MATHEMATICAL TECHNIQUES FOR PROCESSING MIXED-FORMAT DATA AND THEIR VISUALIZATION IN PREDICTIVE MODELING

Xayrullayev Mo‘minbek
Tashkent State University of Economics,
mominbek97@gmail.com
Norqulova Ziyoda
Tashkent State University of Economics,
ziyodanorqulova88@gmail.com
Dinora Gulmurodova
Tashkent State University of Economics, dinoragulmurodova1@gmail.com

Abstract. In today's data times, the processing of text and numerical data in the database is becoming very important. Numerical data is used in accounting, statistics, predictive models and on many fronts that are important to society. For example, financial statements, income and expenses are expressed in a digital format, since this is calculated. But the fact that these data are mixed with other characters in some cases, that is, not in the pure number case, makes the analysis difficult. It is also an important process to transfer textual information to a numerical type to display it in the analysis. Searching and analyzing such data is performed using specialized methods and algorithms, such as full-text search or text indexing.

Keywords: encoding methods, textual data processing, numerical data analysis, data preprocessing, numerical data, missing value imputation.

Introduction

A well-designed database system must utilize optimal data types to improve overall efficiency. The main task of this research is to clean mixed-type data (i.e., data with a mixture of text and numeric characters) and represent them in a single numeric form. This is because the ability to perform visualization operations on numeric or text values accompanied by various characters (for example: %, $, ,, () and others) and to enter this data into artificial intelligence (AI) or machine learning (ML) models is limited [10].

Therefore, this study extensively covers the methods of preliminary data cleaning, including the steps of removing special characters and converting them to float or integer, and methods of converting text columns to numeric values (encoding). Especially since ML and AI models cannot work directly with text values, converting them to a numeric form is of great practical importance [2]. After cleaning the data, the possibilities of visualizing, analyzing and entering them into models increase. However, only the cleaning and encoding processes are

ISSN 3030-3907

considered within the scope of this study. The issue of filling in missing values (NaN) will be analyzed in depth in future research [11-12].

Recent research places emphasis on the importance of numerical and textual data to increase the effectiveness of data analytics and prediction models. Especially for structured and unstructured data synthesis, it plays an important role in the forward development of areas such as financial, healthcare and marketing [13].

For example, a study that proposed a multimodal in-depth learning system that combined textual financial news and digital metrics to predict the company's credit ratings [1].

Similarly, in the field of online reviews, researchers have developed many methods for collaborative analysis of rating scores and textual feedback. This method has been successfully clustered by modeling emotions and themes from the text in conjunction with the distribution of consumer opinion assessments [3]. Another notable contribution was provided by Applied Intelligence, where researchers studied feature selection strategies to identify AI-generated comments [4].

Methodology

The study utilizes the NASDAQ1 dataset, containing stock-related information including names, sectors, industries, and price changes. Initial inspection revealed missing values and mixed-format features that required preprocessing.

Data Cleaning

Columns with over 50% missing values were dropped: earnings_per_share, annualized_dividend, dividend_pay_date, and symbol_yield.

The Unnamed: 0 column was removed as it was not used as an index.

Textual Data Cleaning Textual columns often contained special characters such as $, %, (,). These were removed using the following function:

Let's assume:

$$\sum_{i=1}^{s} - \ all\ set\ of\ sign,$$

$x = s_1, s_2, \ldots, s_n, s_i \in \sum_{i=1}^{s}$ given text values,

$A = \{0,1,2,\ldots,9,\ldots\}$ – only numerical sings and dot.

The cleaning function is:

$$Clean(x) = concat(s_i \in x \mid s_i \in A), A = \{0,\ldots,9,..\} \quad (1)$$

Here

ISSN 3030-3907

concat() is an operator for adding characters one after another (i.e., creating new text),
only characters belonging to the A set (numbers and periods) are retained,
other characters (%, $, (,), /,) are removed.

By applying this function to the corresponding columns, the values of the mixed type column "pricing_percentage_changes", "symbol_yield", "annualized_dividend", "dividend_pay_date", "earnings_per_share" were cleaned. After removing the extra characters with this function, the data can be converted to a numeric type and used in the model.

