: What is data exploration?

by Daniel J. Power

Editor, DSSResources.COM

Explorers venture into the unknown. A modern data explorer wants to find evidence and answers from data to previously unasked and unanswered questions. Exploring is a process of charting or mapping an unknown space. Data exploration is a process where non-technical managers and staff use data visualization and menu-driven query tools for one time or ad hoc analysis and search. Some data exploration may be part of a more systematic analysis. This step is also sometimes termed a descriptive or exploratory analysis. Descriptive analysis is usually a statistical profiling of the data in a data set. The analyst generates statistical measures like the mean, mode, standard deviation, range and other measures to summarize the data set. These measures of central tendency and variability help understand the distribution of values for individual data fields in the data set. With large data sets this step can identify data quality problems like outliers, unexpected values, and missing data. Inferences are also part of data exploration. Inferences may result from visual analysis or from calculating more formal measures of association or both. Measures of correlation or association indicate the statistical strength of a relationship between and among variables of interest.

Data exploration is often informal involving hunches and informal hypotheses rather than formal, testable hypotheses. Managers who interactively explore data sets suspect that a relationship exists and try to confirm or deny that suspicion. The informal nature of data exploration that involves testing inferences is concerning to some who focus on more formal survey research. The mere act of exploring can however generate more testable hypotheses for subsequent testing. The issue is more baout when the proof is adequate to act on the intuitively derived relationship. When is an informal hypothesis confirmed?

Data exploration should begin with understanding a data set and checking to see if it is relevant and useful. The first question is does the data set contain the variables of interest? If our goal is predicting sales, is the sales data usable and are there possible predictor variables in the data set. Is the data set clean and are data types compatible? Are the variables numeric and continuous? Are the variables discrete?

Following a broad overview, examine each individual variable, a univariate analysis. For continuous variables, calculate measures of central tendency and range or spread. For categorical variables, examine frequencies. Part of a univariate analysis is looking for data anomalies, i.e., outliers, possible errors, unexplained results.

Next examine any reasonable relationships that might exist among variables in the data set. If both

variables are continuous, then a simple scatter plot can help to quickly show a relationship (if one exists). If both variables are categorical, a two-way data table can show if a relationship exists. If one variable is continuous and one categorical, it is appropriate to examine the measures of central tendency within a category. For example, calculate average age of male customers and see if there is a difference with female customers.

If there are data anomalies found in data exploration, then the analyst will need to assess the best way to correct the problem. Will observations with missing values be removed? Will an average be used to replace a missing value?

There may be opportunities to transform data to improve or change analyses. For example, a continuous variable may be transformed into a categorical variable to examine "categories of shoppers". Everyone in the data set with a birthday before 1953 would be coded as a "senior citizen shopper". There may be other shopper categories as well, e.g., "youth shopper", "millennial shopper", etc.

Data exploration has an important, possibly even indispensible, role in most analytic tasks and projects (cf., Russom, 2013). Learn to explore data, be comfortable trying to figure out "what the data says" about your questions. Challenge the data and pursue answers to your questions using data and technology. Be inquisitive and persistent. Confirm and then share what you find out.

The more you know about an organization and its data, the more curious you are, and the more comfortable you are with an analysis tool, the more valuable your exploration of a data set. When the data set has been explored, cleaned and transformed, then hypotheses can be tested and more complex analyses are possible.

References

Bertini, E., "What do we talk about when we talk about 'Data Exploration'?" Medium, Oct 15, 2016 at URL
https://medium.com/@FILWD/what-do-we-talk-about-when-we-talk-about-data-exploration-2f82503fb377

Power, D. J., "What is visual analytics?" 6-12-2013 at URL
http://dssresources.com/faq/index.php?action=artikel&id=276

Russom, P., "Data Exploration in the Age of Big Data," TDWI, November 19, 2013 at URL https://tdwi.org/articles/2013/11/19/big-data-exploration.aspx

Author: Daniel Power
Last update: 2018-07-16 02:40