Proactive Identification Of Loss For Company XYZ

Rachel Goldsbury

SNHU: DAT 610 Optimization and Risk assessment

Instructor: Kyle Camac

August 25th, 2024

Introduction

Company XYZ faces operational losses from fraudulent bodily injury claims related to automobile accidents. To address these challenges, the company aims to refine its fraud detection strategies. This paper evaluates analytic tools to enhance fraud identification, specifically by validating the PRIDIT scoring methodology using R Studio. The goal is to improve the accuracy and efficiency of detecting fraudulent claims.

PRIDIT Scoring Method

The PRIDIT (Procedure for the Risk-Adjusted Data for Identification of Transactions) scoring methodology excels in detecting anomalies and assessing risk within datasets. Unlike logistic regression, which requires predefined model structures and may struggle with high-dimensional data, PRIDIT efficiently handles complex datasets without such constraints. It identifies subtle deviations from norms that may indicate fraud.. As noted by (Brockett, 2002), "The PRIDIT methodology does not require a knowledge of which respondents indulged in fraud in order to implement it" (p. 30). This underscores PRIDIT's advantage in environments lacking pre-labeled fraud data, making it a robust tool for analyzing complex datasets.

Comparison of PRIDIT with Other Scoring Methods

When evaluating PRIDIT (Principal Component Analysis of RIDITs) against logistic regression and the Wilcoxon rank-sum test, several advantages of PRIDIT become clear. Logistic regression is effective for predicting binary outcomes based on predictors but can't handle high-dimensional data and requires a predefined model structure (Golden et al., 2019). PRIDIT, however, excels in managing high-dimensional datasets without such constraints and is adept at detecting subtle patterns by identifying deviations from standard norms. This capability makes PRIDIT particularly useful for revealing indicators of fraud that logistic regression might overlook (Golden, 2019).

In comparison to the Wilcoxon rank-sum test, which only compares distributions between two groups and does not provide risk scores or handle multidimensional data, PRIDIT integrates multiple variables simultaneously. This integration allows PRIDIT to deliver a comprehensive assessment of fraud risk, including quantitative risk scores that facilitate a more detailed and actionable evaluation of potential fraud (Golden, 2019).

The benefits of PRIDIT is further supported by its ability to provide metric-level scores, which can be incorporated into further analysis with other variables such as demographics. In contrast, methods like Kohonen's Self-Organizing Feature Maps and cluster analysis either lack this metric-level capability or do not offer the same level of detail and actionable insights (Golden, 2019). PRIDIT's ability to handle complex, high-dimensional data and provide detailed risk assessments highlights its advantages over both logistic regression and other simpler statistical methods.

Application of PCA

Principal Component Analysis (PCA), simplifies complex datasets by transforming original variables into uncorrelated principal components that capture the most variance. This technique aids in identifying patterns and anomalies that could indicate fraud. Applying PCA to RIDIT-transformed variables helps to prioritize suspicious claims and enhances fraud detection capabilities. According to the "*DAT 610 Module Four Exercise"* document, PCA is instrumental in "bringing out the strongest patterns that exist within a data set" and identifying key predictors

.This method effectively reduces dimensionality, making it easier to focus on the most significant features for fraud detection.

PCA Execution

Screenshots of the R Studio input and output are provided below. We begin by standardizing the RIDIT-ized variables to normalize the data. Next, we apply Principal Component Analysis (PCA) to the standardized RIDIT data to extract principal components. We review the PCA output to summarize the variance captured by each component and visualize the results. These visualizations help us interpret the significance of each principal component and understand the underlying patterns in the data.

To confirm the directionality of RIDIT transformations, we import the dataset into R Studio. We inspect the RIDIT variables by comparing them with the raw data through a scatterplot matrix. This visual inspection helps us verify that the RIDIT transformations align correctly with the raw data. Scatterplots are used to confirm the directionality and effectiveness of the RIDIT-ized variables in representing the underlying risk metrics.

