**Project Report on**

**FINANCIAL RISK ASSESSMENT ANALYSIS**

In partial fulfilment for the award   
Of  
**Professional Certification in Data Analysis and Visualization**Year 2024-2025  
Submitted By **Suhaila MP**

Tools & Technologies Used:

**Python, R Programming, Tableau**

Duration:

**15/07/2025 – 30/07/2025**

Submission Date:

**31/07/2025**  
  
**G - TEC CENTRE OF EXCELLENCE PERINTHALMANNA**

**ABSTRACT**

This project, titled Financial Risk Assessment, focuses on identifying and analyzing key factors that contribute to financial risk among loan applicants. Using a dataset of 15,000 individuals, the study employs a mixed-method approach integrating Python for data cleaning and exploratory data analysis (EDA), R programming for statistical hypothesis testing, and Tableau for interactive visualization. The analysis explores how variables such as employment status, credit score, income level, loan purpose, age group, and geographic location influence an individual's risk rating. Key insights reveal that unemployed applicants, certain loan purposes like business and personal, and specific age brackets (40–49) are more prone to high financial risk. Statistically significant differences were found using T-tests, ANOVA, Z-tests, and Chi-square tests. The Tableau dashboard presents these findings in a visually intuitive format for financial institutions and stakeholders. This project demonstrates how a data-driven approach can enhance decision-making in risk evaluation and support more responsible and strategic lending practices.

**INTRODUCTION**

In today’s financial landscape, accurately assessing the risk associated with loan applicants has become more important than ever. With increasing numbers of individuals applying for credit, banks and financial institutions face the challenge of identifying which applicants are most likely to default. This project, titled Financial Risk Assessment, aims to analyze various factors that contribute to an applicant’s financial risk profile using a comprehensive dataset of 15,000 individuals. The focus of this project is on key variables such as employment status, credit score, income level, loan amount, loan purpose, age group, and geographic location. These factors are explored to determine their impact on a person’s risk rating. A mixed-method approach was used for the analysis: Python was applied for data cleaning and exploratory data analysis, R programming was used for performing statistical tests to confirm the significance of observed patterns, and Tableau was used to create interactive dashboards that visualize the results in an accessible way. Together, these tools allowed for a deep understanding of financial behavior and risk patterns. By integrating technical analysis with visual storytelling, this project provides a complete, data-driven view of financial risk. The insights gained can support lending institutions in making more informed, responsible, and strategic credit decisions.

**LITERATURE REVIEW**

**1.Credit Score and Financial Risk**

Credit scores are one of the most important indicators of financial risk. Studies show that individuals with lower credit scores are more likely to default on loans. Many financial institutions rely on credit score models to make lending decisions (Thomas et al., 2002).

**2.Employment Status and Loan Risk**

Research has shown that employment stability has a direct impact on loan repayment. People who are unemployed or have irregular jobs often face difficulties repaying loans, which increases their risk rating (Altman & Sabato, 2007).

**3.Purpose of Loan and Risk**

The type of loan someone applies for can also indicate their risk level. Business and personal loans tend to be riskier compared to home or auto loans. This is because they are often unsecured and depend on variable factors like business success (Khandani et al., 2010).

**4.Income and Debt Burden**

Income level and debt-to-income ratio are strong predictors of financial risk. If a person’s monthly debt is too high compared to their income, they are more likely to default. This is especially important during times of economic uncertainty (Basel Committee, 2019).

**5.Age Group and Region-Based Risk**

Studies have identified differences in financial behaviour based on age and location. Some regions may have higher unemployment or cost of living, which affects repayment ability. Similarly, certain age groups, especially middle-aged individuals, carry more financial responsibility and therefore more risk (Crook & Banasik, 2012).

**6.Role of Data Analytics in Risk Prediction**

Today, data science tools like Python, R, and Tableau are widely used in risk modelling. These tools help clean data, identify patterns, and visualize risk factors effectively. Combining statistical tests with visual dashboards improves understanding and decision-making (Lessmann et al., 2015).

**7. Credit History and Long-Term Behavior**

Borrowers with consistent, responsible credit behavior over time—such as timely payments and low credit utilization—are generally lower risk. Historical patterns provide better prediction than short-term financial snapshots (Avery et al., 2004).

**8. Loan Tenure and Risk Exposure**

Longer repayment periods may lower monthly payments, but they expose lenders to a longer risk window. Bellotti & Crook (2013) found that loan tenure length has a direct impact on default probability due to long-term financial uncertainty.

**RESEARCH GAP**

Many previous studies on financial risk assessment only focus on one or two factors, like credit score or income. They often use a single method such as basic statistics or machine learning, which limits the depth of analysis. These studies also rarely combine tools for better understanding or communication. As a result, important patterns may be missed, and findings are not always clear to decision-makers. This project addresses the gap by using three tools together. Python was used for data cleaning and exploration. R was applied for statistical testing. Tableau was used to create visual dashboards. By combining these tools, the project provides a broader, more practical view of financial risk. This approach supports more informed and accurate lending decisions.

**DATACOLLECTION & PREPROCESSING**

**Data Source and Collection Methods**

The dataset used in this project, financial\_risk\_assessment.csv, was sourced from Kaggle, a widely used open data platform. It contains records of 15,000 individuals, including various features such as employment status, income, credit score, loan amount, loan purpose, number of dependents, and risk rating. The data reflects both personal and financial factors used to evaluate an applicant’s credit risk.

**Data Quality Assessment and Cleaning Procedures**

Initial inspection of dataset was performed using Python, primarily with the pandas and NumPy libraries. The following steps were taken:

* Missing values: Detected and handled by imputing basic values where appropriate.
* Duplicate records: were checked and removed to maintain integrity.
* Data types: Ensured each column had the correct type for further processing (e.g., numerical ratings as float, categorical as factors).

**Feature Engineering and Selection Techniques**

Several new features were created to support deeper analysis and testing:

* **Age Group:** Age values were grouped into categories (18–24, 25–34, 35–44, 45–54, 55–64, and 65–74) to observe risk patterns across age brackets.
* **Credit\_Score\_Category**: Credit scores were grouped into four categories—Excellent (750+), Good (700–749), Fair (650–699), and Poor (below 650)—to simplify analysis and improve visualization.
* **Risk Categories:** Focused analyses were conducted on filtered subsets like "High" and "Low" risk groups for T-tests and proportion analysis.

**Selected Columns for Analysis:**

Employment. Status, Loan. Purpose, Income, Credit. Score,Credit\_Score\_Category, Risk. Rating, Age, Age group, Gender, Assets Value, Loan. Amount, and Previous. Defaults. These variables were central to statistical testing and visualizations performed in R and Tableau.

**METHODOLOGY**

This project follows a multi-tool approach using Python, R programming, and Tableau to analyze financial risk among loan applicants. The methodology involves data preparation, feature engineering, statistical testing, and interactive visualization. This streamlined process ensures accurate, comprehensive insights to support better lending decisions.

**Tools and Technologies Used**

* Python was used for data cleaning, transformation, and exploratory analysis. Key libraries like pandas, matplotlib, seaborn, plotly express enabled efficient data handling, visualization, and the identification of patterns.
* R programming was utilized for statistical hypothesis testing to validate the patterns discovered during EDA. Key packages such as t.test(), aov(), and chisq.test() helped in performing parametric and non-parametric tests.
* Tableau was used for building interactive dashboards and data visualization, presenting the findings to non-technical stakeholders in a clear and intuitive format.

**Exploratory Data Analysis – Python**

After cleaning the data, exploratory analysis was performed using libraries like matplotlib and seaborn to visualize trends and relationships between key variables. These insights helped guide the selection of relevant statistical tests in R for deeper analysis**.**

**Libraries and Tools Used**

Key Python libraries included:

* **pandas** for data manipulation and grouping
* **matplotlib** and **seaborn** for visualizations
* **NumPy** for statistical calculations and feature engineering

**Visualizations**:

* Risk Rating Distribution (Bar chart)

Shows how applicants are distributed across Low, Medium, and High-risk levels.

* Credit score by distribution (Boxplot)

Displays the spread of credit scores and how they relate to financial risk categories.

* Employment Status by Risk Rating by (Stacked bar chart)

Highlights the impact of employment status on risk, with unemployed applicants showing higher risk.

* Loan Purpose Distribution in High-Risk Group (pie chart)

Shows that personal and auto loans are more common among high-risk applicants.

* Loan Amount by Risk Rating and Purpose (Tree map)

Helps identify risk trends across different loan purposes and sizes.

* Risk Rating vs. Years at Current Job (Bar chart)

Indicates that longer job duration tends to be associated with lower risk.

* Average Income by Educational level and Risk Rating (Heatmap)

Reveals that higher education does not always mean lower financial risk.

**Group by and Crosstab:**

* **Group by()**: Used to calculate averages and counts across variables such as Employment Status, Loan Purpose, Age Group, and Risk Rating. For instance, it helped determine the average loan amount by employment status, and the mean income or debt-to-income ratio within each risk category.
* **crosstab ()**: Applied to examine relationships between two categorical variables —such as Employment Status vs Risk Rating and Loan Purpose vs Credit Score Category— helping identify financial behavior patterns and risk profiles among applicant groups.

**Statistical Testing – R Programming**

To validate patterns observed during the exploratory analysis, various statistical tests were performed using R. These tests helped confirm whether differences or relationships between variables were statistically significant:

* **T-test**: Used to examine whether the average asset value differs significantly between employed and unemployed individuals. The results confirmed if employment status influences asset accumulation.
* **Z-test:** Applied to compare the proportion of individuals with previous defaults between high-risk and low-risk groups, revealing whether default history varies by risk level.
* **F-test:** Conducted to check for differences in the variance of loan amounts between male and female applicants. This test assessed whether gender influences variability in the loan amount distribution.
* **ANOVA (Analysis of Variance):** A one-way ANOVA test was performed to examine whether income levels differ significantly across risk rating groups (Low, Medium, High). This test helped identify whether individuals in different risk categories tend to have varying income patterns.
* **Chi-Square Test:** Assessed the relationship between education level and employment status, determining if educational background significantly impacts employment conditions.

.

**Data Visualization and Dashboarding - Tableau**

After completing the data preparation and analysis in Python, Tableau was used to create an interactive dashboard for the Financial Risk Assessment project. The dashboard helps visualize key insights related to credit score, loan purpose, employment status, age group, and state-wise risk distribution.

* **Key Metrics Displayed**:  
  The top panel shows essential summary statistics, including total applicants (15,000), average credit score (699.1), percentage of high-risk individuals (10%), and average loan amount (₹27,450).
* **Dynamic Filtering**:

The dashboard allows users to filter by variables such as credit score group, loan purpose, employment status, and state. This makes it easy to explore financial risk patterns across different applicant segments.

* **Visual Storytelling**:  
  Clear charts like bar graphs, pie charts, and maps were used to highlight important patterns—for example, how unemployment links to higher risk, or how credit score levels affect risk rating. The visuals are designed for clarity, helping even non-technical users grasp complex insights at a glance.

**Interactive Charts Used**:

* **Risk Rating by Credit Score**: Shows how lower credit scores are associated with higher financial risk.
* **Loan Purpose Distribution**: A pie chart displays the proportion of applicants taking personal, business, auto, and home loans.
* **Employment Status vs Risk Rating**: A grouped bar chart compares how employment types (employed, self-employed, unemployed) relate to different risk levels.
* **Average Loan Amount by Employment Status**: Highlights how loan size varies across job categories.
* **Risk Rating by Age Group**: Indicates that age groups like 40–49 and 20–29 have higher proportions of high-risk individuals.

**RESULTS AND ANALYSIS**

This section explains the main findings of the project. It looks at how factors like job status, credit score, loan purpose, income, and location affect the financial risk of loan applicants. Using Python analysis and Tableau dashboards, clear patterns were found that help identify high-risk individuals. These insights can support better lending decisions.

**Python- Based Results**

* Most applicants were classified as Low Risk, followed by Medium and a smaller portion in the High-Risk category.
* Applicants with higher credit scores (Good and Excellent) were mostly in the Low-Risk group. Those with scores below 700 were more commonly associated with Medium or High Risk.
* Unemployed individuals showed the highest proportion of High Risk, while Employed and Self-employed applicants leaned toward Low and Medium Risk.
* Personal, Auto, and Business loans were more frequent among high-risk applicants, indicating that loan purpose plays a key role in risk profiling.
* Higher education did not always correlate with lower risk. In some cases, even graduates showed elevated financial risk, especially when paired with lower incomes.
* Larger loan amounts for business or personal use were more likely to be tagged as risky compared to home or auto loans.
* Applicants with more years at their current job generally showed lower risk, highlighting the role of employment stability.
* The 40–49 age group had the highest concentration of High Risk, followed by 50–59, suggesting financial strain in middle-aged segments.
* Most applicants fell into the Fair or Good credit score categories, with relatively fewer in the Excellent range.
* The filled map chart showed certain states had a higher share of high-risk individuals, useful for regionallending strategies.

**R-Based Statistical Results**

* T-Test:  
  Compared Assets Value between *Employed* and *Unemployed* individuals.  
  Result: Statistically significant difference found (p < 0.05).  
  Interpretation: Employment status influences asset value, likely due to income stability.
* Z-Test:  
  Tested if the proportion of individuals with previous defaults differs between *High* and *Low* risk groups.  
  Result: Statistically significant difference found (p ≈ 0.0442).  
   Interpretation: Default history is associated with risk rating.
* F-Test:  
  Compared variance in Loan Amount between *Male* and *Female* applicants.  
   Result: No significant difference in variance (p = 0.9236).  
  Interpretation: Gender does not affect variability in loan amounts.
* ANOVA:  
  Analyzed whether Income differs across *Risk Ratings* (Low, Medium, High).  
  Result: Statistically significant difference (p = 0.0495).  
   Interpretation: Risk category is related to income levels.
* Chi-Square Test:  
  Assessed relationship between Education Level and Employment Status.  
   Result: No significant association found (p > 0.05).  
   Interpretation: Educational background does not strongly influence employment status in this dataset.

