White Paper

Be a Data Detective

by Ryan McGibony

Data Scientist

June 30, 2015

ELDER RESEARCH

DATA SCIENCE & PREDICTIVE ANALYTICS

Be a Data Detective

Table of Contents

The Investigative Mentality

You've probably heard it before – analytics professionals working directly with data spend as much as 80% of their time on data preparation, leaving only 20% for actual analytics and modeling. There are several common terms for the activities making up this 80%, including data "cleaning," "wrangling," or "munging," with perhaps the highest-profile example being "data janitor work," as discussed in The New York Times. The consensus seems to be that this work is undesirable, a necessary evil we must endure to get to the "cool" parts of data science. The practitioners quoted in the Times article lament the countless hours they pour into data prep, and the author entices the reader with the possibility of automating the process. While anyone who works in predictive analytics would welcome the chance to cut down on prep work, we should consider the downsides of adopting this attitude in the practice of data science.

"Data janitor work" needs to be rebranded. I propose that we, as an industry, endorse a different outlook toward data exploration and transformation and place more emphasis on the former. Why shouldn't learning about your data be a fun and interesting part of data science? Consider the viewpoint of Claudia Perlich, Chief Scientist at the digital advertising firm Dstillery, who describes the process as detective work. In contrast to the common mindset, Perlich has said that there is "no such thing as clean or dirty data, just data you don't understand." Regardless of which term you prefer – cleaning, wrangling, or something else – the investigative effort preceding data changes is critical. First, this investigation determines what those changes will be. And just as importantly, it provides the analyst with the context and understanding necessary to succeed at modeling and drawing conclusions.

Maybe your organization, like many others, is realizing that information you have already collected can be used to uncover valuable new insights? Internal analysts or hired consultants are tasked with deriving these insights, perhaps from a starting point of complete unfamiliarity with the data. How can someone in that position proceed to solve a business problem?
This is largely a matter of mentality. Being naïve about a particular set of data can actually be a good thing, if it allows the analyst to approach the problem with fewer preconceived notions. In fact, adopting a naïve perspective is useful even if the analyst is well-versed in the data sources and problem domain. There are virtually infinite ways that data can be misleading or biased, and even if you think you know what you are dealing with, there are bound to be some surprises.

Let's get more concrete. The following are some steps for data investigation and examples of the types of findings that can be uncovered along the way. It is not an exhaustive list, but rather a set of starting guidelines and suggestions for understanding the data to be analyzed. This assumes a traditionally structured, tabular data set, but the general concepts apply to unstructured data as well.

Getting the Lay of the Land

First, it is important to perform some "sanity checks" and get a feel for how well the data set conforms to your initial expectations.

1. Start with the basics. Understand what the rows (observations) represent. Are they at the level of interest for your analysis, or will you need to group or otherwise reshape them?
    - Most types of modeling and analysis assume independent observations. A data set I worked with described buildings, but it also stored the land surrounding those buildings as separate rows. For our problem, it was necessary to consider a building and its land as a pair.
2. Understand what the columns (variables) represent. Are they named clearly? How can you verify their meanings and understand how the data were generated? Is there a trustworthy data dictionary? (See Table 1.)
    - Looking for contract fraud requires an understanding of how and when contracts were modified. On one project, we encountered a variety of fields holding modification dates, and realized after speaking to subject matter experts in the organization that the appropriate choice for the effective date of modification was not the one we expected based on its name.
3. If multiple tables were joined to create the base table for analysis, either by you or by someone else before you received it, was any duplicated or conflicting information introduced? Did the merge affect which observations are available for analysis?
    - For one project, we had to merge customer survey data with behavioral data. Only a sample of customers took the survey, so we needed to understand whether there were limitations or selection biases. If we had analyzed only customers who completed the survey, we would have also needed to consider whether they were representative of the customer base as a whole.
4. What is <u>not</u> in the data that might be useful for your analytical goals? Do you have what you need to answer your business questions, or should you seek out additional or different data?
    - Once, we were dealing with customer applications for a business process. Completed applications were stored in one database table, while partially completed or rejected applications were stored elsewhere. These data sources had to be combined to obtain the full picture of customer behavior.