At this point, the question arises, what if the data consists entirely of words? In this case, it can be converted to numeric data using Encoding methods according to the data types. In this case, the methods are selected according to the data types. The dataset can have the following data types:

- Categorical data
- Ordinal data
- Numerical data
- Logical data

Also, the available encoding methods are as follows:
One-Hot Encoding: Suitable for unordered categories.
Ordinal Encoding: Suitable for ordered categories (e.g., education levels).

Label Encoding: Each category is assigned a unique integer.

Categorical (nominal) data represents categories or groups and has no order. For example, colors, cities, gender and so on. One-Hot Encoding is a method for converting categorical data into a numerical format in machine learning and data preprocessing. This method is used, in particular, in artificial intelligence and analytical models to create a numerical representation of categorical values [4]. One of the advantages of this method is that it eliminates ordinal dependencies - that is, it prevents the misconception of which of the categories is larger or smaller.

If we have :

$C = \{c_1, c_2, \ldots, c_k\}$ – k unique categories,

$x \in C$ – x chosen category

Then:

$$OneHot(x) = [\delta(x = c_1), \delta(x = c_2), \ldots, \delta(x = c_k)]$$

Here :

$$\delta(x = c_i) = \begin{cases} 1, if\ x = c_i \\ 0, \quad else \end{cases}$$

ISSN 3030-3907

Table 1.

	Name	Sector	Industry
Name	1	0	0
Sector	0	1	0
Industry	0	0	1

Ordinal data is also categorized, but it has an order and is encoded using Ordinal Encoding. Goal is representing internally ordered categories (e.g. "Low" < "Medium" < "High") with ordered integers. Categories are represented by numbers. Ordinal Encoding is a method of converting categorical data into a numeric format by assigning ordinal numbers (0, 1, 2, 3...) to the category values. This method preserves the ranking relationship between the category values. If we have:

$C = \{c_1, c_2, \ldots, c_k\}$ – sorted category set and $x \in C$

$f: C \to Z$ -mapping according to order

Then:

$$Ordinal(x) = f(x), f(c_i) = i - 1$$

For example:

If C = {“low”,” medium”, “high”} then: f (“low”)=0, f (“medium”) =1, f (“high”) =2.

We use Label encoding to assign a numerical index to each of the category values, even if they are not ordered, in which each category is replaced by a single number. As above, if we have:

$$C = \{c_1, c_2, \ldots, c_k\} - unique\ \ category\ set\ and\ x \in C$$

Then:

$$\text{Label(x)} = \text{Index}_c\text{(x)}$$

Here $\text{Index}_c\text{(x)}$ – ordinal number of the value of x within C (usually starting from 0).

For example: if C= {“retail” , ”finance” ,”tech”}, then: Label(“finance”) =1

For the Nasdaq dataset selected above, all three methods were used and the columns “name”, “sector”, and “industry” were converted to numeric format:

Table 2.

	name	sector	industry
0	617	10	30
1	4943	10	31
2	430	10	66
3	429	10	66
…	…	…	…

ISSN 3030-3907

8335	8207	12	141
8336	8208	12	141
8337	8179	12	141
8338	8180	12	141

After all mixed-type columns and text columns are transferred to their numerical appearance, we fill in its missing values (Nan). This filling process will also necessarily depend on the structure of the data. In completing the deficient values, the median and mean were used in our research work as dataset was put on standby:

Mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Here n - number of non-Nan available values,

For skewed distributions, Nan values replaced using median, which is less sensate to outliers:

$$Median = \begin{cases} x_{n+1/2}, if\ n\ is\ odd \\ \dfrac{x_{n/2} + x_{(n/2+1)}}{2}, if\ n\ is\ oven \end{cases}$$

Here x- values in the ordered (incremental) sequence.

After we have cleaned up columns like “pricing_percentage_changes” above using the formula (1), then one can easily fill in Nan values with these mean or median methods. Also if the dataset did not have the need to transfer textual values to the numerical type, filling its Nan values with mode- that is, the most common value, would also have done very well. It can also be easily applied in numbered columns:

$$Mode(X) = arg\ \max_{x_i} f(x_i)$$

Where $f(x_i)$ is the frequency of value x_i in the dataset.

