IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review on Data Exploration and Data Mining Evolution

[1]Soham Navandar, [2]Ganesh Bhutkar

[1]Research Scholar, [2]Associate Professor

[1]Department of Instrumentation and Control Engineering, [2]Department of Computer Engineering

[1]Vishwakarma Institute of Technology, Pune, India, [2]Vishwakarma Institute of Technology, Pune, India

Abstract: This study has been undertaken to investigate the evolution of the technology from static websites to the transformation in peer to peer decentralized technology. The processes involved in processing data that is data exploration and mining. The paper deals with different types of techniques and processes used in the data exploration and data mining. Since the size of data has been growing in terabytes after the technology boom in the early 1990s, as a result of which internet became mainstream for data evolution. The size of data has been growing drastically since then and has become a very important and significant part for technology development. Therefore, data mining and data processing has become the cerebral part of all the tech driven businesses in the 21st century and is estimated to grow at a very steep rate.

***Index Terms* – Data mining, Data exploration, Data mining process, Blockchain technology, Peer to peer network.**

I. INTRODUCTION

In today's world, data has evolved considerably with the revolution of web, from Web 1.0 to Web 2.0 to Web 3.0. With Web 3.0 being the latest hype from digital marketplaces to digital currencies, data can be sent through a peer to peer network without the presence of a third party. How has the web evolved?

1.0.1 WEB 1.0

It alludes to the early development of the World Wide Web. In Web 1.0, the vast majority of users were content consumers and there were very few content creators. Personal websites were widespread and mostly included static pages maintained on free web hosts or hosted by Internet Service Providers (ISP) run web servers. Vital aspects of Web 1.0 include:

1. The server's file system is used to serve content.
2. Web pages are static.
3. Websites are created with the Common Gateway Interface (CGI).
4. The frames and tables are utilized to position and arrange the elements on a page.

1.0.2 WEB 2.0

Web 2.0 refers to the age of dynamic websites making the content more reactive to the user. A user can interact, change, and build any content using simple tools in this era of Web 2.0. Websites that prioritize user-generated content, usability, participatory culture, and interoperability for end users are a part of Web 2.0. Vital aspects of Web 2.0 include:

1. Free information sorting enables users to retrieve and categorize the data as a whole.
2. Content is generally responsive to user input and dynamic.
3. Information is exchanged between site owners and users through evaluation and online comments.
4. Application Programming Interfaces (API) are created that permit self-use, such as by a software program.
5. Due to the flexibility provided by the Web 2.0, the access to the web ranges from the typical internet user base to a wider range of users.

1.0.3 WEB 3.0

Web3 which is commonly referred to as Web 3.0 is a concept for a decentralized internet that is based on open block chains. The idea became well-known in 2020-21 with the rapid increase in interest from crypto currency fans and investments from well-known technologists and businesses. Vital aspects of Web 3.0 include:

1. Artificial Intelligence and Machine Learning (AI / ML),
2. 3D Graphics.
3. Connectivity.
4. Ubiquity.

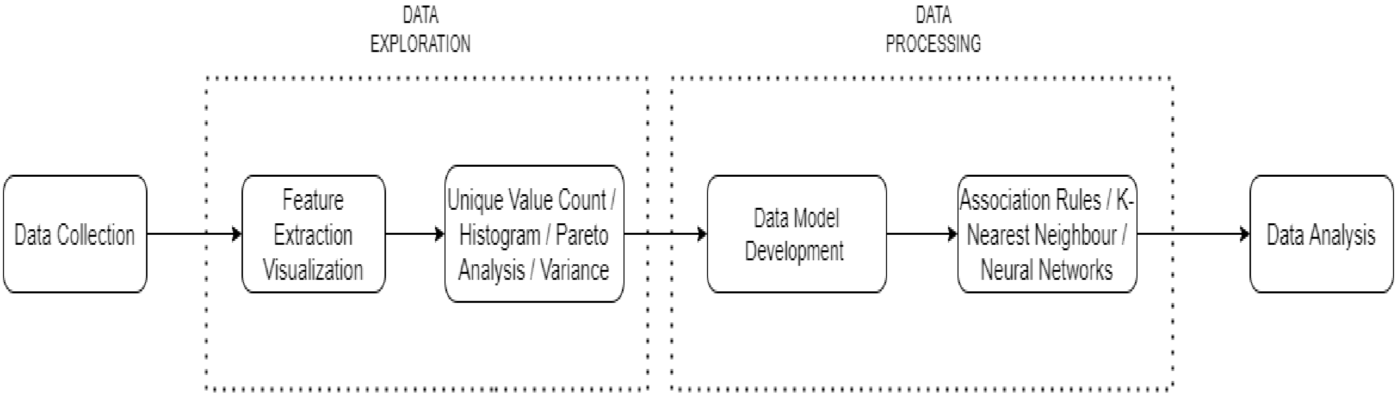

DATA EXPLORATION
DATA PROCESSING
Data Collection
Feature Extraction Visualization
Unique Value Count / Histogram / Pareto Analysis / Variance
Data Model Development
Association Rules / K-Nearest Neighbour / Neural Networks
Data Analysis

Fig. 1: Data Mining Block Diagram

1.2 How is Data mining an important part for WEB 3.0?

The block diagram in Fig. 1 provides a holistic review on the complete data mining process. The diagram depicts the stages through which loads of data is passed before analyzing. Nowadays, a lot of data is being gathered by all software companies, hence data exploration is becoming a cerebral tool day-by-day which helps in data mining and extraction for further analysis. Data analysis's first step is referred to as data exploration. To better comprehend the nature of the data, data analysts employ statistical tools and data visualization to convey dataset characterizations including size, number, and correctness. Large and unstructured amounts of data are frequently obtained from numerous sources. Before extracting pertinent data for additional analysis, such as univariate, bivariate, multivariate, and principal components analysis, data analysts must first comprehend and construct a holistic understanding of the data. Further, if the data collected is unstructured, then self-learning algorithms becomes a ubiquitous tool for analysts.