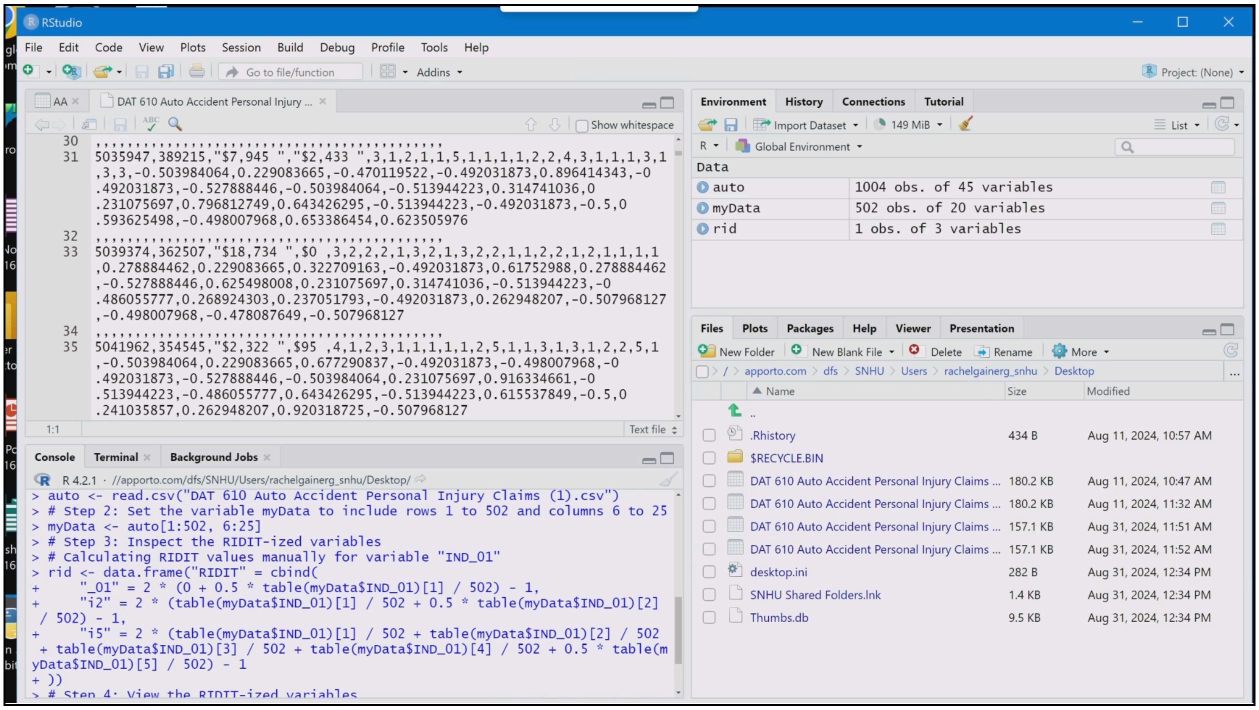

Figure 1: Loading data and calculating RIDIT-ized variables

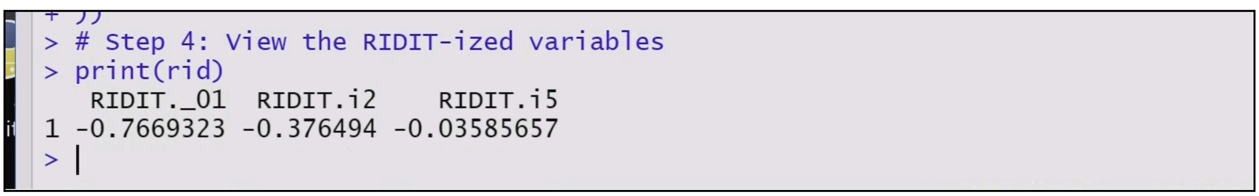

Figure 2: View RIDIT-ized variables.-The RIDIT values for the variables were manually calculated. The calculated RIDIT values were printed. For example, the first category had a RIDIT value of approximately -0.77, indicating a deviation from the norm.

