|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Project Report  GST Hackathon |
| |  |  |  | | --- | --- | --- | | Pulkit Saxena | 9/21/24 | pulkit21aug@gmail.com | |

**Pulkit Saxena**

**Email –** [**pulkit21aug@gmail.com**](mailto:pulkit21aug@gmail.com)

**Working Professional**

**LinkedIn - https://www.linkedin.com/in/pulkit-saxena21/**

Contents

[**Idea/Concept** 2](#_Toc177801963)

[**Project Description** 3](#_Toc177801964)

[**Source Code URL (github.com)** 3](#_Toc177801965)

[**Assumptions** 3](#_Toc177801966)

[**Exploratory Data Analysis** 3](#_Toc177801967)

[**Columns in the dataset:** 3](#_Toc177801968)

[**EDA - ID** 3](#_Toc177801969)

[**EDA-Column0** 4](#_Toc177801970)

[**EDA-Column1** 5](#_Toc177801971)

[**EDA-Column2** 5](#_Toc177801972)

[**EDA-Column3** 6](#_Toc177801973)

[**EDA-Column4** 8](#_Toc177801974)

[**EDA-Column5** 8](#_Toc177801975)

[**EDA-Column6** 9](#_Toc177801976)

[**EDA-Column7** 10](#_Toc177801977)

[**EDA-Column8** 11](#_Toc177801978)

[**EDA-Column9** 12](#_Toc177801979)

[**EDA-Column10** 13](#_Toc177801980)

[**EDA-Column11** 14](#_Toc177801981)

[**EDA-Column12** 15](#_Toc177801982)

[**EDA-Column13** 16](#_Toc177801983)

[**EDA-Column14** 17](#_Toc177801984)

[**EDA-Column15** 17](#_Toc177801985)

[**EDA-Column16** 18](#_Toc177801986)

[**EDA-Column17** 19](#_Toc177801987)

[**EDA-Column18** 20](#_Toc177801988)

[**EDA-Column19** 21](#_Toc177801989)

[**EDA-Column20** 22](#_Toc177801990)

[**EDA-Column21** 23](#_Toc177801991)

[**Co-Relation analysis** 24](#_Toc177801992)

[**Target Class distribution** 25](#_Toc177801993)

[**Data Modelling – Logistic Regression** 26](#_Toc177801994)

[**Model Performance Report** 26](#_Toc177801995)

[**Model performance on training data** 26](#_Toc177801996)

[**Model performance on Test data** 27](#_Toc177801997)

[**Results Summary** 28](#_Toc177801998)

# **Idea/Concept**

Develop a predictive model: Predict the value of target variable using independent variables using efficient algorithm with high prediction accuracy.

# **Project Description**

Design a predictive model aimed at estimating the value of a target variable based on various independent variables. This model should leverage a highly efficient algorithm that can deliver superior prediction accuracy. The focus will be on selecting the most relevant independent variables and optimizing the algorithm to enhance performance. By analyzing the relationships between the independent variables and the target variable, the model will be able to generate reliable forecasts. The ultimate goal is to ensure that the predictions made by the model are as accurate and dependable as possible, enabling informed decision-making based on the insights gained.

# **Source Code URL (github.com)**

<https://github.com/pulkit21aug/GSTHackathonSubmission>

# **Assumptions**

* 1. Domain knowledge about the columns is missing.
  2. Model is based on existing patterns identified during exploratory data analysis.

# **Exploratory Data Analysis**

## **Columns in the dataset:**

Index(['ID', 'Column0', 'Column1', 'Column2', 'Column3', 'Column4', 'Column5',

'Column6', 'Column7', 'Column8', 'Column9', 'Column10', 'Column11',

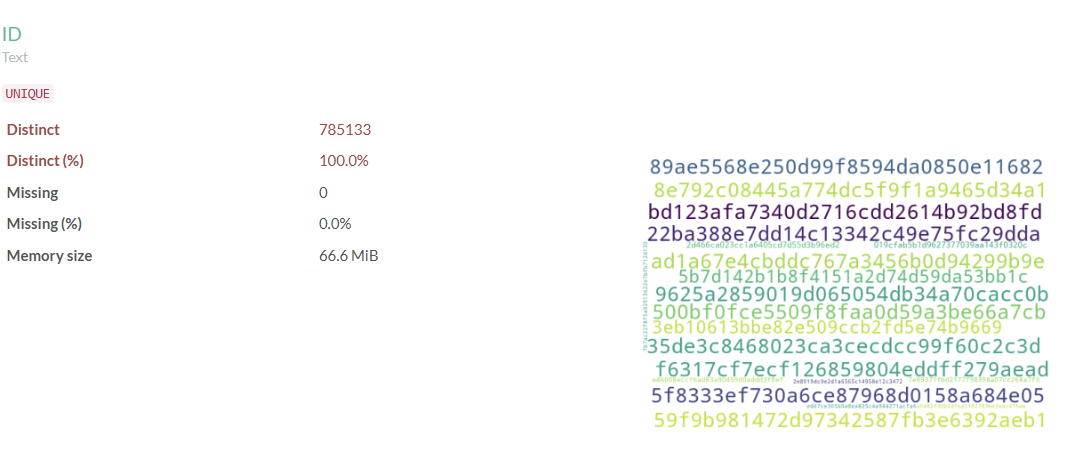
'Column12', 'Column13', 'Column14', 'Column15', 'Column16', 'Column17',

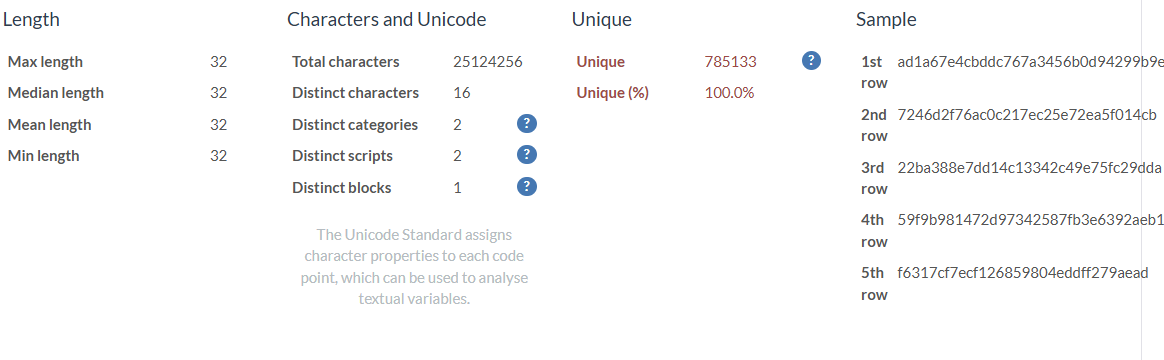
'Column18', 'Column19', 'Column20', 'Column21', 'target'],

dtype='object')

## **EDA - ID**

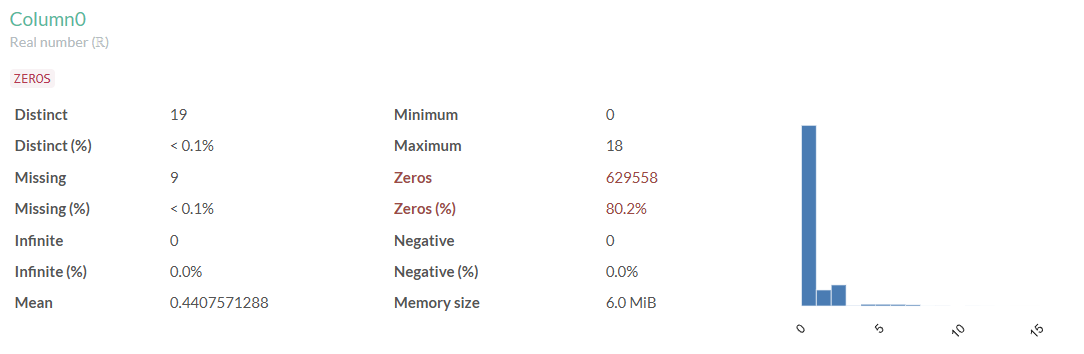
This is the unique identifier column. **It cannot be used as a predictor in the model.**

