International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 4375-4390
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Pre-Processing Techniques applied on mutual funds

Gamini Dhiman[1], Gaurav Gupta[2], Brahmaleen K. Sidhu[3*]

[1] Research Scholar, Department of CSE, Punjabi University, Patiala.
Email: gamini.dhiman99@gmail.com- **ORCID**: 0000-0002-5247-7250

[2] Department of Computer Science & Engineering, Punjabi University, Patiala.
Email: gaurav.shakti@gmail.com- **ORCID:** 0000-0002-5247-7450

[3] Department of Computer Science & Engineering, Punjabi University, Patiala.
*** Corresponding Author Email:** brahmaleen.sidhu@gmail.com- **ORCID:** 0000-0002-5247-7350

Article Info:

DOI: 10.22399/ijcesen.2843
Received : 21 March 2025
Accepted : 14 June 2025

Keywords

Data Pre-Processing
Mutual Funds
Data Mining
Normalization

Abstract:

The raw data obtained from the logs may be noisy, incomplete, and inconsistent that's why data pre-processing is an essential step in data mining. The quality of data plays a vital role during the evaluation process. The results of the evaluation process primarily depend upon the quality of the data input. So, data pre-processing is the primary and most crucial step before knowledge discovery. This paper is based on two main steps- data pre-processing techniques and results after applied data pre-processing on mutual funds' data. Data pre-processing transforms the raw data into a structured, understandable format. Moreover, data pre-processing performs not only the transformation of data but also makes it understandable according to need. It is mainly divided into four steps, i.e., data integration, data cleaning, data transformation, and data reduction. This paper takes the fifteen-year NAV data of twenty mutual funds for analysis propose. This paper explains several techniques of data pre-processing to transform the raw data into an understandable format.

1. Introduction

In today's world, the financial market consists of many securities like debentures, equities, bonds, IPOs, etc. for an individual to invest their money. An individual can invest his money in these securities directly or indirectly through mutual funds. Today, mutual funds are the most dynamic sector in the Indian financial market. This sector has grown many folds in the last few years, mainly due to the increasing complexity of modern investment. Mutual funds ease the work of any investor by providing him with professional expertise. Many techniques are applied to the mutual funds' data to predict the future outcome of any investment by the experts. The mutual fund provides an area in which a large volume of data is created and stored daily. Now, the researcher has to apply the prediction algorithm to this large volume of data to find out the profit of the investment. A frequent challenge in using these algorithms in this dataset isthe large size and the demand to get the results as fast as possible. The performance of any algorithm depends upon the quality of the data. In most of the real-world data, not all points in the dataset are equally important. Sometimes, the real-world data may be incomplete, uncertain, or contaminated by noise. With this incomplete or noisy data, this is not possible to get the correct results in a limited duration. So, the data mining techniques can be applied to the large volume of raw data of mutual funds to get an efficient and unified view of a large dataset. Data mining is applied to the dataset to discover the hidden and principal patterns from the raw data [1]. Data mining is the process of detecting anomalies, trends, and correlations within a large dataset to predict the final results. By using the different data mining techniques, anyone can increase benefits, predict results, build an artificial intelligence model, etc. Before applying the data mining techniques, data pre-processing plays a significant role in the entire process as it can ensure the quality of the data. Data pre-processing is a very essential and primary step in the whole process of data mining which is used to convert the raw data into a useful and efficient format. In data pre-processing, several measures have been followed to create an

understandable and efficient form of raw data. Data pre-processing includes the removal of noisy data, filling in missing values, normalizing the data, and many more. Data pre-processing includes data reduction techniques that reduce the complexity of the data, detecting or removing irrelevant or noisy elements from the data.Mutual fundshave a large volume of data related to each fund with several anomalies. So, the primary goal of this paper is to remove those anomalies from the raw data and make an efficient and understandable useful dataset. This paper is going to preprocess the mutual fund data by following every step of data pre-processing to make the data efficient for further research.

The main structure of this article is: section 1 introduces the need for data pre-processing on the raw data. Then section 2 talks about the collection of data from different sources. After this, section 3 explores and applies the various steps of the data pre-processing technique on the given dataset of mutual funds. Finally, section 4 states the research conclusion and proposed suggestions.

2. Data Collection

The initial step in the data mining process is the collection of data for analysis from different sources. Data collection is also an important step in the research work. Data collection is the procedure of collecting, measuring, and analyzing a large volume of data for research proposed using standard validated techniques. The approach for data collection is different for the different topics of research, depending upon the information needed. There is a fundamental difference between all techniques because of the different qualities of every technique.In this phase, it is to make sure that the collected data is of high quality and collected with a reliable technique because it will directly affect the quality of predicted results or patterns explored.

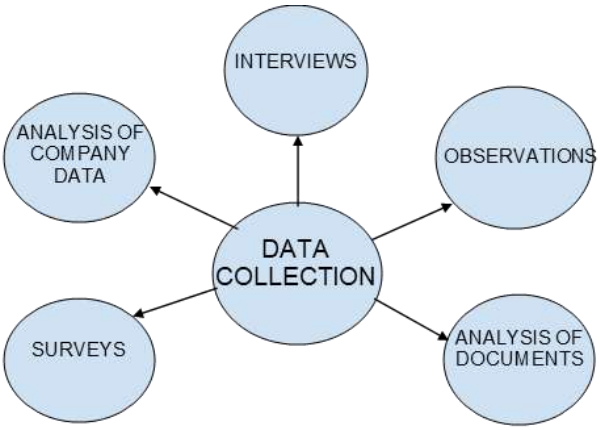


Figure 1. *Types of Data Collection*

Some of the data collection techniques are shown in Fig. 1.

- Interview Method: This method works best in the job analysis sector. This is a good method to collect information about any job or office by interviewing their workers because they are the one who knows their work culture effectively. For any job analysis task, interviewing the candidates, especially questionnaires, is the best method to collect the data. This method has some problems like the ambiguity of language, difficulty in translating scientific terms, etc.
- Observation Methods:The observation method is used to check the reality of people in publicspaces by observing them. In this method, the observer checks the behavior and interaction of people in a public open space. This method is mainly used to avoid the sort of errors that can occur during interview methods. This method results in obtaining more objective data. This method is not dependent only onthe workers who are observed, but also on the observer who observes the whole scenario. This method has its problems like observer biasing nature, some infrequent events, etc.
- Analysis of Company Data: In this method, the company data is analyzed and collected for research work. In this, a lot of data is required for analysislike workplace culture, workplace descriptions, job requirements, etc. This method collects data about every minor detail of the job by analyzing the company data. This method has its problems like no interaction withworkers, missing physical activities, etc.
- Analysis of Documents:In this method, every document is analyzed to collect the data. This method reviews all the documents like file information, filled forms by workers, resumes of every aspirant, etc. This method collects every minor detail by checking the different documents related to that topic. After analyzing all the documents related to that research, data is collected on that topic of research.
- Surveys:The survey is another method to collect information about any topic. Surveys can be conducted in either online or offline mode. In this method, aset of questionnaires is designed and given to the customer to take feedback from them. With this method, data collection is very easy, cheap, and effective. Surveys can be considered either inbound communication mode or outbound communication mode. Surveys are generally considered to be a data collection and analysis source.

Now in this research, different mutual funds' data is collected to carry out the research analysis. The researcher combines two or three data collection techniques to efficiently analyze the data of mutual funds. Firstly, the researcher listed down the top-performing mutual funds to effectively analyze the data. Then to list down the top ten performing mutual funds; moneycontrol [2], and valueresearchonline[3] links are thoroughly analyzed. After this analysis, a list of the top ten growth and top ten dividend mutual funds isanalyzed [closed in Appendix].The next step is to download the historical data of these above-mentioned mutual funds. In this analysis, the Net Asset Value (NAV) of each mutual fund is an essential factor. To carry out the research more efficiently, the last 15 years' NAV of each mutual fund is collected. So, the NAV of each mutual fund is downloaded from the link of the AMFI (Association of MutualFunds of India).

2.1. Data Pre-Processing

The raw data collected in the last stage is highly susceptible to noise, missing values, and inconsistencies. The quality of the raw data affects the predicted results or data mining patterns evaluated.The raw data is pre-processed by following several steps of data pre-processing to improve the quality of data and consequently, the mining results. That can lead to ease of the other analysis process and improve the efficiency of the analysis.

Data pre-processing is a significant and critical step in the data mining process that transforms and prepares the initial raw data into an understandable format for further processing. The main aim of the data pre-processing is to clean the raw data for better quality. Data pre-processing is a collection of activities that are followed to make data more suitable for analysis. Data pre-processing is mainly divided into four steps, i.e.

- Data Integration
- Data Cleaning
- Data Transformation
- Data Reduction

During the data pre-processing process, raw data has to go through all these four steps to transform it into usable and efficient training data for analysis. A model in Fig shows the steps of data pre-processing in Fig 2.

