Data Analytics: Typical Pre-Analytical Mistakes

Author, Michael Baron

A Data Science Foundation Blog

January 2020

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3670 Email:admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

When establishing and implementing standards for data analysis processes, companies appear to be placing a very strong emphasis on following the analytical procedures religiously and ensuring that these procedures are handled in valid, reproducible and accurate ways. All of the data analysis tools used have to be in line with the industry standards and the data scientists behind the analysis are likely to have to produce periodical progress reports to see if the process management benchmarks are being met. While there is no absolute remedy against data processing & analysis errors, overall accuracy of Data Analysis Projects tends to keep increasing over the time. Greater accuracy of the tools multiplied by the greater care and systematism by the analysts does result in analytical errors becoming far less characteristic as compared with the previous decades.

However, the improved analytical practices can still be overshadowed by pre-analytical discrepancies. No matter how accurate the analysis is, it has to be carried out on the basis of verifiable data sets that have been assembled and interpreted in accordance with the project requirements. Some of the most typical (as per my subjective point of view) pre-analytical mistakes are discussed below:

Non-Representative Data Sets

Data Analytics is not going to deliver accurate results if our data is not representative of the parameters required. One classical example of such misrepresentation is when performance of a business outlet is being assessed based on the *number* of transactions but *value* of the transactions is not considered. If this is the case, A Seven-Eleven store that sells AUD$1 coffee is going to show deep levels of market penetration when compared with outlets that specialize in selling products of higher value purely due to the numbers of the transactions completed. Likewise, when collecting data on income levels within a specific region, we should also consider the living costs. If we take representativeness to the next level of complexity to carry out comparative analysis between technologies/business platforms (e.g. comparative analysis for IBM, Oracle, SAP and Peoplesoft platforms usage within an industry), the challenge of having representative data sets is going to be even greater.

Outdated Data

Unfortunately, data lifecycle is usually relatively short. While the 90/90 approach to the diminishing value of the data may be too extreme, the data almost always loses value and relevance over time. In our day and age, the value reduction occurs at a very fast pace. Any analysis that is carried out on the basis of outdated data is clearly going to be of little validity and significance. While even primary data can turn out to be outdated (since it does not take long), secondary data sets are the main ''culprits''. Curiously, I've had instances when I had secondary data provided to me for further analysis and the supplementary documentation failed to include sufficient information on the data currency. I had to have it verified prior to commencement of the analysis.

Overinterpretation of the Qualitative Data

Some analytics projects supplement verifiable Quantitative Data with Qualitative Interpretations. While Data Science would not be complete without incorporation of the Qualitative factors, we need to be extremely careful when mixing factual concrete data with our interpretations of this data. Such overinterpretations often occur due to ''wishful thinking''. We look at the data sets and see what we want to see. As Confucius pointed out centuries ago ''It is very difficult to find a black cat in a dark room particularly if there is no black cat''. Well, some qualitative analysts do imagine ''the cat'' and incorporate it into the analysis. It should also be noted that it is during the pre-analytical stages of the data analysis projects that such overinterpretations are most likely to occur!

Data Set Disparities

Drawing comparisons across a range of data sets is only possible across some common denominators. There are lots of jokes about people carrying out arithmetical tasks of deducting, adding, dividing and multiplying apples by oranges so while we laugh at the jokes wholeheartedly, we should not forget that there is a grain of truth in every joke. Furthermore, some jokes have a grain of…joke and turn out to be realities of our perceptions of the data. Just like school kids need to understand differences between apples and oranges, data analysts need to understand differences between the data sets, namely data collection conditions, environment and representativeness. And mostly importantly, during the pre-analytical stages, the common denominators MUST be identified!

If we go back to the ''apples and oranges'' examples, we can easily see that while we are dealing with 2 different kinds of fruit, *we can still carry out a comparison*! However, we need

to be very clear about what exactly we are going to compare. For instance, we could compare relative nutritional value, production opportunity costs or any other verifiable values that are equally applicable to the both commodities.

To sum up, it is evident from the discussion above that the nature of the pre-analytical mistakes has little to do with technologies or tools utilized by the data scientists. The core cause is human (aka ours) failures to understand the analytics requirements and to ensure that the analysis parameters are accurate and the so is the data collected!

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email: admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation