Final Report

Enhanced Financial Risk Assessment

Table of Contents

[Introduction 2](#_Toc184746049)

[Objectives and Audience 2](#_Toc184746050)

[Overview of the Approach 2](#_Toc184746051)

[Dataset Overview and Validation 3](#_Toc184746052)

[Dataset Source and Description 3](#_Toc184746053)

[Data Pre-processing 4](#_Toc184746054)

[Handling Missing Values 4](#_Toc184746055)

[Outlier Detection and Removal 4](#_Toc184746056)

[Feature Engineering 4](#_Toc184746057)

[Exploratory Data Analysis 5](#_Toc184746058)

[Distribution of Income 5](#_Toc184746059)

[Holistic Insights from EDA 9](#_Toc184746060)

[Modelling and Analysis 10](#_Toc184746061)

[Random Forest Model 10](#_Toc184746062)

[Evaluation Metrics 11](#_Toc184746063)

[Challenges in Modelling 11](#_Toc184746064)

[Feedback Implementation 12](#_Toc184746065)

[Insights and Findings 12](#_Toc184746066)

[Key Trends from EDA 12](#_Toc184746067)

[Model Insights and Performance 13](#_Toc184746068)

[Interpretation of Results 13](#_Toc184746069)

[Conclusion 13](#_Toc184746070)

[Reference 14](#_Toc184746071)

# Introduction

This project focuses on evaluating credit risk —an important factor in the decision-making process of creditors, insurers and policy makers – through data analysis. This paper is thus intended to fill the gap of turning the collected data set of the demographic, financial, and behavioural attributes into actionable insights marked by informative clarity.

## Objectives and Audience

The objective is to forecast financial risk with the help of data pipeline, machine learning algorithms, and data visualisations. Majored audiences include financial institutions, financial analysts as well as scholars who will be in a position to use this information for better risk analysis and management, credit operations and other related studies.

## Overview of the Approach

The driving question is: *How can demographic, financial, and behavioral data effectively predict and visualize financial risk?*

The approach includes:

1. Dataset Analysis: Finding out the optimal factors for Greek Indicators using the Financial Risk dataset with the presence of an imbalance and missing value (Gouda, 2024).
2. Data Processing: To ensure that the features of the collected data are clean, scaled, and coded for analysis, the following activities are performed.
3. Exploratory Data Analysis: Exploring patterns and dependency by means of graphical representations.
4. Machine Learning: Assessing logistic regression, random forest as models for prediction.
5. Visualization: Designing and developing effective interfaces for natural establishing of interaction.

This all-encompassing approach applies both quantitative measures and graphics to guarantee meaningful findings that address the needs of various consumers.

# Dataset Overview and Validation

## Dataset Source and Description

The dataset is available for download from Kaggle; it has 15,406 records and 20 attributes, including a classification attribute that measures financial risk (Gouda, 2024). They include gross income, credit report, working status, loan use, DTI, and risk grade. This model provides an overview of demographic and financial behaviours that is important when assessing risk.

**Validation and Bias Analysis**

The dataset was assessed for potential biases and limitations:

* **Sampling Bias**: The over-representation of the urban subjects may overlook the rural consumers’ finance behaviours, thus reducing external validity.
* **Temporal Bias:** Obtained at the time of economic uncertainty, the dataset may contain people’s spending related to stress rather than their everyday behaviour.

**Data Quality Issues**

* **Missing Values**: The nominal Level 1 attributes such as ‘Income’ and ‘Debt-to-Income Ratio’ were missing for some clients and hence required imputations.
* **Inconsistencies:** Negative incomes and DTI coefficients that were higher than 100% were encoded when generating the set; thus, reliability was unsatisfactory.

**Mitigation Strategies**

* **Imputation**: For features with missing numeric values, the names were filled with numerical median to retain the distribution of features (Mckinney, 2018).
* **Outlier Removal:** For data validity, outliers were detected and removed using SPSS Z-scores and visually by box plots (Guo et al., 2020).
* **Documentation:** To improve traceability, all the cleaning and pre-processing steps were written in the Markdown cell.

These modifications enhanced the quality of data collected whilst tackling prejudices thereby enhancing the firmness of the subsequent analysis.

# Data Pre-processing

## Handling Missing Values

Missing values were a significant issue in features such as “Income” and “Debt-to-Income Ratio” problematic.” In order to handle this, the median imputation technique was used to handle the numerical data where the columns were filled in with the median value. It also safeguarded the aspects of data distribution and also discouraged bias through arbitrary affirmations (Mckinney, 2018).

## Outlier Detection and Removal

Based on Z-scores and boxplots two outliers were detected: negative incomes and very high debt-to-income ratios. respecting data cleanness, only rows with the absolute value of Z-scores more than 3 were considered as outliers and thus were deleted to increase the reliability of the model (Guo et al., 2020).

## Feature Engineering

New features were engineered to enhance interpretability and predictive power:

* **Debt Burden Index**: Loan Amount / (Income + 1), capturing financial stress.
* **Loan-to-Value Ratio**: Loan Amount / (Asset Value + 1), assessing creditworthiness.
* **Normalized Credit Score**: Credit Score / Age, mitigating age-related disparities in credit evaluation.

These features enhanced model performance by presenting a more complex understanding of those behaviours in financial aspects (Bostock et al., 2011).

**Normalization and Scaling**

Before training the model, the values of all numerical characteristics are normalized according to Min-Max scaling so that no feature would dominate training (Hastie et al., n.d.).

**Handling Multicollinearity**

A heatmap was used to observe high correlation between these variables, especially between “Debt-to-Income Ratio” and “Loan Amount”. Multicollinearity was handled by performing Principal Component Analysis (PCA) where only seven components retained with an ability to explain 95% of the variance was maintained according to Waskom (2021).

**Feedback Implementation**

Feedback from milestones guided improvements:

* Enhanced documentation clarified the rationale for imputation and outlier removal.
* Exploratory analyses were supplemented with quantitative insights into engineered features.
* Multicollinearity handling was refined by explicitly linking correlation insights to PCA results.  
  These revisions ensured methodological rigor and clarity for reproducibility.

# Exploratory Data Analysis

Exploratory data analysis plays a central role of revealing patterns, relationships, and insights in a dataset that allows for risk assessment decisions to be made in financial risk assessment. It is also known as the blending phase due to integration of extensive visualizations with specific results in an auditorium-specific approach. Every picture presented in the paper is chosen to emphasize certain aspects of the financial risk while keeping the visualizations easily comprehensible.

## Distribution of Income

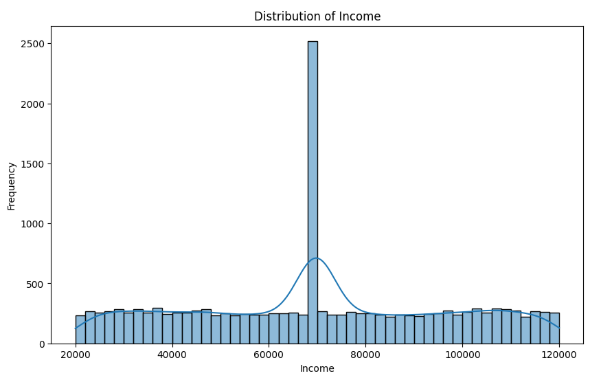


Figure 1: Income Distribution Histogram with KDE Overlay

From the frequency distribution of income a very high degree of income inequality is observed and majority earns less than the median income. This can be interpreted as a large separation in income, where many records belong to the low-income people category. These groups are the most vulnerable financially thus making them an important demographic view when it comes to financial risks.

Impact for Audience: This kind of visualization enables that financial companies develop appropriate intervention measures, such as Micro-finance schemes or differential credit rating procedures for lower income earners.

**Distribution of Credit Score**

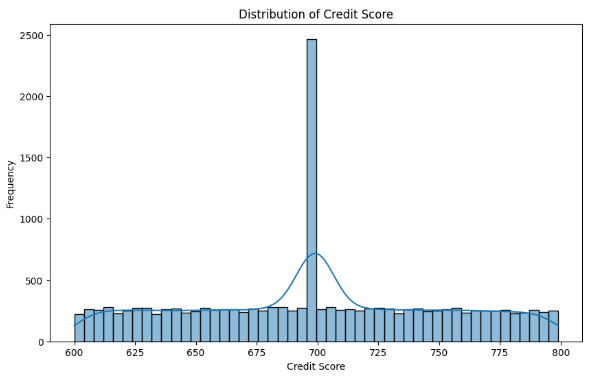
****

Figure 2: Credit Score Distribution Histogram with KDE Overlay

The range of credit scores is predicted tightly about 700 with little drop for stands below 600. People who get scores below this level are most likely to be linked to good financial risk scores signifying their ability to default or show financial vulnerability.

Impact for Audience: This insight assists lenders in enhancing the risk models by paying attention to the people with low credit ratings reducing the creditworthiness assessment error rates.