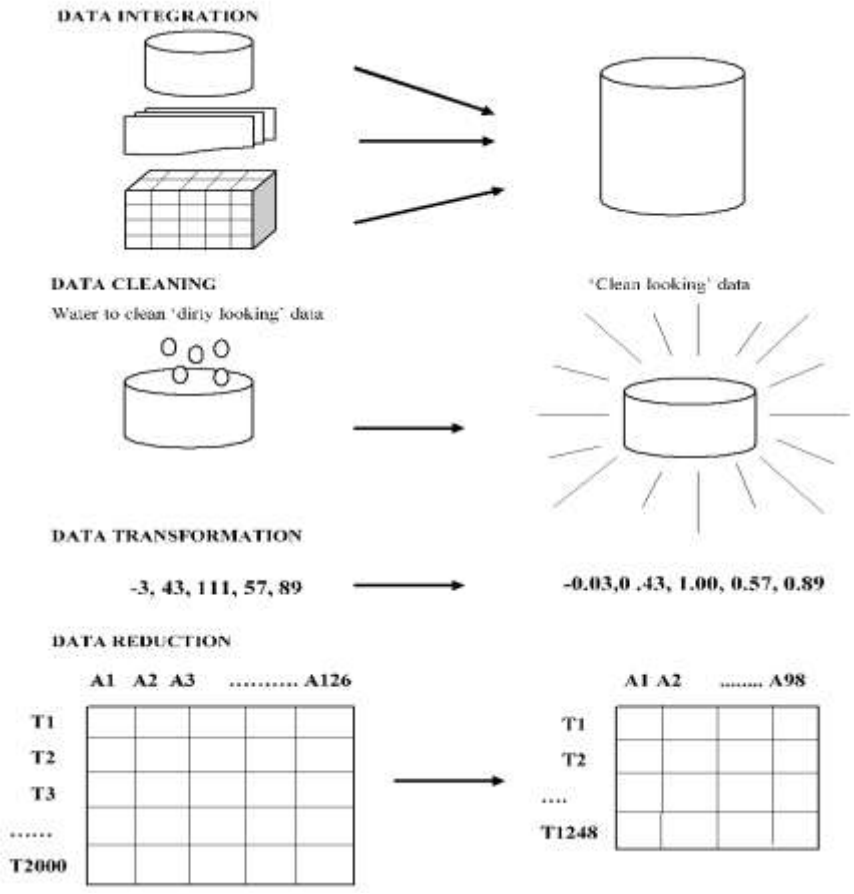


Figure 2. *Steps of Data Pre-Processing*

2.2. Data Integration

In data pre-processing, the primary step is to integrate the whole raw data into a single dataset that is clean, understandable, and consistent. This whole process is named data integration. Data integration is the practice of consolidating data from multiple source systems into a single unified set of information for both analytical and operational uses [4]. To get a better understanding of data integration, there are five different types to integrate the dataas shown in Fig. 3.

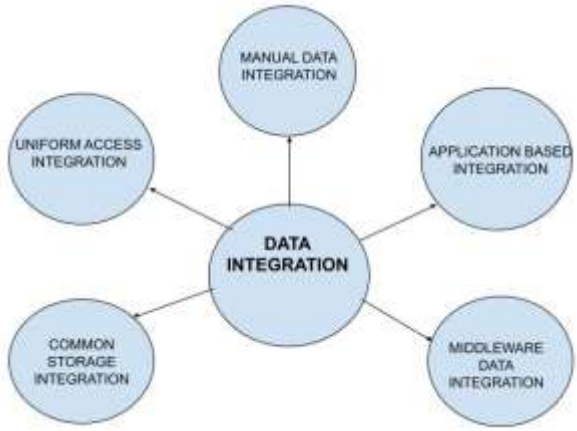


Figure 3. *Types of Data Integration*

The primary objective of this step is to get clean and redundant data from the raw dataset. This redundant data should fulfill the information needs of different end-users in an organization.

- Manual Data Integration: This technique is usedby small-scale companies to avoid the excess cost. In this, the data analyst himself collects the data, cleans it, and integrates it with custom code. This process has its benefits with less cost and freedom to do anything with the handwrittencode. This process quickly becomes untenable for complex and large datasets due to its manual process. There is a lot of space for error due to the custom coding system.
- Application-Based Integration: This approach is perfectly software-based integration. In this process, software is required to do all the work. A software application is used to extract information from different sources, clean it, and integrate it according to the user'sneeds. In this, the compatibility of different data sources is also done by software. Due to this compatibility transformation, it is easy to move the data from one source to another. This process is common in enterprises because of their hybrid cloud environments. This process compatibles the data and workflows between these environments.
- Middleware Data Integration: This approach acts as the middle layer between old and new integrated systems. This middleware approach is used when the user wants to integrate data from a legacy system into a modern system. At the time of integration, this middleware layer acts as an interpreter between two systems. This approach is ideal for businesses that want to integrate from a legacy system to an advanced system. But this only actsas a communication tool with limited functionality.
- Uniform Access Integration: This approach is optimal for businesses that wantto collect data from more discrepant sources. In this, the data is only shown in a unified view to the user. The location of the data is not changed because the consolidated view of the data is not stored in a separate place. In this process, less storage space is required with easier access to the data. The main drawback of this process is the compromise of data integrity.
- Common Storage Integration: Common storage integration is the most promising and sophisticated integration approach. In this, the consolidated and unified view of the whole data is stored in different spaces for better access. This approach needs more storage space, with more management costs. However, this allows the analyst to handle more complex queries. The most well-known example of this approach is data warehousing.

Now in this research work, the data is related to 20 mutual funds of different companies [listed in Appendix]. The most promising and useful approach of data integration i.e., common storage integration, approach is used for this research analysis. The Net Asset Value (NAV) of each mutual fund is collected and stored in a file for further processing. Each mutual fund's data is stored in different files. So, the next step is to integrate these files to form a single unified file. From the link as mentioned above, i.e. AMFI3 one can download the NAV data of any mutual fund for three months only at a time [5]. To carry out the research more efficiently, and effectively, the 15 years of data of one mutual fund is downloaded and collected from this link. This whole data of different mutual funds is in the form of a collection of 60 files as each file contains the three-month NAV data of each mutual fund. Now, the first step is to append all these files into one single file so that the single file contains the whole 15 years of NAV data of one mutual fund. After this append operation, one single file contains 15 years of historical data of one mutual fund. Now in this

research analysis, there are 20 files of each mutual fund in which the historical NAV data of each mutual fund is collected and stored. After this, the next step is to join these all files into one unified file view so that the research analysis can take place in an efficient and organized way. All the functions of join and append are done in Python.

2.3. Data Cleaning

In data pre-processing, the next step is to clean the integrated data by removing all outliers, filling missing values, etc. Data cleaning is the core step of any analysis process because the research data directly affects the result of the analysis part. Data cleaning is the process of preparing the integrated data for analysis by removing the noisy data, filling the missing values, or modifying the inconsistent data [6]. This process is not only about removing the noisy data or making space for new data but also finding a way to maximize the accuracy of the dataset without deleting the information. Most importantly, the ultimate goal of the data cleaning step is to generate a standardized and uniform dataset so that data analytics tools can easily access and discover the right path for each query. In this research analysis, we have two large data sets in which the historical data of mutual funds are stored. One data set contains the data of all growth option mutual funds, and the other data set contains the data of dividend option mutual funds. Now, in these two data sets, there may be some errors like existing noisy data, having missing values, etc. So, the data cleaning process is an essential step before going to the next analysis process[7]. In data cleaning, there are mainly three steps that should be followed, as shown in Fig. 4.

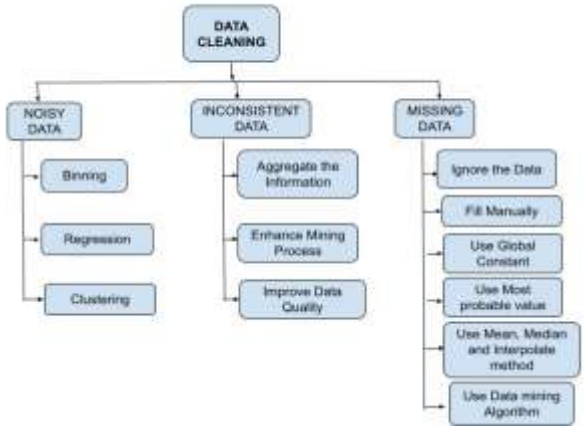


Figure 4. *Types of Data Cleaning*

- Noisy Data: Noisy data is a random error also called a variance in the integrated data. The raw data is meaningless and cannot be interpreted or understood by the data analytics tools efficiently. The noisy data need extra space for storage and can adversely affect the result of the data analysis [8]. This data may be prompted by various reasons like faulty data collection, limited buffer size, data transfer errors, data entry problems, etc. Mainly three procedures are followed to handle this type of data:
  - ❖ Binning: In the binning method, sorted values of any data are divided into several segments to smooth it. During the binning process, all the segments are treated independently. All segments are smoothed locally by checking their mean value or median value. In smoothing by bin mean, all values are changed to the mean value of that segment. In smoothing by bin median, all values are changed to the median value of that segment. In smoothing by bin boundaries, maximum and minimum boundary values are searched and then all values are changed to the closest boundary value of that segment.
  - ❖ Regression: In this, data can be smoothed by fitting it to a regression function. The regression function is of two types: linear regression function and multiple regression function. In the linear regression function, the best line is searched that can fit two attributes. So that one attribute can be used at the time of prediction. In multiple regression functions, data tries to fit in multidimensional space with more than two attributes involved.
  - ❖ Clustering: In the clustering process, several different clusters are formed with data that have similar features. Every cluster has different data according to its features. The values that are not in any cluster or fall outside all the clusters are called outliers.
- Inconsistent Data: Inconsistent data is the dataset that contains variation between different data items with the same name [9]. This problem is typically referred to as the content of the database, not the design and structure of the database.Sometimes the data is inputted from different sources for the same concept, and that data leads to inconsistencies in the database. Some methods can be used to handle this type of data, i.e., aggregate the information, enhance the mining process, and improve data quality.
- Missing Data: in the real world, every dataset has some instances where a particular data item is missing its value because of various reasons

such as incomplete extraction, the event did not happen, failure to load information, corrupt data, etc. due to the presence of missing values, there are many problems occurred during analysis like efficiency loss, the complication in analyzing the data, incomplete or incorrect results, etc. So, the foremost challenge for any analyst or researcher is to handle the missing value. Missing values are often encoded as NaNs, blanks, and any placeholders [10]. Now, the missing data can be handled in various ways, i.e.

