

School of Computing, Creative Technology and Engineering

|  |  |
| --- | --- |
| Student ID | C7356850 |
| Student Name | Suyog Sitaula |
| Module Name & CRN | Applied Machine Learning |
| Level | 5 |
| Assessment Name & Part No. | Assessment 2 part 1 |
| Project Title | Case Study 2- Fraudulent Companies |
| Data of Submission | 29th March 2024 |
| Course | Computing |
| Academic Year | 2024 |

Contents

[Abstract 3](#_Toc162821057)

[Introduction 3](#_Toc162821058)

[Literature Review 3](#_Toc162821059)

[Exploratory Data Analysis 4](#_Toc162821060)

[Data Pre-processing 8](#_Toc162821061)

[Missingness 8](#_Toc162821062)

[Outliers 13](#_Toc162821063)

[Multicollinearity 17](#_Toc162821064)

[Scaling 18](#_Toc162821065)

[Bibliography 20](#_Toc162821066)

# Abstract

This report addresses the ways to work around the variables of given dataset and gives us the insight to make some decisions regarding multiple cases such as missingness, outliers, etc. In this report we have used R language as our primary language to interact with dataset. The dataset consisted total of 27 columns and 775 rows. In this report we have used multiple library such as mice and ggplot2. This report gives us the visual representation of the data and the reason of taking various decisions.

# Introduction

Fraud is considered as a major issue worldwide. Companies that take part in this activity affects negatively in world economy. Fraudulent companies’ classification helps the auditors by identifying the fraudulent companies from the given dataset. As the evolution of technology, the ways of fraud have also evolved we cannot always rely on traditional rule-based system. In real world this is important and used in various sector such as bank, ecommerce platforms etc to detect fraudulent activities and many other crimes related to fraud. The models are trained to identify the fraudulent activities from the given set of data that filters them with the genuine companies.

# Literature Review

This review aims to present a systematic literature review (SLR) that systematically reviews and synthesizes the existing literature on machine learning (ML)-based fraud detection (Abdulalem Ali 1, 2022). Its purpose is to find the effective fraud detection technique in the case of financial sector. There are some methods that are being used like manual verification but they are high cost and time consuming.

This case study aims to explore the usefulness of machine learning in improving the quality of an audit work (Nishta Hooda, 2018). To tackle this problem, they collect the annual data from 777 different firm from 14 different sectors. The Particle Swarm Optimization algorithm is used as selection method. In the overall case study, it highlights the importance of machine learning in audit work.

In this article, various machine learning and deep learning approaches are used for detecting frauds in credit cards (Abolfazl Mehbodniya, 2021). This article focuses on the detection of credit card fraud and in healthcare sector where such fraud can cause huge financial damages.

This paper aims to understand how deep learning (DL) models can be benevolent in detecting fraudulent transactions with high accuracy (Aji Mubarek Mubalaike, 2018). In this paper we can see both traditional machine learning algorithms and deep learning algorithms being used. The traditional machine learning includes decision trees (EDT) and deep learning includes Stacked Auto Encoders and Restricted Boltzmann Machines.

The purpose of this journal is to evaluate the possibility of rating the credit worthiness of a firm’s quarterly financial report using a dynamic anomaly detection method (Mark Lokanan, 2019). It uses the 937 Vietnamese listed firms’ data as its dataset and among the 24 are chosen as control variable.

It uses Mahalanobis distance to identify the values that has unusual behaviour compared to the other value of the dataset.

This journal proposes an approach for detecting statement fraud through the combination of information from financial ratios and managerial comments within corporate annual reports(Patricia Craja, 2020). It uses the Management Discussion and Analysis (MD&A) section of annual reports. The relevant values are extracted using Hierarchical Attention Network (HNA).

This paper aims to establish a rigorous and effective model to detect enterprises’ financial statements fraud for the sustainable development of enterprises and financial markets (Jan, 2018). It takes 160 companies listed on either Taiwan Stock Exchange or Taipei Exchange as sample where 40 of them are fraud. It uses artificial neural network and support vector machine to know the most important variable from given dataset and uses 4 different decision trees for classification.

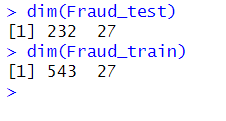
This review presents a comprehensive review of the literature on the application of data mining techniques for the detection of financial accounting fraud and proposes a framework for data mining techniques-based accounting fraud detection (Anuj Sharma, 2013). It addresses the importance of Financial Accounting Fraud Detection (FAFD). The review proposes a framework for FAFD that is based on datamining techniques such as logistic model, neural model, decision trees and Bayesian belief networks. This review gives us the insight of most commonly used methods in this particular field.

This paper gives a comprehensive revision of the state-of-the-art research in detecting financial fraud from 2009 to 2019 inclusive and classifying them based on their types of fraud and data mining technologies utilized in detecting financial fraud (Khaled Gubran Al-Hashedi, 2021). The review has shown that total of 34 different data mining techniques have been used to detect fraud on many financial sectors. Support Vector Machine is one of the main methods used to do so.

This paper investigates the performance of naïve bayes, k-nearest neighbor and logistic regression on highly skewed credit card fraud data (John O. Awoyemi, 2017). The study focuses on the performance of mentioned three classification techniques on dataset of 284807 transactions. The results of those 3 classification technique for optimal accuracy was Bayes=97.92% , KNN=97.69% and logistic regression=54.86%.

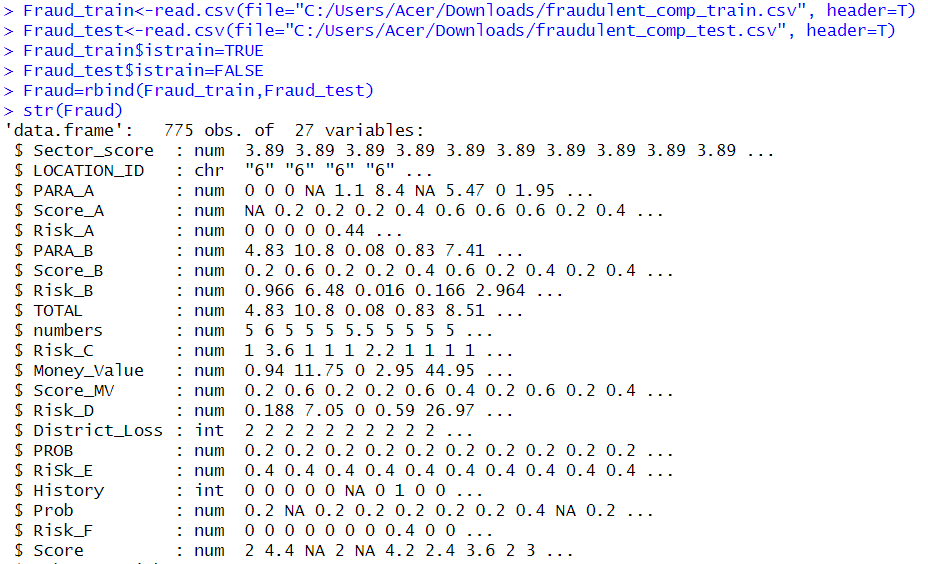
# Exploratory Data Analysis

At first let’s look at the dimension of the two data set.



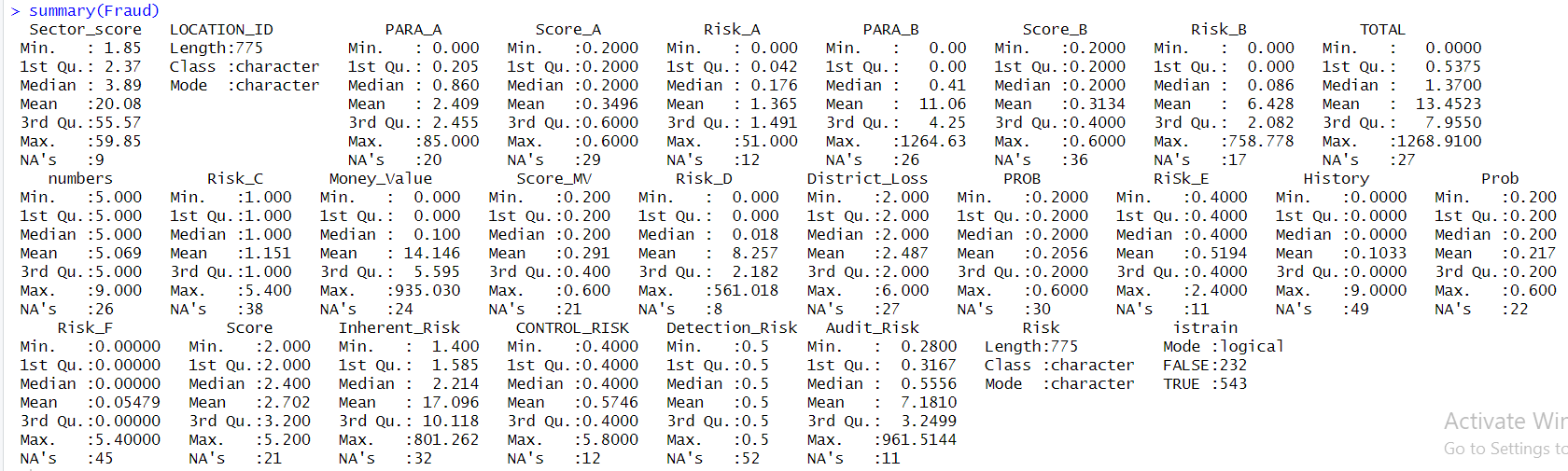
It shows the column and row of the two datasets. Test dataset have 232 rows and 27 columns and train dataset have 543 rows and 27 columns.