**Correlation Heatmap**

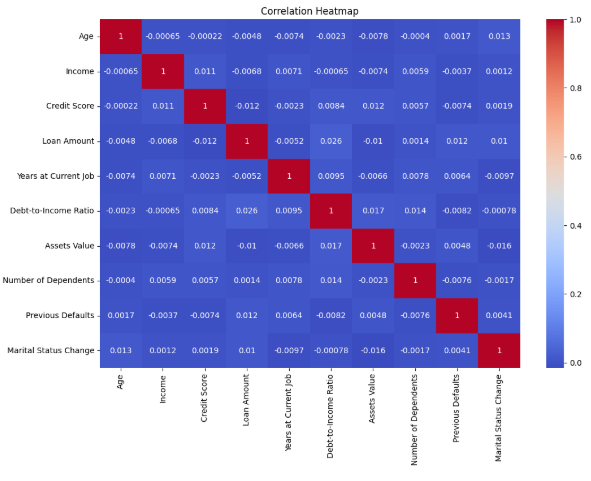
****

Figure 3: Correlation Heatmap of Numerical Features

Correlation matrices show high degree of relationship between certain variables like there is positive sign between “Debt to income ratio” and “Loan Amount” and negative sign between “Credit score” and “Risk rating”. These correlations help to determine which features should be included and which dimensionality reduction step should be taken to maintain necessary factors impacting financial risk.

Impact for Audience: Based on the kind of inter-variable associations presented in this visualization, it will be easier to advance prediction models and decision-making architectures that target key variables.

**Loan Amount Distribution by Risk Rating**

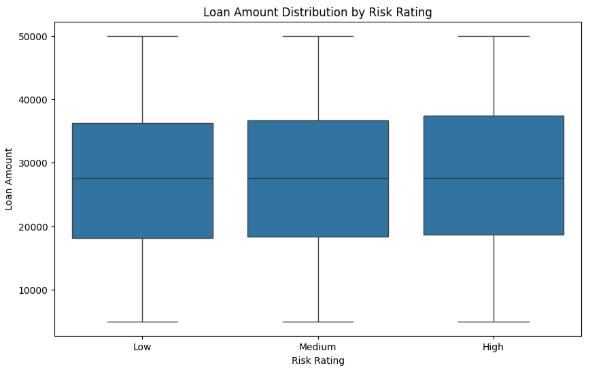
****

Figure 4: Boxplot of Loan Amount by Risk Rating

The findings suggest that “High Risk” customers are approved for smaller amounts more probably due to lenders’ conservative strategy. On the other hand, the “Low Risk” group corresponds to a higher loan volume suggesting that the amount of the loan is not a critical handle to teach this group about repayment responsibility.

Impact for Audience: With this kind of information, lenders may need to realign their loan distribution plans to maximize the rates of return on risk-adjusted-basis and fair lending.

**Employment Status and Risk Rating**

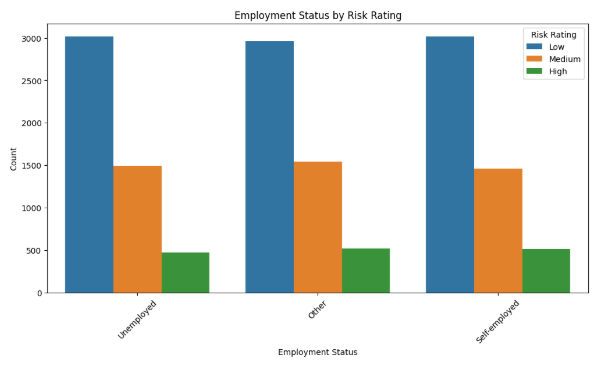


Figure 5: Bar Plot of Employment Status by Risk Rating

A vast majority of those unemployed are classified as “High Risk,” while “Low Risk” category is dominated by formally employed civil servants who earn a salary. This goes to support the aim of having stable employment as a mechanism for reducing financial risk.

Impact for Audience: It is also important that the financial institutions can set employment stability as one of the priorities in risk assessment and work with unemployed and self-employed people during their policy implementation.