- Ignore the missing value: In this method, the entire tuple related to that missing value is ignored or deleted. In this way, the impact of that missing value is negligible on the whole data.
- Fill manually: By following this approach, all the missing values are handled manually by checking their history. All missing values are filled in by checking their all details manually.
- Use a global constant to fill: In this approach, the missing value of any data is filled with a global constant. A global constant is taken according to the research topic and the missing value is filled with that global constant.
- Use the most probable value to fill: In this approach, a most probable value is searched within the whole data and that value is used to fill all the missing values.
- Use a mean, median, or interpolate method to fill: This is a very popular and effective approach to fill the missing value in any data. In this approach, the missing value is filled with the mean value or median value. The interpolation method can also be used to fill the missing value. Interpolation is a mathematical method that modifies a function according to the given data and uses this function to extrapolate the missing data.
- Use data mining algorithm to fill: In this approach, data mining algorithms are used to fill the missing values. To predict the missing value, various data mining algorithms can be used like k-mean, Bayesian, etc. With the help of these algorithms, the missing value is predicted and filled at a specific position.

Now in this research analysis, NAV data is missing regarding some dates of some mutual funds in the dataset. While joining all the mutual funds' data, there are missing values in the data set for some mutual funds. In the financial market, the value of any mutual fund for each day is a very crucial factor. Before the research analysis, these missing values should be handled efficiently. In any research analysis, no one cannot merely ignore missing values in the data set. The missing data must be handled in some significant way for the practical reason that most algorithms do not accept missing values. As mentioned above, there are many methods like mean, median, mode, linear regression, interpolation, etc. to handle the missing values. In this research work,the interpolate method is used to handle the missing values. Interpolation is a mathematical method that modifies a function to the given data and uses this function to extrapolate the missing data. There are various ways in which interpolation is done. The most used method of interpolation is linear interpolation which creates a mean value between the values before and the value after the missing one.After applying the interpolation method to the dataset, the missing value is filled by the average of its previous and next-day values. After this cleaning process, two data sets of the growth option and dividend option are free from any type of error.

2.4. Data Transformation

In the data pre-processing process, the next step is data transformation. In the data transformation step, the integrated data is consolidated or transformed into a specific format that is understandable by the data mining algorithm easily [11]. In this process, the data is transformed from one format to another format so that the result of any analysis may be more efficient and faster. The basic need for data transformation is to convert the unstructured data into homogeneous and structured data. Due to this homogeneous and structured data, it is easy to access the data efficiently and perform any analysis quickly. The data transformation process has the following strategies which are shown in Fig 5:

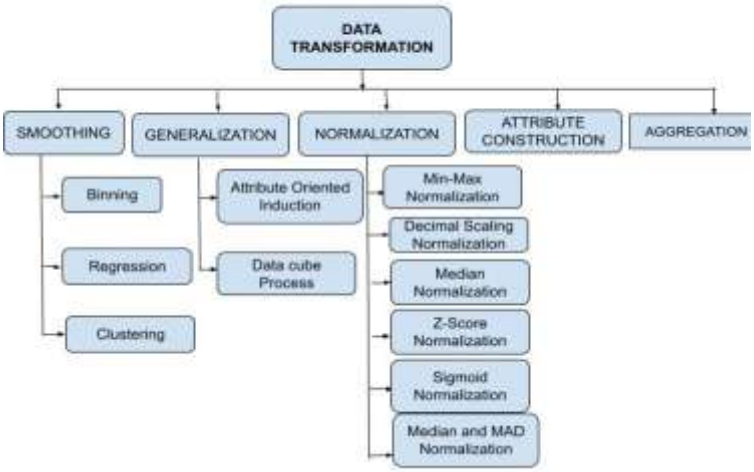


Figure 5. *Types of Data Transformation*

- Smoothing: This process is to remove the noise from the integrated data by using some specific algorithms. Noise is present in the form of meaningless, corrupted, or distorted data. The main idea to use the smoothing process is to highlight the unique features of the dataset. This highlighting process makes the prediction or analysis process easy. It saves much time for analysts or researchers by highlighting notable trends or patterns. The following techniques can be followed to accomplish the smoothing task, i.e. binning, clustering, and regression.
- Aggregation: Data aggregation is the process of presenting data in a summarised form. This process is an essential step in data pre-processing because the accuracy of the data depends on the quantity and quality of the collected data.The data is collected from different sources for analysis, so it is imperative to aggregate the data carefully. To get efficient and proper results, the collected data should be of high quality and large quantity. The aggregation of data is essential in every aspect like the daily sale data must be aggregated to compute the monthly or annual amount.
- Attribute construction: In this process, a new set of attributes is formed from the given set of attributes. This process is used to simplify the given dataset and make the analysis process faster. This process helps in the reconstruction of a new dataset with few new attributes easily and quickly.
- Generalization: In this process, the low-level or primitive data is replaced by high-level data by following the concept hierarchy. The generalization process is useful to get a clearer or more concise picture of any dataset. The categorical or lower-level concepts like street name, and colony name can be generalized to higher-level concepts like a town, country, etc. The data generalization can be divided into two approaches:
  - ❖ Attribute-Oriented Induction Approach: This method is specially used to generalize or summarize the data on a character basis. This induction approach is a relational database query-oriented process. It is an online data analysis technique. In this technique, the generalization is performed based on attribute values within the relevant dataset. Later on, some tuples are merged and some are aggregatedto perform summarization. This induction technique is implemented by two methods: A) Attribute Removal and B) Attribute Generalization.
  - ❖ Data Cube Process: This approach is also known as the online analytical processing (OLAP) approach. This approach can be considered a data warehouse-based precomputation-oriented materialized view approach because the precomputed data is stored in different data cubes in a data warehouse. These materialized views of different cubes are mainlyused for different business skills like knowledge discovery, decision support, etc. This data cube process is done using twomain operations: drill down and roll up operations. These operations involve different types of functions like count, sum, aggregation, mean, median, etc.
- Normalization: Data normalization is the process in which the integrated data is scaled into a specific smaller range [12]. It is used to scale all the attributes into the same range of values to minimize the bias at the time of analysis from one attribute to another. This

process helps in applying the data mining algorithm easier and faster. The data normalization process is especially useful in data modeling techniques in which most attributes are on different scales. In this process, data should be converted into a smaller range like -1.0 to 1.0 or 0.0 to 1.0. Various methods [13] usedforthe normalization process are:

- Min-Max Normalization: In this technique, all the dataset's attributes are scaled to a new range of values, i.e. [0,1] or [-1,1]. Most importantly, the Min-Max normalization technique preserves the original relationship between the values of the dataset. This technique uses the maximum and minimum values of the dataset to normalize the original values to a predefined range of values. It normalizes the value xof any attribute to in the predefined range by computing the following formula:

$$\hat{x} = low + \frac{(x - \min X)}{(\max X - \min X)} * (high - low) \quad (1)$$

Where x is the original value of the attribute; low is the minimum value of the predefined range; high is the maximum value of the predefined range; minX is the minimum value of the attribute and maxX is the maximum value of the attribute.

- Decimal Scaling: This method normalizes the value of the attribute by moving the position of the decimal point in the value. The movement of the decimal point in the attributedepends upon the absolute maximum value of that attribute. In this way, the value x of any attribute is normalized to$\hat{x}$by using the following formula:

$$\hat{x} = \frac{x}{10^d} \quad (2)$$

Where d is the smallest integer such that max ($|\hat{x}|$)<1

This method also depends upon the minimum and maximum values of the attributes. The process of normalization is somewhat similar to the process of min-max normalization. This normalization procedure is not suitable for time series data.