The two data sets that are provided to us are training set and testing set. It is preferred to combine two data set and make it one when doing exploratory data analysis. The R code to combine two data set is:

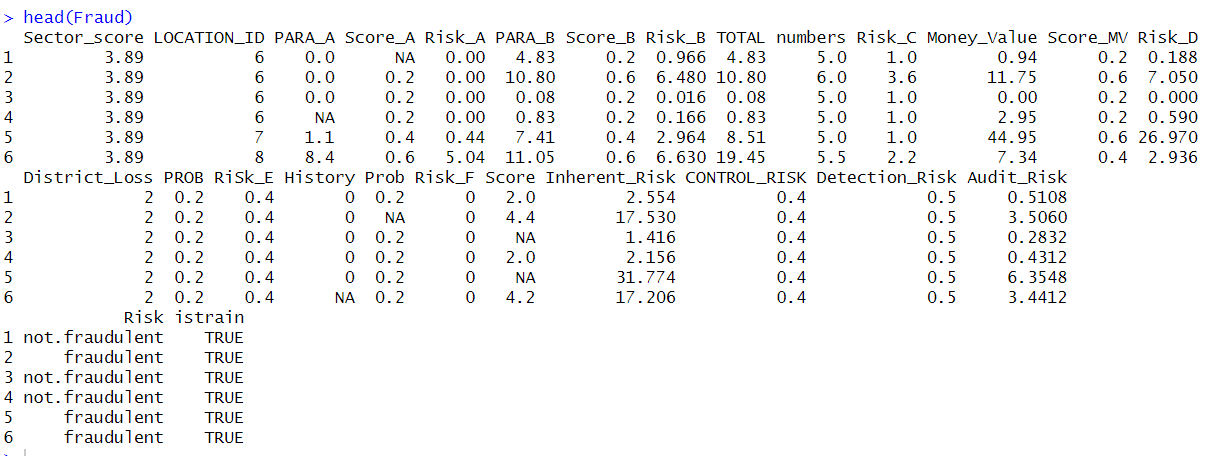


Str() shows the structure of the given dataset.

Lets look at the summary of dataset after it is combined.

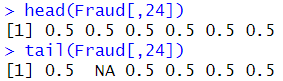


Let’s look at the first 6 value of every column.



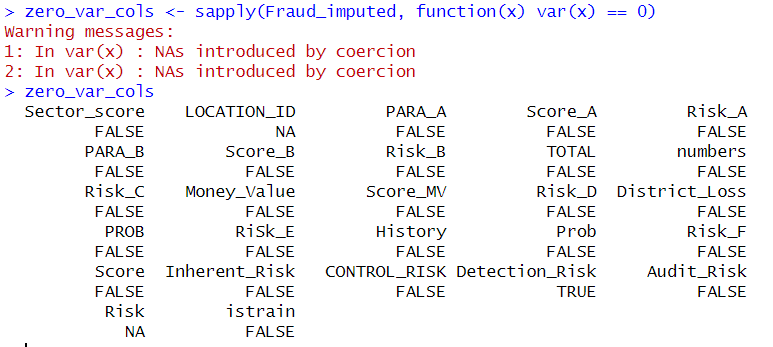
We can see that there are some missing values in some of the column. The column Risk says its status after all the evaluation and istrain column is added to help separate training and testing dataset if needed.

Let’s look at the Detection\_Risk column’s first 6 and last 6 values:



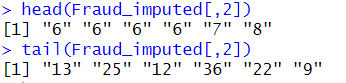
We can see first 6 and last 6 values of a Particular column.

Let’s check the zero variance variable.



Here var() checks the variance of individual column and then checks if the variance is zero. Here Detection\_Risk is considered as zero variance variable. Here LOCATION\_ID shows NA because due to its value type(char) we cannot calculate its variance.

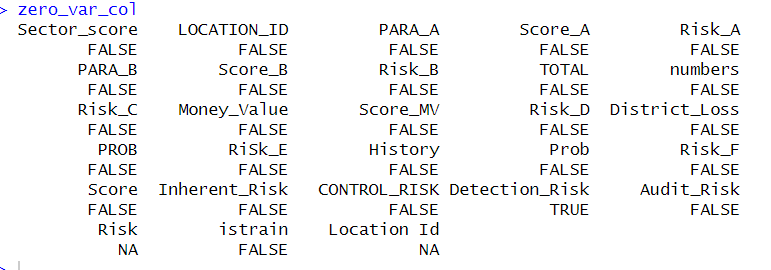
As all the value of LOCATION\_ID is convertible to numeric



Let’s convert it and check whether it is zero variance variable or not. While doing so we can get a warning



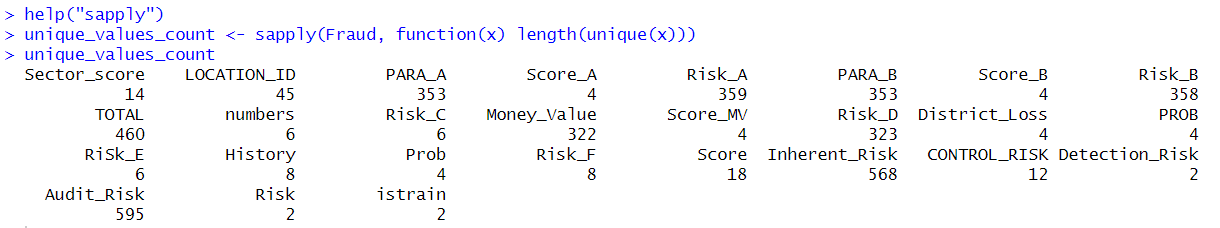
So, we need to impute the missing data which we will do in later. So the result after imputing missing values and checking zero variance variable is



Here we can see LOCATION\_ID is not zero variance variable.

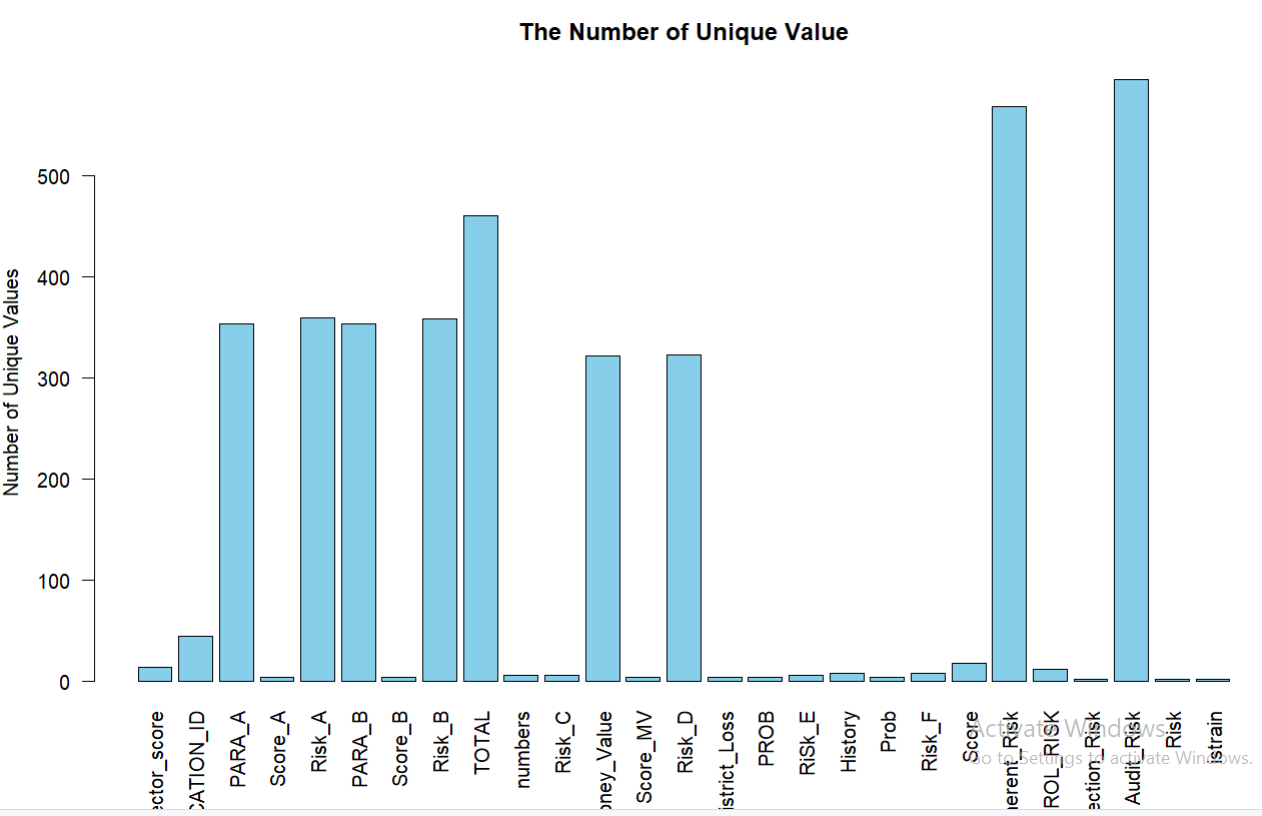
(note: zero variance variable is calculated after imputing the missing values)