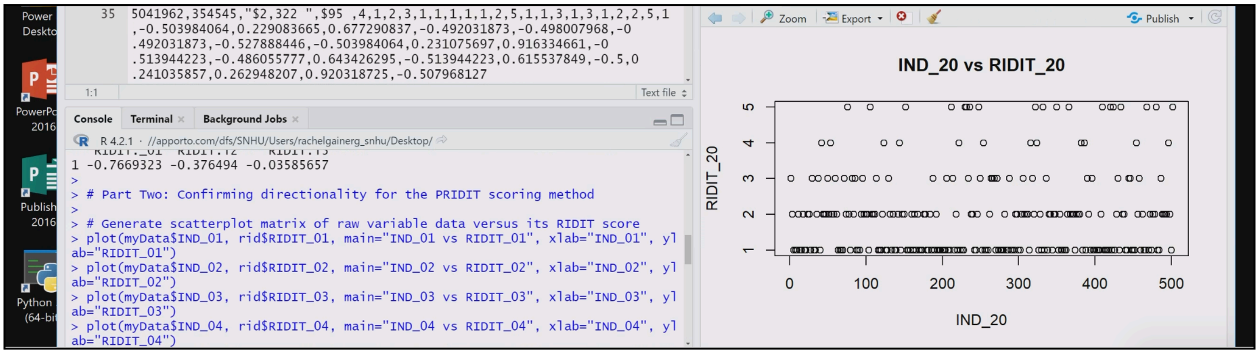

Figure 3: Confirming directionality for PRIDIT scoring method

```
> # Step 2: View the summary of PCA analysis
> summary(myRiditPCA)
Importance of components:
                           Comp.1     Comp.2     Comp.3     Comp.4
Standard deviation     0.65317538 0.64119157 0.62385077 0.59004304
Proportion of Variance 0.07628443 0.07351093 0.06958854 0.06225062
Cumulative Proportion  0.07628443 0.14979535 0.21938389 0.28163451
                           Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     0.58632769 0.56811051 0.55584653 0.54321523
Proportion of Variance 0.06146913 0.05770878 0.05524412 0.05276187
Cumulative Proportion  0.34310364 0.40081242 0.45605654 0.50881841
                           Comp.9    Comp.10    Comp.11    Comp.12    Comp.13
Standard deviation     0.54004099 0.5306189 0.52496576 0.51754342 0.49262861
Proportion of Variance 0.05214705 0.0503433 0.04927632 0.04789276 0.04339259
Cumulative Proportion  0.56096546 0.6113088 0.66058508 0.70847784 0.75187043
                          Comp.14    Comp.15    Comp.16    Comp.17
Standard deviation     0.47681805 0.47480309 0.46041410 0.44860120
Proportion of Variance 0.04065197 0.04030912 0.03790299 0.03598298
Cumulative Proportion  0.79252240 0.83283153 0.87073452 0.90671750
                          Comp.18    Comp.19    Comp.20
Standard deviation     0.42717170 0.42451091 0.39877124
Proportion of Variance 0.03262731 0.03222211 0.02843308
Cumulative Proportion  0.93934481 0.97156692 1.00000000
>
```

Figure 4: Summary of PCA analysis - PCA was conducted on the RIDIT-ized variables. The summary of the PCA revealed that the first few components captured a significant proportion of the variance in the data, with Component 1 accounting for about 7.6% of the variance. The cumulative proportion of variance explained by the first four components was approximately

28.2%, indicating that these components together summarize a substantial amount of the information in the data.