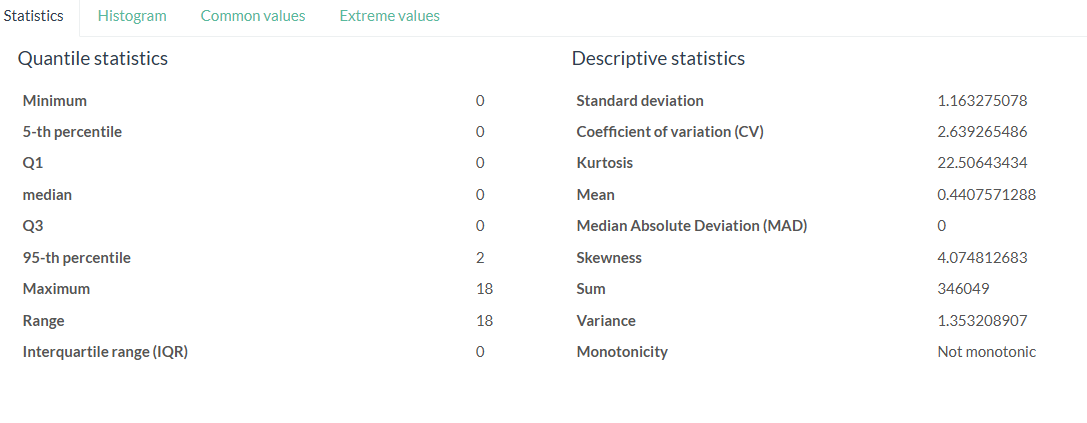




## **EDA-Column0**

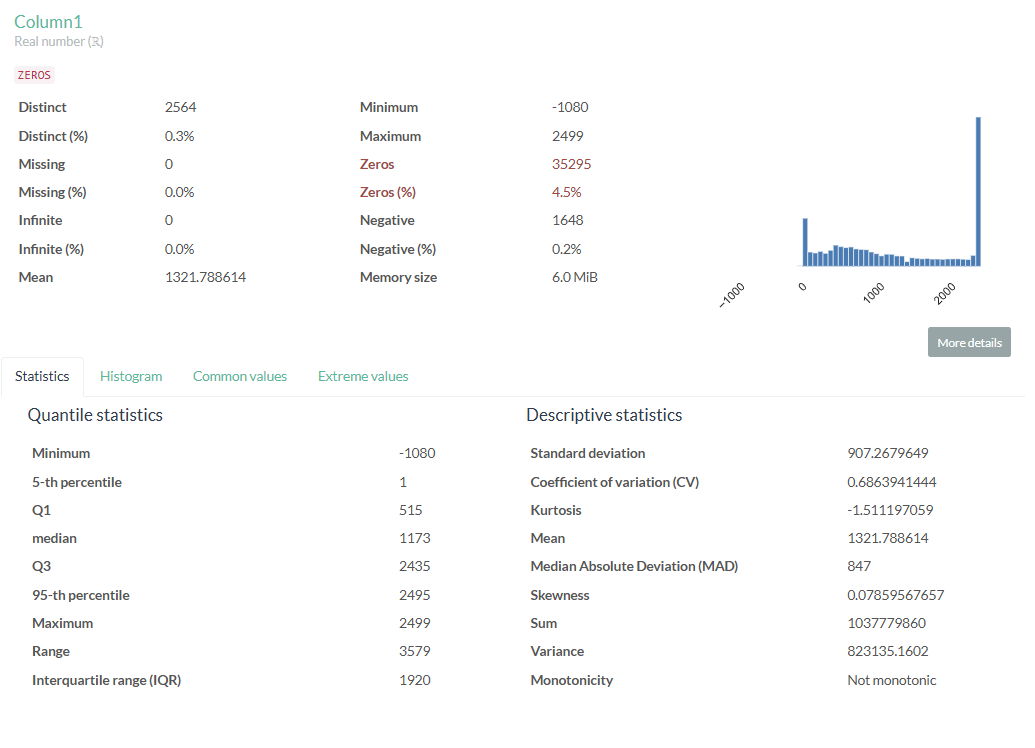
This column has **missing** values. The data consists of real numbers, but there are some missing values. To address this, the strategy is to replace the missing values with zero. This approach is based on the assumption that when no data is available, it can reasonably be interpreted as zero or null.