Self-learning algorithms like **K-Nearest Neighbor (KNN)** and **Neural Networks** help to process data more efficiently. Generally, data exploration is the pre-processing step before developing comprehensive and convoluted data mining modules. This step helps to clean and visualize data for better mining results and efficacy. Since, data sets exceed more than terabytes, pre-processing helps to reduce the size of data and hence reduce the load on the self-learning autonomous models and the computational power required for the same. An articulate description of the process is further reviewed in the paper.

II. LITERATURE REVIEW

Shija Mirza et.al [1] in their paper – 'A review of data mining literature' describe a disparate data mining techniques to a complete unequivocal data mining architecture which is made up of data mining engine, pattern evolution module, user interface and data sources. They also describe various tasks and complexities along with security and performance issues which are faced during the process of data mining. Fayyad et.al [2] in their paper 'From data mining to knowledge discovery in databases' describe the definition of data that is any set of valid facts that are available in an electronic form. Patterns are models that are stated in a language as a subset of data. The patterns must be true and able to be modelled for any new data in order to be valid. The process consists of several sub processes ranging from data preparation through knowledge augmentation. All of which are repeated until the desired results are obtained. Charles et.al [3] has proposed data mining is an efficacious tool for marketing which helps us to improve product marketing in this world of technology where provincial means of marketing such as mass marketing is decreasing its popularity day-by-day. By using data mining, we may identify buying trends from a client list and identify potential buyers. Data mining as a direct marketing tool has proven to be more profitable than conventional mass marketing strategies because it only targets potential customers.

Michael Goebel et.al [4] in their paper - 'A survey of data mining and knowledge discovery tools' offers a broad overview of typical knowledge discovery challenges and comprehends many approaches used to tackle them. It has been suggested to utilize a feature categorization technique to examine knowledge and data mining applications. For the knowledge discovery software is to be used efficiently and to solve more difficulties that haven't been thoroughly investigated. They listed a few key aspects that must be considered essential. Today, many businesses throughout the globe have enormous databases that are constantly growing. Millions of records per day are added to these databases as new information. These kinds of databases provide fresh difficulties and uncommon chances to explore these data streams. Ragavi. R et.al [5] have discussed data mining system from multiple angles, including data, knowledge, technology, and application. They have provided a succinct overview of the data mining system's approach and user engagement. The influence of data mining systems on decision-making in future generations is significant.

Furthermore, Ghimire et.al [6] in their paper 'Analysis of bitcoin crypto currency and its mining techniques' provide a detailed information about crypto currencies by taking bitcoin as an example and the underlying technology used to secure it autonomously and thus achieving a complete peer to peer decentralized digital currency. This makes bitcoin the top crypto currency in today's world with over 20,000 cryptos. Hence, seeing at the steep rise in the crypto world, there is a high demand of data mining techniques and related applications are growing day-by-day rigorously.

Lastly, Venkatadri.M et.al [7] has discussed that the importance in decision-making, data mining and knowledge discovery applications have gained a lot of attention over the past two decades and have become a crucial part of many businesses. With numerous integrations and developments in the domains of statistics, databases, machine learning, pattern reorganization, artificial intelligence, and computational capabilities among others. The field of data mining has flourished and been brought into new spheres of human life. Rutuja et.al [8] in their paper 'What is data exploration? and its importance in data analytics' discuss the different steps and methods involved in data pre-processing known as data exploration. Using these techniques, we can take raw data and identify patterns to draw out important information or insights. It assists both individuals and businesses in making sense of the data gathered. It helps to prepare data for self-learning modules and improve the performance of the complex algorithms.

III. DATA EXPLORATION

The initial phase in data analysis is called data exploration, and it involves looking at and visualizing the data to find insights right away or point out regions or patterns that need further investigation. Users may gain quick insights by using interactive dashboards and point-and-click data exploration techniques to better understand the broader picture. Data can be of any type because it is a rather generic idea. Therefore, data exploration helps us to investigate the data to obtain a better understanding of it.

3.1 Types of Data:

Following are the different types of data, dealt during data exploration:

3.1.1 Structured Data:

Structured data is a standardized format for describing a page's content and categorizing it. For an instance, on a recipe page, structured data may include information on the ingredients like the cooking time, cooking temperature, calories and so forth. It is data that follows a pre-established data model and is easy to analyze. A tabular format with relationships between the various rows and columns is what structured data follows. Excel spreadsheets and Structured Query Language (SQL) databases are typical instances of structured data. Each of them has structured, sortable rows and columns.

Furthermore, a data model is a representation of how data can be stored, processed, and accessed. It is necessary for the existence of structured data. Each field is distinct and can be accessed alone or combined with information from other fields because of a data model. Since it is feasible to swiftly aggregate data from many areas in the database, structured data is incredibly powerful. Since the first **Database Management Systems** (**DBMS**) could store, handle, and access structured data, structured data is regarded as the most "conventional" kind of data storage.

3.1.2 Unstructured Data:

Unstructured data is information that is either not organized in a predefined way or does not have an established data model. Unstructured data can also include facts like dates, numbers, and figures, but is often text-heavy. In contrast to data stored in organized databases, this produces anomalies and ambiguities that makes it challenging to understand using conventional algorithms. Unstructured data is frequently found in audio, video, or NoSQL databases. The ability to store and process unstructured data has significantly improved in recent years, due to the development of a number of new tools and technologies that can store specific kinds of unstructured data. For example, MongoDB is designed to store documents efficiently. Given that a sizable portion of the data in the organizations is unstructured, the capacity to analyze unstructured data is particularly important in the context of **Big Data**. For example, images, video snippets or Portable Document Format (PDF) files are one of the largest resources of unstructured data. One of the key factors influencing the rapid expansion of big data is the capacity to derive value from unstructured data.