Naturally, these methods are selected based on the column values in the dataset. All mixed-type columns in the dataset were cleaned, text columns were transferred to the numerical type, and all Nan values were supplemented by the above methods.

Results

Visualization of Preprocessed Data

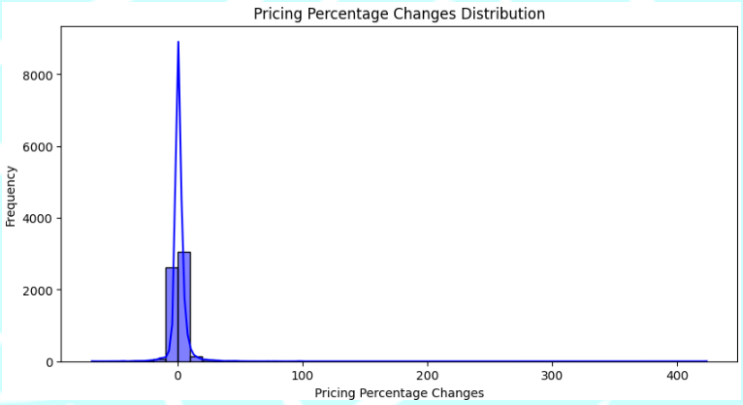
So, now it is possible to carry out an analysis on it, visualizing the dataset, which has come to a ready-made state. That all the data has been converted to a

ISSN 3030-3907

numeric type, it is time to analyze the results visually and apply models to it [5]. To understand and analyze the data, it is important to display it in a graphical form, which can be histograms, charts, and network graphs [6].

Some values in the dataset were originally represented by special characters, and a pre-cleaning operation was performed to convert them to numeric float type. This cleaning method allowed for numerical analysis of the data, training model and its visualization.

Values in the "pricing_percentage_changes" column are mostly concentrated around 0%, but there is a significant density in the range of ±5%, in figure 1. The distribution is right-skewed, indicating that extremely high percentages of price changes are rare.



).

Figure 2.

The boxplot diagram identified outliers. For example, some companies had price changes below -20% or above +20%. This indicates the need to consider or transform these values separately during statistical analysis. These unusual values can be the result of possible technical errors, market reactivity, or inconsistencies in data entry.

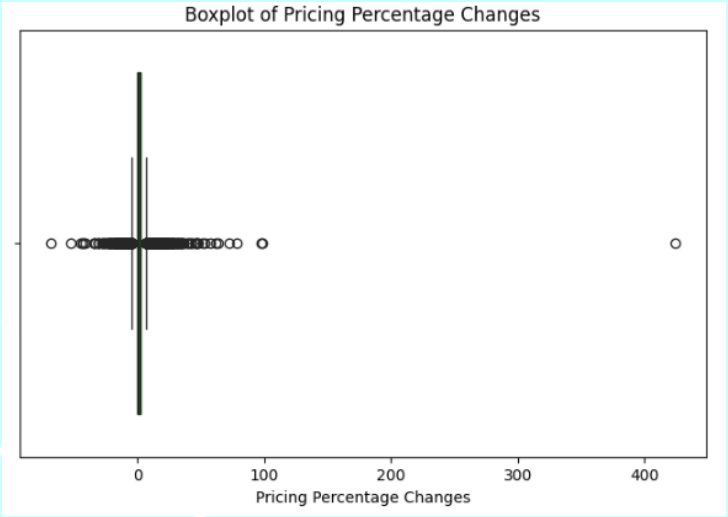


Figure 3.

ISSN 3030-3907

Based on the refined values of the "pricing_percentage_changes" column, average price changes across different industries were analyzed. The visualization shown in Figure 3. reveals differences across sectors.