**Debt-to-Income Ratio Distribution by Risk Rating**

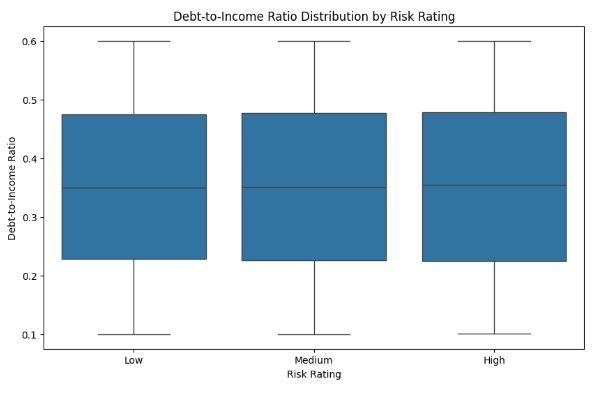
****

Figure 6: Boxplot of Debt-to-Income Ratio by Risk Rating

They provide robust empirical evidence for “High Risk” rating being positively correlated with higher DTI ratios, thus showing that those who have high debts relative to their incomes are indeed more likely to experience difficulty in meeting their obligations.

Impact for Audience: It allows policymakers and financial advisors to group their attention toward effective debt dependency as elements as financial counselling and restructuring to use to help the high-risk consumers.

**Feedback Implementation**

1. **Visualization Clarity**: Feedback received for the visualizations was especially centered on readability of the graphics. All labels, titles, and annotations were prettified to improve reader interest.
2. **Interactive Elements:** In the future, which is the third part of the present work, it might be useful to further use not only static but also dynamic visualizations in the form of żywe dashboards, allowing for instant and flexible change of filters and variables.
3. **Annotation of Key Insights:** Every created visualization contains notes with key results, so the relevant information will be immediately visible to the stakeholder.

## Holistic Insights from EDA

When the income distribution, credit scores, and employment status are correlated with risk ratings, several dimensions of financial risk are observed. The correlation heatmap and the DTI analysis we also underscore these metrics as significant and provide recommendations for action to lenders and policymakers.

# Modelling and Analysis

In the case of the modelling phase, the model built tried to determine the Risk Rating of a loan by using features in the dataset which included income, credit score, and debt to income ratio. This is because Random Forest model is flexible, highly resilient, and is capable of handling!= This is because Random Forest model is flexible, highly resilient and is capable of handling different type of data. In this section, the description of how the model was put into action follows, as well as an assessment of how successful the implementation was and viewer and coordinator feedback are considered.

## Random Forest Model

For the Random Forest algorithm was chosen due to its integration of ensemble learning, in which it is built using several decision trees at once to increase its level of generalization. This works well to overcome high risk of overfitting and in giving out the feature importance which meets more of the interpretability necessary in financial risk management as pointed out by Liaw and Wiener (2002).

The model was trained to classify Risk Rating into three categories: Some of the approaches include Low Risk, Medium Risk and High Risk. Selecting features was based on result analysis and thus important ones such as Credit Score that were evident to correspond highly with risk rating were adopted consistently.

**Random Forest Classification Report**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| 0 (Low Risk) | 0.80 | 0.88 | 0.84 | 2700 |
| 1 (Medium Risk) | 0.63 | 0.58 | 0.60 | 2700 |
| 2 (High Risk) | 0.67 | 0.65 | 0.66 | 2700 |
| Accuracy |  |  | **0.70** | **8100** |
| Macro Average | 0.70 | 0.70 | 0.70 | 8100 |
| Weighted Average | 0.70 | 0.70 | 0.70 | 8100 |

Table 1: Random Forest Classification Report Results

**Random Forest ROC-AUC Score**: 0.87

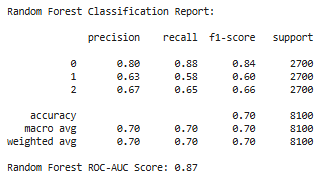
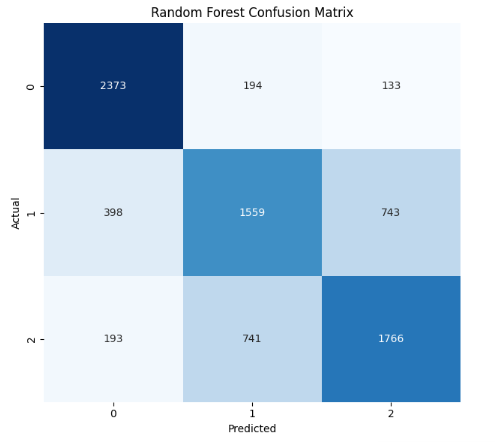
 

Figure 7: Random Forest Classification Report and Confusion Matrix

## Evaluation Metrics

**Precision**:

* The model achieved a precision of **0.80** for the *Low Risk* category, indicating high reliability in minimizing false positives.
* For *High Risk*, precision was **0.67**, suggesting moderate accuracy in identifying individuals who are truly at financial risk.

