IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

Importance of Data Exploration in Data Analysis A Review Paper

Abhishek Sheshnath Jaiswal
Department of Information Technology
Sir Sitaram and Lady Shantabai Patkar College of Arts and Science, Mumbai, Maharashtra, India
abhishekjaiswalsheshnath@gmail.com

Abstract: *Exploration, one of the first steps in data processing, is a way to know the data before working with it. Through research and investigation, large data sets are prepared for in-depth, systematic analysis. Data Analytics refers to the process of analyzing data collected from a variety of sources in order to reach a meaningful conclusion. This process enables us to capture raw data and reveal patterns to extract important information or details from it. It helps organizations and individuals to make sense of the data collected. There are various tools and strategies that help organizations make decisions and succeed in them. Once the data has been collected it is important to process that data. In Data Analytics, Data Exploration is the first or main step used to understand, evaluate and visualize data to obtain important information from the beginning or to identify patterns or key areas that you can dig deeper. It Uses a combination of automated tools and manual methods such as charts, visuals, and reports. In this case, we get a lot of details from the data, reveal its basic structure, detect any external, error data, and confusion if there is data, evaluate the basic assumptions, and determine the appropriate feature settings. Using data exploration tools and methods such as dashboards, reports, and point-to-point data test users can understand the big picture and can find information on it easily.*

Keywords: Data Analysis, Data Exploration, Data Management

I. Introduction

The word “data” is derived from the Latin word dare, which means “something given”—an observation or a fact about a subject. (Interestingly, the Sanskrit word dAta also means “given”). Data science helps to clarify the useful hidden relationships between data. Before embarking on any advanced data analysis using statistics, machine learning, and algorithmic techniques, it is important to perform a basic data analysis to learn the basic features of the dataset. Data exploration helps to better understand data, process data in a way that makes advanced analysis possible, and sometimes obtain the required data in data faster than using advanced analysis techniques.

Data is often collected in large, unstructured volumes from a variety of sources and data analysts must first understand and develop a broader view of data before releasing relevant data for further analysis, such as univariate, bivariate, multivariate, and key component analysis.

Data exploration definition: Data exploration refers to the first step in data analysis where data analysts use data visualization and statistical techniques to define dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

Data exploration techniques include both manual analysis and automated data exploration software solutions that visually explore and identify relationships between different data variables, the structure of the dataset, the presence of outliers, and the distribution of data values in order to reveal patterns and points of interest, enabling data analysts to gain greater insight into the raw data.

II. Objectives of Data Exploration

In the data science process, data exploration uses many different steps including preprocessing or data preparation, modeling, and interpretation of the modeling results

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

1. **Data understanding:** Data exploration provides a high-level overview of each attribute (also called variable) in the dataset and the interaction between the attributes. Data exploration helps to answer the questions such as what is the typical value of an attribute or how much do the data points differ from the typical value, or presence of extreme values.
2. **Data preparation:** Before using the data science algorithm, the dataset should be configured to manage any possible anomalies in the data. These anomalies include outliers, missing values, or highly correlated attributes. Some data science algorithms do not work well when input attributes are correlated with each other. Therefore, correlated attributes need to be identified and removed.
3. **Data science tasks:** Exploring basic data can sometimes replace the entire data science process. For example, scatterplots can identify clusters in low-dimensional data or can help to develop regressionr or classsification models with simple visual rules.
4. **Interpreting the results:** Finally, data exploration is used to understand the prediction, classification, and clustering of the results of the data science process. Histograms help to comprehend the distribution of the attribute and can also be useful in visualizing numeric prediction, error rate estimation, etc.

III. Important Steps in Data Exploration

After data preparation step data exploration is required. The modified dataset is analyzed to enable queries from the data preparation section.Data exploration plays an important role because the quality of input is directly proportional to quality of output. In data exploration large amount of project time is spent on cleaning and preparation of the data for further deep analysis. The following are the steps involved in preparing, understanding and cleaning data for predictive modelling :