- *Median Normalization:* This method normalizes the value of any attribute by calculating the median of all the values of that attribute. This normalization method is used when there is a need to get the ratio of two hybridized attributes. This method is also used to perform the distribution process. By calculating the median of any attribute, the value x of any attribute is changed to$\hat{x}$by using the below formula:

$$\hat{x} = \frac{x}{median(x)} \quad (3)$$

- *Z- Score Normalization*: This normalization method is the most effective method for the stationary environment because the maximum and minimum value of an attribute is not needed in this method. This method normalizes the data with the help of the mean and standard deviation of the attribute. In this method, the mean and standard deviation of each attribute are computed and then the value x of any attribute is changed to$\hat{x}$by using the below formula:

$$\hat{x} = \frac{x - \mu(x)}{\partial(x)} \quad (4)$$

Where μ(x) is the mean value and $\partial(x)$ is the standard deviation of the attribute X respectively. This method is very effective and balancedbecause it reduces the effect of outliers that dominate the min-max normalization results.

- *Sigmoid Normalization:* The biggest advantage of this normalization method is that it does not depend upon the distribution of data as sometimes distribution is not known at the time of training. This is the simplest method to calculate the normalized value of any attribute. In this, the exponential power of the value is used to compute the normalized value. The value x of any attribute is changed to$\hat{x}$by the following formula:

$$\hat{x} = \frac{1}{1 + e^x} \quad (5)$$

- *Median and Median Absolute Deviation (MAD) Normalization:* In this method, median and MAD values are used to normalize the value of any attribute. Median Absolute Deviation (MAD) is the measure of statistical dispersion. The median and MAD of all attributes are computed first to normalize the data. These two values are the robust measure of the variability of all the attributes of

quantitative data. The value x of any attribute is changed to $\hat{x}$ by the following formula:

$$\hat{x} = \frac{x - median(x)}{MAD} \quad (6)$$

$$MAD = median((abs(\{x_k\} - median(x))) \quad (7)$$

Now in this research analysis, the NAV data of all mutual funds are on widely different scales. Their range of input data is very different from one another. It is challenging to analyze the data with these different scales. So, the data normalization is done on the dataset. From all of the above-mentioned normalization techniques, the Z-Score Normalisation technique is applied to this dataset. In this technique, all the dataset's attributes are scaled to a new range of values by using the values of the mean and standard deviation of that attribute. The mean and standard deviation value of all the attributes is computed and then normalize the data by using them. In this research work, the mean and standard deviation value of each mutual fund's NAV value is computed differently and then the NAV value of each mutual fund is normalized by using these values.

2.5. Data Reduction

Data reduction is the process of reducing the dataset from a large one to a smaller form without any information loss. Complex and large datasets are challenging to handle and analyze. There are many of data related to any research work on the internet. A Dataware house can store a terabyte of information related to the work. So, it may take much time to perform data mining and analysis tasks on such a massive amount of data. The data reduction process is done on the dataset to reduce the computation time [14]. The data reduction process reduces the size of the dataset without compromising the integrity of the data and yet producing accurate results of the analysis. There are many procedures to reduce the volume of the dataset efficiently with a promise of useful knowledge. Some data reduction methods are shown in Fig. 6.

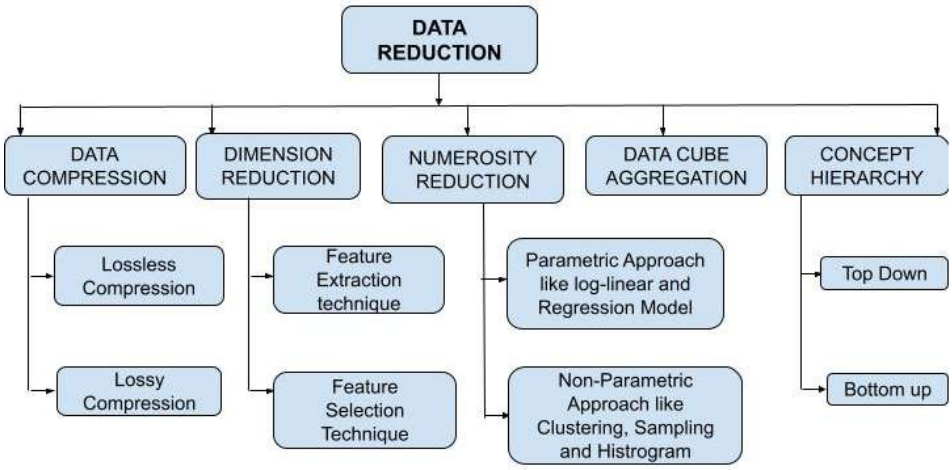


Figure 6. *Types of Data Reduction*

- Data Cube Aggregation: A data cube is the representation of the dataset in a more straightforward form by constructing a cube. The aggregation operation is used to form a data cube. By using the aggregation operator, the data cube is formed that contains all the aggregated values in the summarised way as shown in Fig. 7.

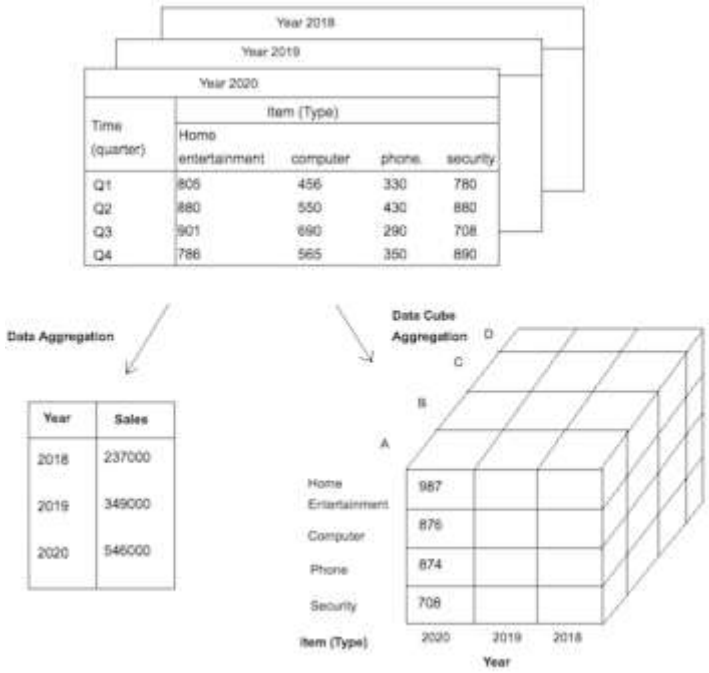


Figure 7. *Representation of Data Cube Aggregation with Example*

In this example, the data aggregation and data cube aggregation represent the random sales data. In data aggregation, there are only two attributes after the reduction of data. So, there is aloss of data in data aggregation. But data cube aggregation covers all attributes with minimum space and without the loss of data. So that, actual data can recover when the user needs that.

- Dimension Reduction: In this process, redundant or weakly relevant attributes are detected and removed from the database to reduce the size of the data without any loss of information [15]. When any attribute seems to be less informative and irrelevant to the analysis part, then it is better to reduce that attribute from the dataset by merging or combining the data. Dimension reduction is mainly divided into two categories:
  - ❖ Feature Selection Technique: This technique is used in the case of high-dimensional data. When there are many features or attributes related to any data that are irrelevant, misleading, and redundant, then this feature selection method is used. These extra irrelevant features take extra space and time to execute the processing. Soin this method, a subset of features is selected to make a new database with relevant and useful features and used to discriminate classes. There are three main classes of feature selection algorithms: a) Wrapper Methods b) Filter Methods c) Embedded Methods.
  - ❖ Feature Extraction Technique: This method is used when there is redundant data in the database with extra features. This technique transforms the original attributes to form another attribute by combining the more significant attributes. The Feature extraction technique is mainly used to reduce the complexity of the database and give a simple and unified view of data. With the help of this technique, each variable of any data is represented as a linear combination of the original output variable. This technique is mainly implemented by three extraction approaches: a) Performance Measure) b) Transformation c) Number of New Features.
- Data Compression: Data compression is the process in which encoding schemes are used to reduce the size of the dataset. Various encoding schemes for data compression are wavelet transform analysis, Huffman encoding, principal component analysis, run-length encoding, etc. Mainly, data compression is divided into two categories:
  - ❖ Lossy Compression: This compression technique is used when the user can tolerate some information loss. This method reduces the data by compressing it and removing some of the bits of data without noticing it. This method takes less time and is cheaper than all the other compression techniques. After compression by this method, the original file can't be constructed back. This method is specially used for videos, graphics, audio, and images where the loss of some data at the time of compression is tolerated. There are

a few methods to implement lossy compression: a) JPEG (Joint Photographic Expert Group) b) MPEG (Motion of Pictures Expert Group) c) MP3 (MPEG audio layer 3).

