

Exploratory Data Analysis and its applications in Data science and Supply Chain Management

Submitted by,
Feljose Shristeca

Table of Contents

Introduction.....	3
Applications of EDA in Data Science	4
Applications of EDA in SCM	5
EDA of Supply Chain Management with Power BI	5
Conclusion	5
Reference	6

Table of Figures

Figure 1: Fundamental steps of Exploratory Data Analysis	3
Figure 2: Crisp DM methodology	4

Introduction

The significance of data in the current world is high that it can put together as “The new oil”. Like oil, data can also be used effectively only when it is refined. Hence it is obvious to analyse and explore the data. This step generally known as the Exploratory Data Analysis, helps to perform a research on the data to discover patterns, uncover underlying insights, find out key variables, identify outliers, test our assumption about the dataset etc. EDA is cross classified into two ways where each method is a combination of,

- Graphical or non-graphical
- Univariate or multivariate

Graphical is always visualisations done with the dataset and non-graphical involves statistical calculations done with the dataset. Univariate is where the data is being analysed over one variable whereas analysis done over two or more variables is multivariate. Some classic visuals used for performing univariate analysis are box plot and histogram and for multivariate analysis bar chart and scatter plot are the most common visuals. The basic quantitative calculations (non-graphical) are mean, median, min, max, standard deviation and quartile.

In addition to the above methods, exploratory data analysis methods include,

Dimensionality reduction: Technique used to reduce the number of features by removing columns that are unwanted, that has only a single value, that are highly correlated etc that increase the volume of the data and results in slow refreshes.

Clustering: Grouping of similar features in the dataset into smaller groups.

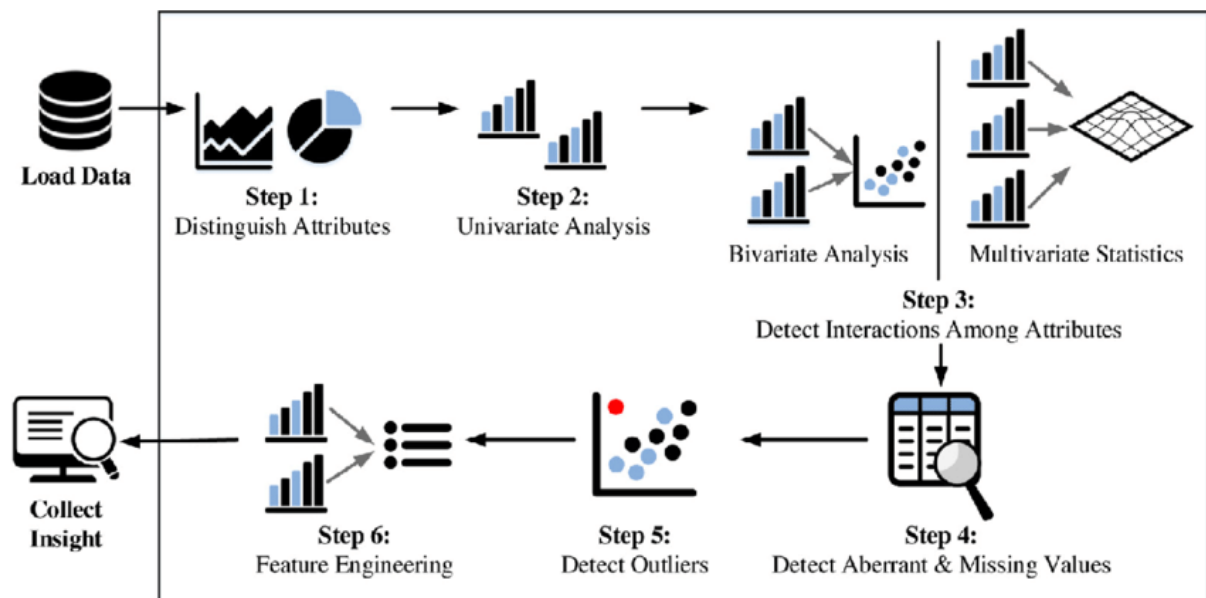


Figure 1: Fundamental steps of Exploratory Data Analysis

Applications of EDA in Data Science

When it comes to Data Science, Exploratory Data Analysis is an important step before diving into machine learning or statistical modeling. It enables them to gain in depth knowledge of the variables in the datasets and their relationships. The insights obtained through EDA will help to develop an appropriate model and produce accurate results.

The common methodology followed by data scientist to achieve their result is the CRISP DM methodology. It gives a structured approach in planning a data science project. The third step, data preparation is where the data is explored and analysed to gain insights and further move on the next process which is the data modeling process.