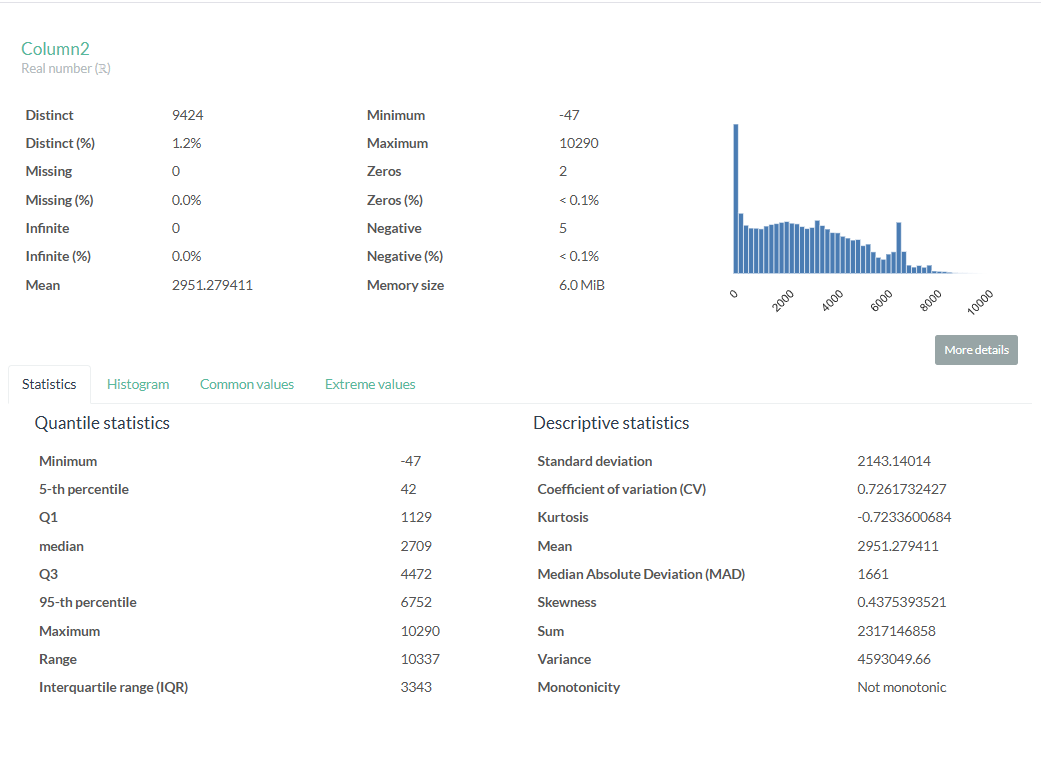
## **EDA-Column1**

No special handling required



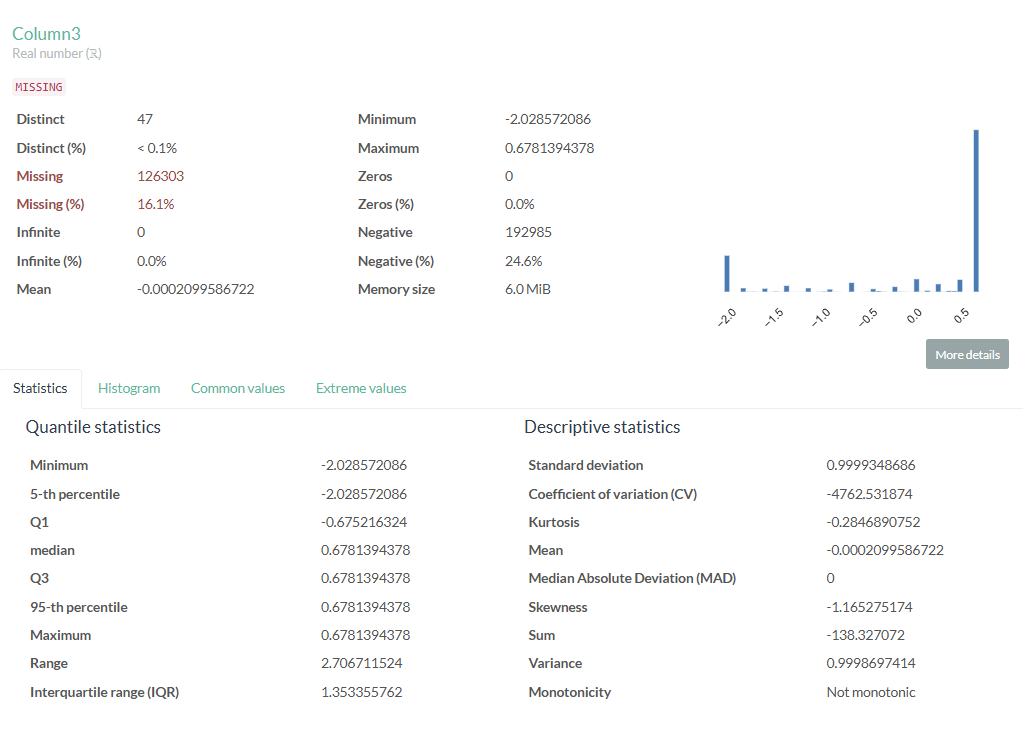
## **EDA-Column2**

No special handling required

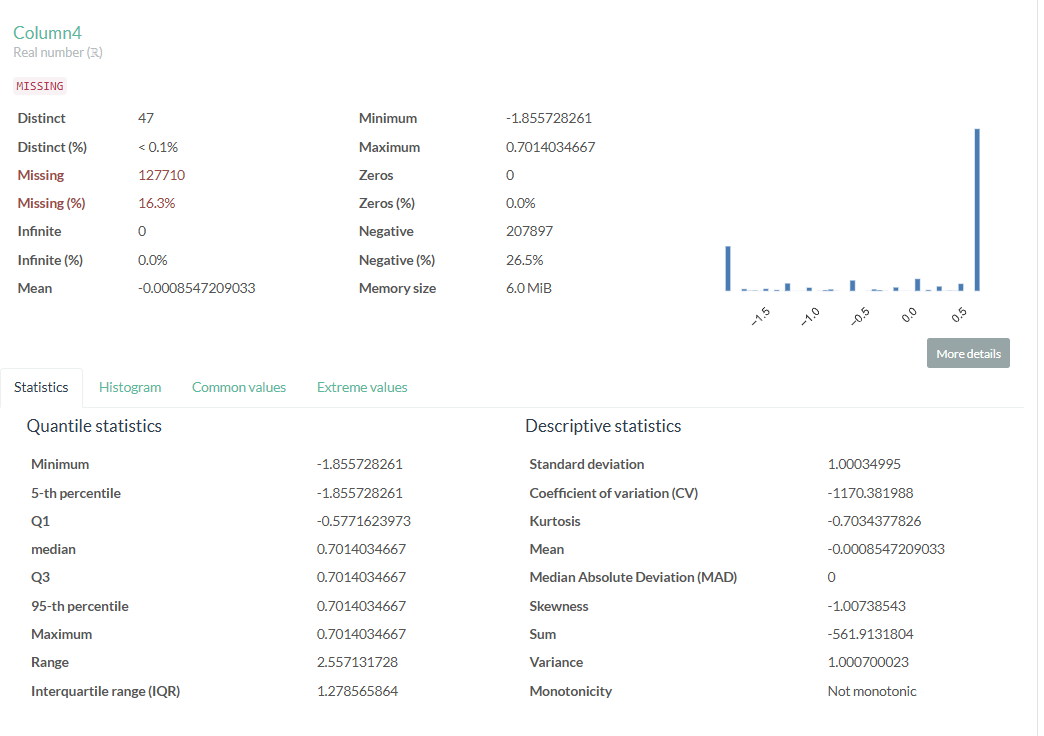


## **EDA-Column3**

This column has **missing** values. The column is densely corelated to Column4 as mentioned in Co-relation analysis section of this report .This column will be dropped from data modelling to avoid data redundancy.

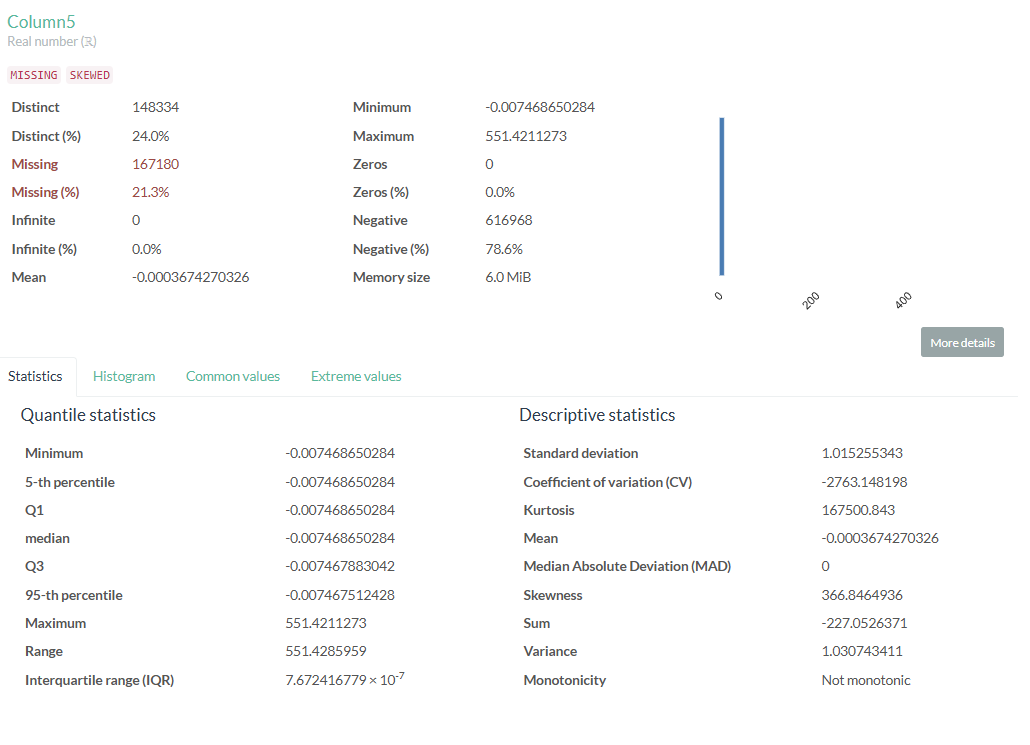


## **EDA-Column4**

This column contains missing values and, as noted, its values are highly correlated with those in column3. The strategy will be to replace the missing values with the median of the column

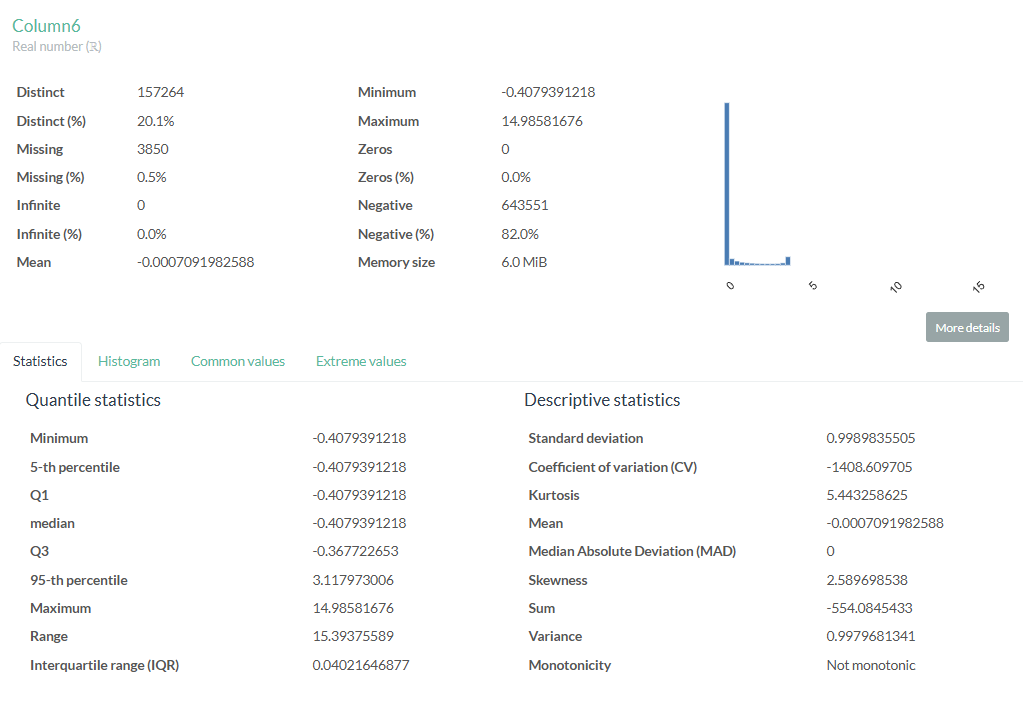
## **EDA-Column5**

This column contains missing values and exhibits skewness. Transforming skewed data is crucial for enhancing the performance of statistical models and ensuring that normality assumptions are satisfied. Data cleaning will be necessary for the model, especially since a significant amount of data is missing. Given that this column consists of real numbers, we can replace the missing values with zero. Additionally, since the data is skewed and includes negative values, the Yeo-Johnson transformation can be applied.