Visualization of Results

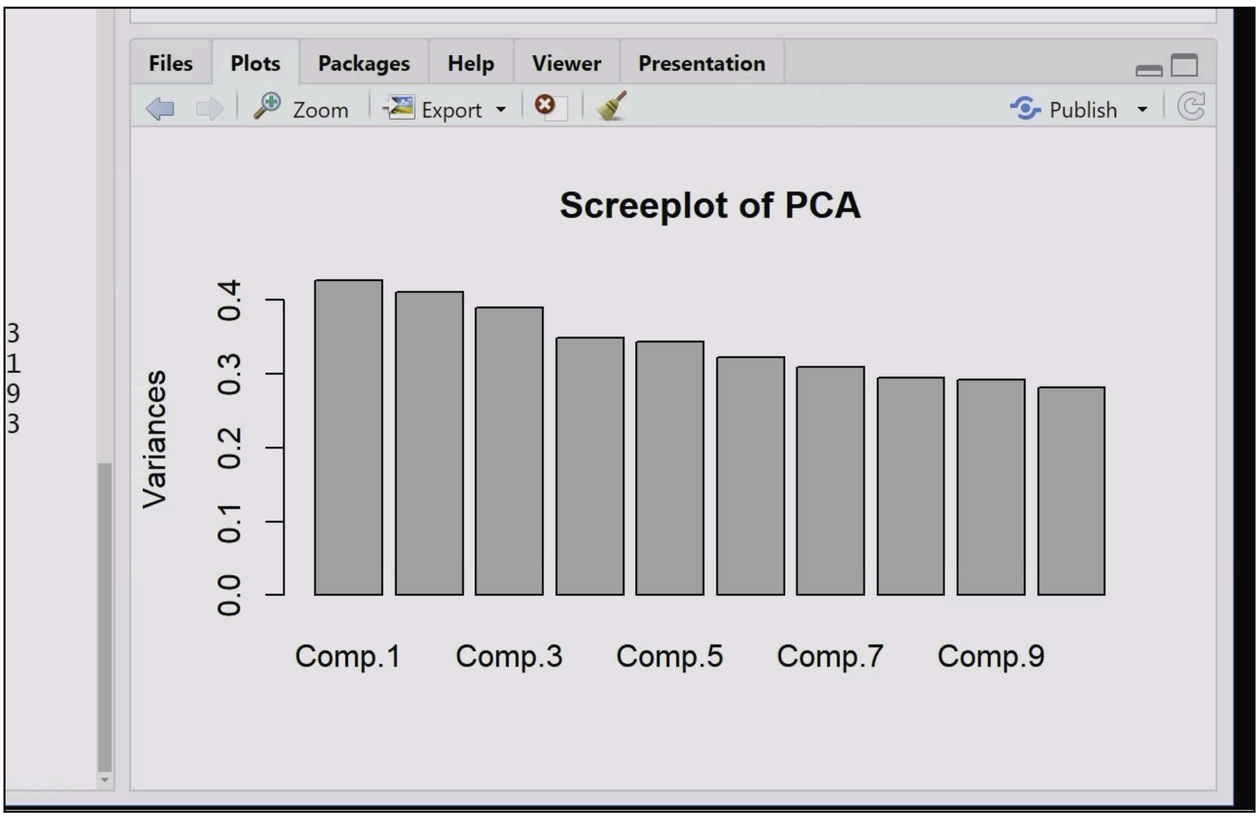

Figure 5: Screeplot of PCA- A scree plot was produced to visualize the importance of each principal component. The plot shows a decreasing trend, indicating that the first few components are the most informative. This helps in deciding how many components to retain for further analysis.