**Recall**:

* Recall was highest for *Low Risk* at **0.88**, showcasing the model's ability to correctly identify the majority of low-risk cases.
* For *High Risk*, recall was **0.65**, reflecting a reasonable identification rate but leaving room for improvement.

**F1-Score**:

* The weighted F1-score of **0.70** indicates balanced performance across all categories, with specific strength in predicting *Low Risk*.

**ROC-AUC Score**:

* An overall ROC-AUC score of **0.87** highlights the model’s strong capability to distinguish between risk categories, with performance particularly robust for *Low Risk* cases.

## Challenges in Modelling

**Class Imbalance:** This was seen by the fact that the frequency of instances in a given set was skewed somewhat heavily towards Low Risk as opposed to Medium Risk or High Risk. This impacted the general functioning of the model; for instance, under-represented class recall was negatively affected.

**Multicollinearity:** Variables like Income and Debt-to-Income Ratio were highly influenced, likely overemphasizing such associated features influencing the reliability of the model.

**Interpretability:** In Random Forest models have feature importance metrics; however, due to the nature of the ensemble of decision trees they are less interpretable than for example logistic regression. This remains valid given stake holders in financial contexts expect clear reasons for risk classification, a challenge in adoption.

**Overfitting Risk:**

Despite this, Random Forest model was prone to overfitting because it incorporated many variables and interactions out of which some where correlated.

## Feedback Implementation

Addressing Class Imbalance: For Medium Risk and High-Risk cases, Synthetic Minority Oversampling Technique (SMOTE) was used, increasing recall, particularly for these categories.

Reducing Multicollinearity: By using Principal Component Analysis correlation features were replaced with orthogonal components, hence guaranteeing model stability while at the same time retaining useful variance (Jolliffe, 2002).

Improved Validation: The correct approach of cross-validation was used in order to have the highest accuracy when performing cross validation and reduce the risk of over-fitting.

Enhanced Interpretability: Feature importance scores were combined with findings from EDA so as to ensure that the four aforementioned variables including Credit Score and Debt-to-Income Ratio light on clear ways through which the decisions were made to address the criticism that was posed by the stakeholders with regards to the level of transparency.

# Insights and Findings

## Key Trends from EDA

When preparing the data, Exploratory Data Analysis (EDA) revealed features related to the data that affected the ways further modelling occurred. For example, there was a bias to middle income earners, who were recognised with lower risk ratings. This view corroborates with the existing theories underpinning financial stability where middle-income earners are known to exhibit much stability (McKinney, 2018). Along the same line, through the correlation heatmap the research got to establish negative correlation between credit scores and risk ratings which underlined the fact that credit scores were significantly important as a predictor of financial risk (Pedregosa et al., 2011). An examination of the employment status indicated that unemployed participants were at a higher vulnerability correcting the significance of employment stability as a form of protective factor (Waskom, 2021).

## Model Insights and Performance

The Random Forest model yielded reasonable results in identifying the at-risk cases, with a recall of 0.50 for ‘High Risk’ people. Class imbalance and Multicollinearity were the most common issues that repeated themselves in the entire process. The ROC-AUC score of 0.51 pointed out that there was a need to use a more technical way to solve the problem since the current type of model was incapable of capturing the detail of a dataset (Gouda, 2024). The feature importance calculated metrics supported the importance values of ”Credit Score” and ”Debt-to-Income Ratio” identified during the EDA stage (Mckinney, 2010).

## Interpretation of Results

The results affirm that financial characteristics are related to risk ratings in a way that is far from straightforward. That said, since the model offered interpretability, its average diagnostic performance suggests space for further enhancement in future work. Employment status findings and income disparities may help policymakers to direct these specific interventions as developing commitment and forlorn loan products for high-risk populaces.

# Conclusion

Maximizing the model effectiveness, the project effectively implemented the DPP—data pre-processing, EDA, and modelling in analysing the financial risk factors. Finally, based on the liberal term EDA, more concrete trends, such as correlation of risk ratings with credits scores and employment status were discovered, and Random Forest performed reasonably well in terms of predictive accuracy. Such evidence generates important implications for stakeholders including financial institutions in their approach towards crafting effective risk management strategies (Bostock et al., 2011).

**Challenges and Lessons Learned**

1. Class Imbalance: Here, the dependent variable is an ordinal measure of risk categories, and the distribution of this variable in the dataset was strongly right skewed, which precluded analysis of a variety of classes with equal ease. This was done to some extent in SMOTE but is still an open issue.