3.1.3 Semi-Structured Data:

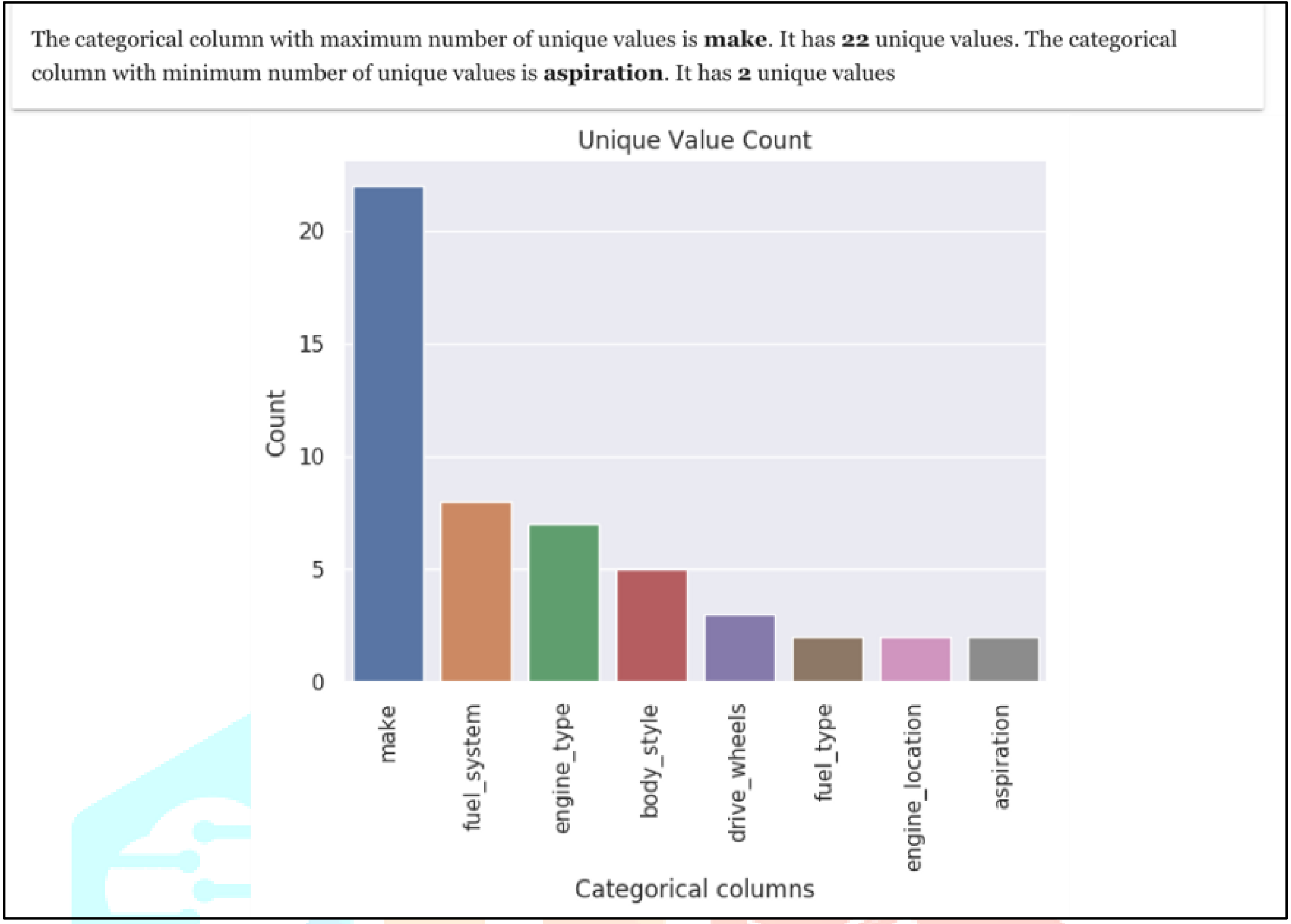
Semi-structured data is that type of structured data which does not adhere to the organization of data models linked to relational databases or other types of data tables, but still contains tags or other markers to enforce hierarchies of records and fields within the data and to separate semantic elements. Hence, it is known for its self-describing structure. **Java Script Object Notation (JSON)** and **Extensible Markup Language (XML)** are two types of semi-structured data that serve as examples. Semi-structured data is much simpler to analyze than unstructured data, which is why there is a third category between structured and unstructured data. The capacity to 'read' and process either JSON or XML is a feature of many **Big Data** solutions and technologies. Compared to unstructured data, this makes the analysis of semi-structured data less challenging.

3.2 Techniques for Data Exploration

Depending on the needs and other criteria, we conduct exploration using a variety of various sorts of exploration approaches, different forms of data call for different types of research strategies. A car dataset is taken for generating graphs for different techniques for data exploration [9].

3.2.1 UNIQUE VALUE COUNT

Counting the number of distinct values in categorical columns is one of the first things that can be helpful while exploring data. This provides insight into the data's subject matter. The Fig. 2 divides the car into subsets of different systems installed in a car. Their unique value count gives us the insights of the distinct values used in that particular sub system.