We can also see the unique value present in each column.

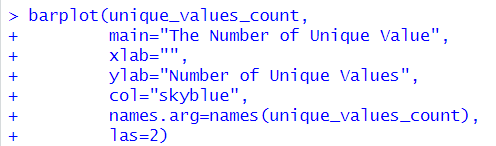


This shows the number of unique values present in each column. Here we use sapply to take dataframe as a parameter and return the number and column name in matrix form.

Lets compare the number of unique value using bar plot



The code for this plot is

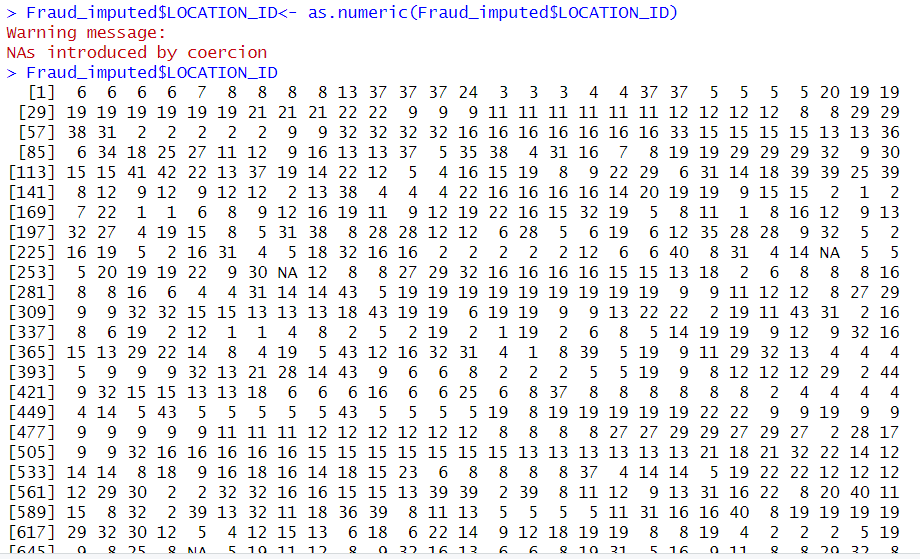


# Data Pre-processing

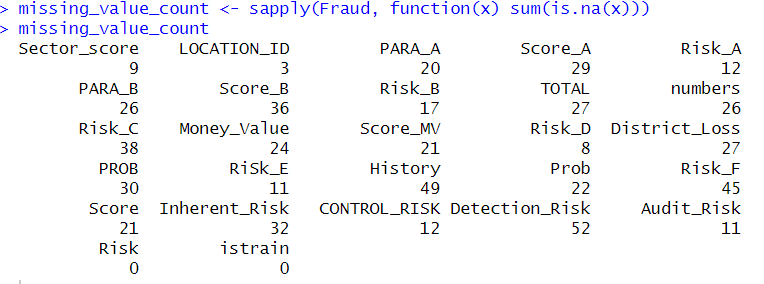
## Missingness

Occurrence of missing value is every problematic thing for a dataset. So here we will see the total amount of missing value per column in combined dataset.

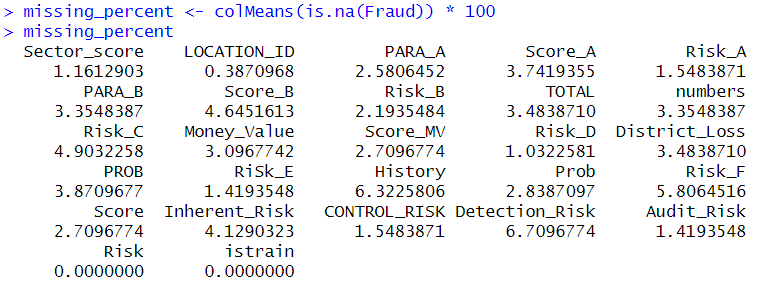
Before that we need to covert LOCATION\_ID to numeric data



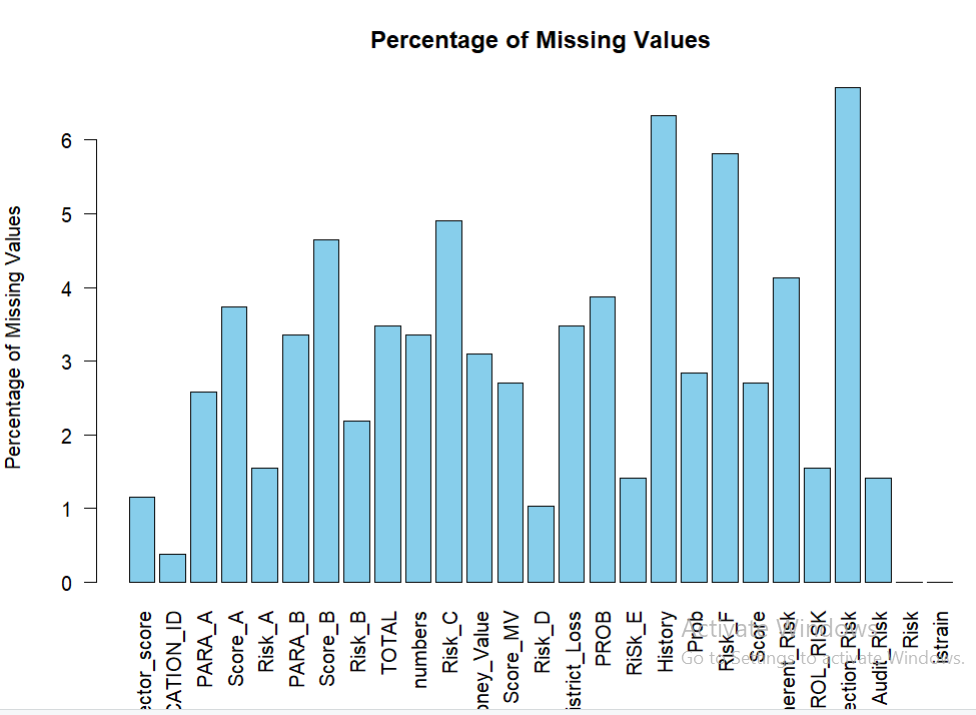
The warning is because there were some value that we were not able to convert into numeric value.



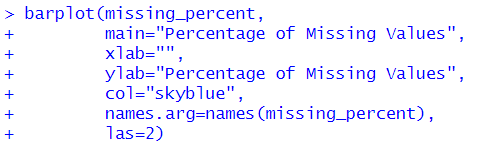
We need to convert it in percentage so that we can make the decision to whether to delete or impute the missing values. So the missing value percentage of each column is:



Lets create a bar plot to visualize the data.