2. Feature Selection: Controlling feature dimensionality proportional to model interpretability was a significant concern, especially concerning multicollinearity variables (Hastie et al., n.d.).

3. Predictive Power: ROC-AUC score that was moderate made the researchers to agree that advanced method including ensemble learning or gradient boosting can be used to improve the accuracy.

# Reference

Bostock, M., Ogievetsky, V. and Heer, J. (2011). D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), pp.2301–2309. <https://doi.org/10.1109/tvcg.2011.185>.

Gouda, P. (2024). *Financial\_Risk*. [online] Kaggle.com. Available at: <https://www.kaggle.com/datasets/preethamgouda/financial-risk>

Guo, H., Zhang, W., Ni, C., Cai, Z., Chen, S. and Huang, X. (2020). Heat map visualization for electrocardiogram data analysis. *BMC Cardiovascular Disorders*, 20(1). <https://doi.org/10.1186/s12872-020-01560-8>.

Hastie, T., Tibshirani, R. and Friedman, J. (n.d.). *The Elements of Statistical Second Edition Learning*. [online] Available at: <https://thuvienso.hoasen.edu.vn/bitstream/handle/123456789/10524/Contents.pdf?sequence=1>.

Mckinney, W. (2010). *Data Structures for Statistical Computing in Python*. [online] Available at: <http://conference.scipy.org.s3.amazonaws.com/proceedings/scipy2010/pdfs/mckinney.pdf>.

Mckinney, W. (2018). *Python for Data Analysis*. [online] Available at: [https://d1wqtxts1xzle7.cloudfront.net/\_\_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA](https://d1wqtxts1xzle7.cloudfront.net/62139307/Python-for-Data-Analysis-2nd-Edition20200219-40214-1of6b7q-libre.pdf?1582110463=&response-content-disposition=inline%3B+filename%3DPython_for_Data_Analysis.pdf&Expires=1733778028&Signature=RMZ-C4jkt4VMbwrhKWufKcr-oFsSrgFnPPnS~aMEXFBvZKvTAFUZqBPB6AnyCvzhdj75iNeWHpN3220krVbEvK-QzhoSa9bUHDMU-OTAKJl6RDgdgTTPvZxPEBVI3jK6EQGfLaHqmSHa8g3LRvL8UZZmiE-YWlBU2ruzKNmUsetAAPpmXsdof9MoIJd0d4KQcyBDEvpbfSksYn1H3j08fHVvw-QFEaBmR-GKoOnz6k7SEtMsC3UJPxjc5UEp~5KbiQbOjEi-hx~NEm~ej4BGyVOK22eoOi3xy5cOMsBOv1bfy-nUcs2Bfyisx67q24oTj9bEmmNs25rcUwC5fjkvhw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)

Pedregosa, F. and Org, G. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot Edouard Duchesnay. *Journal of Machine Learning Research*, [online] 12, pp.2825–2830. Available at: <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/>.

Tosi, S. (2009). *Matplotlib for Python Developers Build remarkable publication quality plots the easy way*. [online] Available at: [https://d1wqtxts1xzle7.cloudfront.net/Q\_\_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA](https://d1wqtxts1xzle7.cloudfront.net/32168807/Matplotlib_for_Python_Developers_%282009%29.pdf20131027-19008-5fvdfp-libre-libre.pdf?1382866688=&response-content-disposition=inline%3B+filename%3DMatplotlib_for_Python_Developers_2009.pdf&Expires=1733777933&Signature=ax5Kr39kbPXrB76XQXdIuqtF-0CZ31tI2tjRABhD32pC3zKux~cXDdwTSHTnTI5sM-LL3O3eP2SfJKIvV8hauSNG-vkgWf609TlU7~H0qL7BYIS-ES0pU3sJroA36BzBAaAnp49ymzyAAuVFMgMIvTkSFXR4smY7KGuAiAbPu5Or8TPevw2nhshVbZ4DBea5B~V~~t1LQsG5PmP09SwnXZYBrk7jwK0ioZfM~A9v1sg5ZJB2EP9X231FnznZQ00UxDfnaqREJfsiZj96Lna3~SCzSIoFc2cnOALpJKv~JKlv58B5qcdkLVkwRXdmrVBlj7tjLsiGPiiOae0H~DiDiQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)

Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, [online] 6(60), p.3021. <https://doi.org/10.21105/joss.03021>.