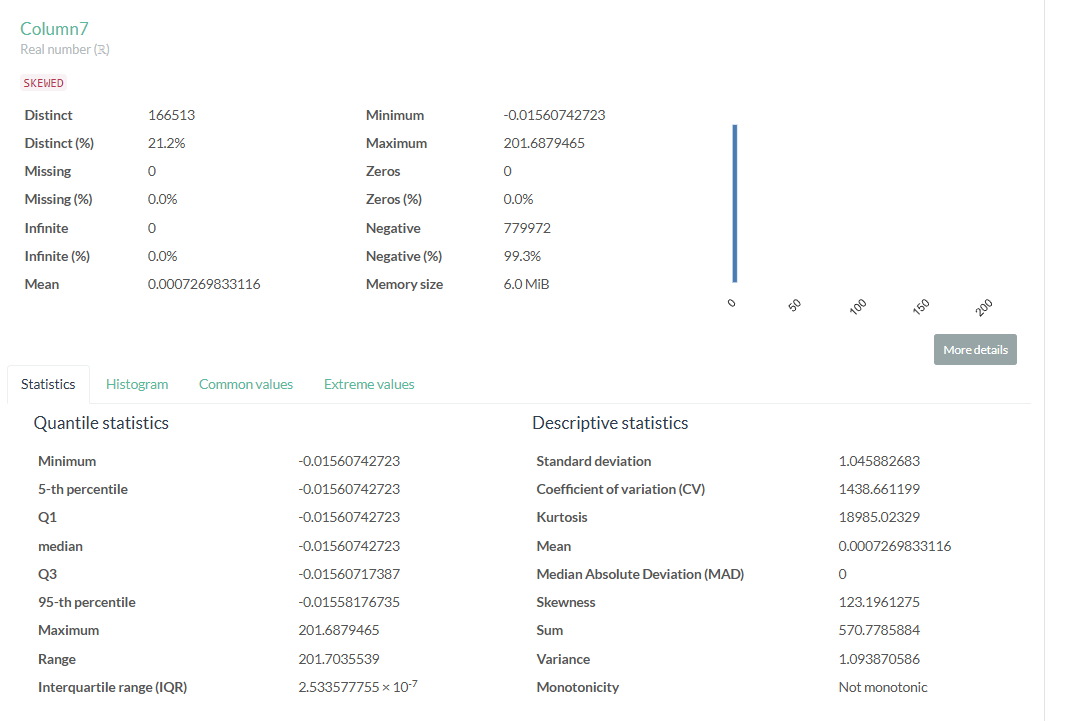
## **EDA-Column6**

This column has **missing** values. The data consists of real numbers, but there are some missing values. To address this, the strategy is to replace the missing values with zero. This approach is based on the assumption that when no data is available, it can reasonably be interpreted as zero or null.



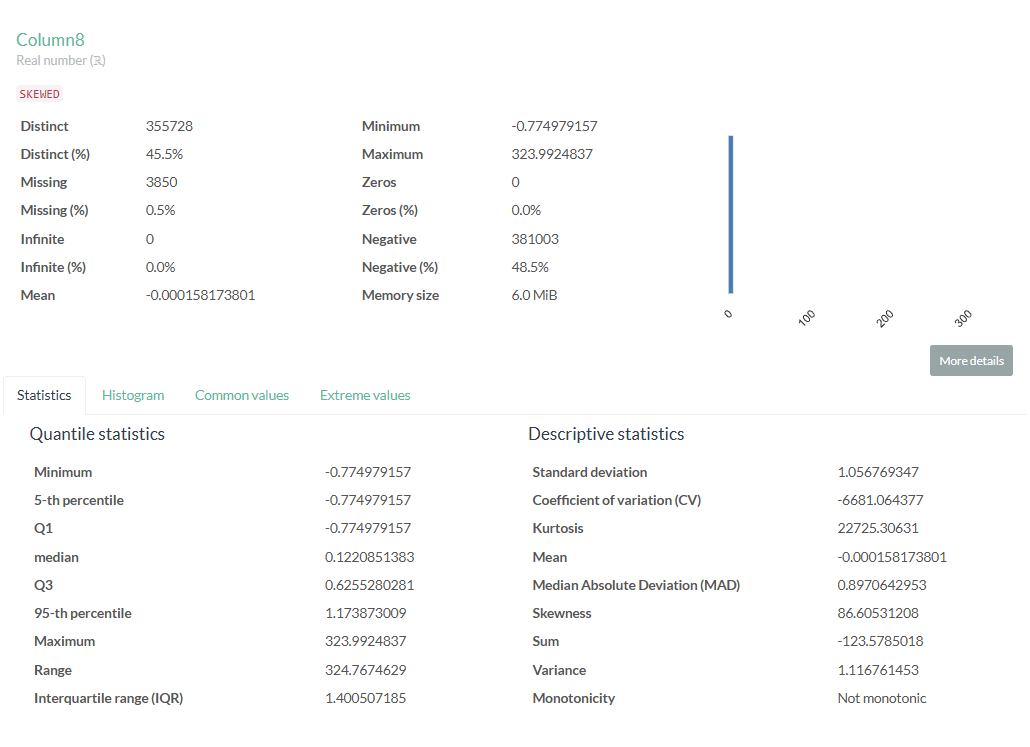
## **EDA-Column7**

This column exhibits skewness. Transforming skewed data is crucial for enhancing the performance of statistical models and ensuring that normality assumptions are satisfied. Data cleaning will be necessary for the model, especially since a significant amount of data is missing. Given that this column consists of real numbers, we can replace the missing values with zero. Additionally, since the data is skewed and includes negative values, the Yeo-Johnson transformation can be applied.



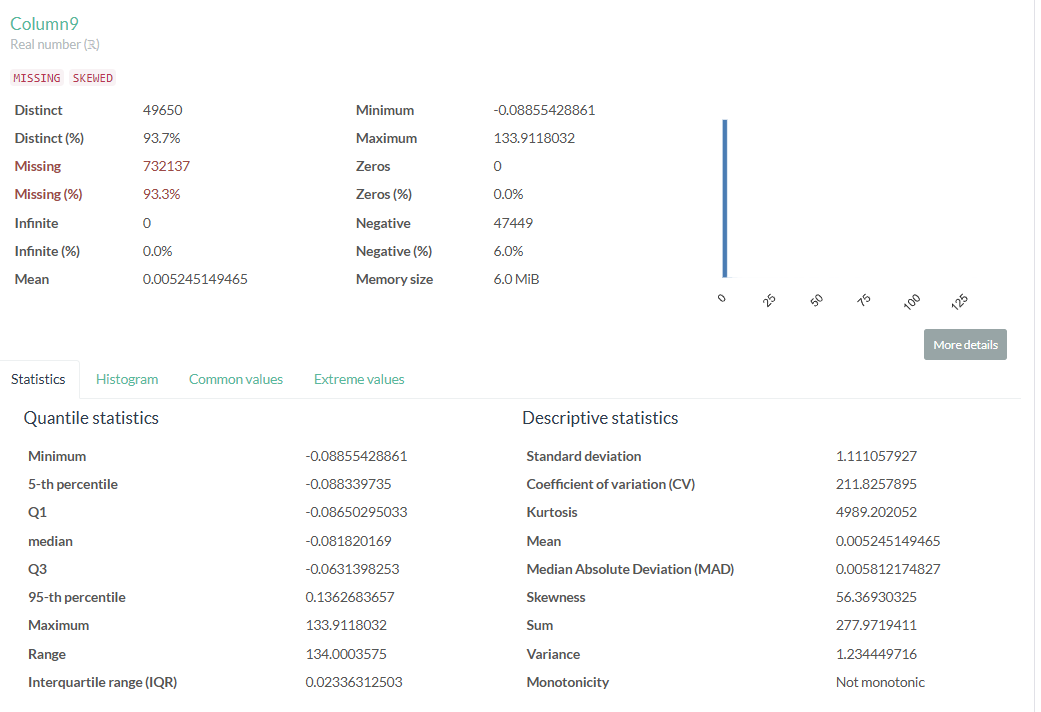
## **EDA-Column8**

This column contains missing values and exhibits skewness. Transforming skewed data is crucial for enhancing the performance of statistical models and ensuring that normality assumptions are satisfied. Data cleaning will be necessary for the model, especially since a significant amount of data is missing. Given that this column consists of real numbers, we can replace the missing values with zero. Additionally, since the data is skewed and includes negative values, the Yeo-Johnson transformation can be applied.