- ❖ Lossless Compression: This compression technique provides the capability to uncompress the file without the loss of a single bit of any data. The user can retrieve the original file from a compressed file without any loss. This compression technique is used for executable files, text files, and spreadsheet files. When there is data redundancy then this compression technique is used. The integrity of the data is preserved with the use of a lossless compression technique. There are a few methods to implement this compression technique: a) Run-Length encoding b) Huffman Encoding c) Lempel Ziv Encoding

- Numerosity Reduction: In this process, the dataset is replaced by some mathematical model or a smaller representation model. There are many ways to replace the dataset with smaller alternatives or estimates. This technique is used to represent the original data in a much smaller form by using models. This method uses two types of approaches for implementation:
  - ❖ Parametric Approach: In this approach, a model is used to store the data. In this model, only the data parameters must be stored not the actual data. In this way, the redundancy of the data is removed by using different models. To implement this parametric approach, there are mainly two methods: a) Regression Models and b) Log-Linear models.
  - ❖ Non-Parametric Approach: This method is used to store the reduced representation of the original data. In this approach, the redundant data is first reduced to smaller data then that reduced data is stored by these non-parametric models. This non-parametric approach is implementedmainly by three models: a) Histograms b) Clustering c) Sampling
- Discretization or Concept Hierarchy Generation: Data discretization is the transformation of the continuous dataset of any attribute into a discrete or smaller set of intervals without the loss of any information. This method is beneficial for analysis because discrete data is much easier to analyze than a continuous one. In this, many constant values of the attributes are replaced by labels of smaller intervals. This method allows the mining of data at multiple levels of abstraction. This method is implemented using two approaches:
  - ❖ Top-Down Approach: This approach traverses the data from the generalized form to the specialized form. This method starts to find one or a few attributes and then split that attribute according to some parameters. This process of splitting is performedrecursively until the final specialized attributes are found.
  - ❖ Bottom-Up Approach: This approach traverses the data from the specialized form to the generalized form. This process starts with all the continuous attributes as potential split points, then some attributes are reduced by combining or aggregating them. This process of merging is performed recursively until the final aggregated database is found.

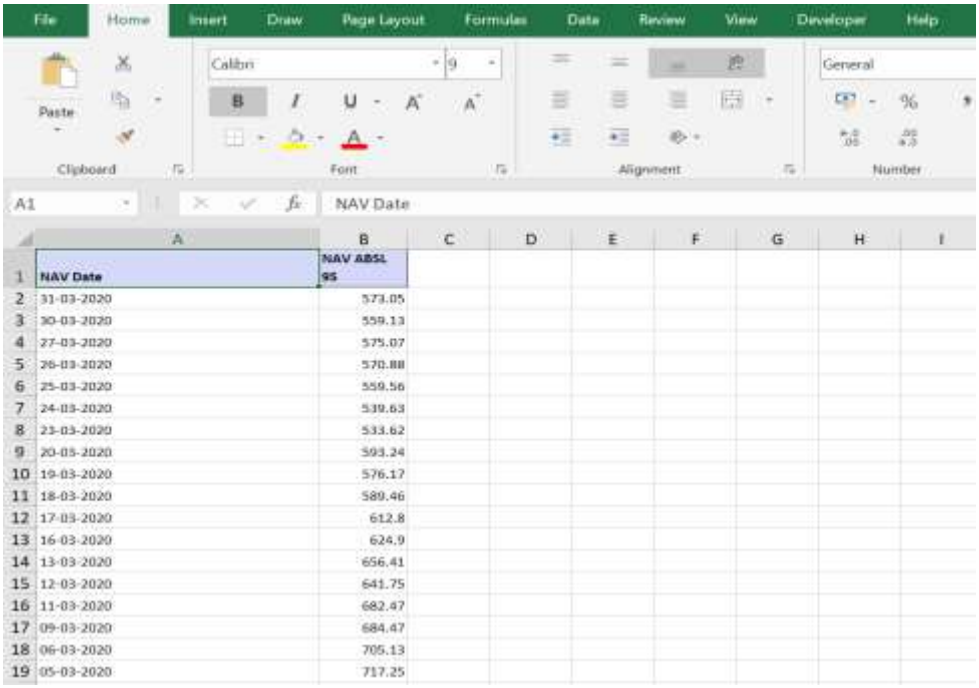
In this research analysis, the NAV of mutual funds is time-based data. The value of NAV is changed with the phase of time, and it is numeric in form. This data is already in a very concise and understandable format. So, there is no need for any reduction of data on this dataset.

3. Results And Discussions

This study presents the application of different data pre-processing techniques on mutual funds' data. In this section, the results of all the data pre-processing techniques are thoroughly analyzed and discussed pointwise.

3.1. Data Integration

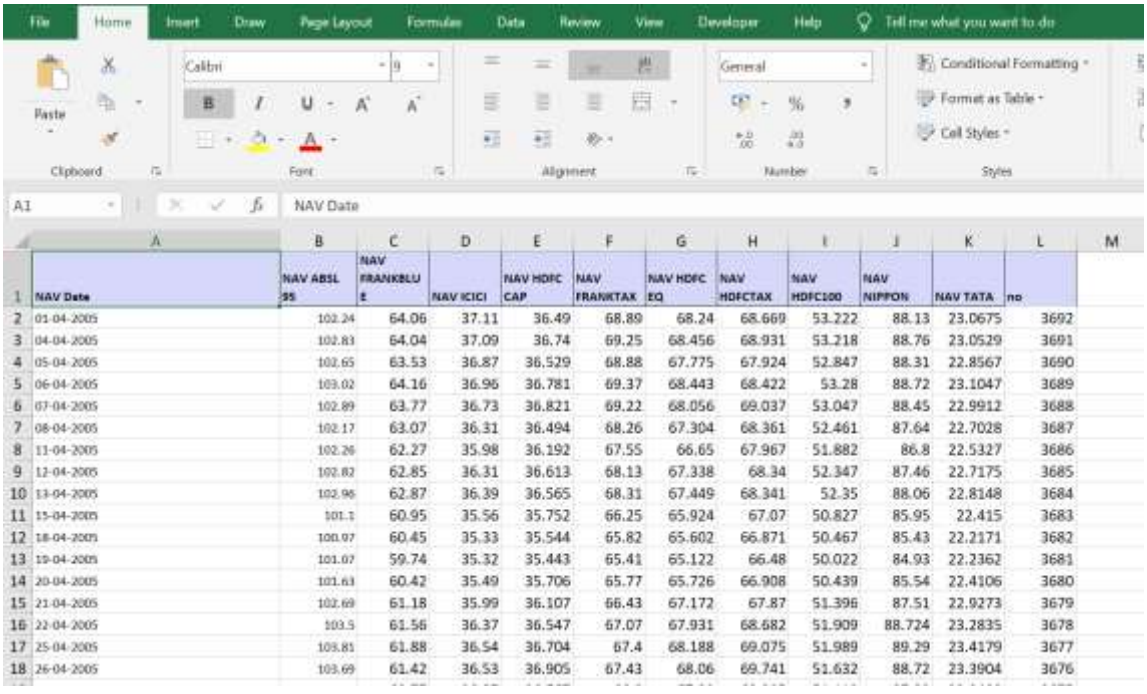
In this study, there are 20 files of each mutual fund with the NAV data of each mutual fund as shown in Fig. 8.



	NAV Date	NAV ABSL 95
2	31-03-2020	573.05
3	30-03-2020	559.13
4	27-03-2020	575.07
5	26-03-2020	570.88
6	25-03-2020	559.56
7	24-03-2020	539.63
8	23-03-2020	533.62
9	20-03-2020	593.24
10	19-03-2020	576.17
11	18-03-2020	589.46
12	17-03-2020	612.8
13	16-03-2020	624.9
14	13-03-2020	656.41
15	12-03-2020	641.75
16	11-03-2020	682.47
17	09-03-2020	684.47
18	06-03-2020	705.13
19	05-03-2020	717.25

Figure 8. *Single Mutual Fund view*

The next step is to join these all files into one unified file view so that the analysis can take place in an efficient and organized way as shown in Fig. 9. All the functions of join and append are done with the help of pandas in Python.