	mpg	cyl	disp	hp	wt	am	time	timemin
Mazda RX 4	21.0	6	160	110	2.62	1	16.46	0.27
Datsun 710	22.8	6	108	93	2.32	1	18.61	0.31
Hornet 4 Drive	21.4	6	258	110	3.21	0	19.44	0.32
Hornet Sportabout	18.7	8	360	175	3.44	0	17.02	0.28

Table 1. Cars data: Do we have a data dictionary that explains these variable names? We want evidence rather than assumptions. "timemin" and "time" sound suspiciously similar. Do we need both?

Digging In

With some initial understanding under your belt, it is time to dig into the specifics of the data and become comfortable with its contents and structure. There is much more to these steps, so I'll only touch on a few key points.

1. Perform univariate analysis to gain an understanding of each variable. Look at summary statistics and frequency tables, and visualize that information through histograms or other charts. How can you identify outliers, and how should you handle them? Are the variables distributed in a way you would expect? How common are missing values? How might that affect the analysis?

    - One data set commonly used for teaching data mining contains attributes of home equity loan applicants and the result of the loan (whether the applicant defaulted). In building a credit scoring model on this data, it turns out that an important predictive feature is whether the applicant's debt-to-income ratio is missing. Practical meaning aside, that fact might be missed without careful consideration of how to structure the data for modeling.

2. Perform bivariate or multivariate analysis to further explore how the variables are related. Scatterplots, correlation tables, and cross-tabulation are useful tools. Is there redundant information that could cause issues with modeling? (See Figure 1.)

    - Anyone who has studied regression analysis will be familiar with problems that arise from multicollinearity, where predictor variables are highly correlated. This affects the stability and interpretability of the model, and investigative work is needed to avoid that problem.