## **EDA-Column9**

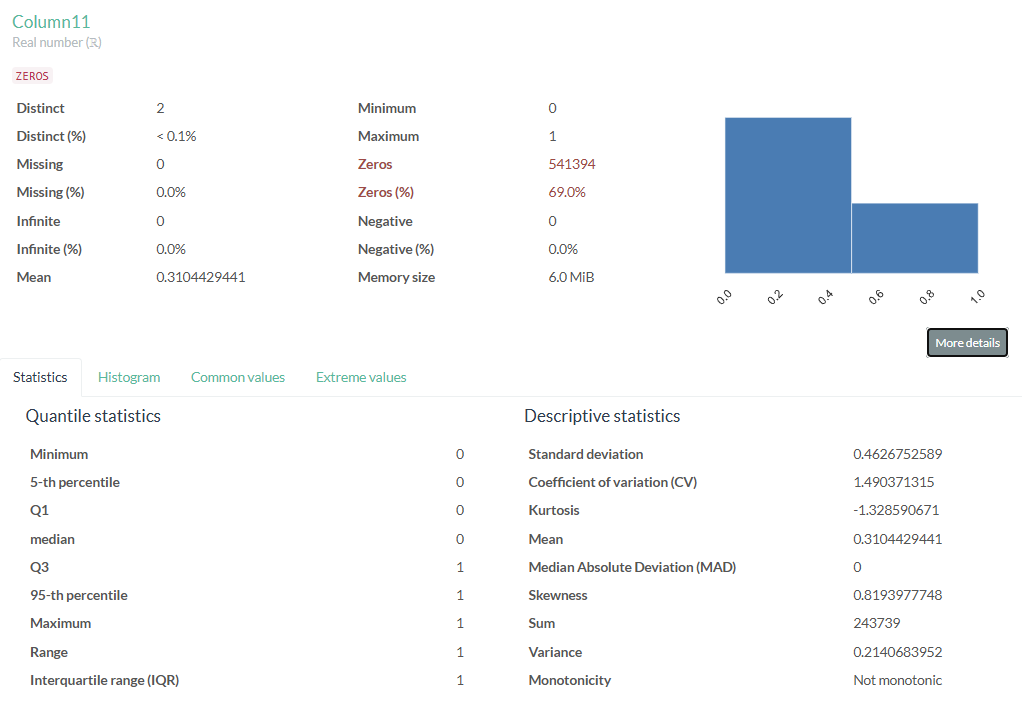
This column contains missing values and exhibits skewness. Transforming skewed data is crucial for enhancing the performance of statistical models and ensuring that normality assumptions are satisfied. Data cleaning will be necessary for the model, especially since a significant amount of data is missing. Given that this column consists of real numbers, we can replace the missing values with zero. Additionally, since the data is skewed and includes negative values, the Yeo-Johnson transformation can be applied.



## **EDA-Column10**

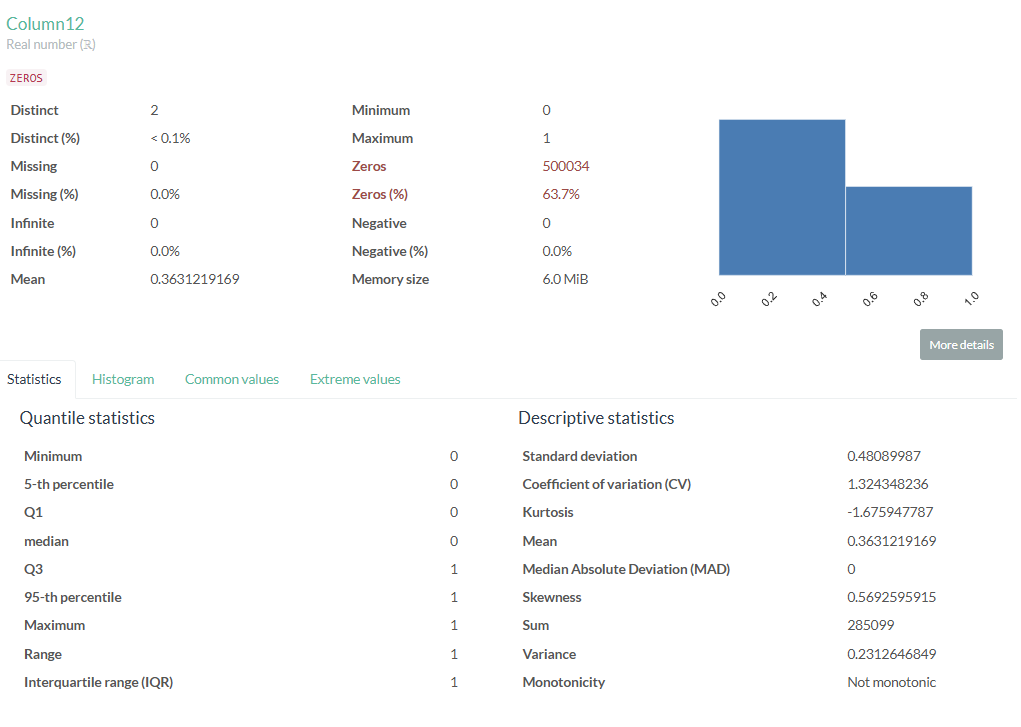
This column contains two distinct values, 0 and 1, which likely represent class labels or indicate Yes/No type data. No transformation is required, as most machine learning models can directly work with this column.

## **EDA-Column11**

This column contains two distinct values, 0 and 1, which likely represent class labels or indicate Yes/No type data. No transformation is required, as most machine learning models can directly work with this column.

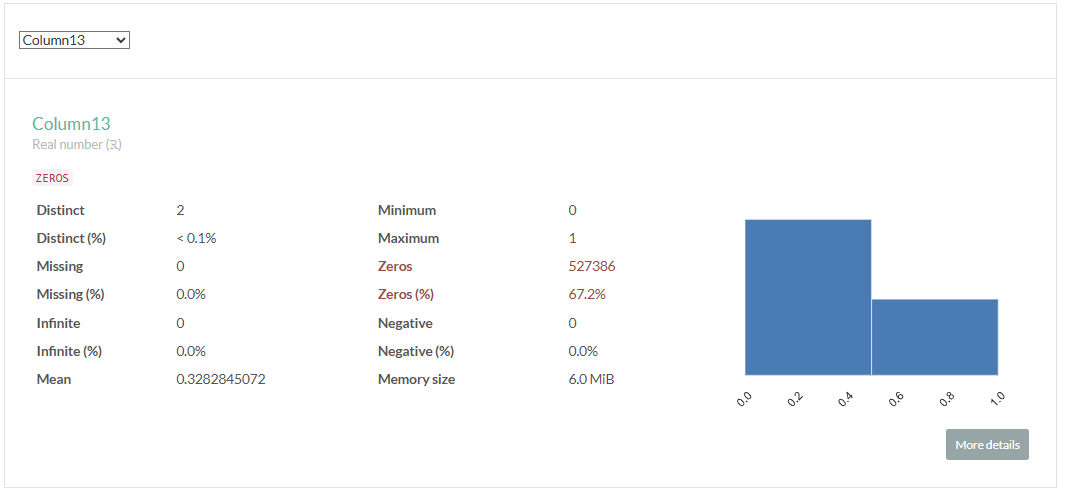
## **EDA-Column12**

This column contains two distinct values, 0 and 1, which likely represent class labels or indicate Yes/No type data. No transformation is required, as most machine learning models can directly work with this column.



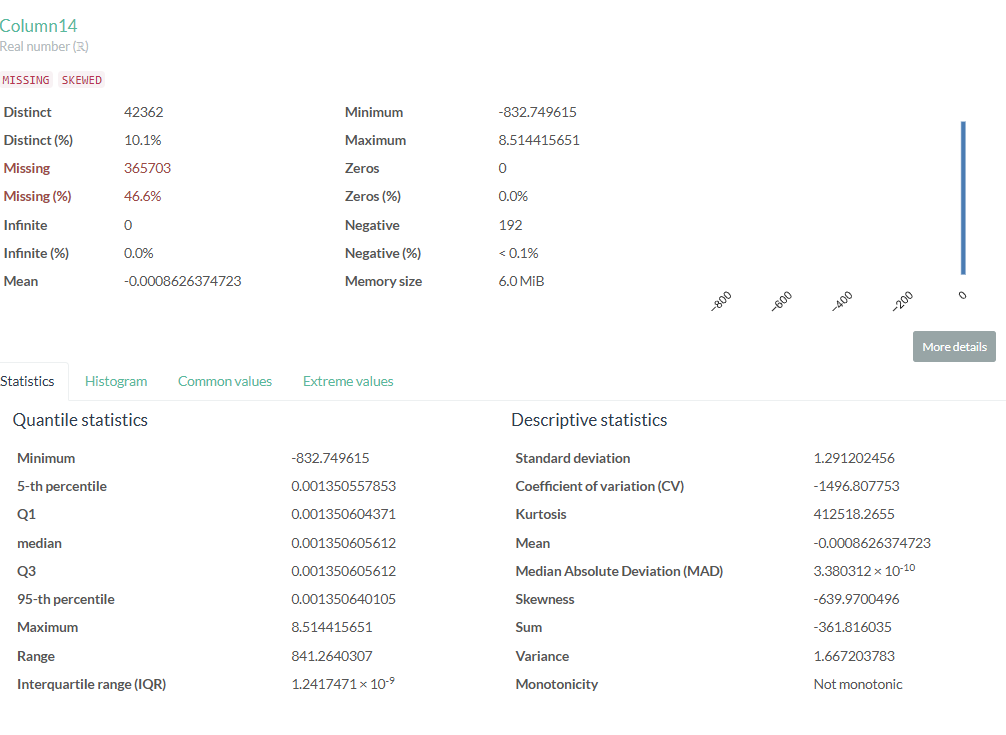
## **EDA-Column13**

This column contains two distinct values, 0 and 1, which likely represent class labels or indicate Yes/No type data. No transformation is required, as most machine learning models can directly work with this column. As mentioned in the Co-Relation analysis section this column is highly co-related to column10, column11 and column12. From data modelling those columns will be dropped.