1. **Variable Identification:** For variable identification, we need to identify predicator that is input variable and output variable in order to further evaluate the data . Based on our needs we can change the data type of the variable.
2. **Univariate Analysis:** In the univariate analysis, we need one to explore the variable one after another for performing univariate analysis it depend on variable type, that is if variable is continuous or categorical .
3. **Bi-variate Analysis:** Bi-variate analysis helps to find the relationship between two variables. We can apply this analysis to any kind of combination of categorical and continuous variables. There are several kinds of methods used to tackle this kind of combination of variables during the analysis process. The possible combinations of variables are categorical and categorical, categorical and continuous & continuous and continuous.
4. **Treatment of missing values:** Defective amounts in training data that need to be managed cause if we do not correct them in a timely manner will result in a misinterpretation of the predictions later. There are a few ways to manage these missing amounts in the data such as the removal of pairs or a list containing missing values, the standard mode and the average rating this fills the missing values with limited values, the prediction model is one of the more complex to use. and using non-data values, the KNN calculation is also used for the treatment of missing values, in which case the missing attribute values are calculated using a given number of adjectives such as the attribute to their missing values in the database.
5. **External treatment:** Abnormal monitoring of data may result in external data. Data analysts and scientists need to identify these outsiders before they can come up with the most accurate estimates. There are different types of external product such as input errors, measurement errors, intentional builders, external testers, sample error, data processing error, and external environmental factors. Outliers can be obtained using Box-plot, histogram, and scatterplot during viewing. To exclude outsiders from data it uses certain methods such as removing views, setting, adjusting and consolidating values, and managing separately
6. **Variable transformation:** This refers to replacing variables with the function. There are three types of variable transformation Logarithm, Binning, and Square or Cube root. The variable transformation changes the relationship or distribution of the variable with the others. This is used when we need to change the scale of a variable or standardize the variables for good understanding, when we can transform the complex non-linear relationship into linear ones, symmetric distribution is favored over the skewed distribution as it is easier to generate inference, interpret, and variable transformation is also done from the implementation viewpoint.
7. **Variable or Feature creation:** This is the process to generate new variables from existing or old variables as an input variable in the data set. This is used to highlight the relationship between the hidden variables. There are

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

different techniques to create the variables or generate new features such as creating derived variables and creating dummy variables.

8. So, from all these steps in data exploration, we get a better and deeper understanding of the dataset, which it makes easier to navigate and use the data later easily.

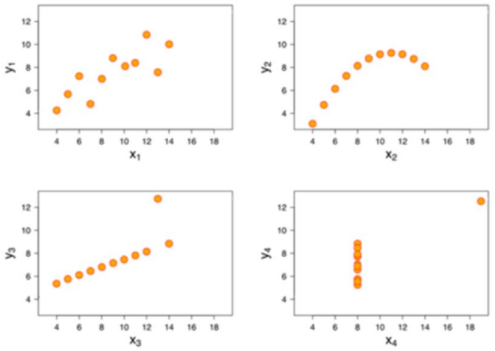
IV. Data Exploration Necessity

The two most important benefits of data analysis are ...

1. To enable unexpected discoveries in the data
2. To promote a deeper understanding of the data as an essential fundament for successful and efficient data science projects.

A simple and popular example of mathematician F. J. Anscombe illustrates these features very well. Displays four different data sets, each data set with two variables and eleven data points.

Although the distribution and correlation of data sets looks very different, all four data sets with accurate values have the same meaning and standard deviations for both the dynamics and the corresponding coefficients and the linear polynomials.



Mean X: 9
Mean Y: 7,5

Variance X: 11
Variance Y: 4,125

Correlation: 0,816

Regression: y = 3 + 0,5x

V. How Data Exploration Works

Data without a question is just information. Asking a question of data turns it into an answer. Data with the relevant questions and exploration can provide a deeper understanding of how things work and even enable predictive abilities.

R and Python are the most common languages used for exploration; the first one works best for statistical learning while the latter lends itself well to machine learning. Coding is not required for data exploration through no-code platforms. The exploration process is also increasingly important to working with Geographic Information Systems (GIS) since so much of today's data is location-enriched.

Data exploration typically follows three steps:

1. **Understand the Variables:** The basis of any data analysis begins with an understanding of variables. A quick read of column names is a good place to start. Examining data catalogues, field descriptions, and metadata can offer insight into to what each field represents and help discover missing or incomplete data.
2. **Detect Any Outliers:** Outliers or anomalies can derail an analysis and distort the reality of a dataset, so it is important to identify it in advance.Data visualization, numerical methods, interquartile ranges, and hypothesis testing are the most common ways of detecting outliers. A boxplot, histogram, or scatterplot, for example, makes it easy to spot points far outside the standard range, while a z-score informs how far from the mean a data point is. Once found, an analyst can investigate, adjust, omit, or ignore the outliers. No matter the choice, the decision should be noted in the analysis.
3. **Examine Patterns and Relationships:** Plotting a dataset in a variety of ways makes it easier to identify and examine the patterns and relationships among variables. For example, a business exploring data from multiple stores may have information on location, population, temperature, and per capita income. To estimate sales for a new location, they need to decide which variables to include in their predictive model.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

VI. Importance Of Data Exploration

Exploration allows for deeper understanding of a dataset, making it easier to navigate and use the data later. The better an analyst knows the data, the better their analysis will be. Successful exploration start with an open mind, reveals new paths for discovery, and helps identify and refine future analytics questions and problems.