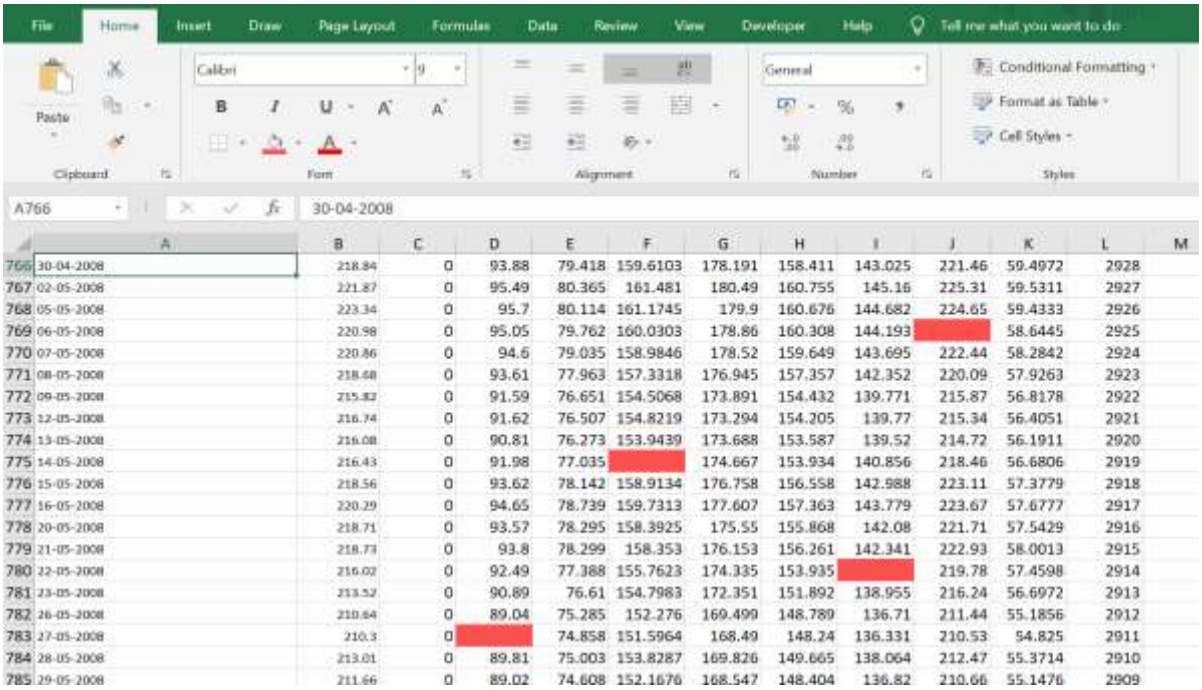


	NAV Date	NAV ABSL 95	NAV FRANKBLUE	NAV ICICI	NAV HDFC CAP	NAV FRANKTAX	NAV HDFC EQ	NAV HDFCTAX	NAV HDFC100	NAV NIPPON	NAV TATA	no
2	01-04-2005	102.24	64.06	37.11	36.49	68.89	68.24	68.669	53.222	88.13	23.0675	3692
3	04-04-2005	102.83	64.04	37.09	36.74	69.25	68.456	68.931	53.218	88.76	23.0529	3691
4	05-04-2005	102.65	63.53	36.87	36.529	68.88	67.775	67.924	52.847	88.31	22.8567	3690
5	06-04-2005	103.02	64.16	36.96	36.781	69.37	68.443	68.422	53.28	88.72	23.1047	3689
6	07-04-2005	102.89	63.77	36.73	36.821	69.22	68.056	69.037	53.047	88.45	22.9912	3688
7	08-04-2005	102.17	63.07	36.31	36.494	68.26	67.304	68.361	52.461	87.64	22.7028	3687
8	11-04-2005	102.26	62.27	35.98	36.192	67.55	66.65	67.967	51.882	86.8	22.5327	3686
9	12-04-2005	102.82	62.85	36.31	36.613	68.13	67.338	68.34	52.347	87.46	22.7175	3685
10	13-04-2005	102.96	62.87	36.39	36.565	68.31	67.449	68.341	52.35	88.06	22.8148	3684
11	15-04-2005	101.1	60.95	35.56	35.752	66.25	65.924	67.07	50.827	85.95	22.415	3683
12	18-04-2005	100.97	60.45	35.33	35.544	65.82	65.602	66.871	50.467	85.43	22.2171	3682
13	19-04-2005	101.07	59.74	35.32	35.443	65.41	65.122	66.48	50.022	84.93	22.2362	3681
14	20-04-2005	101.63	60.42	35.49	35.706	65.77	65.726	66.908	50.439	85.54	22.4106	3680
15	21-04-2005	102.69	61.18	35.99	36.107	66.43	67.172	67.87	51.396	87.51	22.9273	3679
16	22-04-2005	103.5	61.56	36.37	36.547	67.07	67.931	68.682	51.909	88.724	23.2835	3678
17	25-04-2005	103.81	61.88	36.54	36.704	67.4	68.188	69.075	51.989	89.29	23.4179	3677
18	26-04-2005	103.69	61.42	36.53	36.905	67.43	68.06	69.741	51.632	88.72	23.3904	3676

Figure 9. Unified View of All Mutual Funds

3.2. Data Cleaning

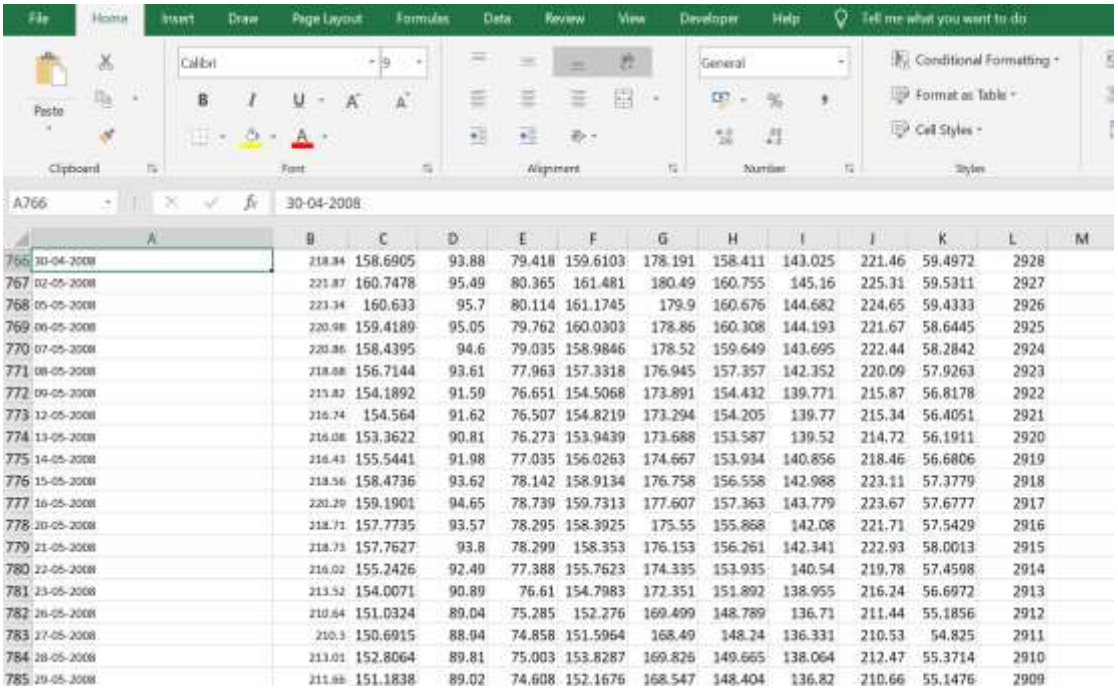
In the cleaning process, NAV data of mutual funds is missing regarding some dates in the dataset. While joining all the mutual funds' data, there are missing values in the dataset for some mutual funds which are shown in red color in Fig. 10.



	A	B	C	D	E	F	G	H	I	J	K	L	M
766	30-04-2008	218.84	0	93.88	79.418	159.6103	178.191	158.411	143.025	221.46	59.4972	2928	
767	02-05-2008	221.87	0	95.49	80.365	161.481	180.49	160.755	145.16	225.31	59.5311	2927	
768	05-05-2008	223.34	0	95.7	80.114	161.1745	179.9	160.676	144.682	224.65	59.4333	2926	
769	06-05-2008	220.98	0	95.05	79.762	160.0303	178.86	160.308	144.193		58.6445	2925	
770	07-05-2008	220.86	0	94.6	79.035	158.9846	178.52	159.649	143.695	222.44	58.2842	2924	
771	08-05-2008	218.68	0	93.61	77.963	157.3318	176.945	157.357	142.352	220.09	57.9263	2923	
772	09-05-2008	215.82	0	91.59	76.651	154.5068	173.891	154.432	139.771	215.87	56.8178	2922	
773	12-05-2008	216.74	0	91.62	76.507	154.8219	173.294	154.205	139.77	215.34	56.4051	2921	
774	13-05-2008	216.08	0	90.81	76.273	153.9439	173.688	153.587	139.52	214.72	56.1911	2920	
775	14-05-2008	216.43	0	91.98	77.035		174.667	153.934	140.856	218.46	56.6806	2919	
776	15-05-2008	218.56	0	93.62	78.142	158.9134	176.758	156.558	142.988	223.11	57.3779	2918	
777	16-05-2008	220.29	0	94.65	78.739	159.7313	177.607	157.363	143.779	223.67	57.6777	2917	
778	20-05-2008	218.71	0	93.57	78.295	158.3925	175.55	155.868	142.08	221.71	57.5429	2916	
779	21-05-2008	218.73	0	93.8	78.299	158.353	176.153	156.261	142.341	222.93	58.0013	2915	
780	22-05-2008	216.02	0	92.49	77.388	155.7623	174.335	153.935		219.78	57.4598	2914	
781	23-05-2008	213.52	0	90.89	76.61	154.7983	172.351	151.892	138.955	216.24	56.6972	2913	
782	26-05-2008	210.64	0	89.04	75.285	152.276	169.499	148.789	136.71	211.44	55.1856	2912	
783	27-05-2008	210.3	0		74.858	151.5964	168.49	148.24	136.331	210.53	54.825	2911	
784	28-05-2008	213.01	0	89.81	75.003	153.8287	169.826	149.665	138.064	212.47	55.3714	2910	
785	29-05-2008	211.66	0	89.02	74.608	152.1676	168.547	148.404	136.82	210.66	55.1476	2909	

Figure 10. *Missing Data*

In the interpolation method, the missing value in the dataset is filled by the average of its previous and next-day values as shown in Fig. 11. The datasets are now free from any error after the cleaning process.

