International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Data Analytics Scope in bigdata - Data premutation

Vidya Vijayan[1], Deepthy S[2],

[1]FComputer Science and Engineering Department, ITM Vocational University, Waghodia
[2]Computer Science and Engineering Department, Baselios Mathews II College of Engineering, Sasthamcotta

---

Abstract - MentionIn the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyze some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains.

Key Words: Data Analysis, Data Preparation, Data Analysis Methods, Data Analysis Types, Descriptive Analysis, Explanatory Analysis

1.INTRODUCTION

Data analysis is simply the process of converting the gathered data to meaningful information. Different techniques such as modeling to reach trends, relationships, and therefore conclusions to address the decision-making process are employed in this process . However, the data needs to be prepared before being used in the data analysis process. Data preparation is the process in which data is converted to the numerical format which is machine readable to be used in specific analyzing programs such as SAS or SPSS.

The steps to follow for the data preparation process are data coding, data entry, missing values, and data transformation. These steps are described briefly here:

Data Coding: Converting data to numerical values happens during the data coding process. It uses a codebook which is a document including different information such as an explanation of the variables, measures, and format of variables, the response, and finally codding them. In this process response means determining the types of scales for instance, whether the scale is chosen as nominal, ratio, ordinal, or interval; whether the scale is five-point, seven-point, etc. For example, to code the industry

type, we can use a numerical form, and the coding scheme can be considered as 1 for healthcare, 2 for manufacturing, 3 for retailing, and 4 for financial.

Data entry: In this process, the coded data from the previous step is entered into text files or spreadsheets. It also can be directly added to the statistical program.

Missing data: As some respondents may not answer all the questions because of different reasons, a method should be used to face these missed values. For example, you need to add the value -1 or 999 in some programs, some of them automatically address the missed values, and others use a listwise deletion technique facing the missing values which drop all the answers even with a single missed value.

Data transformation: Transforming data is needed before interpreting them in some cases. Reverse coded items can be considered as an example that should be transformed before comparing or combining with not reversed ones. This concept is used where the meaning of the item is opposite to their underlying construct .

2. BIG DATA ANALYTICS

Imagine a world without data storage; a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc. has become essential for day-to-day continuity. Data is the building block upon which any organization thrives. Now think of the extent of details and the surge of data and information provided nowadays through the advancements in technologies and the internet. With the increase in storage capabilities and methods of data collection, huge amounts of data have become easily available. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data has become cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data. The size, variety, and rapid change of such data require a new type of big data analytic, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.

3. STRUCTURE OF DATA ANALYSIS REPORT

The Primary audience: A primary collaborator or client. Reads the Introduction and perhaps the Conclusion to find out what you did and what your conclusions were, and then perhaps fishes/skims

through the Body, stopping only for some additional details on the parts that he/she thought were interesting or eye-catching. Organize the paper around an agenda for a conversation you want to have with this person about what you've learned about their data: e.g., from most general to most specific, or from most important to least important, etc. Provide the main evidence from your analysis (tabular, graphical, or otherwise) in the Body to support each point or conclusion you reach, but save more detailed evidence, and other ancillary material, for the Appendix.

• Secondary Audience: An executive person. Probably only skims the Introduction and perhaps the inclusion to find out what you did and what your conclusions are. Leave signposts in the Introduction, Body and Conclusion to make it easy for this person to swoop in, find the "headlines" of your work and conclusions, and swoop back out.

• Secondary Audience: A technical supervisor. Reads the Body and then examines the Appendix for quality control: How good a job did you do in (raising and) answering the interesting questions? How efficient were you? Did you reach reasonable conclusions by defensible statistical methods? Etc. Make specific cross-references between the Body and specific parts of the Appendix so that this person can easily find supporting and ancillary material related to each main analysis you report in the Body. Add text to the technical material in the Appendix so that this person sees how and why you carried out the more detailed work shown in the Appendix.

The data analysis report has two very important features:

• It is organized in a way that makes it easy for different audiences to skim/fish through it to find the topics and the level of detail that are of interest to them.

• The writing is as invisible/unremarkable as possible, so that the content of the analysis is what the reader remembers, not distracting quirks or tics in the writing. Examples of distractions include:

– Extra sentences, overly formal or flowery prose, or at the other extreme overly casual or overly brief prose.

– Grammatical and spelling errors.

– Placing the data analysis in too broad or too narrow a context for the questions of interest to your primary audience.