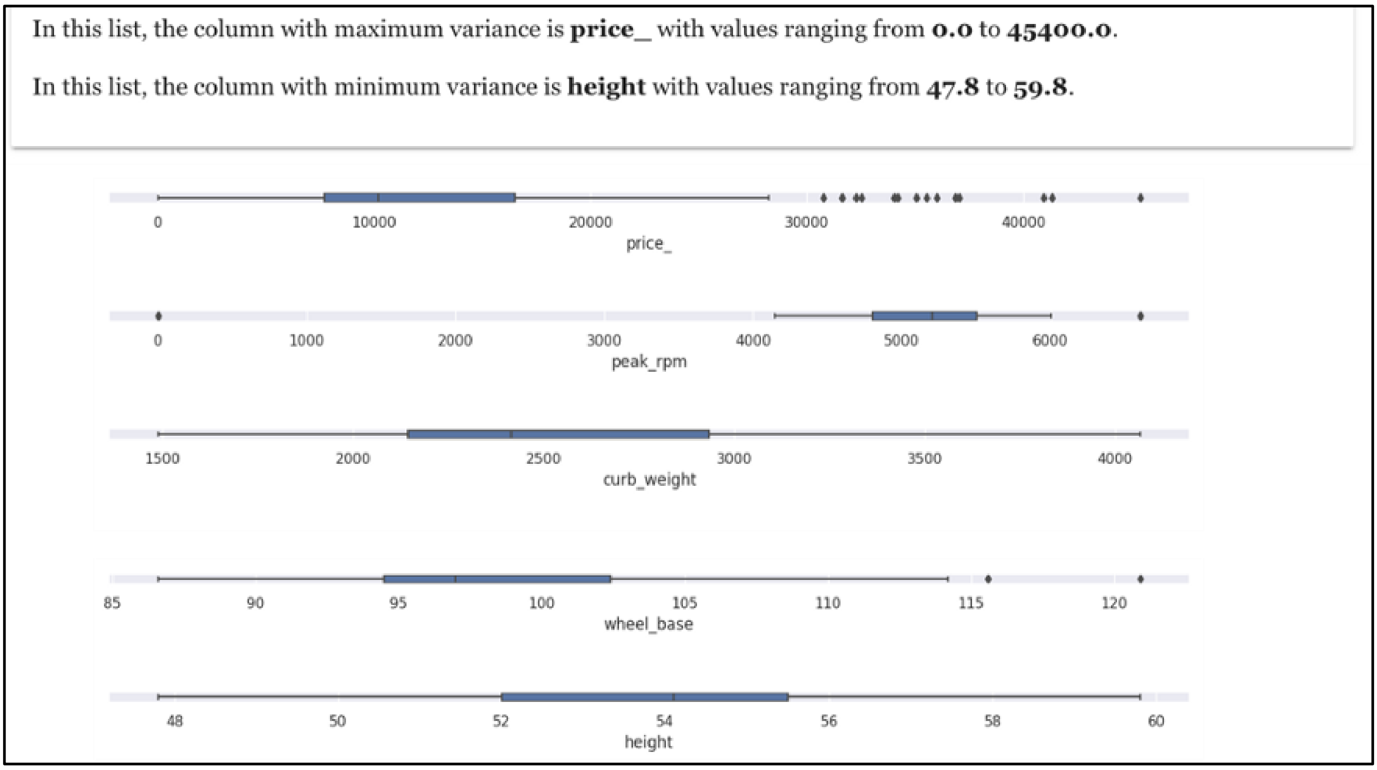


The categorical column with maximum number of unique values is **make**. It has **22** unique values. The categorical column with minimum number of unique values is **aspiration**. It has **2** unique values

Fig. 2: Unique Value Count [10]

3.2.2 VARIANCE

Some fundamental details, such as minimum, maximum, and variance, are highly helpful when analyzing numerical data. Variance provides a clear picture of how the values are distributed. The Fig. 3 describes the inter-quartile ranges for different features of a car like price, peak rpm, curb weight, wheel base and height.



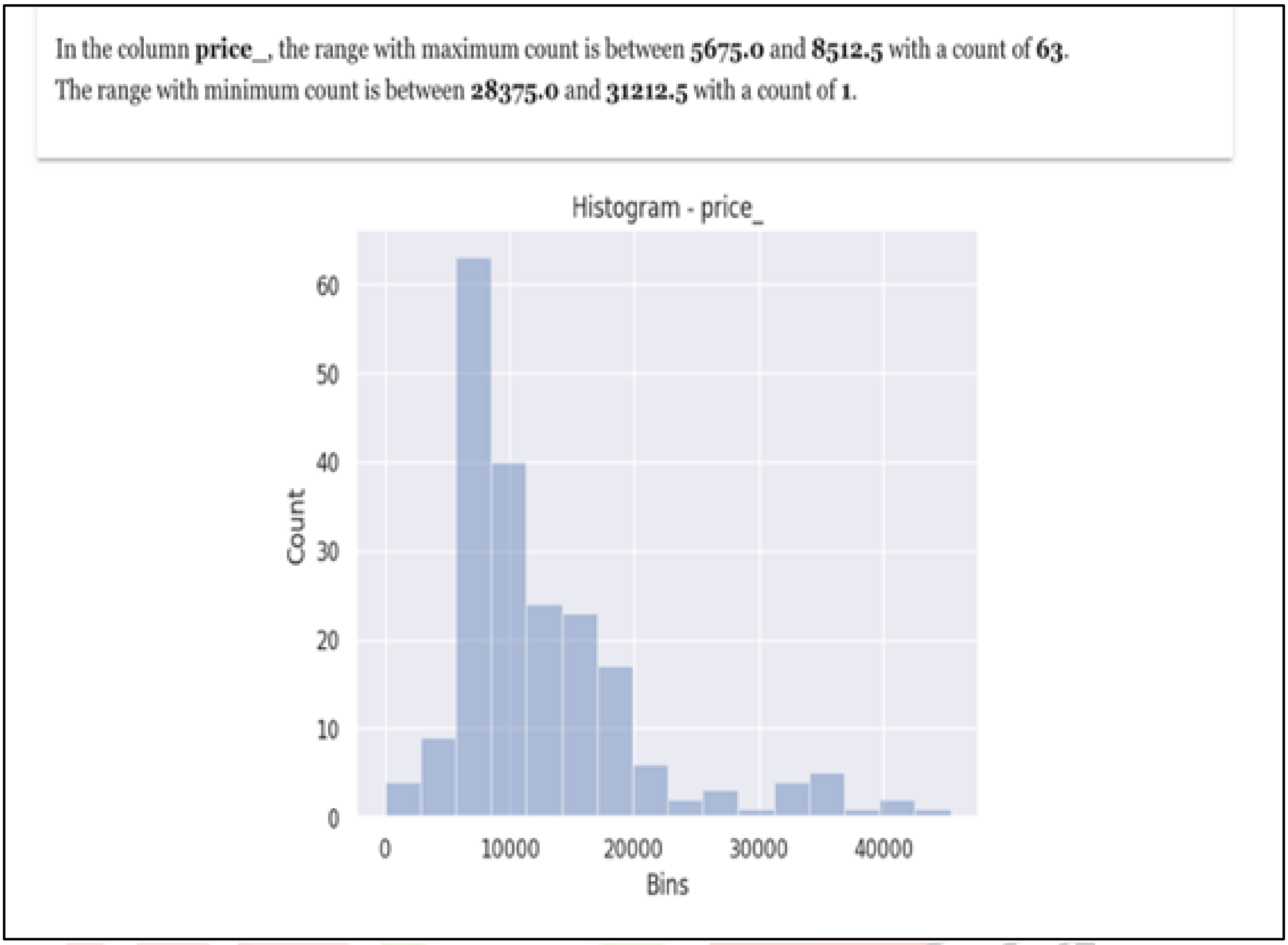
In this list, the column with maximum variance is **price_** with values ranging from **0.0** to **45400.0**.

