the coming model year, this cleansing activity can have the consequence of removing a very popular color from a dealer's inventory. Thus, the analyst needs to consider the business need for the data before cleansing the data.

Color	Frequency (original)	Frequency (cleansed)
blue	10	11
red	4	4
green	3	3
purple	2	2
white	1	1
orange	1	1
maroon	1	1
yellow	1	1
grey	0	1
gold	1	1
n =	24	26

Table 2
Favorite Color

Here are a couple of rules to follow when cleansing a data set:

- Maintain an original copy of the data.
- Label all pre-conditioning or cleansing activities (i.e. tell the reader what you have done to the data).
- Discuss the cleansing activities with your team to be sure that the business consequences for data cleansing have been addressed.

A fun exercise is to compute the average height of a student in the class. This is a seemingly straight-forward calculation, but if you ask Excel to compute the average from the data as it stands, it yields a #DIV/0! error. The original data and the cleansed data are shown in Table 3, where the cleansed data has been converted to a numerical value (from feet and inches) for Excel computations, units have been added in the heading, and the average has been computed.

	Height (original)	Height (inches) (cleansed)
	6'1" 73"	73
	6'4"	76
	5'11"	71
	6'2"	74
	5'7"	67
	5'9"	69
	5'8"	68
	5'7"	67
	5'11"	71
	6'4"	76
	6'3"	75
	6'0"	72
	5'5	65
	5'11"	71
	5'4"	64
	5'11"	71
	6'5"	77
	6'0"	72
	5'4"	64
	5'11	71
	5'6	66
	6'3"	75
	6'2	74
	6'5"	77
	5'8"	68
	5'10"	70
Average	#DIV/0!	70.92308
n =	26	26

Table 3
Height

Note also that upon cleansing the data, one should add the data units to the column heading for clarification purposes. Students also need to be careful when converting from the original to the cleansed form, as this is a manual operation, and errors can arise! Every time a human "touches" the data, errors can enter into the data set.

Another seemingly straight-forward calculation is to compute the average shoe size of a person in the data set. One could even compute the average size by gender, which makes more sense from a retail perspective. The sorted data is shown in Table 4.

	Gender (m/f)	Shoe Size (original)	Shoe Size (US size) (cleansed)
	f	8.5	8.5
	f	9	9
	f		
	f	7.5	7.5
	f	10	10
	m	12	12
	m	12	12
	m	9	9
	m		
	m		
	m		
	m		
	m	10 in	10
	m	12	12
	m	9	9
	m	11.5	11.5
	m	10 1/2	10.5
	m	10	10
	m	13	13
	m	11	11
	m	10.5-11	10.75
	m	12	12
	m	13	13
	m	12	12
	m	10.5	10.5
	m	11	11
n =	26	21	21

Table 4
Shoe Size

Table 4 illustrates for the student other issues that come with open ended survey questions. While 10 ½ is a valid shoe size, 10.5 is more appropriate for computational purposes (Excel readability – i.e. no text characters). Other questions can arise as well, such as:

- Should 10.5-11 be recorded as 10.75, the arithmetic average? or 10.5? or 11?
- What should we do about missing values?
- What does 10 in mean as a shoe size?

Finally, keep the units in the header row, not associated with the individual data values, again for computational purposes, because this is how Excel requires data to be formatted.

Many other ideas and examples can be created from a small data set such as this one, which illustrates for the students how easily data flows can be contaminated, inadvertently, by humans, machines, or software.

5. CONCLUSIONS

In coursework covering quantitative methods, spreadsheets or data sets come to the student pre-conditioned, or cleansed, ready for analysis. However, in business applications, the data might come in to the analyst in a raw form and need to be cleansed. Some of the issues that should be addressed include:

- units and unit conversions
- missing values
- extra text characters
- unclear answers (survey responses)
- non-printing characters

This case illustrated a simple but effective method to show students some of the issues that arise with data cleansing and how to address these issues in order to obtain a data set ready for analysis. Further, this data set can be used to explore different data types (NOIR), graphical representations of the various data types, and many concepts in data analysis from descriptive statistics to hypothesis testing…once the data is cleansed!

6. REFERENCES

Boyno, E. (2003). Extraction, transformation, and loading in a data warehouse course. Information Systems Education Journal 1(10). p. 3-10.

Corrales, D., Corrales, J., and Ledezma, A. (2018). How to address the data quality issues in regression models: a guided process for data cleaning. Symmetry 10(99). p. 1-20.

Echavarria, T., Mitchell, B., Newsome, K., Peters, T., and Wentz, D. (1995). Encouraging research through electronic mentoring: A case study. College and Research Libraries, 56 (4): 352-361. doi: 10.5860/crl_56_04_352

Excel. (2016). Computer software. Redmond, WA: Microsoft.

Experian. (2018). Data Steward. Retrieved from: https://www.edq.com/uk/glossary/data-steward/

Fitzsimmons, J. (1985). Consumer Participation and Productivity in Service Operations. Interfaces, 15(3), p. 1-146.