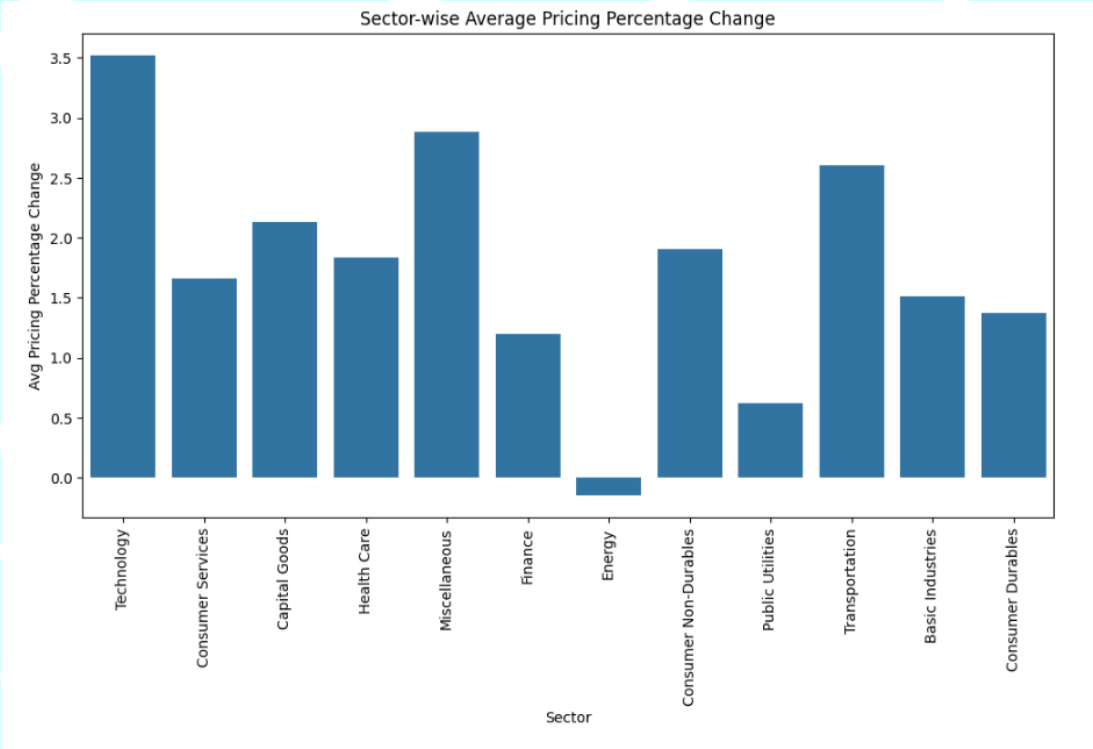


Figure 3.

The graph shows that some industries (e.g. energy, financial services) were more affected by average price changes than others. In contrast, price changes in some other industries were minimal and stable. These differences may be due to different market conditions or risks specific to the industry. Such a group visualization is important for sectoral analysis and shows the need to consider the sector factor as a parameter in building a model [15].

Discussion

Working with numerical and textual data simultaneously in a single analytical workflow has proven to be both a challenge and an opportunity in data science. The results of this study point out the importance of preprocessing, encoding, and visualization techniques to transform raw, mixed-format datasets into structured formats that can be effectively analyzed and modeled. Preprocessing significantly affects the predictive power of deep learning systems, particularly when multiple data modalities are combined [8].

The use of encoding techniques such as One-Hot Encoding, Ordinal Encoding, and Label Encoding enabled the transformation of categorical text data into a machine-readable format. These methods differ in terms of complexity and suitability depending on whether the data is ordinal or nominal. In this study, the use of all three methods showed how hybrid encoding strategies can be tailored to

ISSN 3030-3907

suit the data characteristics. Proper encoding improves the detection of AI-generated content when using both numeric and textual features [9].

In addition, textual data cleaning significantly improved the quality of the dataset. By removing special characters and redundant symbols, the data became suitable for numeric transformation and integration into predictive models. Furthermore, the visualization techniques applied (histogram, boxplot, and bar chart) played a crucial role in understanding data distribution and identifying anomalies. For example, the boxplot effectively revealed outliers in price percentage changes, which may indicate anomalous stock behavior. Such graphical visualizations are not only useful for exploratory data analysis (EDA) but are also fundamental in communicating insights to stakeholders and guiding model selection.