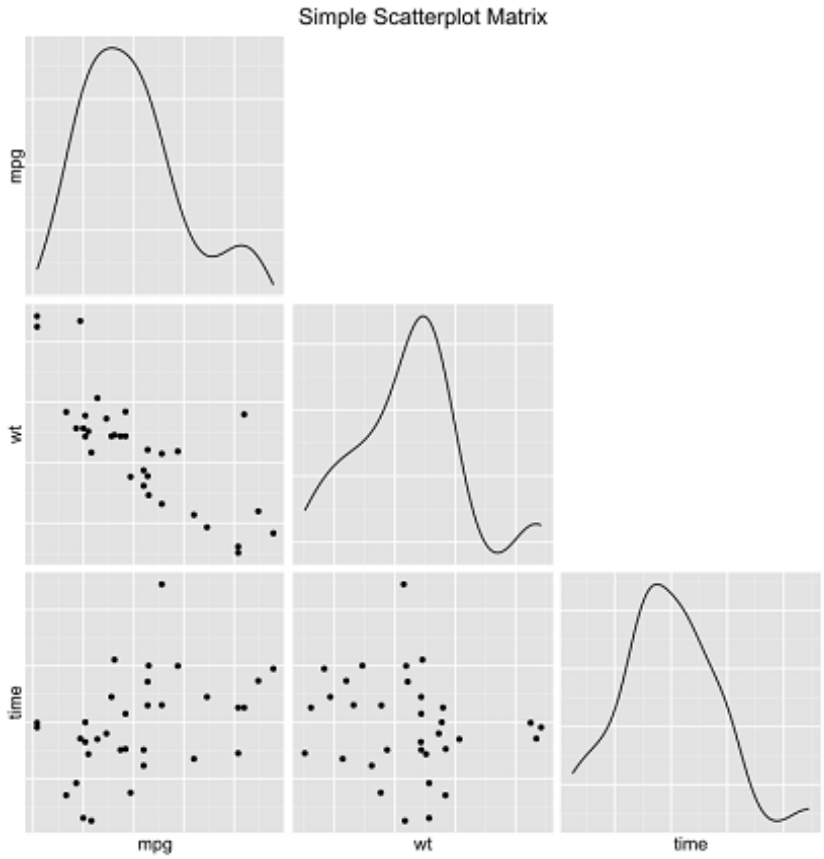

Figure 1. Cars data: miles per gallon, weight, quarter-mile time. There is a negative correlation between weight and MPG. Some potential outlier points are apparent.

Transforming the Data

Now that you have a good handle on the data, you're ready to put away your magnifying glass, grab your sponge (or other metaphorical cleaning tool), and start making changes. A few points to consider as you prepare for modeling and analysis:

1. If you identified outliers during your exploration, did you determine their validity? Is there an explanation for their occurrence? If you plan to exclude any cases, there should be a good reason to do so. "It's making my model insignificant" doesn't count!
2. Similarly, are transformations necessary in order to attempt the analytical techniques you have in mind? For example, a valid, but outlying point could be overly influential on a model due to the variable scale. Maybe a logarithmic transformation would be appropriate, or perhaps all input variables need to be standardized for a cluster analysis. Be aware of the assumptions of the statistical techniques you are applying. For instance, as John Elder points out as part of his PAW talk this year, it is often very useful to transform variables to be more Gaussian in shape – especially for squared-error methods like regression and neural networks.
3. Would combining information from different variables or restructuring them yield more useful inputs for modeling? This is known as "feature creation," and while it could be an

article in itself, the point is that your data is unlikely to arrive wrapped in a bow and ready for predictive modeling; there could be some re-mapping required to extract the particular information relevant to your problem.

This is only a starting point! Depending on the structure, size, and complexity of the data at hand, there could be many more investigative paths to pursue. It's easy to see how data exploration time adds up. It should also be clear why this time is such a wise investment: without a solid understanding of the data, and without making it suitable for the modeling step following, there are many possible ramifications that could call into question the validity of modeling results.

There often won't be a right or wrong way to handle these questions, so the most important thing is to make your investigative work repeatable and understandable by documenting the decision process so that others in the organization can understand why certain choices or assumptions were made. Deciding to take a particular analytical path does not necessarily validate or invalidate analytical conclusions, but such decisions certainly do influence how those conclusions should be interpreted. Moreover, findings from the data investigation could reveal the need to adjust data generation processes upstream.

The overall purpose of being a data detective is to gain confidence in your understanding of the data and develop a firm foundation for drawing inferences from that data. No matter how extensive your technical knowledge or ambitious your analytical goals may be, arming yourself with a healthy skepticism and a curious attitude will help to insure you against improper conclusions.

Let's consider the steps of data understanding and preparation to be an opportunity, not a burden. Just like the later phases of an analytics project, performing data detective work is about discovering truth, and failing to pursue that truth from the beginning of the process introduces unnecessary risk. The next time you find yourself in position to do some data sleuthing, try to challenge your beliefs about your data and collect the evidence you need so you can be confident that you truly understand it. The most exciting parts of data science will always be the flashy visualizations and impressive model results, of course. But we should give proper credit to the detective work that helps us get there.

About the Author

Data Scientist Ryan McGibony came to Elder Research after earning a Master of Science in Analytics from North Carolina State University. In his previous work, he conducted and analyzed custom marketing research for corporate and non-profit clients at a full-service research firm. While there, he gained experience in customer segmentation and predictive modeling techniques. Ryan also spent two years in Mongolia as a Peace Corps volunteer, supporting the staff of a local chamber of commerce in improving their service delivery to member businesses. Throughout his career, Ryan has enjoyed working with a wide variety of clients, taking care to understand their needs, and finding solutions to their problems.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673