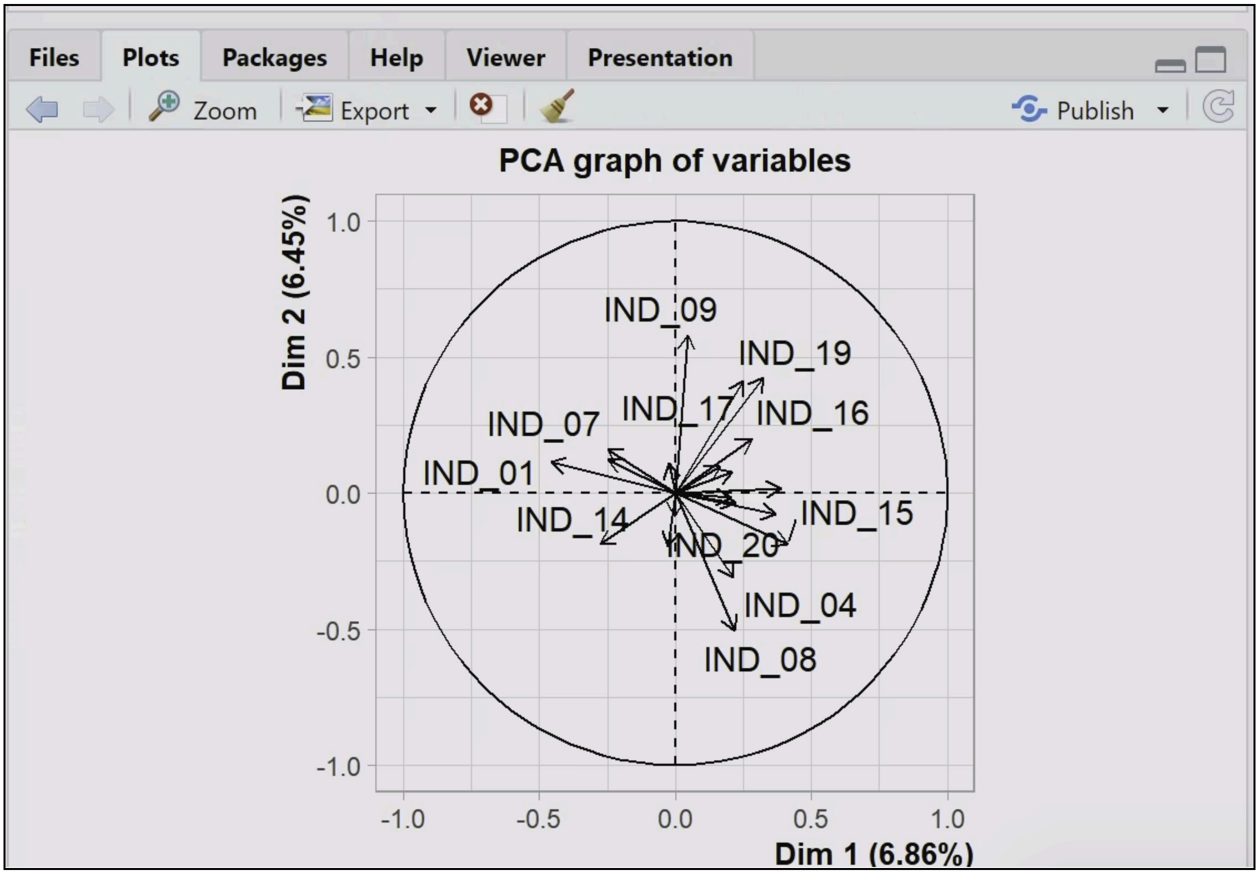

Figure 6: Variables Factor Map

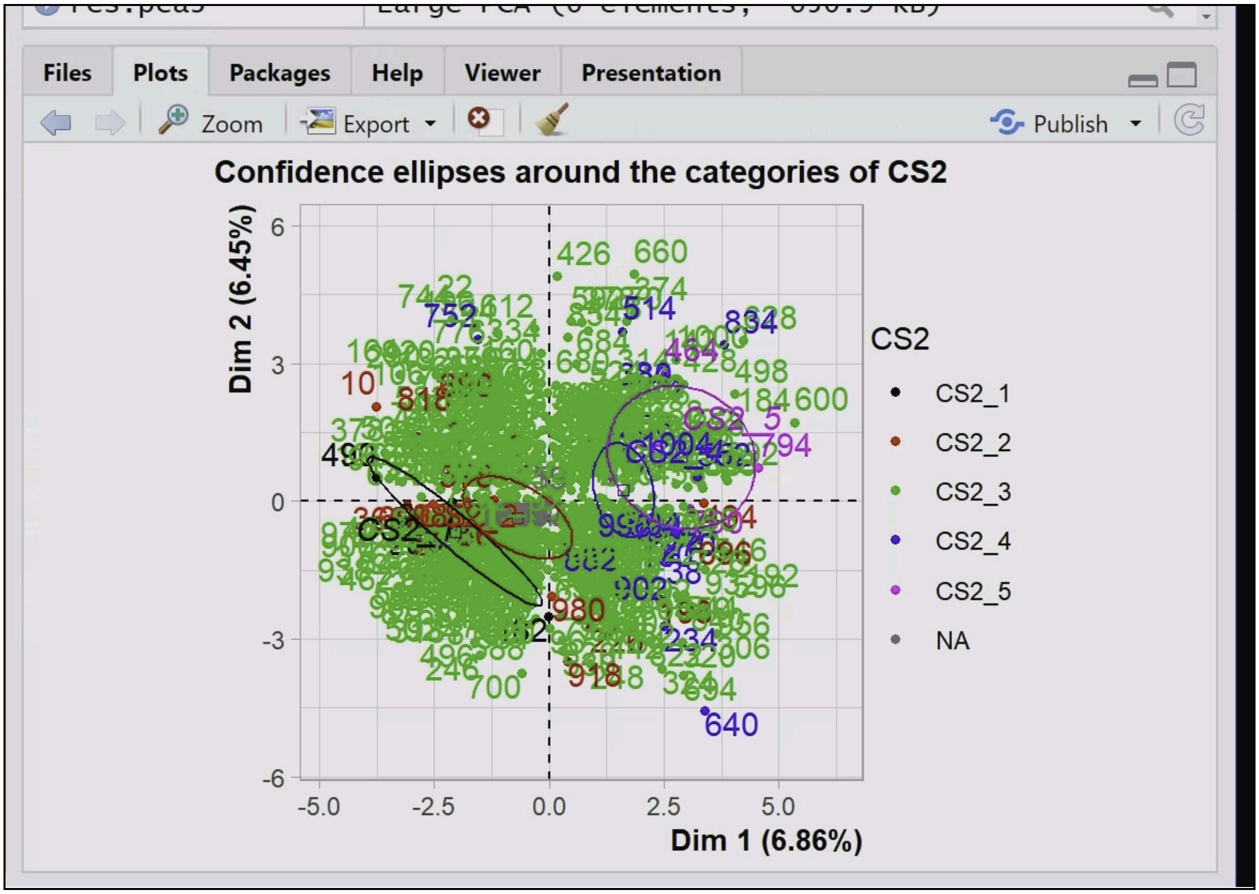

Figure 7: Scatter Plot of PCA Scores- The scatterplot was generated to visualize the relationship between the raw variables and their corresponding RIDIT scores. These plots help in confirming the directionality for the PRIDIT scoring method, which is crucial for identifying patterns or outliers that may suggest fraudulent activity.

Enhancement of Claims File with Additional RIDIT-ized Variables

To enhance fraud detection, we can improve the claims file by incorporating additional RIDIT-ized variables. We identify variables that may impact fraud risk, such as claim amount and paid amount. Then, we apply the RIDIT transformation to compute scores for these new variables, integrating them into the analysis. As noted by Brockett, Derrig, and Golden (2003), “Principal Component Analysis of RIDIT scores (PRIDIT) allows for reducing uncertainty and targeting the appropriate claims, making it possible to allocate investigative resources more efficiently” (p. 342). We then update the Principal Component Analysis (PCA) by incorporating these new RIDIT-ized variables to refine the results and enhance the accuracy of the fraud detection process.

Applying PC Scoring to Future Claims

To apply Principal Component (PC) scoring to future claims, we first develop a PC scoring model using the results from the Principal Component Analysis (PCA). This model is based on the principal component scores derived from the analysis. We then integrate these PC scores into the fraud detection system to evaluate future claims effectively.

Incorporating Other Features into Visualization

To enhance the visualizations of fraud detection results, we start by selecting features that provide a greater context. Next, we prepare the data by integrating these features into the dataset, which enriches the visual analysis. Then we create visualizations that incorporate these enhanced features, offering a more comprehensive view of the fraud detection results and enabling a clearer understanding of the underlying patterns.