In this list, the column with minimum variance is **height** with values ranging from **47.8** to **59.8**.

Fig. 3: Variance [11]

3.2.3 HISTOGRAM

One of the data scientists' preferred methods of data investigation is the histogram. It provides details on the range of values that the most values fall within. Additionally, it provides information about data skew. This technique divides the number of cars based on their prices and gives us a histogram view to analyze the data. This helps in understanding the mean of the prices in a better way.

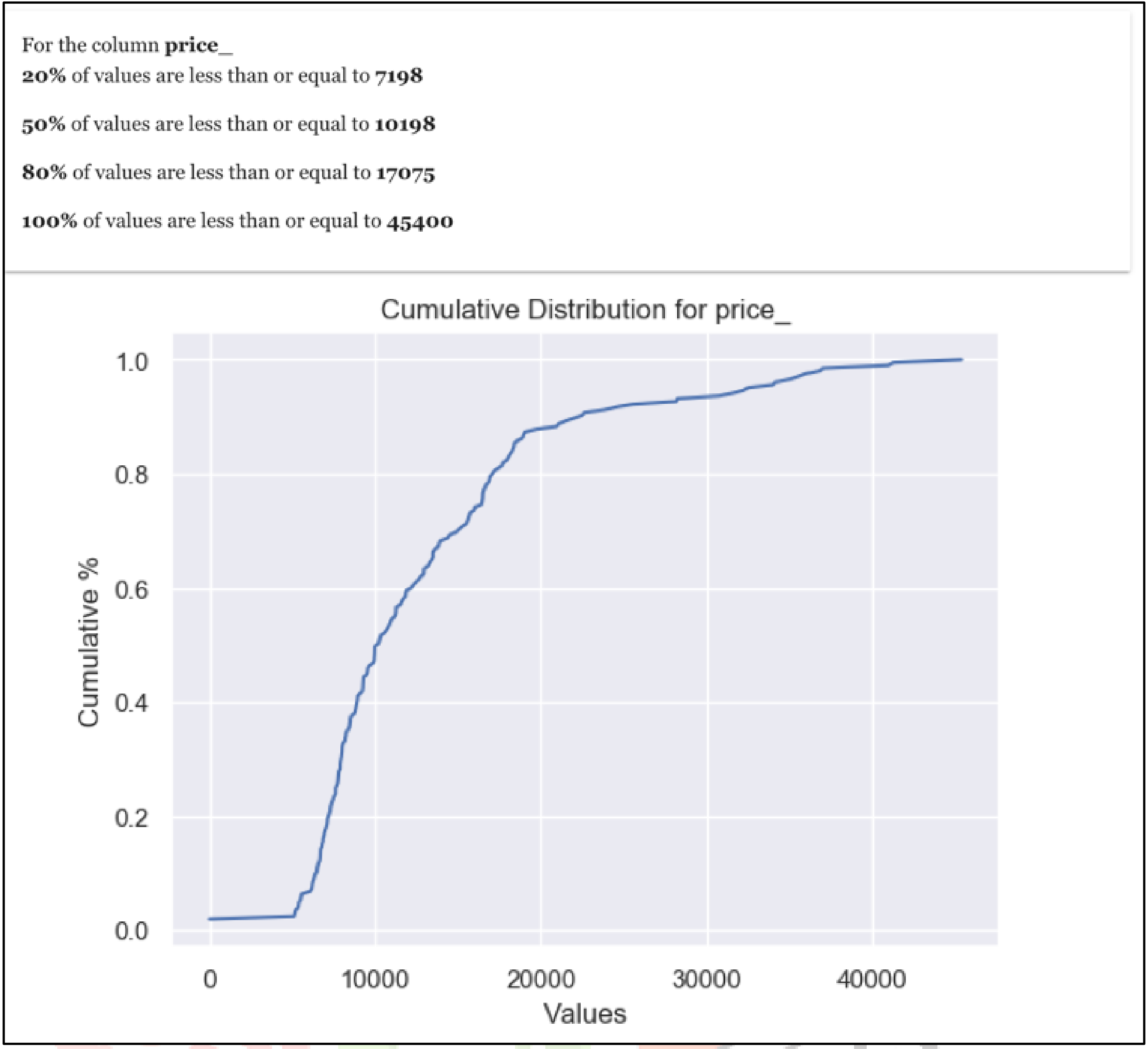


In the column **price_**, the range with maximum count is between **5675.0** and **8512.5** with a count of **63**.
The range with minimum count is between **28375.0** and **31212.5** with a count of **1**.