This graph shows the column with its respective missing value percentage. The code for this bar plot is:

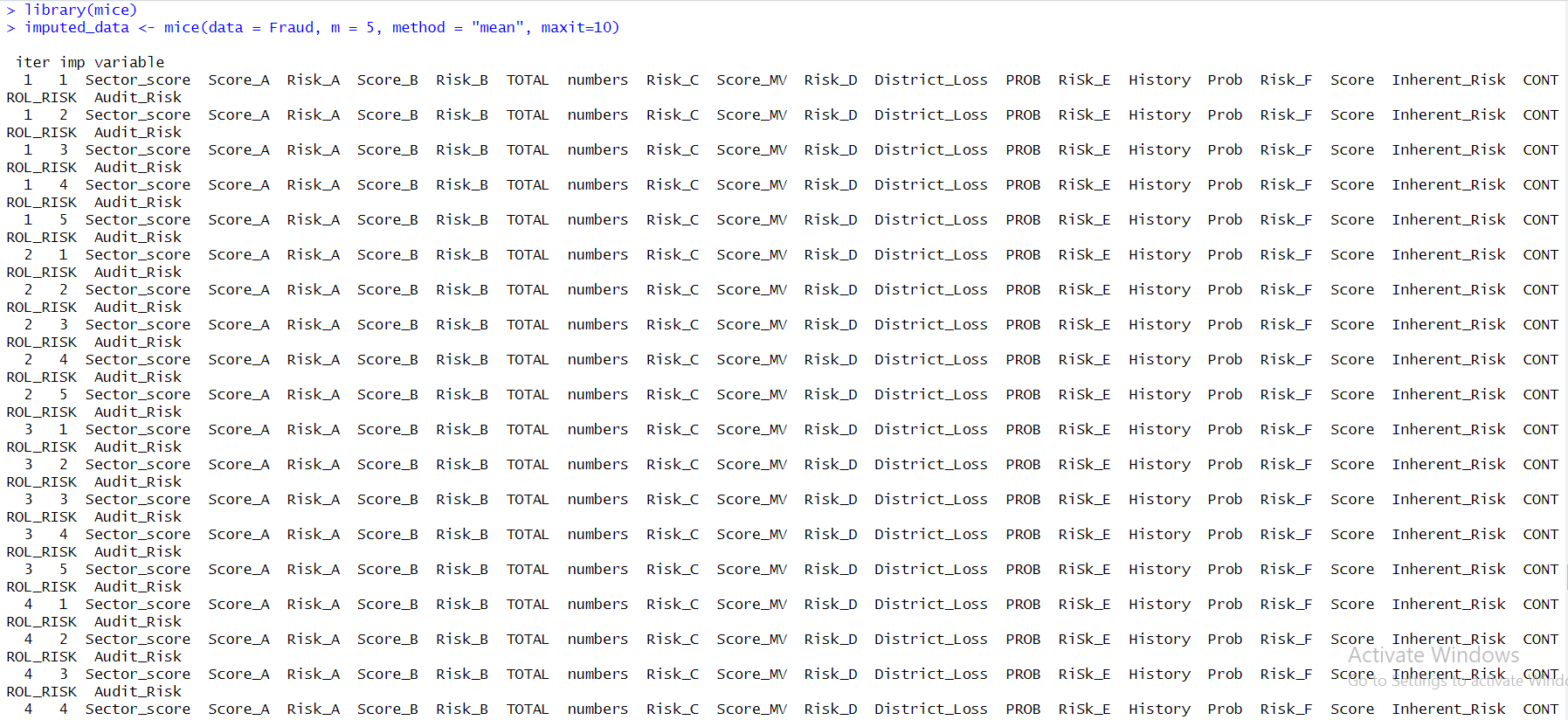


Lets see the percentage of missing value in overall dataset:

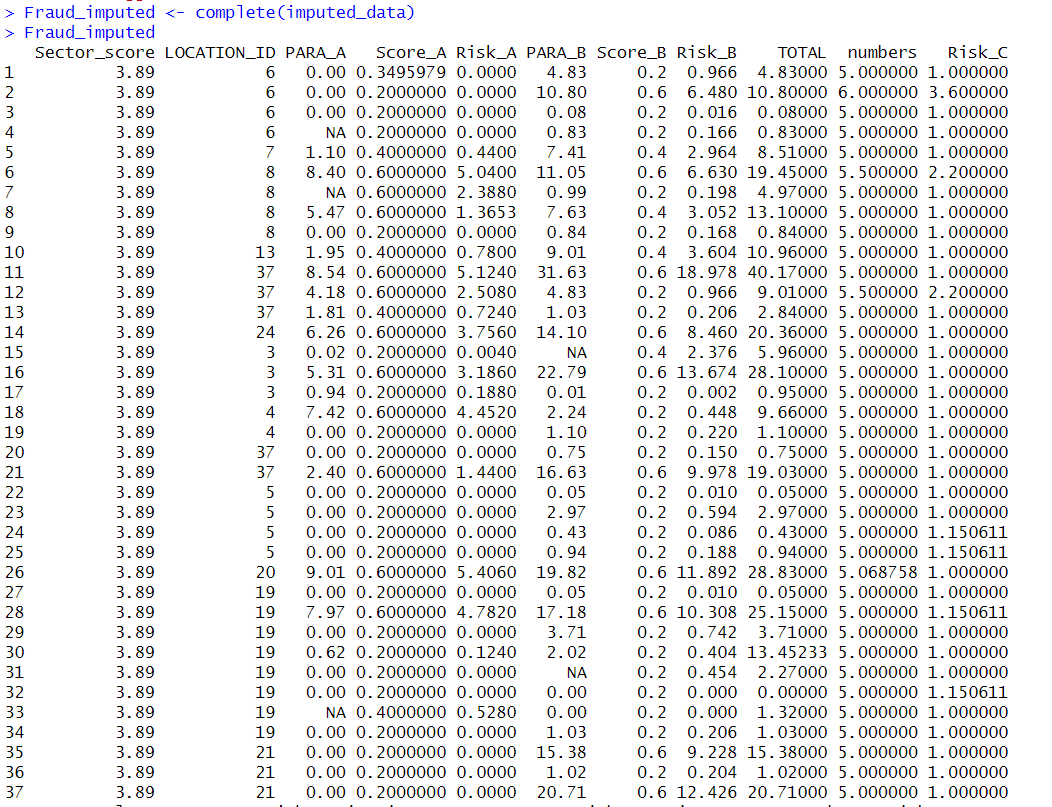


Here unlist is used to convert the dataset into a single vector and is necessary because is.na works on vector and cannot handle dataset directly.

Now its time to impute the missing data as the missing percentage is more than 1%.



Here we are using mice to impute the missing data.

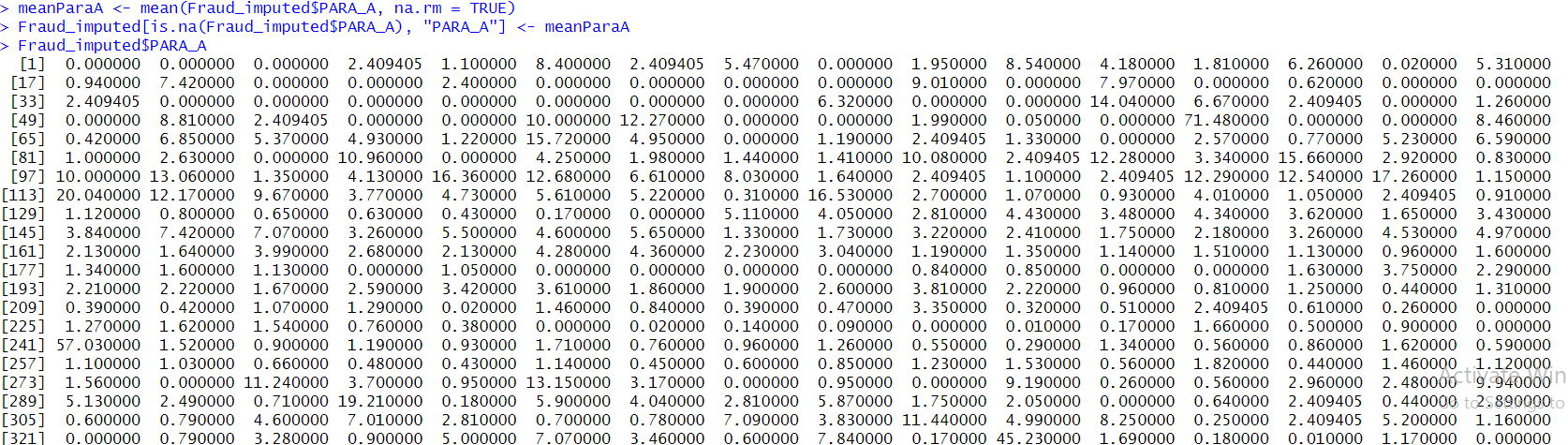


This is the dataset after removing missing values. But some columns still have the missing values. That is because they are collinear columns and constant columns.