People process visual data better than numerical data, so it is a big challenge for data scientists and data analysts to provide meaning to thousands of lines and columns of data points and articulate that meaning without any visual components. Data visualization in data exploration leverages familiar visual cues such as shapes, dimensions, colors, lines, points, and angles so that data analysts can effectively visualize and define the metadata, and then perform data cleansing. Performing the initial step of data exploration enables data analysts to better understand and visually identify anomalies and relationships that might otherwise go undetected.

Visualizing data is much easier for humans rather than just mathematical data,so it is quite challenging for data analysts or data scientists to provide significantly large amounts of rows and columns of data and get information from it without any visual parts. Exploratory data analysis gives utmost value to any business by helping analysts or scientists to understand if the results that they have obtained are correct. The following are the some importance of data exploration in data analysis :

1. Identifying missing and incorrect data in a data set.
2. Significant and critical dynamics of your database
3. Understanding and Mapping the key dynamics that are fundamental to your database
4. Exploring the considerations or exploring the theory of a particular model
5. Creating a low-level model, a model that can interpret your data using variable variables
6. Finding error parameters and measurement parameters
7. Data analysis provides the context needed to develop an accurate and relevant model for accurate and effective data interpretation.
8. It enables us, in unexpected data acquisition Provides a deeper understanding of data as a key element in successful and efficient data science projects.
9. With easy-to-use interaction, anyone throughout the organization can familiarize themselves with the database, generate meaningful questions that can go deeper, discover patterns or trends, and gain critical analysis to make decisions later.
10. Enables users to view data from any view. It speeds up response time and deepens users' understanding by laying more foundation in less time. Evaluation is required in decisions, who receive information from data that was previously difficult to obtain and view.

VII. Conclusion

Data exploration in data analysis is obviously one of the most important steps in the whole process of data analysis and getting insights from it. By laying a solid foundation for the further analysis process data exploration plays a crucial role that you should focus for the strength. The main use of data exploration is, to assist in the analysis of the data prior to making any assumption or decision regarding something important. Most data analysts and data scientists employ data exploration in order to ensure that the results they produce or obtain are accurate and acceptable for any desired business goals and outcomes. The better an analyst or scientist knows the data they are operating on or working with, the better their analysis will be. Successful exploration begins with an open and clear mind, reveals better insights and different path for discovery, and help us to identify and refine future analytics problems and questions.

Though data exploration might take some significant amount of effort that is it might involve large datasets of the data that are being identified and sorted using various tools and techniques, these techniques may require a lot of effort and time to understand and adopt. But this surely results in the good model than bad ones. In the whole world a significantly large amount of data is accumulated, structured and unstructured volumes from the sources across the whole globe so it is necessary for us to understand and comprehensive, complete view of the data. Such kind of correct and comprehensive view is essential to be able to use the data collected from various sources for further analysis. Successfully extracting the data will ensure organizations or businesses will not miss out on any opportunities to leverage web data and will not be left behind due to incomplete data access, erroneous data, poor quality data, unreliable data, out of date data, high costs, or any

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

uncertain business risks. A lot of hard work goes into extracting, exploring, and transforming data into a usable format, but once it is done it can provide users or customers with greater insights into the business and industry they are working in.
All in all, in this way all research and developments, engineering, and data science are those fields that can benefit a lot from the data exploration during the data analysis process. In today's world with computing power and modern analytics support, interactive data exploration and engaging experience for everyone to discover and unfold value in large amounts of complex data. Though a lot of hard work and effort goes into cleaning, preparing, transforming, and extracting data once it is complete it gives better insights to data analysts for decision making.

VIII. ACKNOWLEDGEMENT

I would like to express my sincere gratitude towards the Information Technology Department of Patkar- Varde College. I give my special thanks and sincere gratitude towards the In-Charge Principal Dr. Trisa Joseph, Chief Co-ordinator Ms. Ruchita Rane and Co-ordinator of IT Department Mr. Chayan Bhattacharjee. I owe my sincere thanks to Mr. Sujal Shah Sir for constant support, encouragement and for guiding me.

REFERENCES

[1]. A. Bagozi, D. Bianchini, V. De Antonellis, A. Marini, D. Ragazzi, Summarisation and Relevance Evaluation Techniques for Big Data Exploration: the Smart Factory case study. Proc. of 29th Int. Conference on Advanced Information Systems Engineering (CAISE'17), pp. 264–279, 2017
[2]. M. Golfarelli, S. Rizzi, Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill, 2009.
[3]. L. Orr, D. Suciu, M. Balazinska, Probabilistic Database Summarization for Interactive Data Exploration. Proc. of the VLDB Endowment 10, pp. 1154–1165, 2017.
[4]. https://www.alteryx.com/glossary/data-exploration
[5]. https://www.heavy.ai/learn/data-exploration
[6]. https://visplore.com/benefits-of-data-exploration/
[7]. 2 Miles M, Huberman A. Qualitative data analysis. London: Sage, 1984.