Fig. 4: Histogram [12]

3.2.4 PARETO ANALYSIS

An innovative method for concentrating on what matters is Pareto analysis. The **Pareto - 80-20 rule** is a useful tool for data research. We may do Pareto analysis on the price column in the automobiles dataset, as demonstrated in Fig. 5.



For the column **price_**

20% of values are less than or equal to **7198**

50% of values are less than or equal to **10198**

80% of values are less than or equal to **17075**

100% of values are less than or equal to **45400**

Fig. 5: Pareto Analysis [9]

The Fig. 5 shows 80% of prices, according to the study, are less than $17075 [9]. Knowing this information, it can help you understand what price level is deemed to be high.

IV. DATA MINING

Finding patterns and other important information from huge data sets is a technique known as data mining, commonly referred to as **Knowledge Discovery in Data (KDD)**. The usage of data mining techniques has surged over the past two decades due to the development of data warehousing technologies and the rise of big data and by helping businesses by converting their raw data into useful knowledge. Despite the fact that technology is constantly evolving to manage data on a huge scale, large scale IT industries and businesses still struggle with scalability and automation.

In addition to this, through smart data analytics, data mining has improved corporate decision-making. The two primary goals of the data mining techniques used to support these studies are to either characterize the target dataset or forecast results using **machine learning** algorithms. The most interesting information including **fraud detection, user habits, bottlenecks and even security breaches** are surfaced using these approaches for organizing and filtering data.

Exploring the world of data mining has never been simpler or more efficient when combined with data analytics and visualization tools such as Apache Spark, Apache Hadoop, Google BigQuerry, Lumify etc. Artificial intelligence advancements only serve to speed up adoption across sectors.

4.1 DATA MINING PROCESS

Data mining is a multi-step process that starts with data collection and ends with visualization to clean useful information from massive data sets. As it is discussed, descriptions and forecasts regarding a given data set are produced using data mining techniques. Data scientists use their observations of patterns, relationships and correlations to describe data. Additionally, they use classification and regression techniques to classify and cluster data as well as identify outliers for applications like spam and fraud detection, security breaches, user habits etc.

Data mining generally consists of four main steps:

4.1.1 Set the Business Objectives:

Many businesses under-invest in this vital stage of the data mining process which can be challenging. The business problem must be defined by data scientists and business stakeholders in order to guide the data queries and project specifications. In order to properly comprehend the company business, analysts might also need to conduct further research.

4.1.2 Data Preparation:

It is simpler for data scientists to determine which collection of data will aid in addressing the essential concerns for the business after the problem's scope has been established. After gathering the necessary information, the data will be cleaned to eliminate noise like duplicates, missing values and outliers. Depending on the dataset, another step to minimise the number of dimensions may be necessary because too many features can make any subsequent computation take longer. To achieve the highest level of accuracy in any models, data scientists will try to keep the most crucial predictors.

4.1.3 Model Building and Pattern Mining:

Data scientists may look at any intriguing data linkages such as sequential patterns, association rules or correlations depending on the sort of study they are performing. While high frequency patterns have broader applicability, occasionally the aberrations in the data might be more fascinating and revealing areas of probable fraud.

4.1.4 Evaluation of Results and Implementation of Knowledge:

The outcomes of data aggregation need to be assessed and interpreted. Results should be valid, original, applicable and comprehensible when they are finalized. When this criterion is satisfied, businesses can use this information to put new strategies into practice and accomplish their intended goals.

4.2 DATA MINING TEHNIQUES

Following are vital data mining techniques:

4.2.1 Association Rules:

A rule-based approach for identifying connections between variables in a particular dataset is called an association rule. Market basket analysis usually employs these techniques which help businesses comprehend the connections between various products. Businesses may create more effective cross-selling techniques and recommendation engines by better understanding the consumer consumption patterns.

4.2.2 Neural Networks:

Neural networks handle training data by simulating the connectivity of the human brain using layers of nodes, which is mostly used for deep learning algorithms. Inputs, weights, a bias (or threshold) and an output make up each node. If the output value exceeds a predetermined threshold, then it ‘fires’ or activates the node and sends the data to the following layer in the network. This mapping function is learned by neural networks through supervised learning with gradient descent adjustments made in response to the loss function. We can be sure in the model's accuracy to produce the right answer when the cost function is at or close to zero.

4.2.3 K-Nearest Neighbour:

The K-nearest neighbor (KNN) is a non-parametric algorithm that groups data points according to their proximity and correlation with other pieces of accessible information. This approach makes the assumption that related data points can be discovered close to one another. It then assigns a category based on the most prevalent category or average after attempting to determine the distance between data points (typically by calculating Euclidean distance).

V. APPLICATIONS OF DATA MINING

Some of the data mining applications are:

5.1 Education:

Educational institutions have begun to gather data to better understand their student populations and the conditions that are conducive to success. They can use a range of dimensions and metrics such as keystrokes, student profiles, classes, universities, and time spent, to observe and evaluate performance as courses continue to move to online platforms.