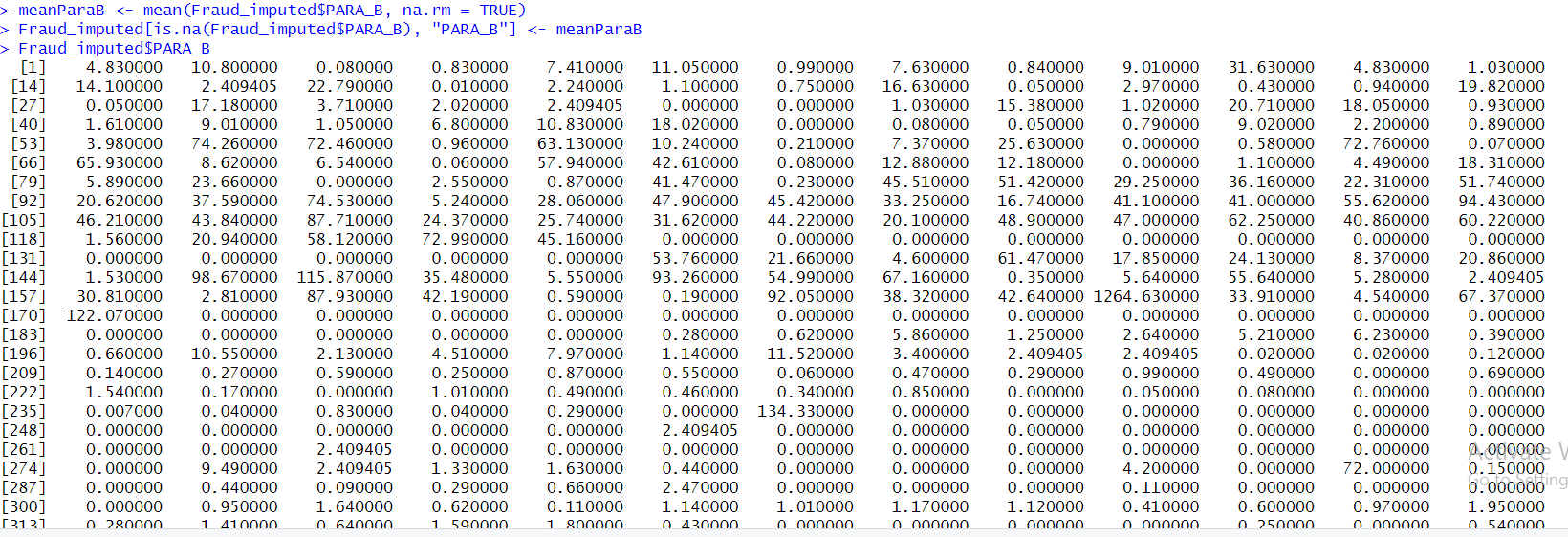




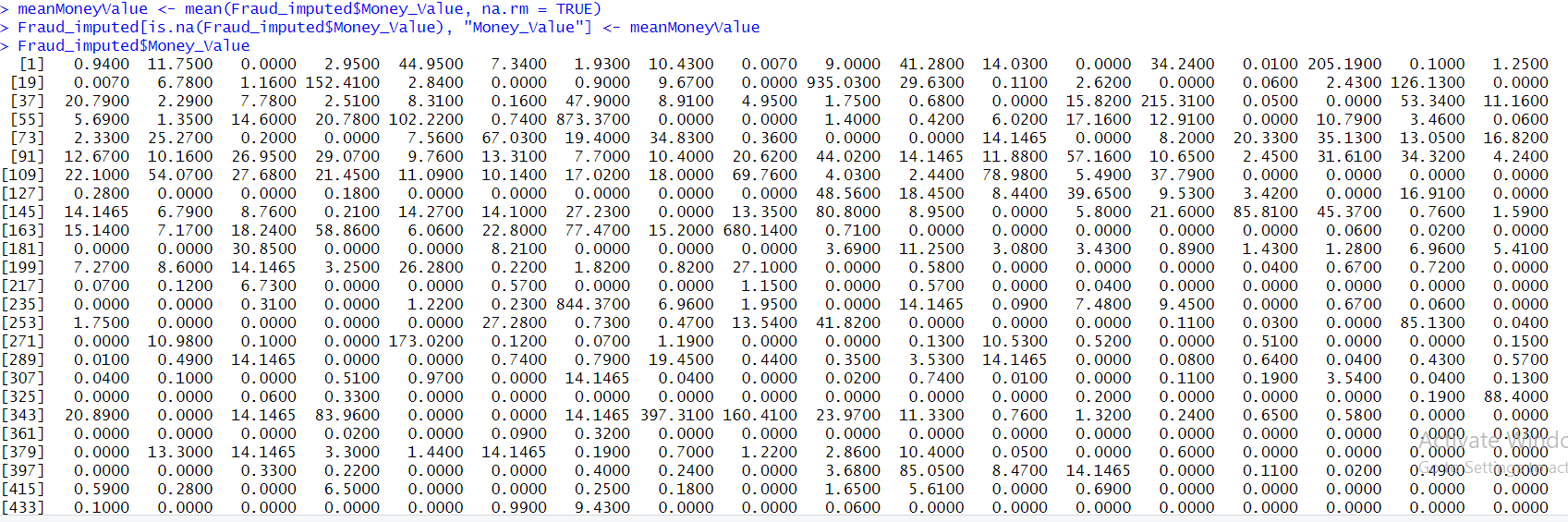
So we need to manually impute data in those columns.