**Tableau- Based Results**

* The dataset contains information on 15,000 loan applicants across various employment, demographic, and financial backgrounds.
* The average credit score among applicants is 699.1, placing many individuals near the boundary between fair and good credit.
* Approximately 10% of applicants fall into the high-risk category, indicating a smaller but critical group that requires close monitoring.
* The average loan amount requested is $27,450, with only slight variation across employment types.
* Applicants with lower credit scores are significantly more likely to be classified as high risk, confirming credit score as a strong risk predictor.
* Loan purpose distribution is balanced among key categories:
  + Personal Loans: 25.1%
  + Home Loans: 25.1%
  + Auto Loans: 24.8%
  + Business Loans: 24.9%
  + These amounts are relatively similar, but unemployed individuals carry higher risk with comparable loan volumes.
* Unemployed individuals represent the highest proportion of high-risk applicants, followed by self-employed and then employed individuals.
* The 40–49 age group shows an alarming 100% high-risk classification. Other age groups follow a pattern of increasing risk with age:
  + 60–69: 75% high risk
  + 50–59: 50% high risk
  + 20–29: 25% high risk
* Certain U.S. states show 100% of applicants labeled as high risk, indicating potential regional financial stress or economic vulnerability.

**CONCLUSION**

This project offered a detailed analysis of financial risk patterns among loan applicants by combining Python, R programming, and Tableau. With data from 15,000 individuals, the study explored how factors like employment status, credit score, income, loan purpose, age group, and geographic location influence financial risk classification.

The findings revealed that applicants with low credit scores and unemployed individuals are more likely to be categorized as high risk. Interestingly, personal and auto loans were more common among high-risk groups, while business and home loans showed more balanced risk levels. Age-wise, the 40–49 group showed the highest risk, while younger applicants were generally lower risk. Additionally, some U.S. states had significantly higher risk distributions, suggesting possible regional economic challenges.

Statistical tests (T-test, Z-test, F-test, ANOVA, and Chi-square) confirmed the significance of these patterns, while Tableau dashboards translated the insights into visual stories for easier interpretation. The credit score categorization and age group segmentation supported clearer analysis and better communication of results.

Overall, this project highlights the value of a multi-tool approach in financial analytics. It not only identifies who is at risk but also uncovers why, enabling financial institutions to make more accurate and responsible lending decisions. The approach used here is scalable and adaptable to broader financial datasets or policy frameworks.

**FUTURE WORKS**

* Build Prediction Models:

Use machine learning to predict if a person is high, medium, or low risk based on their data.

* Use Real-Time Data:

Connect the system to live financial data (like credit score updates or bank records) for better accuracy.

* Create a Web Dashboard:

Make an online dashboard using tools like Streamlit or Dash so banks can view and filter risk profiles easily.

* Collect More Data:

Include more people from different places and time periods to get a broader view of financial risk patterns.

* Group Similar Applicants:

Use clustering techniques to group people by financial behaviour. This helps in giving more personalized loan decisions**.**

* Automate Reports:

Set up automatic weekly or monthly reports showing trends in high-risk applicants and loan defaults.

**REFERENCES**

* Altman, E. I., & Sabato, G. (2007). *Modeling credit risk for SMEs: Evidence from the U.S. market*. Abacus, 43(3), 332–357. <https://doi.org/10.1111/j.1467-6281.2007.00234.x>
* Avery, R. B., Bostic, R. W., Calem, P. S., & Canner, G. B. (2004). *An overview of consumer data and credit reporting*. Federal Reserve Bulletin, 89(2), 47–73. <https://www.federalreserve.gov/pubs/bulletin/2003/0203lead.pdf>
* Basel Committee on Banking Supervision. (2019). *Credit risk and expected credit loss*. Bank for International Settlements. <https://www.bis.org/bcbs/publ/d505.pdf>
* Bellotti, T., & Crook, J. (2013). *Forecasting and stress testing credit card default using dynamic models*. International Journal of Forecasting, 29(4), 563–574. <https://doi.org/10.1016/j.ijforecast.2012.10.005>
* Crook, J. N., & Banasik, J. (2012). *Measuring and explaining the marginal contribution of credit scorecards*. Journal of the Operational Research Society, 63(6), 782–789. https://doi.org/10.1057/jors.2011.88
* Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). *Consumer credit-risk models via machine-learning algorithms*. Journal of Banking & Finance, 34(11), 2767–2787. <https://doi.org/10.1016/j.jbankfin.2010.06.001>

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>

* homas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit scoring and its applications. SIAM Monographs on Mathematical Modeling and Computation. <https://epubs.siam.org/doi/book/10.1137/1.9780898718072>