Despite the successful outcomes, certain drawbacks exist. The dataset, while suitable for demonstrating methods, was relatively small in scale. In larger datasets, One-Hot Encoding can lead to the "high dimensional complexity", making memory usage and computation time significant bottlenecks. In future studies, we will explore methods to not only prepare the dataset for model, but also to fill in missing values, therefore, should investigate dimensionality reduction techniques such as PCA or feature selection heuristics to optimize such encoding results.

Conclusion

This article examines a practical approach to textual and numerical data processing methods, their mathematical foundations, and visualization. Also, data visualization was performed to understand and analyze the data. Data distribution and anomalies were analyzed using graphical tools such as histograms, boxplots, and bar charts. Therefore, there are special algorithms for filling Nan values in the world, there are many special algorithms for calculating nan values, in our future work we will focus on them in detail, so far we have limited ourselves to simple mathematical methods.

This proposed methodology will help to obtain accurate analytical conclusions from the dataset and make decisions, simplify the workflow, as well as expand the possibilities of forecasting. The article outlines the fundamentals and practical methods of managing text and numerical data using Python and late distributed data science libraries. In general, this strategy is a valuable resource for effective processing and analysis.

References

1.Mahsa Tavakoli, Rohitash Chandra, Fengrui Tian, Cristián Bravo. Multi-modal Deep Learning for Credit Rating Prediction Using Text and Numerical Data (2025).

ISSN 3030-3907

2.Joshi, A. P., & Patel, B. V. (2019). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. Oriental Journal of Computer Science and Technology.
3.Elena Barzizza, Nicola Biasetton, Marta Disegna, Luigi Salmaso. Combining Textual and Rating Data from Online Reviews to Cluster Customer Feedback. Springer (2025).
4. Seung-Cheol Lee, Dong-Gun Lee, Yeong-Seok Seo. Determining the Best Feature Combination Through Text and Numerical Data. Springer (2023).
5. Norqulova, Ziyoda. "DATA VISUALIZATION STEPS TO SORT DATA BY REMOVING MISSING VALUES AND CONVERTING DATA TO NUMERIC TYPE." DTAI–2024 1.DTAI (2024): 37-39.
6.Muminov, Bahodir, and N. Z. Norqulova. "VIZUALIZATSIYA MASALALARIDA SUN'IY INTELLEKTNING ROLI." DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE 1.1 (2023): 7-14.
Tomas Hrycej, Bernhard Bermeitinger, Matthias Cetto, Siegfried Handschuh. Mathematical Foundations of Data Science. Springer, 2023.
7.Beknazarova, S. S., and M. Sh Bekmirzayeva. "Study of the Degree of Influence of Fog on Images in Visualization Systems." (2024).
8.Mahsa Tavakoli, Rohitash Chandra, Fengrui Tian, Cristian Bravo. Multi-modal Deep Learning for Credit Rating Prediction Using Text and Numerical Data Streams. Applied Soft Computing. 2025.
9.Jooyoung Lee, Thai Le, Jinghui Chen, Dongwon Lee. Identifying Artificial Intelligence–Generated Content in Online Q&A Platforms. Journal of Information Science. 2023.
10.Kovalerchuk, B., & McCoy, E. Explainable Machine Learning for Categorical and Mixed Data with Lossless Visualization. 2023.
11.Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D., Khomh, F. Data Cleaning and Machine Learning: A Systematic Literature Review. 2023.
12.Ahmad, A. F., Sayeed, M. S., Alshammari, K., Ahmed, I. Impact of Missing Values in Machine Learning: A Comprehensive Analysis. 2024.
13.Kaur, S., Sharma, R. Structured vs. Unstructured Data in Machine Learning Prediction of Suicide Risk: A Systematic Review and Meta-Analysis. Frontiers in Digital Health.2022.
14.Maćkowiak, B., Moench, E., Wiederholt, M. Sectoral Price Data and Models of Price Setting. 2009.

ISSN 3030-3907