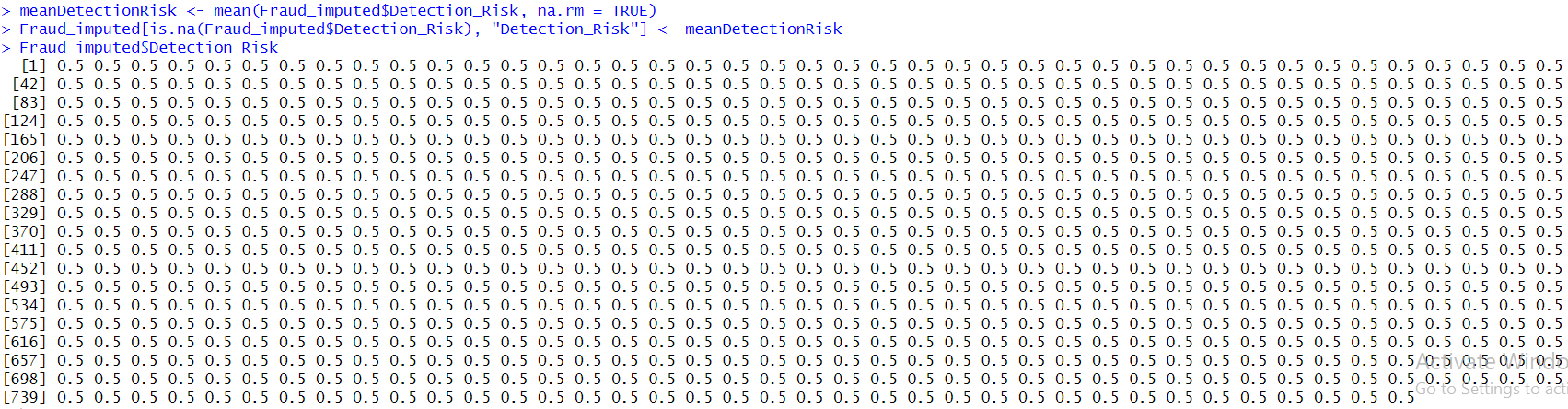
This is for PARA\_A column



This is for PARA\_B column

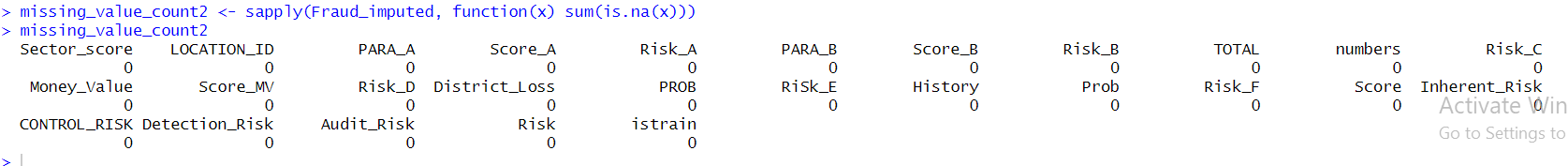


This is for Money\_Value column



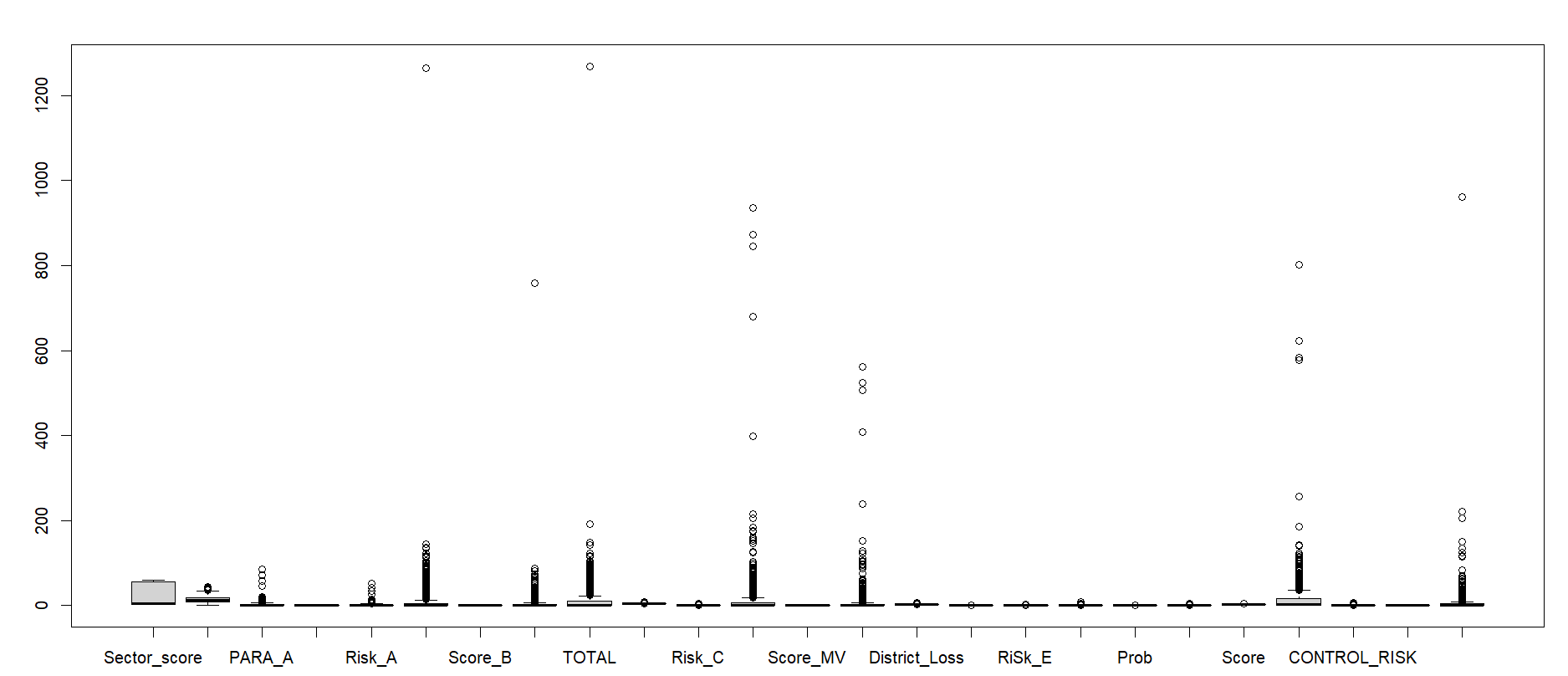
This is for Detection\_Risk column.

So again the null value of the dataset is:

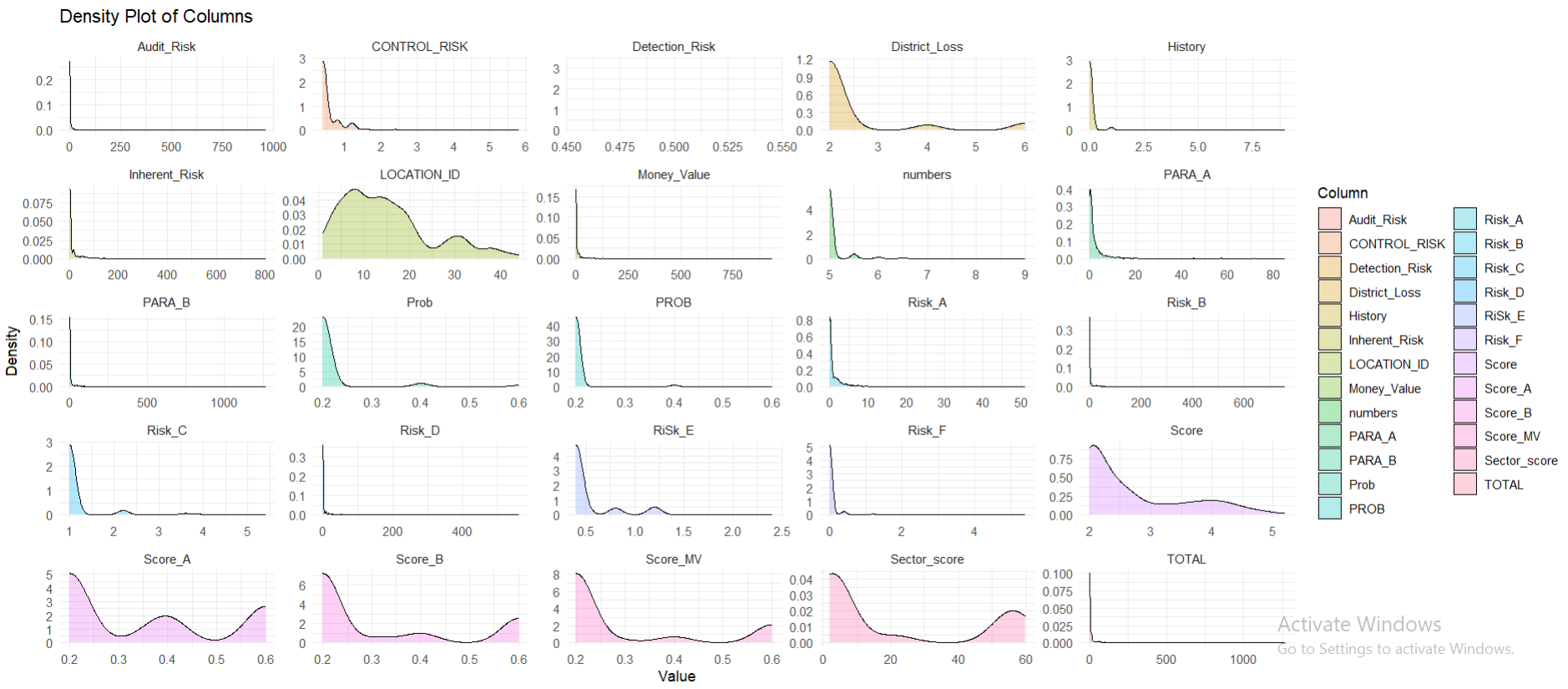


## Outliers

Outliers are values that are vastly different from their population. When operating with datasets we need to know whether to remove them or not. So let’s look at the outliers through box plot:

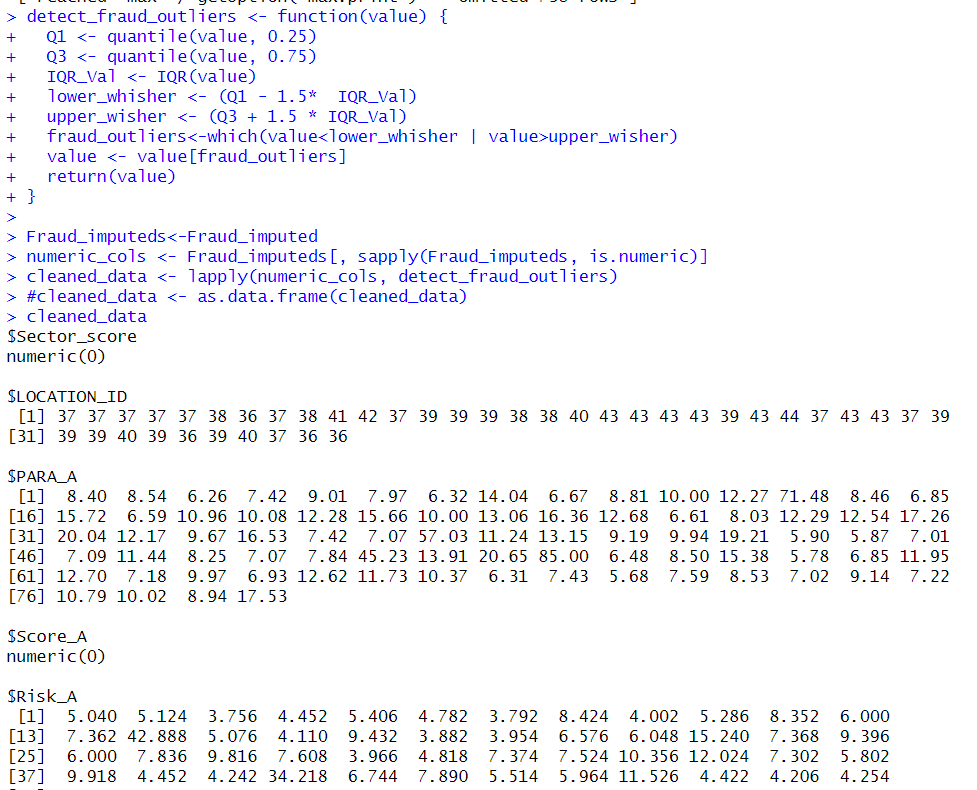


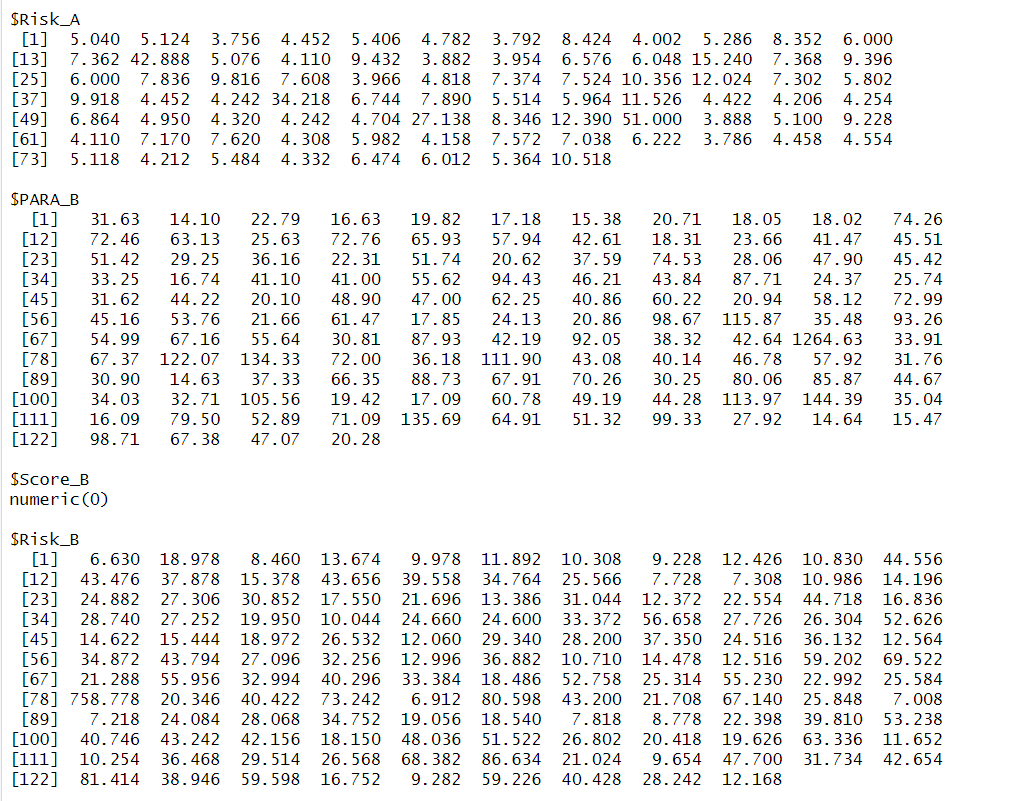
As we can see we can’t do anything with this plot so we are going to check skewness to decide which method to use(zscore or IQR).



This is the skewness of every column.

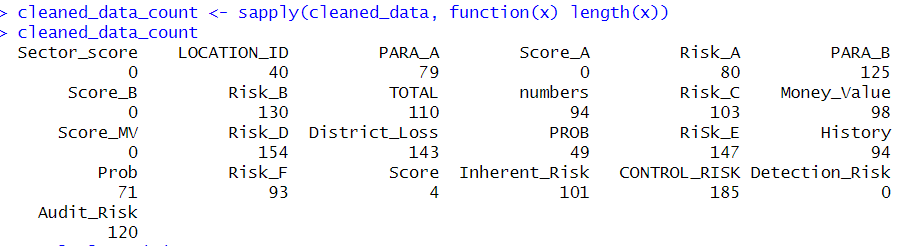
If it was a normal distribution, it would have been better to use Z-Score but it is skewed data so IQR would be less affected by it. So, we will be using IQR to detect outliers.



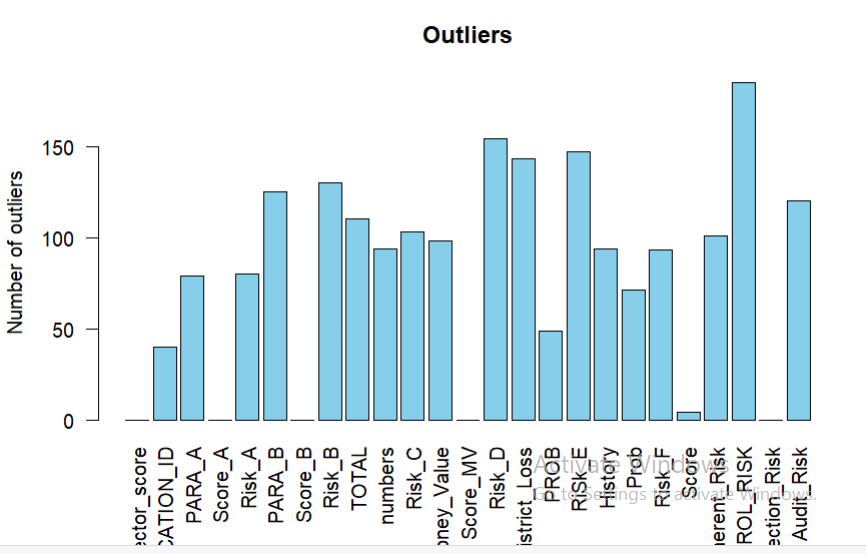


These all are the outliers of their respective column.

The number of outliers present in columns are:

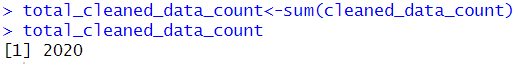


So lets plot the number of the outliers present in columns



These are the number of outlier present in diff column.

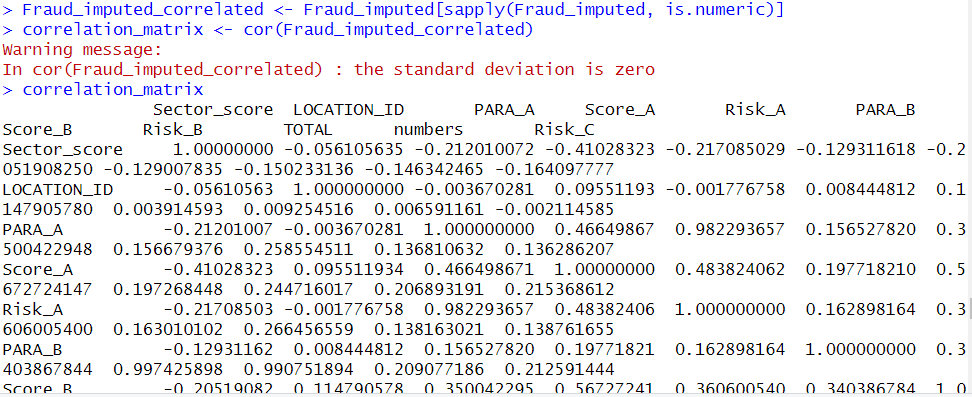
The total number of outliers present in a dataset is:



I decided not to delete any of them as they will remove all the rows present in a dataset and it won’t be useful to do it.