Groebner, D., Shannon, P., and Fry, P. (2014). Business Statistics: A Decision Making Approach (9th edition). Boston, Mass: Pearson.

Hellerstein, J. (2008). Quantitative data cleaning for large databases. Retrieved from: http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf

Hill, S. and Scott, R. (2017). Developing an approach to harvesting, cleaning, and analyzing data from Twitter using R. Information Systems Education Journal 15(3). p. 42-54.

Koch, R. (1998). The 80/20 Principle: The Secret to Achieving More with Less. New York: Doubleday.

Larson, (2018). Using the 80/20 Rule to Improve Quality in Auto and Aerospace Manufacturing. Retrieved from: https://www.beaconquality.com/blog/using-the-80/20-rule-to-improve-quality-in-auto-and-aerospace-manufacturing

Mar, A. (2013, May 31). 13 examples of the Pareto Principle. Retrieved from: https://management.simplicable.com/management/new/examples-of-the-pareto-principle

Monster (2018). Data steward jobs. Retrieved from: https://www.monster.com/jobs/search?q=data-steward&jobid=2bc7ea0b-0fcb-4aa1-a026-db37660cb157

Pareto, V. (1971), Translation of Manuale di economia politica (Manual of political economy), Page, A.

Pressman, R. (2010). Software Engineering: A Practitioner's Approach (7th ed.). Boston, Mass: McGraw-Hill.

Render, B., Stair, R., Hanna, M., and Hale, T. (2018). Quantitative Analysis for Management (13th edition). Boston, Mass: Pearson.

Ruiz, A. (2017, Sept. 26). The 80/20 data science dilemma. Retrieved from: https://www.infoworld.com/article/3228245/data-science/the-80-20-data-science-dilemma.html

Trueswell, R. (1969). Some behavioral patterns of library users: The 80/20 Rule. Wilson Libr Bull, 43(5), p. 458-461.

Van den Broeck, J., Cunningham, S., Eeckels, R. and Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2(10).

Weinberg, M. (2009, July 27). In health-care reform, the 20-80 solution. Retrieved from: http://www.projo.com/opinion/contributors/content/CT_weinberg27_07-27-09_HQF0P1E_v15.3f89889.html

Yue, K. (2012). A realistic data cleansing and preparation project. Journal of Information Systems Education, 23(2). p. 205-21

Appendix A – Data gathering survey

Class Survey – first data set!

Demographic Information
Gender: □ Male □ Female Year of birth:
Height: Shoe Size:
Number of: Brothers _________ Sisters _________ - you have
Favorite color: _______________________
I am looking forward to this class: (circle on the next line)
Strongly agree = 1 2 3 4 5 = Strongly disagree
Type of PC Security Software you use: _______________________
Major area of study:

Appendix B – The data set….as respondents answered

Class Data										
Survey Number	Gender	Year of Birth	Height	Shoe Size	Brothers	Sisters	Favorite Color	Looking Forward	Security Software	Major Area of Study
1	m	1997	6'1" 73"	12	0	1	red	4	microsoft windows	marketing
2	m	1998	6'4"	12	0	1	maroon	3	slim cleaner +	accounting
3	m	1997	5'11"	9	1	1	green	3	?	business
4	m	1998	6'2"		1	1	blue/black	2	norton safe security	business admin
5	m	1997	5'7"		0	2	blue/black	2	?	accounting
6	m	1998	5'9"		1	1	blue	3	none	business
7	m	1997	5'8"		4	2	yellow	2	none (self watched)	cis
8	f	1997	5'7"	8.5	1	2	purple	3	don’t know	accounting
9	m	1995	5'11"	10 in	6	0	navy blue	3	Macfee	management
10	m	1997	6'4"	12	0	1	gold	2	Microsoft	finance
11	m	1997	6'3"	9	2	0	grey	2	mac	management
12	m	1997	6'0"	11.5	0	2	green	3	Norton	marketing
13	f	1996	5'5	9	1	0	green	3	I don't know	accounting
14	m	1996	5'11"	10 1/2	1	3	blue	2	none	economics
15	f	1999	5'4"		0	1	red	2	McAfee	accounting
16	m	1979	5'11"	10	0	1	blue	2	Mac	cis
17	m	1998	6'5"	13	1	0	blue	1	Apple	management
18	m	1997	6'0"	11	1	0	orange	3		management
19	f	1997	5'4"	7.5	0	1	blue	2	Mac	culnary arts/hospitality management
20	m	1998	5'11	10.5-11	1	4	blue	4	McAfee	marketing
21	f	1998	5'6	10	1	2	red	3	McAfee	cis
22	m	1998	6'3"	12	1	0	red	2	Mcafee	entreprenuership
23	m	4/22/1998	6'2	13	1	1	white	3	none	business
24	m	1998	6'5"	12	2	1	blue	3	Microsoft	finance
25	m	1996	5'8"	10.5	1	3	purple	4	Microsoft	administration
26	m	1996	5'10"	11	0	1	blue	2	Norton, Homebuilt	cis