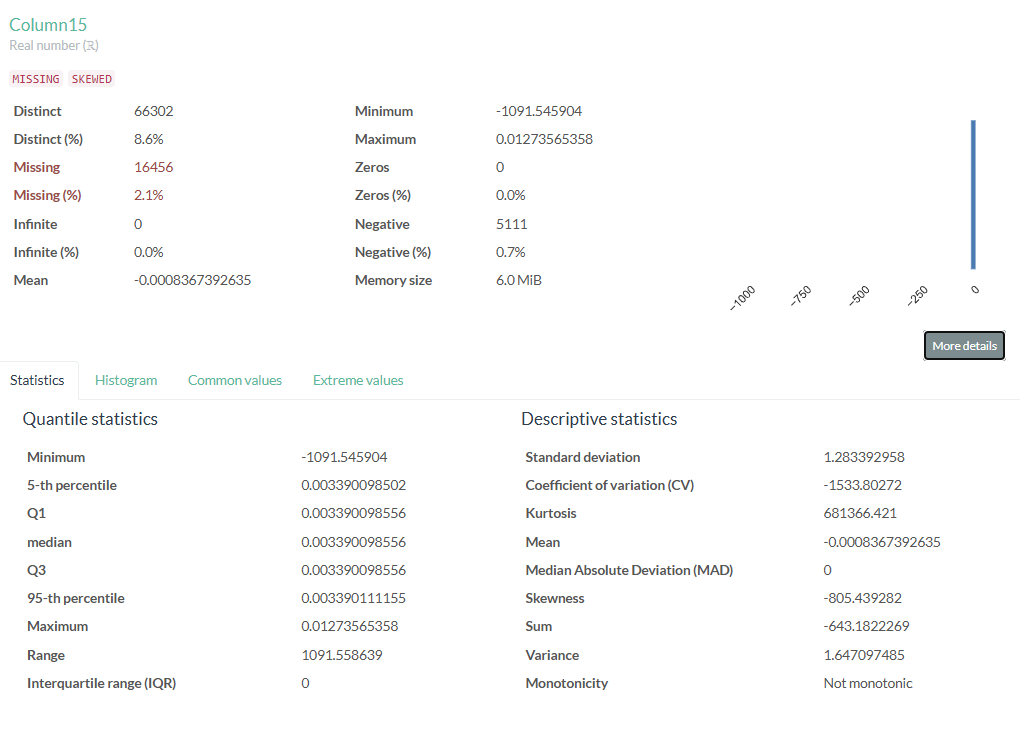
## **EDA-Column14**

This column contains missing values and exhibits skewness. Transforming skewed data is crucial for enhancing the performance of statistical models and ensuring that normality assumptions are satisfied. Data cleaning will be necessary for the model, especially since a significant amount of data is missing. Given that this column consists of real numbers, we can replace the missing values with zero. Additionally, since the data is skewed and includes negative values, the Yeo-Johnson transformation can be applied.



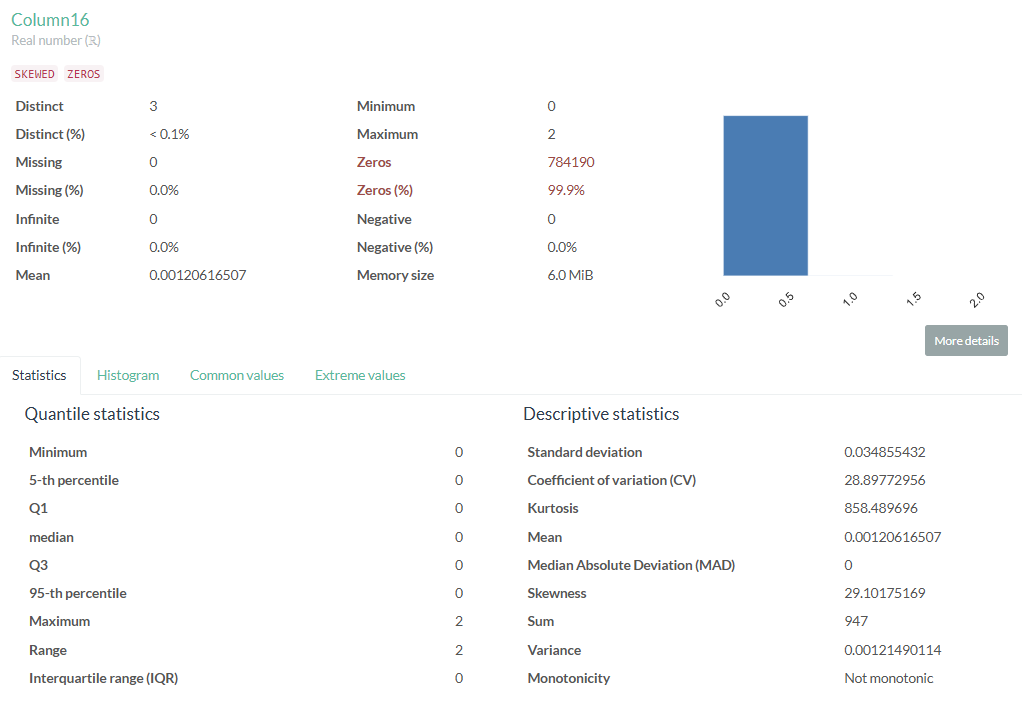
## **EDA-Column15**

This column contains missing values and exhibits skewness. Transforming skewed data is crucial for enhancing the performance of statistical models and ensuring that normality assumptions are satisfied. Data cleaning will be necessary for the model, especially since a significant amount of data is missing. Given that this column consists of real numbers, we can replace the missing values with zero. Additionally, since the data is skewed and includes negative values, the Yeo-Johnson transformation can be applied.



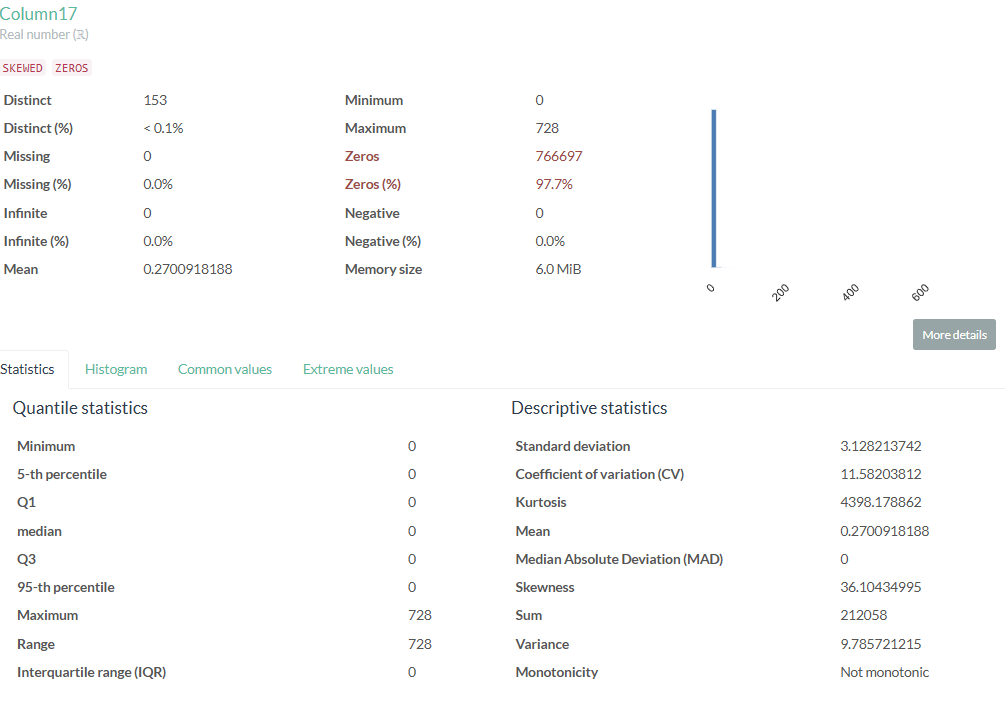
## **EDA-Column16**

This column data is **SKEWED**. However there are only 3 distinct values i.e 0,1,2. It seems data is already transformed for some class values. There is no need for further transformation