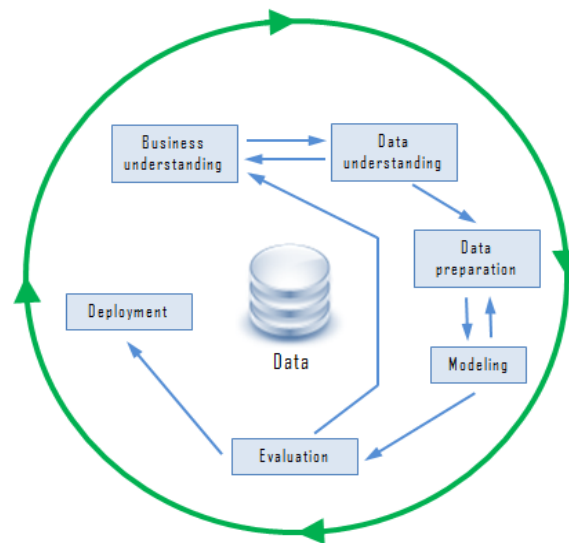


Figure 2: Crisp DM methodology

Before building visuals to gain knowledge over the dataset, there are few steps that has to be explored to get ready with the data. Some of the steps include checking,

- The null values
- The data type of each column (if any mismatches)
- Maximum and minimum value of each column
- Highly correlated features
- Single value columns
- Variations in the values
- If any extra columns needed (that can be derived for other columns)

Clear evaluation of the above steps will help in gaining more understanding over the data. In this digital age, organizations have realized the vital role of data to improve its efficiency and productivity. The growth of data science and its innovative analytical technologies has made EDA no longer a strategic plan. Instead, it has gained importance as one of the most significant elements in optimizing the core business process.

Supply Chain Management, healthcare, banking, retail as some of the sectors where use of exploratory data analysis is very much helpful.

Applications of EDA in Supply Chain Management

Supply chain management is the management of the flow of goods and services and includes all processes from acquiring the raw materials to the final product delivery stage. It attempts to centrally control the production, distribution and shipping of a product. This process helps the companies to cut excess cost by keeping tighter control over the inventories, production distribution and sales. Though SCM have existed for years, most companies have recently paid attention to them as a value-add to their operations. The competition among the enterprises and their supply chains have increased which has resulted in huge volume and different type of data. This has created a need to rapidly analyse the data and this where EDA comes into play.

By exploring the data, the stakeholders can make sure that the right question is asked by providing a well-defined context around the problem and maximize the potential value of the output. EDA can also lead to insights we would not even think to investigate but can be hugely informative to the business. These analysis attempt to establish metrics and goals to control cost, optimize resources and ensure customer partner satisfaction.

EDA of Supply Chain Management with Power BI

EDA is generally performed using popular scripting languages like R and python or Business Intelligence tools like Qlik sense, tableau etc. For this analysis purpose power BI will be a good option.

Power BI lets you analyse and create reports, KPI's, dashboards that easily visualize results with clear outcomes, manage alerts and use auto-recommend and auto-suggestion tools to prepare forecasts and predictions and establish objective goals for supply chain management. Features like drill up, drill down, drill through ensure that the supply chain runs smoothly and the issues that threaten business are noted immediately. The drill down function, helps to dive deep and identify the root cause of the issue and adapt to the changing requirements to revert back to the supplier, step up production, add resources of shifts, monitor customer satisfaction and appropriate decisions.

The functional language DAX (Data Analysis Expression) in power bi is used for data analysis calculations. They contain inbuilt functions that can be used to create new columns and measures that help to know more about the dataset. Power BI also has the capability to connect with scripting languages like R and python to make use of their unique features. Overall, the benefits of using Power BI includes in better understanding of customer behaviour, higher supply chain transparency and visibility and higher operational efficiency.

Conclusion

EDA is an important step when it comes to data science. Skipping this step might lead to missing out valuable insights that can be identified to enhance the business. Hence, it is better to keep in mind the following steps while performing EDA, that includes:

- Asking the right question
- Set clear objectives
- Gain knowledge on the problem domain.

Reference

- 1) IBM.com. What Is Exploratory Data Analysis?. [online] Available at: <<https://www.ibm.com/cloud/learn/exploratory-data-analysis>>.
- 2) Investopedia. Supply Chain Management (SCM): What You Need To Know. [online] Available at: <<https://www.investopedia.com/terms/s/scm.asp>>.
- 3) Medium. Guide To Exploratory Data Analysis For Data Science. [online] Available at: <<https://medium.com/swlh/guide-to-exploratory-data-analysis-for-data-science-294baff8b741>>.
- 4) PhD, J., Significance Of Exploratory Data Analysis (EDA). [online] Jeannjoroge.com. Available at: <<http://www.jeannjoroge.com/significance-of-exploratory-data-anaysis/>>.
- 5) Simplilearn.com. Exploratory Data Analysis: Techniques, Best Practices & Applications. [online] Available at: <<https://www.simplilearn.com/exploratory-data-analysis-article>>.