5.2 Operational Optimization:

By utilizing data mining approaches, process mining enables firms to operate more effectively by lowering costs across operational functions. This procedure has enhanced organizational decision making and has assisted in locating expensive bottlenecks.

5.3 Fraud Detection:

Observing data anomalies can help businesses spot fraud while often occurring patterns in the data can provide corporates an insightful information. Although this is a common application in banks and other financial institutions, **Software as a Service (SaaS)** based businesses have begun to employ these techniques to remove fake user accounts from their datasets.

5.4 Cryptocurrency:

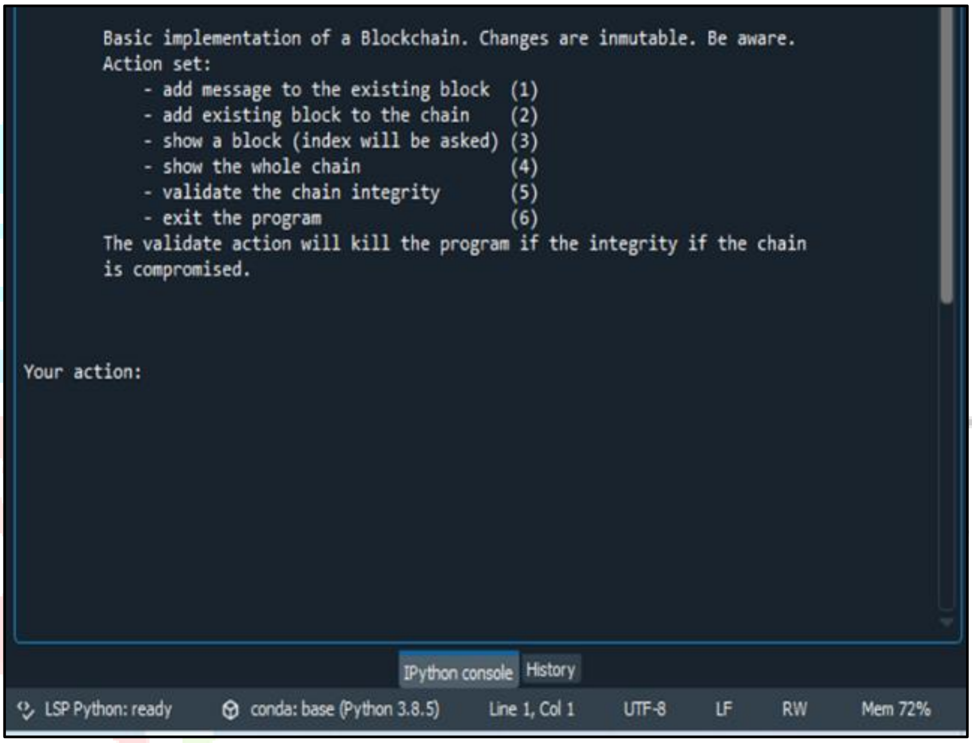
Cryptography protects the digital or virtual currency known as cryptocurrency, making it nearly hard to forge or double spend. Blockchain technology, a distributed ledger enforced by a dispersed network of computers is the foundation of many cryptocurrency decentralised networks. Cryptocurrencies are based on blockchain technology, so it is necessary to analyse the methodology and security in order to verify their integrity.

Two of the most popular cryptocurrencies in the world are Bitcoin and Ethereum. Since Ethereum is an open-source tool, numerous coins like Ocean Protocol Coin are built using its technology. In the Web 3.0 era, blockchain technology is seen as the most important component.

5.4.1 Review of Cryptocurrency Project:

A complete review of cryptocurrency is presented here based on one of our projects.

- We built a decentralized peer to peer simple blockchain, developed completely with python language.
- We used fingerprinting and the **Secure Hash Algorithm 256-bit (SHA 256)** hashing technique to develop blockchain technology. The data will be kept in **Java Script Object Notation (JSON)** format which is simple to use and easy to understand. Each block is encrypted and connected to one another using hashing technology, therefore protecting it from being altered by an unauthorised person. The data are kept in blocks, each of which has several data. Multiple blocks are added every minute and we will employ fingerprinting to distinguish one from another. The project snapshot is provided in Fig. 6.



```
        Basic implementation of a Blockchain. Changes are inmutable. Be aware.
        Action set:
            - add message to the existing block  (1)
            - add existing block to the chain    (2)
            - show a block (index will be asked) (3)
            - show the whole chain               (4)
            - validate the chain integrity       (5)
            - exit the program                   (6)
        The validate action will kill the program if the integrity if the chain
        is compromised.


Your action:
```