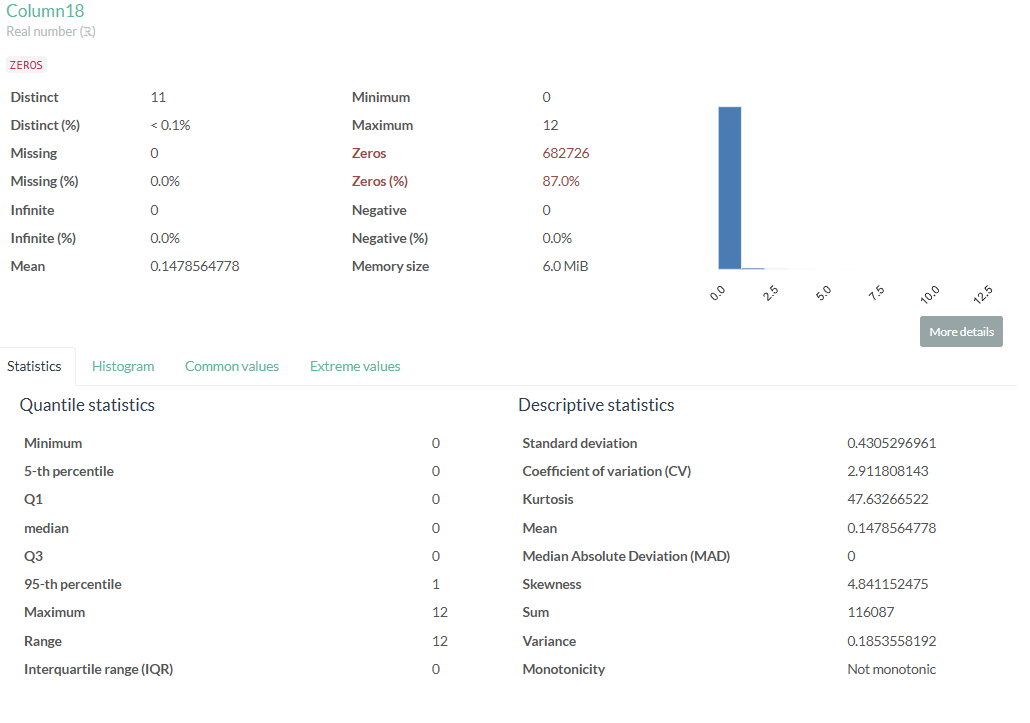
## **EDA-Column17**

This column exhibits skewness. Transforming skewed data is crucial for enhancing the performance of statistical models and ensuring that normality assumptions are satisfied. Data cleaning will be necessary for the model, especially since a significant amount of data is missing. Given that this column consists of real numbers, we can replace the missing values with zero. Additionally, since the data is skewed and includes negative values, the Yeo-Johnson transformation can be applied.



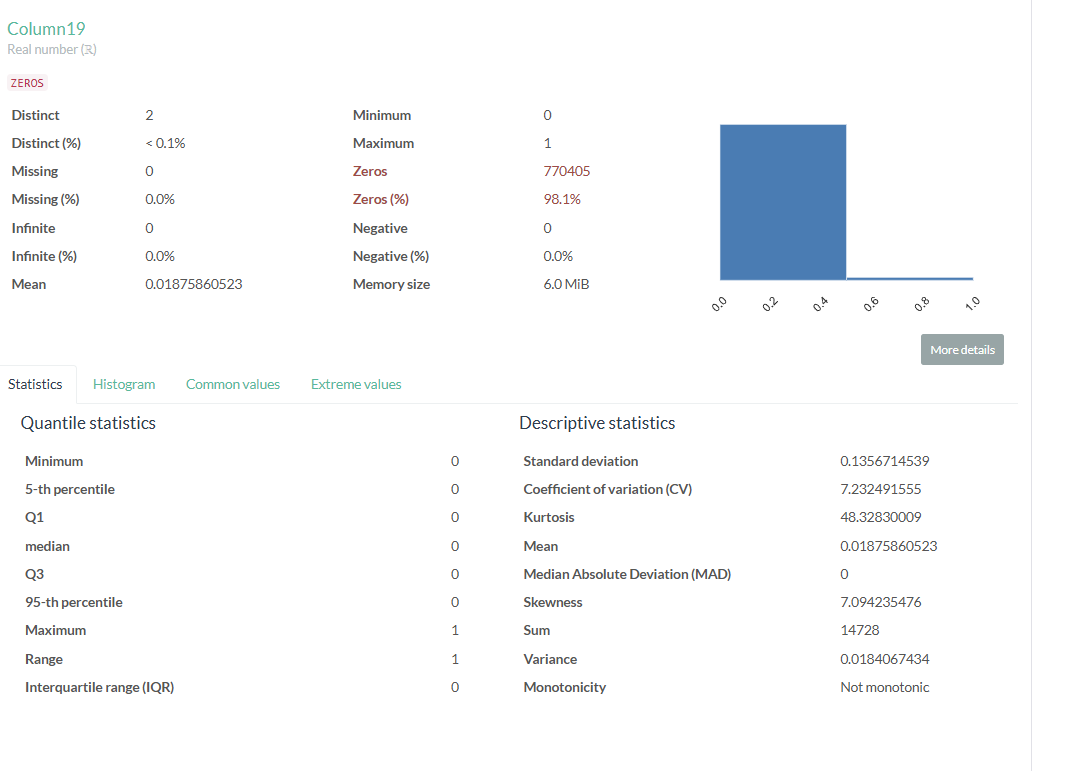
## **EDA-Column18**

The column predominantly contains zeros, indicating that the data is skewed. However, it includes only three distinct values ranging from 0 to 9, suggesting that the data has already been transformed for certain class values. No additional transformation is necessary.



## **EDA-Column19**

The column predominantly contains zeros, indicating that the data is skewed. However, it includes only three distinct values ranging from 0 to 1, suggesting that the data has already been transformed for certain class values. No additional transformation is necessary.



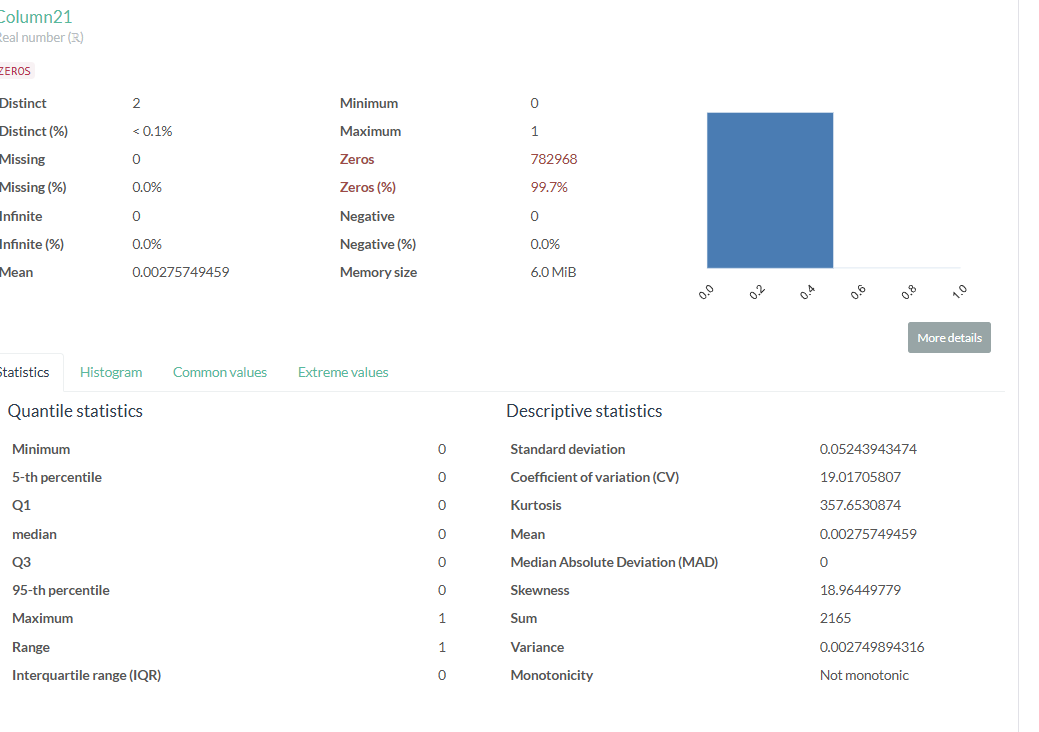
## **EDA-Column20**

The column predominantly contains zeros, indicating that the data is skewed. However, it includes only three distinct values ranging from 0 to 1, suggesting that the data has already been transformed for certain class values. No additional transformation is necessary.



