What is Data Exploration? and its Importance in Data Analytics

Rutuja Magdum[1]

[1]*Student , B.Tech In Information Technology , DKTE Ichalkaranji (Maharashtra),India*

***

Abstract – *Data Analytics refers to the process of examining the data collected from various sources to draw a meaningful conclusion from it. This technique enables us to take raw data and uncover patterns to extract valuable information or insights from it. It helps organizations and individuals make sense of the data collected. There are various tools and techniques that help organizations for making decisions and succeed in it. Once the data is collected it is important to process that data.*

In Data Analytics, Data Exploration is the first or primary step used to understand, explore and visualize the data to gain valuable insights from the beginning or identify the pattern or important areas to dig in deeper. It uses the combination of automated tools and manual methods such as charts, visualizations, and reports. From this, we get maximum insights from the data, uncover its underlying structure, detect any outliers, erroneous data, and anomalies if any are present in the data, test underlying assumptions, and determine the optimal factor settings. Using data exploration tools and techniques such as dashboards, reports, and point-to-point data exploration users can understand the bigger picture and can gain insights from it easily.

Key Words: Data Analysis, Data Exploration, Data Management

1.INTRODUCTION

Data is often gathered in large and in unstructured format or volumes from various sources and data analysts must first view , understand and develop a comprehensive view of the data before extraction of the data for later analysis . Data explorationinclude both manual as well as automated tools and technologies for data exploration. Manual data exploration tools entail either writing scripts to analyze the raw data or manually filtering data into spreadsheets. Automated data exploration tools, such as data visualization software, help data scientist easily perform big data exploration monitor data sources and monitor data sources on otherwise overwhelmingly large data. Graphical representation of data, such as scatter plots and bar charts, are valuable tools in visual data exploration. This enables us deeper analysis of trends and patterns that need to be identified .Data Exploration create straightforward and clear view of the data.

In data exploration, the most essential steps are variable identification, univariate and bivariate analysis, missing values treatment, outlier treatment, variable transformation and variable creation etc. Variable identification , univariate and bivariate analysis or tools can be used to begin to gain the knowledge or information of the data . By skipping this step data scientists are not be able to readily understand the main or key issue in the data or be able to guide into deeper analysis in correct direction. Understanding data and interpreting from large volume data can be really challenging. It is truly difficult to understand the data and make conclusions or gain valuable insights without looking through the entire dataset. This means that we have to spend more time on exploring the sample of the data to get better representation of the data. By using the different exploratory data analysis techniques, methods and visualizations will ensures that we have best understanding of our data. Then, Once data exploration has uncovered the connections within the dataset and formed into different variables, it is easier to interpret the data into charts , reports , dashboards or visualizations.

2. IMPORATANT STEPS IN DATA EXPLORTION

After data preparation step data exploration is needed. The prepared dataset is analyzed to enable questions arising from the data preparation stage. Steps in 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.

Following are the steps involved in preparing, understanding and cleaning data for predictive modelling :

1] Variable Identification:

In variable identification, we need to identify predicator that is input variable and output variable for the further data exploration . Based on our needs we can change the data type of the variable.

2] Univariate Analysis:

In the univariate analysis, we need to explore the variable one after another . In order to perform univariate analysis it depend on variable type, that is if variable is continuous or categorical .

3] Bi-variate Analysis:

The bi-variate analysis helps find the relationship between two variables. We can use this analysis for 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] Missing values treatment :

The missing values in the training data need to be treated cause if we do not correct them at the right time it will result in wrong classifications and predictions later. There are several methods to treat these missing values in the data such as deletion of pairs or list that contain missing values, mean mode and median imputation this method fills the missing values with the estimated values, prediction model is one of the sophisticated methods for using and operating the missing values in the data, KNN imputation is also used for missing values treatment, in this method missing values of an attribute are imputed using the given number of attributes that are similar to the attribute whose values are missing in the dataset.

5] Outlier treatment :

Abnormal observations in the data can cause outliers in the data. Data analysts and scientists need to identify these outliers before they will result in severely wrong estimations. There are different types of an outlier such as data entry errors, measurement errors, intentional outliers, experimental outliers, sampling error, data processing error, and natural outliers. Outliers can be detected using Box-plot, histogram, and scatterplot during visualization. To remove the outliers from the data some techniques are used such as deleting the observation, imputing, transforming and binning the values, and treating 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 different techniques to create the variables or generate new features such as creating derived variables and creating dummy variables.

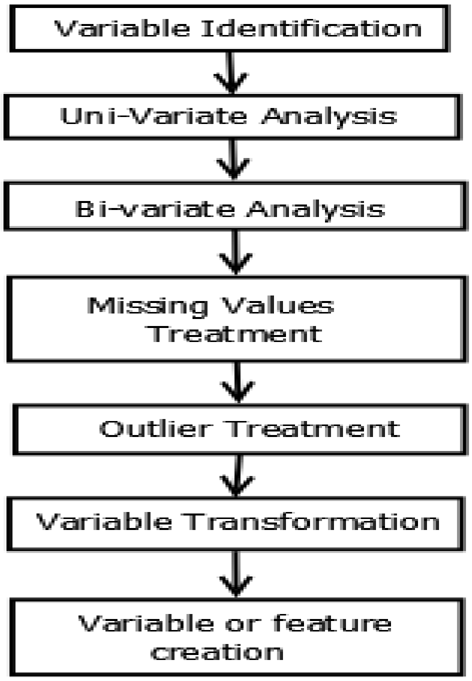

fig. Important steps in Data Exploration

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.