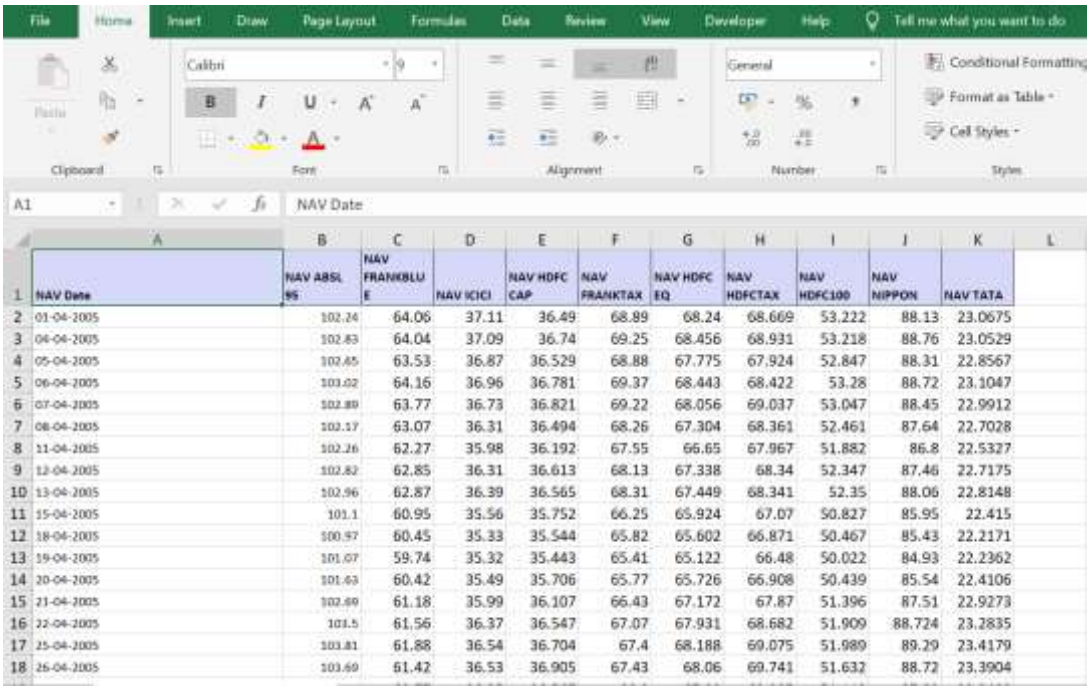


	A	B	C	D	E	F	G	H	I	J	K	L	M
766	30-04-2008	218.84	158.6905	93.88	79.418	159.6103	178.191	158.411	143.025	221.46	59.4972	2928	
767	02-05-2008	221.87	160.7478	95.49	80.365	161.481	180.49	160.755	145.16	225.31	59.5311	2927	
768	05-05-2008	223.34	160.633	95.7	80.114	161.1745	179.9	160.676	144.682	224.65	59.4333	2926	
769	06-05-2008	220.98	159.4189	95.05	79.762	160.0303	178.86	160.308	144.193	221.67	58.6445	2925	
770	07-05-2008	220.86	158.4395	94.6	79.035	158.9846	178.52	159.649	143.695	222.44	58.2842	2924	
771	08-05-2008	218.68	156.7144	93.61	77.963	157.3318	176.945	157.357	142.352	220.09	57.9263	2923	
772	09-05-2008	215.82	154.1892	91.59	76.651	154.5068	173.891	154.432	139.771	215.87	56.8178	2922	
773	12-05-2008	216.74	154.564	91.62	76.507	154.8219	173.294	154.205	139.77	215.34	56.4051	2921	
774	13-05-2008	216.08	153.3622	90.81	76.273	153.9439	173.688	153.587	139.52	214.72	56.1911	2920	
775	14-05-2008	216.43	155.5441	91.98	77.035	156.0263	174.667	153.934	140.856	218.46	56.6806	2919	
776	15-05-2008	218.56	158.4736	93.62	78.142	158.9134	176.758	156.558	142.988	223.11	57.3779	2918	
777	16-05-2008	220.29	159.1901	94.65	78.739	159.7313	177.607	157.363	143.779	223.67	57.6777	2917	
778	20-05-2008	218.71	157.7735	93.57	78.295	158.3925	175.55	155.868	142.08	221.71	57.5429	2916	
779	21-05-2008	218.73	157.7627	93.8	78.299	158.353	176.153	156.261	142.341	222.93	58.0013	2915	
780	22-05-2008	216.02	155.2426	92.49	77.388	155.7623	174.335	153.935	140.54	219.78	57.4598	2914	
781	23-05-2008	213.52	154.0071	90.89	76.61	154.7983	172.351	151.892	138.955	216.24	56.6972	2913	
782	26-05-2008	210.64	151.0324	89.04	75.285	152.276	169.499	148.789	136.71	211.44	55.1856	2912	
783	27-05-2008	210.3	150.6915	88.94	74.858	151.5964	168.49	148.24	136.331	210.53	54.825	2911	
784	28-05-2008	213.01	152.8064	89.81	75.003	153.8287	169.826	149.665	138.064	212.47	55.3714	2910	
785	29-05-2008	211.66	151.1838	89.02	74.608	152.1676	168.547	148.404	136.82	210.66	55.1476	2909	

Figure 11. *Filling Missing Values Using Interpolation Method*

3.3. Data Transformation

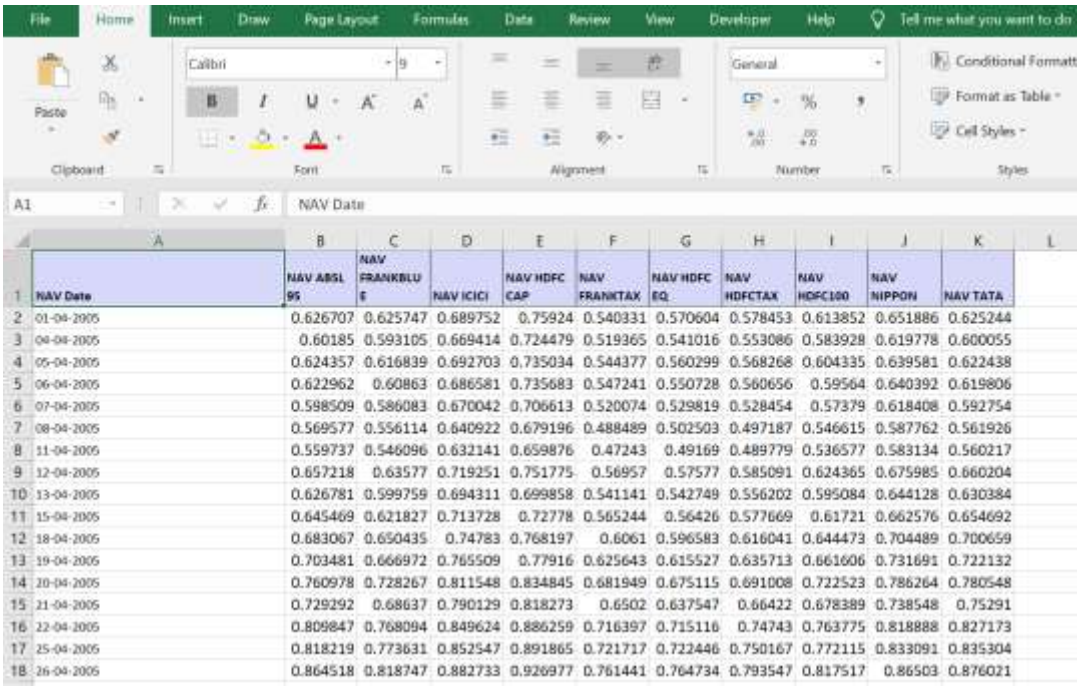
The NAV data of all mutual funds are on widely different scales. Their range of NAV data is very distinct from each other as shown in Fig. 12. It is highly demanding to analyze this data with these different scales.