Conclusion

This analysis highlights the effectiveness of PRIDIT and PCA as key tools for enhancing fraud detection at Company XYZ. The PRIDIT scoring method excels at identifying deviations from standard norms, providing a nuanced approach to fraud identification, while PCA simplifies complex data, making it easier to uncover patterns. By comparing PRIDIT with other methods like logistic regression and the Wilcoxon rank-sum test, and by enhancing the claims file with additional RIDIT-ized variables, Company XYZ can significantly refine its fraud detection strategies. Applying PC scoring to future claims and incorporating additional features into visualizations will further strengthen fraud detection efforts, offering deeper insights and improving overall risk management practices.

References

1. Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A., & Alpert, M. (2002). Fraud classification using principal component analysis of RIDITs. *Journal of Risk and Insurance*, 69(4), 271-312. https://doi.org/10.1111/1539-6975.00027
2. Brockett, P. L., Derrig, R. A., & Golden, L. (2003). Fraud Classification Using Principal Component Analysis of RIDITs. *Journal of Risk and Insurance, 69*(3), 341-371. http://silver.ima.umn.edu/industrial/2003-2004/derrig/OfficialReprint.pdf
3. DAT 610 Module Four Exercise. (n.d.). *Introduction and Additional Resources*. Retrieved from DAT 610 Module Four Exercise
4. Girling, P. X. (2022). *Operational Risk Management* . Wiley Professional Development (P&T). https://mbsdirect.vitalsource.com/books/9781119836056
5. Golden, L. L., Brockett, P. L., Guillén, M., & Manika, D. (2019). aPRIDIT Unsupervised Classification with Asymmetric Valuation of Variable Discriminatory Worth. *International Journal of Forecasting, 35*(2), 685-703. https://doi.org/10.1080/00273171.2019.1665979
6. YouTube. (August, 2024.). *Loading data into R* . YouTube. https://www.youtube.com/watch?v=VLtazaiYo-c