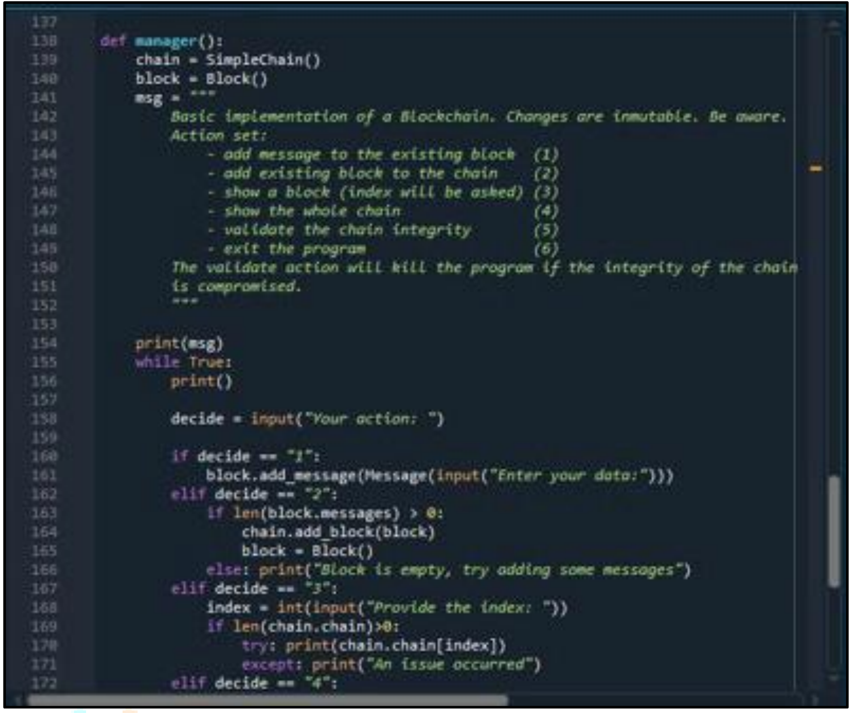
Fig. 6: Project Snapshot

- The underlying framework is completely implemented in python and Java Script Object Notation. Fig. 6 shows a screenshot of the project output. After running the project in the Integrated Development Environment (IDE) prompt, 6 options are provided to choose from, and these include various features like adding a message to the existing block, add a block to the chain, show specific details of the block (block index will be asked further for this), a brief overview of all the blocks, validate and check the authenticity of the existing block chains and finally to exit the program.

- When a transaction is successful, a block is created and added to the block chain in the distributed network. It is a collection of 'n' blocks that are linked to one another to form a distributed block chain. The next block's hash values are used to link them together. The preceding blocks are stored in each new block. Block chain is spread as a result.

- Basically, a block represents a transaction which takes place between two peers without knowing their profiles. The transaction is recognized by the foreign key which is given to each user and used for the hashing algorithm in the blockchain. There is a primary key and a foreign key for every user which keeps the blockchain autonomous, decentralized and secured.

- The essential thing to remember is that the produced hash key links the two blocks and keeps track of the previous block by storing the information for the current block and the address of the following block.



```
def manager():
    chain = SimpleChain()
    block = Block()
    msg = """
        Basic implementation of a Blockchain. Changes are inmutable. Be aware.
        Action set:
            - add message to the existing block  (1)
            - add existing block to the chain    (2)
            - show a block (index will be asked) (3)
            - show the whole chain               (4)
            - validate the chain integrity       (5)
            - exit the program                   (6)
        The validate action will kill the program if the integrity of the chain
        is compromised.
        """

    print(msg)
    while True:
        print()

        decide = input("Your action: ")

        if decide == "1":
            block.add_message(Message(input("Enter your data:")))
        elif decide == "2":
            if len(block.messages) > 0:
                chain.add_block(block)
                block = Block()
            else: print("Block is empty, try adding some messages")
        elif decide == "3":
            index = int(input("Provide the index: "))
            if len(chain.chain)>0:
                try: print(chain.chain[index])
                except: print("An issue occurred")
        elif decide == "4":
```

Fig. 7: Sample Code

- Fig. 7 shows a screenshot of a sample project code which gives a brief description of the working that takes place in a block. There is a payload along with every block which helps in identifying the index and adding a message to existing block.

VI. CONCLUSION

In the paper, a review of the literature on data mining and a method for identifying hidden patterns in huge datasets was presented. These identified trends assist manufacturers in forecasting future consumer or product behaviour. The study provides information on numerous data mining steps, techniques, applications as well as certain data mining-related issues. We are likely to evaluate and examine the various results from data mining algorithms in the future.

In addition, the paper provides information on data exploration which is unquestionably one of the most crucial aspects in the process of analysing data and drawing insights from it. Data exploration plays a significant function in the analysis process by providing a solid foundation and therefore, you should concentrate on it for the strength. Data exploration is mostly used to aid in data analysis before making any assumptions or decisions on something crucial. The majority of data scientists and analysts use data exploration to make sure that their results are precise and appropriate for any desired business goals and outcomes.

VII. REFERENCES

[1] Shuja Mirza, Sonu Mittal and Majid Zaman. A Review of Data Mining Literature: International Journal of Computer Science and Information Security (IJCSIS), 14(11), Nov. 2016.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases: AI Magazine, 17(3), 1996.

[3] Charles Ling and Chenghui Li. Data Mining for Direct Marketing: Problems and Solutions: Association for the Advancement of Artificial Intelligence (AAAI), Knowledge Discovery and Data Mining, 98(11), 1998.

[4] Michael Goebel and Le Gruenwald. A Survey of Data Mining and Knowledge Discovery Software Tools: ACM SIGKDD Explorations Newsletter 1.1, 1999.

[5] R Ragavi, B Srinithi and Anitha Sofia. Data Mining Issues and Challenges: A Review: International Journal of Advanced Research in Computer and Communication Engineering, 7(11), Nov. 2018.

[6] Suman Ghimire. Analysis of Bitcoin Cryptocurrency and Its Mining Techniques: UNLV Theses, Dissertations, Professional Papers, and Capstones 3603, 2019.

[7] Venkatadri Marriboyina and Lokanatha Reddy. A Review on Data Mining from Past to the Future: International Journal of Computer Applications, 15(7), Feb. 2011.

[8] Rutuja Magdum. What is Data Exploration? and Its Importance in Data Analytics: International Research Journal of Engineering and Technology (IRJET), 09, Jan. 2022.

[9] https://miro.medium.com/max/875/1*OO33WPY_SLhH7-ccIz9xMg.png - Accessed on 5th Sept. 2022.

[10] https://i.stack.imgur.com/OQTU9.png - Accessed on 5th Sept. 2022.

[11] https://miro.medium.com/max/3000/1*XYvrUI5pWk3h1OnzhwegsQ.png - Accessed on 10th Sept. 2022.

[12] https://miro.medium.com/max/875/1*j334Fz1sffl7BrQPBx0Dqg.png - Accessed on 10th Sept. 2022.