– Focusing on process rather than reporting procedures and outcomes.

– Getting bogged down in technical details, rather than presenting what is necessary to properly understand your conclusions on substantive questions of interest to the primary audience. It is less important to worry about the latter two items in the Appendix which is expected to be more detailed and process-oriented. However, there should be enough text annotating the technical material in the Appendix so that the reader can see how and why you carried out the more detailed work shown there. The data analysis report isn't quite like a research paper or term paper in a class, nor like a research article in a journal. It is meant, primarily, to start an organized conversation between you and your client/collaborator. In that sense it is a kind of "internal" communication, sort of like an extended memo. On the other hand it also has an "external" life, informing a boss or supervisor what you've been doing. Now let's consider the basic outline of the data analysis report in more detail:

1. Introduction. Good features for the Introduction include:

• Summary of the study and data, as well as any relevant substantive context, background, or framing issues.

• The "big questions" answered by your data analyses, and summaries of your conclusions about these questions.

• Brief outline of remainder of paper.

The above is a pretty good order to present this material in as well.

2. Body. The body can be organized in several ways. Here are two that often work well:

• Traditional. Divide the body up into several sections at the same level as the Introduction, with names like:

– Data

– Methods

– Analysis

– Results

This format is very familiar to those who have written psych research papers. It often works well for a data analysis paper as well, though one problem with it is that the Methods section often sounds like a bit of a stretch: In a psych research paper the Methods section describes what you did to get your data. In a data analysis paper, you should describe the analyses that you performed. Without the results as well, this can be pretty sterile sounding, so I often merge these "methods" pieces into the "Analysis" section when I write.

• Question-oriented. In this format there is a single Body section, usually called "Analysis", and then there is a subsection for each question raised in the introduction, usually taken in the same order as in the introduction (general to specific, decreasing order of importance, etc.).

Within each subsection, statistical method, analyses, and conclusion would be described (for each question). For example:

2. Analysis

2.1 Success Rate

Methods

Analysis

Conclusions

2.2 Time to Relapse

Methods

Analysis

Conclusions

32.3 Effect of Gender

Methods

Analysis

Conclusions

2.4 Hospital Effects

Methods

Analysis

Conclusions

Etc. . .

Other organizational formats are possible too. Whatever the format, it is useful to provide one or two well-chosen tables or graphs per question in the body of the report, for two reasons: First, graphical and tabular displays can convey your points more efficiently than words; and second, your "skimming" audiences will be more likely to have their eye caught by an interesting graph or table than by running text. However, too much graphical/tabular material will break up the flow of the text and become distracting; so extras should be moved to the Appendix.

3. Conclusion(s)/Discussion.

The conclusion should reprise the questions and conclusions of the introduction, perhaps augmented by some additional bservations or details gleaned from the analysis section. New questions, future work, etc., can also be raised here.

4. Appendix/Appendices. One or more appendices are the place to out details and ancillary materials.

These might include such items as

• Technical descriptions of (unusual) statistical procedures

• Detailed tables or computer output

• Figures that were not central to the arguments presented in the body of the report

• Computer code used to obtain results.

In all cases, and especially in the case of computer code, it is a good idea to add some text sentences as comments or annotations, to make it easier for the uninitiated reader to follow what you are doing. It is often difficult to find the right balance between what to put in the appendix and what to put in

the body of the paper. Generally you should put just enough in the body to make the point, and refer the reader to specific sections or page numbers in the appendix for additional graphs, tables and other

details.

4. CONCLUSION

This article provided a summary of the most common data analysis techniques. It first describes data preparation methods which are an essential process in analyzing data. Then, common methods are reviewed, and the tools for the most important techniques are discussed. Qualitative data analysis and its strategies are also discussed more specifically in the final section.

REFERENCES

1.Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research
52(1), 11–19 (2010)
2.Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International
Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499
(2010)
3. Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of
the IEEE Aerospace Conference, pp. 1–7 (2012)
4. Cebr: Data equity, Unlocking the value of big data. in: SAS Reports, pp. 1–44 (2012)
5. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analy
sis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492
(2009)
6. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data:
The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data
Warehousing and OLAP, pp. 101–104 (2011)
7. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In:
Capgemini Reports, pp. 1–24 (2012) Big Data Analytics: A Literature Review Paper 227
8. Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis,
German University in Cairo, p. 164 (2013)
9. EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508