	NAV Date	NAV ABSL 95	NAV FRANKBLUE	NAV ICICI	NAV HDFC CAP	NAV FRANKTAX	NAV HDFC EQ	NAV HDFCTAX	NAV HDFC100	NAV NIPPON	NAV TATA
2	01-04-2005	102.24	64.06	37.11	36.49	68.89	68.24	68.669	53.222	88.13	23.0675
3	04-04-2005	102.83	64.04	37.09	36.74	69.25	68.456	68.931	53.218	88.76	23.0529
4	05-04-2005	102.65	63.53	36.87	36.529	68.88	67.775	67.924	52.847	88.31	22.8567
5	06-04-2005	103.02	64.16	36.96	36.781	69.37	68.443	68.422	53.28	88.72	23.1047
6	07-04-2005	102.89	63.77	36.73	36.821	69.22	68.056	69.037	53.047	88.45	22.9912
7	08-04-2005	102.17	63.07	36.31	36.494	68.26	67.304	68.361	52.461	87.64	22.7028
8	11-04-2005	102.26	62.27	35.98	36.192	67.55	66.65	67.967	51.882	86.8	22.5327
9	12-04-2005	102.82	62.85	36.31	36.613	68.13	67.338	68.34	52.347	87.46	22.7175
10	13-04-2005	102.96	62.87	36.39	36.565	68.31	67.449	68.341	52.35	88.06	22.8148
11	15-04-2005	101.1	60.95	35.56	35.752	66.25	65.924	67.07	50.827	85.95	22.415
12	18-04-2005	100.97	60.45	35.33	35.544	65.82	65.602	66.871	50.467	85.43	22.2171
13	19-04-2005	101.07	59.74	35.32	35.443	65.41	65.122	66.48	50.022	84.93	22.2362
14	20-04-2005	101.63	60.42	35.49	35.706	65.77	65.726	66.908	50.439	85.54	22.4106
15	21-04-2005	102.69	61.18	35.99	36.107	66.43	67.172	67.87	51.396	87.51	22.9273
16	22-04-2005	103.5	61.56	36.37	36.547	67.07	67.931	68.682	51.909	88.724	23.2835
17	25-04-2005	103.81	61.88	36.54	36.704	67.4	68.188	69.075	51.989	89.29	23.4179
18	26-04-2005	103.69	61.42	36.53	36.905	67.43	68.06	69.741	51.632	88.72	23.3904

Figure 12. *Unified View of all Mutual funds before Normalization*

The data normalization is done on the dataset as shown in Fig. 13. To normalize the dataset, the Min-Max normalization technique is used in which all the dataset's values are scaled to a new range of values, i.e. [0,1]. This technique preserves the original relation of the values of the dataset with each other.



	NAV Date	NAV ABSL 95	NAV FRANKBLUE	NAV ICICI	NAV HDFC CAP	NAV FRANKTAX	NAV HDFC EQ	NAV HDFCTAX	NAV HDFC100	NAV NIPPON	NAV TATA
2	01-04-2005	0.626707	0.625747	0.689752	0.75924	0.540331	0.570604	0.578453	0.613852	0.651886	0.625244
3	04-04-2005	0.60185	0.593105	0.669414	0.724479	0.519365	0.541016	0.553086	0.583928	0.619778	0.600055
4	05-04-2005	0.624357	0.616839	0.692703	0.735034	0.544377	0.560299	0.568268	0.604335	0.639581	0.622438
5	06-04-2005	0.622962	0.60863	0.686581	0.735683	0.547241	0.550728	0.560656	0.59564	0.640392	0.619806
6	07-04-2005	0.598509	0.586083	0.670042	0.706613	0.520074	0.529819	0.528454	0.57379	0.618408	0.592754
7	08-04-2005	0.569577	0.556114	0.640922	0.679196	0.488489	0.502503	0.497187	0.546615	0.587762	0.561926
8	11-04-2005	0.559737	0.546096	0.632141	0.659876	0.47243	0.49169	0.489779	0.536577	0.583134	0.560217
9	12-04-2005	0.657218	0.63577	0.719251	0.751775	0.56957	0.57577	0.585091	0.624365	0.675985	0.660204
10	13-04-2005	0.626781	0.599759	0.694311	0.699858	0.541141	0.542749	0.556202	0.595084	0.644128	0.630384
11	15-04-2005	0.645469	0.621827	0.713728	0.72778	0.565244	0.56426	0.577669	0.61721	0.662576	0.654692
12	18-04-2005	0.683067	0.650435	0.74783	0.768197	0.6061	0.596583	0.616041	0.644473	0.704489	0.700659
13	19-04-2005	0.703481	0.666972	0.765509	0.77916	0.625643	0.615527	0.635713	0.661606	0.731691	0.722132
14	20-04-2005	0.760978	0.728267	0.811548	0.834845	0.681949	0.675115	0.691008	0.722523	0.786264	0.780548
15	21-04-2005	0.729292	0.68637	0.790129	0.818273	0.6502	0.637547	0.66422	0.678389	0.738548	0.75291
16	22-04-2005	0.809847	0.768094	0.849624	0.886259	0.716397	0.715116	0.74743	0.763775	0.818888	0.827173
17	25-04-2005	0.818219	0.773631	0.852547	0.891865	0.721717	0.722446	0.750167	0.772115	0.833091	0.835304
18	26-04-2005	0.864518	0.818747	0.882733	0.926977	0.761441	0.764734	0.793547	0.817517	0.86503	0.876021

Figure 13. *Applying Min-Max Normalization*

3.4. Data Reduction

In this analysis, the mutual funds' data is time-based. All the data in the dataset is in numeric form and change with the phase of time. The data related to mutual funds is already very limited and standardized form. So, there is no need for any data reduction on this dataset.

4. Conclusion

Data pre-processing is an essential issue for both data warehousing and data mining. In the real

world, the raw data, which is collected from different sources, is incomplete, noisy, or inconsistent. The primary goal of this paper is to pre-process the mutual funds' data and make it easily accessible for analysis. Firstly, in this paper, all the pre-processing techniques are explained thoroughly. After this, the focus is on the technique, which follows in this research work, to pre-process the dataset. In this research work, the dataset contains data related to mutual funds for the last 15 years. All the pre-processing techniques are applied to that dataset and make it efficient for the analysis propose. In this paper, the overall impact of data pre-processing on the raw data of mutual funds has been systematically evaluated and discussed in the result section. Data pre-processing gives high-quality data that leads to efficient results and less cost on data mining.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Appendix

After the analysis of different pages on various links, a list of the top ten growth and top ten dividend mutual funds is listed and the list of mutual funds that are used for the research work is as follows:

- Aditya Birla Sun Life Equity Hybrid 95 Fund - Regular Plan-Dividend
- Aditya Birla Sun Life Equity Hybrid 95 Fund - Regular Plan-growth
- Franklin India Bluechip Fund-Dividend
- Franklin India Bluechip Fund- growth
- Franklin India Taxshield-Dividend
- Franklin India Taxshield- growth
- HDFC Capital Builder Value Fund - Dividend Option
- HDFC Capital Builder Value Fund - Growth Option
- HDFC Top 100 Fund - Dividend Option
- HDFC Top 100 Fund - Growth Option
- HDFC Equity Fund - Dividend Option
- HDFC Equity Fund - Growth Option
- HDFC TaxSaver-Dividend Plan
- HDFC TaxSaver- Growth Plan
- ICICI Prudential Multicap Fund – Dividend
- ICICI Prudential Multicap Fund – growth
- Tata Ethical Fund Regular Plan - Dividend
- Tata Ethical Fund Regular Plan - growth
- Nippon India Vision Fund-Dividend Plan-Dividend option
- Nippon India Vision Fund- growth Plan- growth option

References

[1] Gupta, G., Aggarwal, H., & Rani, R. (2016). Segmentation of retail customers based on cluster analysis in building successful CRM. *International Journal of Business Information Systems*, 23(2), 212–228.

[2] Qamar, H., & Singh, S. (2016). Mutual fund performance prediction. *2016 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics, DISCOVER 2016 - Proceedings*, 1, 185–189.

[3] Anish, C. M., & Majhi, B. (2016). Prediction of mutual fund net asset value using low complexity feedback neural network. *2016 IEEE International Conference on Current Trends in Advanced Computing, ICCTAC* 2016, 0–4.

[4] Abasova, J., Janosik, J., Simoncicova, V., & Tanuska, P. (2018). Proposal of Effective Pre-processing Techniques of Financial Data. INES 2018 - *IEEE 22nd International Conference on Intelligent Engineering Systems, Proceedings*, 293–298.

[5] Anish, C. M., & Majhi, B. (2015). An ensemble model for Net asset value prediction. 2015 IEEE Power, *Communication and Information Technology Conference, PCITC 2015 - Proceedings*, 392–396.

[6] Samsani, S. (2016). An RST based efficient pre-processing technique for handling inconsistent data. 2016 *IEEE International Conference on Computational Intelligence and Computing Research, ICCIC* 2016.

[7] Gupta, S., & Gupta, A. (2019). Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, 161, 466–474.

[8] Aljuaid, T., & Sasi, S. (2016). Proper imputation techniques for missing values in data sets. Proceedings of the *2016 International Conference on Data Science and Engineering*, ICDSE 2016.

[9] Wang, X., Ma, Y., & Li, X. (2010). Data mining in inconsistent data. *International Conference on Internet Technology and Applications, ITAP 2010 - Proceedings.*

[10] Kumar, M., & Kalia, A. (2012). Pre-processing and symbolic representation of stock data. Proceedings - 2012 *2nd International Conference on Advanced Computing and Communication Technologies, ACCT 2012*, 83–88.

[11] Pan, J., Zhuang, Y., & Fong, S. (2016). The impact of data normalization on stock market prediction: Using SVM and technical indicators. *Communications in Computer and Information Science*, 652, 72–88.

[12] Nayak, S. C., Misra, B. B., & Behera, H. S. (2014). Impact of Data Normalization on Stock Index Forecasting. *International Journal of Computer Information Systems and Industrial Management Applications*, 6, 257–269.

[13] Wei, J. (2010). Research on data pre-processing in supermarket customers data mining. *2nd International Conference on Information Engineering and Computer Science - Proceedings, ICIECS 2010.*

[14] Chen, A., Liu, F. H., & Wang, S. De. (2019). Data reduction for real-time bridge vibration data on edge. Proceedings - *2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA* 2019, 602–603.