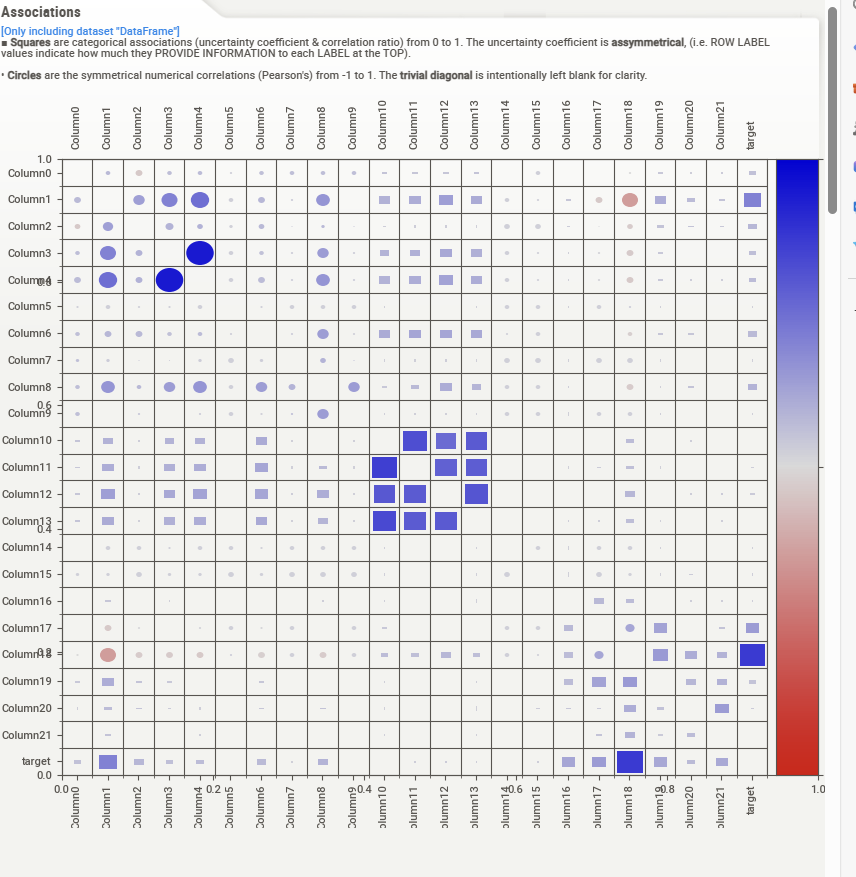
## **EDA-Column21**

The column predominantly contains zeros, indicating that the data is skewed. However, it includes only three distinct values ranging from 0 to 1, suggesting that the data has already been transformed for certain class values. No additional transformation is necessary.



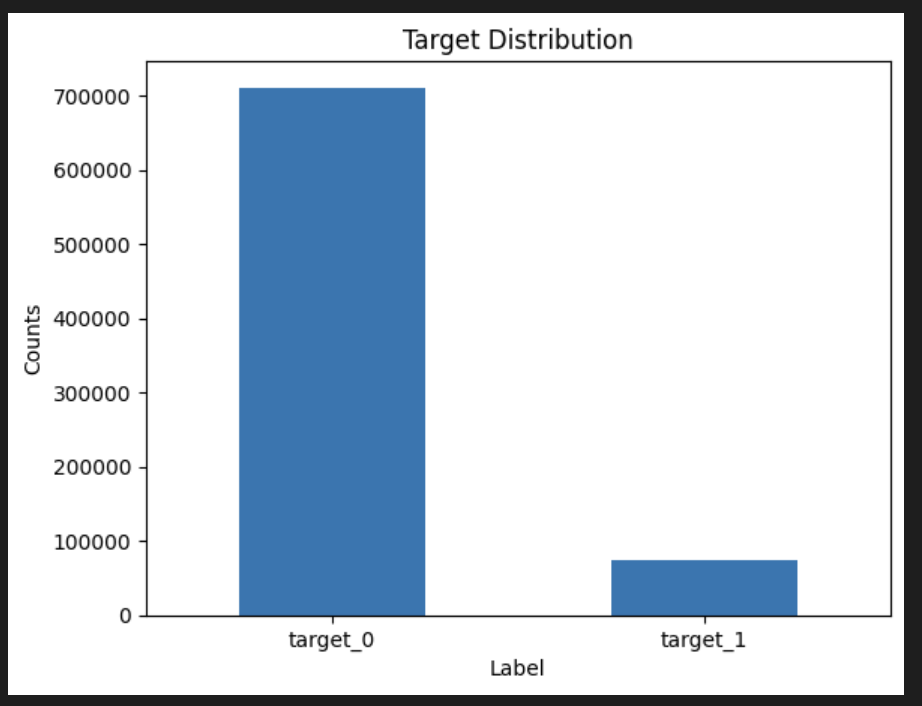
## **Co-Relation analysis**

Certain columns exhibit a high degree of correlation with one another and should be removed from the modeling process to reduce data redundancy. For instance, column3 and column4 are significantly correlated. Additionally, columns 10 through 13 also show a strong correlation.



## **Target Class distribution**

As evident from the graph below dataset is highly imbalanced for target variable class\_0



# **Data Modelling – Logistic Regression**

Since the target variable has binary data , Logistic regression is good algorithm for prediction.

All the independent variables will be used to predict the value of target variable

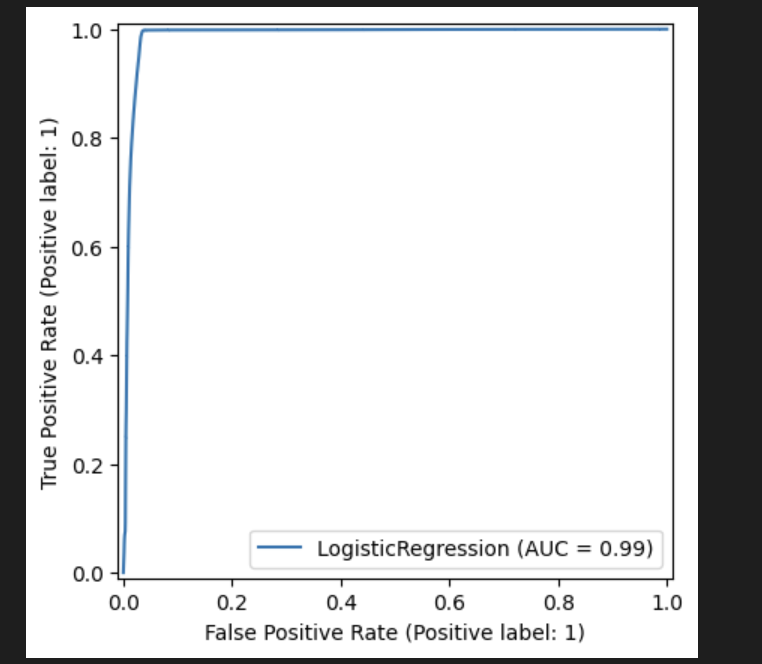
## **Model Performance Report**

### **Model performance on training data**

**Auc Score: 92.892%**

**Eval Accuracy: 96.890%**

**ROC Curve**

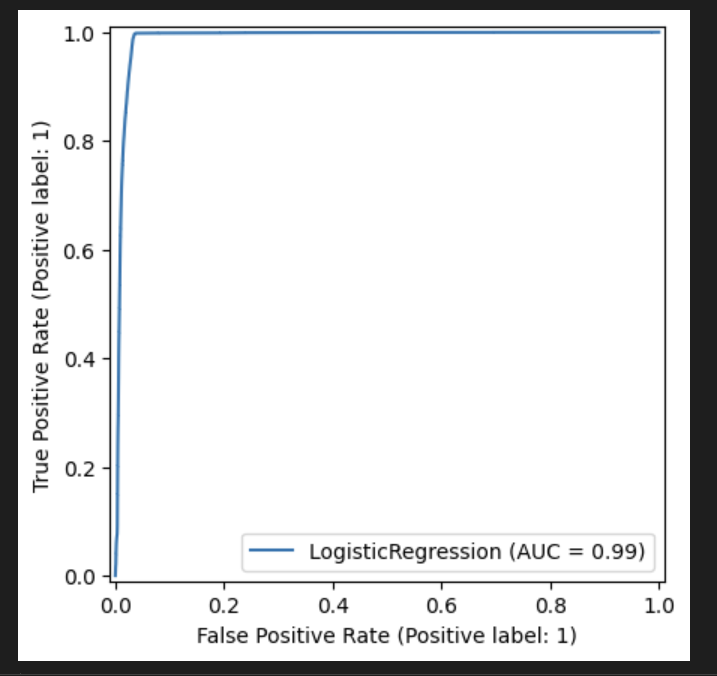


### **Model performance on Test data**

**Auc Score: 93.334%**

**Eval Accuracy: 96.970%**

**ROC Curve**

****

# **Results Summary**

**Logistic Regression Model Summary**

* **Model Type**: Logistic Regression
* **AUC (Area Under ROC Curve)**: 93.334%
  + The AUC score represents the ability of the model to distinguish between classes. A score of 93.334% indicates excellent model performance, as values closer to 100% show better classification.
* **Evaluation Accuracy**: 96.970%
  + This accuracy indicates that 96.970% of the predictions made by the model were correct. It's a strong indicator of the model's overall performance on the evaluation dataset.

**Key Insights:**

* The high **AUC score** suggests that the model effectively distinguishes between the positive and negative classes.
* The **Evaluation Accuracy** reflects the overall proportion of correctly predicted labels, with nearly 97% of predictions being correct.

**Conclusion:**

The Logistic Regression model shows strong performance with a high AUC score and evaluation accuracy, indicating that it is well-suited for the current dataset.