## Multicollinearity

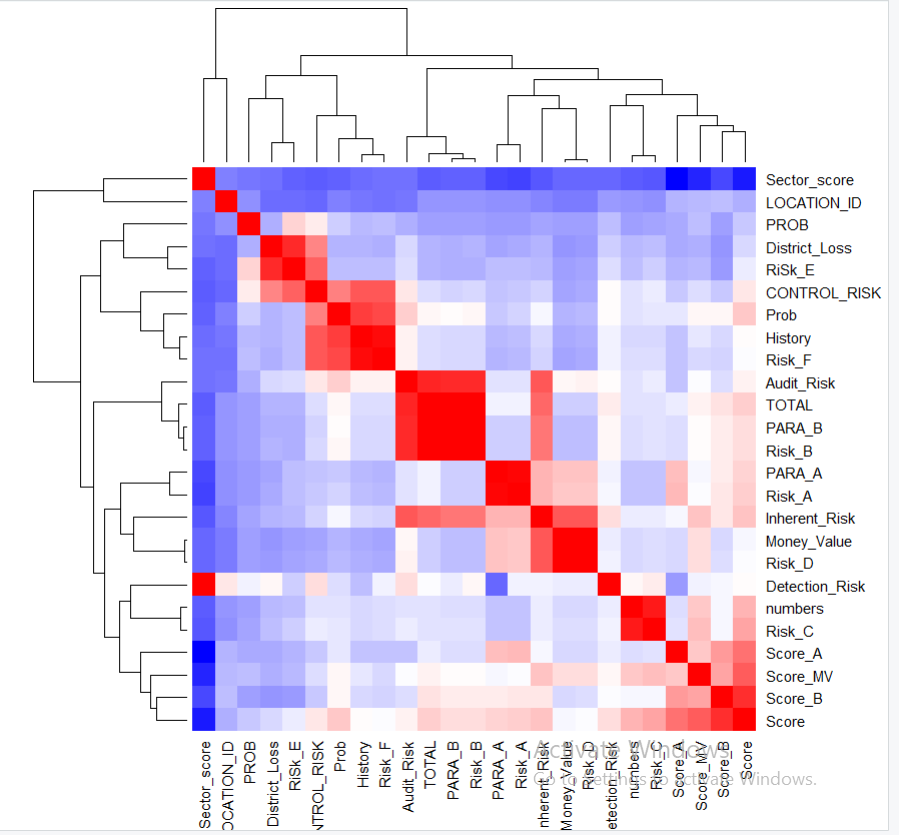
In R it is easier to calculate correlation. We can use the following code to do that:



Here we have used cor() to compute the correlation coefficient of a column in data frame . it returns 25x25 matrix

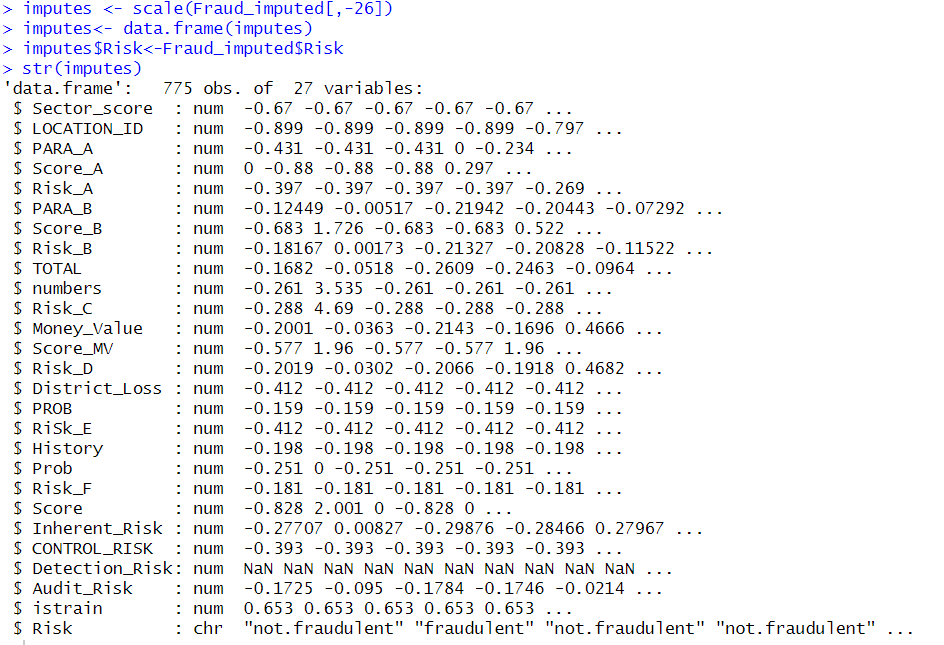


Now lets make a heatmap to visualize the given matrix:

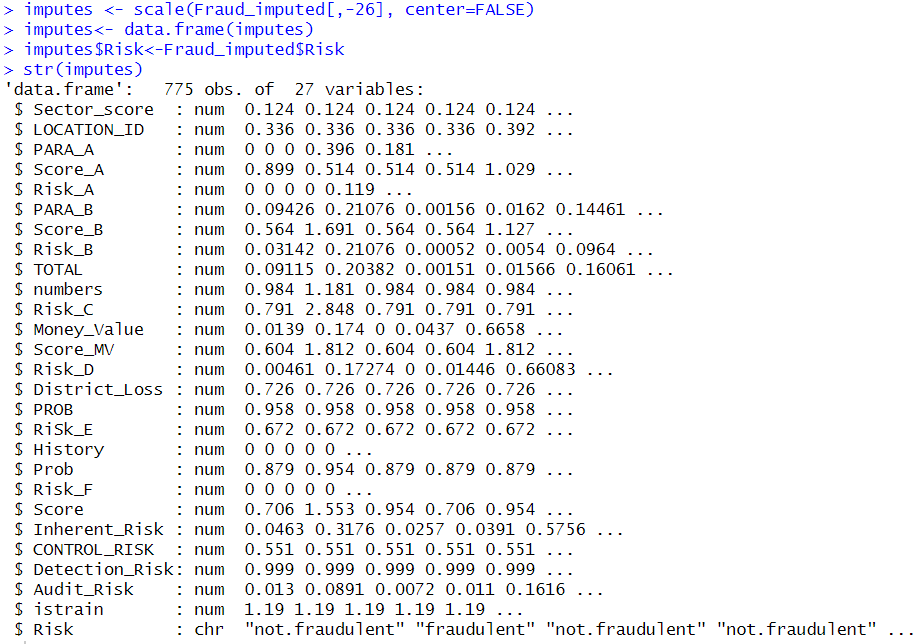


## Scaling

Scaling is done to standardize numeric data. We will be doing it in our dataset and it is very easy to do with R.



Here we used scale() to do scaling but we had to del 26th column because it was char datatype. Here you can see NaN value in Detection\_Risk. It is because it is an constant variable and the sd of constant variable is 0. If the scale is true then the scaling is done by dividing the centred column by their sd and as we know anything divided by zero is imaginary. So for that we need to do “center=False”



Here Detection-Risk is a valid value.

# Bibliography

Abdulalem Ali 1, S. A. R. ,. H. O. A. E. E. ,. A.-D. ,. N. ,. E. ,. E. a. A. S., 2022. *mdpi.* [Online]   
Available at: https://www.mdpi.com/2076-3417/12/19/9637  
[Accessed 3 March 2024].

Abolfazl Mehbodniya, I. A. P. N. P. R. S. V. M., 2021. *hindawi.* [Online]   
Available at: https://www.hindawi.com/journals/scn/2021/9293877/  
[Accessed 31 march 2024].

Aji Mubarek Mubalaike, E. A., 2018. *IEEE.* [Online]   
Available at: https://ieeexplore.ieee.org/abstract/document/8566574  
[Accessed 31 march 2024].

Anuj Sharma, P. K. P., 2013. *arxiv.* [Online]   
Available at: https://arxiv.org/abs/1309.3944  
[Accessed 31 march 2024].

Jan, C.-l., 2018. *MDPI.* [Online]   
Available at: https://www.mdpi.com/2071-1050/10/2/513  
[Accessed 31 March 2024].

John O. Awoyemi, A. O. A. S. A. O., 2017. *IEEE Xplore.* [Online]   
Available at: https://ieeexplore.ieee.org/abstract/document/8123782/authors#authors  
[Accessed 31 March 2024].

Khaled Gubran Al-Hashedi, P. M., 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *ScienceDirect,* Volume 40.

Mark Lokanan, N. H. V. T., 2019. Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. *Asian Jounal of Accounting Research,* 4(2), p. 20.

Nishta Hooda, S. B. P. S. R., 2018. *Taylor&FrancisOnline.* [Online]   
Available at: https://www.tandfonline.com/doi/full/10.1080/08839514.2018.1451032  
[Accessed 3 march 2024].

Patricia Craja, S. L. A. K., 2020. Deep learning for detecting financial statement fraud. *ScienceDirect,* Volume 139.

Yessy Yuliana, A. H. F. M. A. D., 2022. A LITERATURE STUDY OF FINANCIAL STATEMENT FRAUD DETECTION. *International Journal of Social Science,* 2(3).