4. DATA EXPLORATION TECHNIQUES

- There are so many different techniques are used in data exploration some of which are mentioned as follows:

- To identify how frequently individual values occur in the column. This will give an insight into the content of categorical variables.

- Correlation

- Heat-map between numeric columns is a great way to understand the relationship between the various types of variables that are present.

- Use of unique count value of categorical columns.

- Cramer V is an effective data exploration that correlated between every categorical variable.specialized visualizations from scatter plot and bar charts to radar charts , Sankey charts and neural network visualization.

- Pareto analysis is also an effective data exploration technique as well.

- Numeric values, the minimum, maximum, and variance of the data values in data analysis provide a good insight or indication of the spread of the values.
- Histograms are also used to get information about the range of values falling into the majority sectors. It can point out any skew in the data as well as the maximum and minimum values of the data.
- The Pearson correlation method is used to understand the trend between two numeric columns.
- Cluster size analysis, grouping things together gives us a high-level perspective or insights.
- In segmentation, after identifying the number of clusters it is important to divide all data into a specific number of segments.
- Outlier analysis for multiple columns is one of the most important steps in exploratory data analysis which is based on finding outliers in the data based on multiple columns.

From this, we can see that data exploration is the initial step in data analysis that involves both the manual and automated software that visualizes and identify the relationship between the different variables or features, the dataset structure, presence of outliers in the data, and distribution of values in order to reveal the points and the pattern of our interest enable analysts to gain better insights from it.

IMPORTANCE OF DATA EXPLORATION IN DATA ANALYSIS

Visualizing data is quite easier for humans rather than only just mathematical data, therefore it is quite challenging for data analysts or data scientists to allocate 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.

Following are the some importance of data exploration in data analysis :

1) Spotting missing and erroneous data in the data set.

2) Identifying the valuable and important variables in your dataset

3) Understanding and Mapping the underlying important variables in your dataset

4) Checking assumption or testing the hypothesis of the specific model

5) Creating a parsimonious model, the model that can explain your data using minimum variables

6) Figuring the margins of errors and estimating parameters

7) Data exploration provides the context needed to develop a correct and appropriate model to interpret the insights correctly and efficiently.

8) It enables us, to unexpected discoveries in the dataset

Gives a deeper understanding of the data as an important fundamental thing for successful and efficient data science projects.

9) With the user-friendly interface, anyone across an organization can familiarize themselves with the dataset, generate thoughtful questions that may spur on deeper, discover the patterns or trends, gain valuable analysis in order to make decisions later.

10) It empowers users to explore data in any visualization.

Speeds up a time to answers and deepens understanding of users by covering more ground in less time.

Exploration is necessary for decisions, who obtain information from data that was previously hard to obtain and perceive.

5. CONCLUSIONS

Data exploration in data analysis is clearly one of the most important steps during the whole process of data analysis and getting insights from it. By setting a strong foundation for the further analysis process data exploration plays a crucial role that you should focus for the strength. The primary 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 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.

6. REFERENCES

1) L. Orr, D. Suciu, M. Balazinska, Probabilistic Database Summarization for Interactive Data Exploration. Proc. of the VLDB Endowment 10, pp. 1154–1165, 2017.

2) Data analytics made accessible by Anil Maheshwari Roberts P. Practical issues in 'writing up' a research thesis. Nurse Res 2000;7:14-23.

3) Kallestinova ED. How to write your first research paper. Yale J Biol Med 2011;84:181-90.

4) Roberts P. Practical issues in 'writing up' a research thesis. Nurse Res 2000;7:14-23.

5) Kallestinova ED. How to write your first research paper. Yale J Biol Med 2011;84:181-90.

6) 2 Miles M, Huberman A. Qualitative data analysis. London: Sage, 1984.

7) Kelle U, ed. Computer-aided qualitative data analysis: theory, methods and practice. London: Sage, 1995.

8) https://www.jigsawacademy.com/blogs/business-analytics/data-exploration/ 10 Lee R, Fielding N. User's experiences of qualitative data analysis software. In: Kelle U, ed. Computer aided qualitative data analysis: theory, methods and practice. London: Sage, 1995.

9) https://dl.acm.org/doi/abs/10.1145/2723372.2731084

10) htts://blog.ml.cmu.edu/2020/08/31/2-data-exploration/

11) htts://searchbusinessanalytics.techtarget.com/definition/data-exploration

12) Miles, M., & Huberman, M. (1994). Qualitative data analysis: A sourcebook of new methods (2nd ed.). Newbury Park, CA: Sage.

13) Wouters, L., Gohlmann, H. W., Bijnens, L., Kass, S. U., Molenberghs, G. and Lewi, P. J. 2003. Graphical exploration of gene expression data: a comparative study of three multivariate methods. Biometrics 59: 1131–1139.

14) Burnard P. A pragmatic approach to qualitative data analysis. In Newell R, Burnard P (eds). Research for evidence based practice. pp 97–107. Oxford: Blackwell Publishing, 2006.

15) Cutcliffe J R, McKenna H P. Establishing the credibility of qualitative research findings: the plot thickens. J Adv Nurs 1999; 30: 374–380.