

Financial Risk for Loan Approval

Ivan Filipchuk DS-1

ivan.filipchuk@student.pk.edu.pl
153461

Yurii Maisuradze DS-2

yurii.maisuradze@student.pk.edu.pl
153464

18 listopada 2024

1 Abstract

2 Wprowadzenie

Celem niniejszej pracy jest opracowanie modelu regresyjnego, który na podstawie dostępnych danych o wnioskodawcach będzie przewidywał Risk Score – wskaźnik ryzyka kredytowego. W oparciu o zmienne takie jak wiek, dochód, scoring kredytowy, historia zadłużenia, czy też wartość majątku, model ten pozwoli określić indywidualne ryzyko finansowe każdego wnioskodawcy. Przewidywany wynik może być następnie wykorzystany do oceny wiarygodności kredytowej oraz podjęcia decyzji o przyznaniu bądź odrzuceniu wniosku kredytowego.

3 Zbiór danych

3.1 Opis zbioru danych

Zbiór danych Financial Risk for Loan Approval, źródło: <https://www.kaggle.com/datasets/lorenzozoppeletto/financial-risk-for-loan-approval/data?select=Loan.csv>, zawiera informacje o wnioskach o pożyczki oraz szczegółowe dane dotyczące sytuacji finansowej i kredytowej wnioskodawców. Dane te mogą być wykorzystane do analizy ryzyka kredytowego, przewidywania, czy pożyczka zostanie zatwierdzona, oraz oceny czynników wpływających na decyzje o przyznaniu pożyczki.

ApplicationDate - Data złożenia wniosku o pożyczkę.

Age - Wiek wnioskodawcy.

AnnualIncome - Roczny dochód wnioskodawcy.

CreditScore - Wynik oceny zdolności kredytowej.

EmploymentStatus - Status zatrudnienia.

EducationLevel - Najwyższy poziom wykształcenia.

Experience - Długość doświadczenia zawodowego.

LoanAmount - Kwota żądanej pożyczki.

LoanDuration - Okres spłaty pożyczki (w miesiącach).

MaritalStatus - Stan cywilny wnioskodawcy.

NumberOfDependents - Liczba osób na utrzymaniu.

HomeOwnershipStatus - Status posiadania nieruchomości.

MonthlyDebtPayments - Miesięczne zobowiązania dłużne.

CreditCardUtilizationRate - Wskaźnik wykorzystania karty kredytowej.

NumberOfOpenCreditLines - Liczba aktywnych linii kredytowych.

NumberOfCreditInquiries - Liczba zapytań kredytowych.

DebtToIncomeRatio - Proporcja długu do dochodów.

BankruptcyHistory - Historia upadłości.

LoanPurpose - Cel pożyczki.

PreviousLoanDefaults - Informacje o wcześniejszych zaległościach w spłacie pożyczek.

PaymentHistory - Historia spłat.

LengthOfCreditHistory - Długość historii kredytowej.

SavingsAccountBalance - Stan konta oszczędnościowego.

CheckingAccountBalance - Stan konta bieżącego.

TotalAssets - Calkowita wartość aktywów.

TotalLiabilities - Całkowita wartość zobowiązań.

MonthlyIncome - Miesięczny dochód.

UtilityBillsPaymentHistory - Historia płatności rachunków za media.

JobTenure - Staż pracy w obecnym miejscu zatrudnienia.

NetWorth - Wartość netto majątku.

BaseInterestRate - Podstawowa stopa procentowa.

InterestRate - Zastosowana stopa procentowa.

MonthlyLoanPayment - Miesięczna rata pożyczki.

TotalDebtToIncomeRatio - Całkowity dług w stosunku do dochodu.

LoanApproved - Status zatwierdzenia pożyczki (tak/nie).

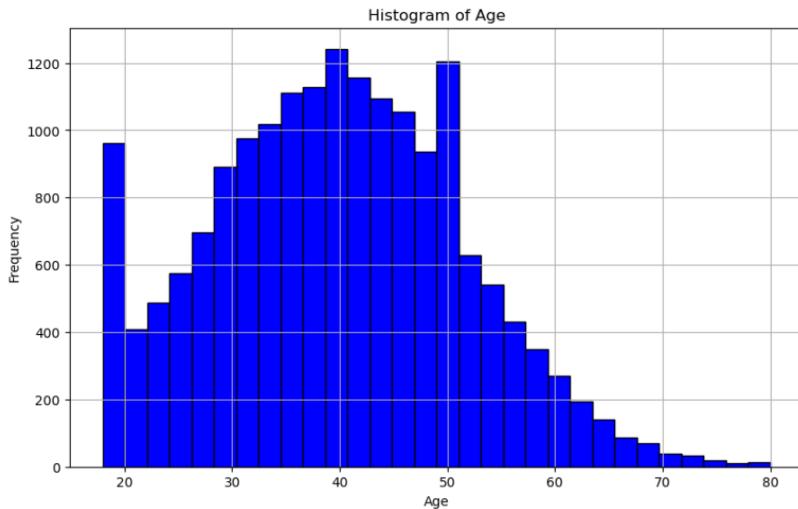
RiskScore - Wynik oceny ryzyka.

Zbiór danych jest reprezentowany poniżej:

	ApplicationDate	Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	MaritalStatus
0	2018-01-01	45.0	39948.0	617.0	Employed	Master	22.0	13152.0	NaN	Married
1	2018-01-02	38.0	39709.0	628.0	Employed	Associate	15.0	NaN	48.0	Single
2	2018-01-03	47.0	40724.0	NaN	Employed	Bachelor	26.0	17627.0	NaN	Married
3	2018-01-04	58.0	69084.0	545.0	Employed	High School	34.0	37898.0	96.0	Single
4	2018-01-05	37.0	103264.0	594.0	Employed	Associate	17.0	9184.0	36.0	Married

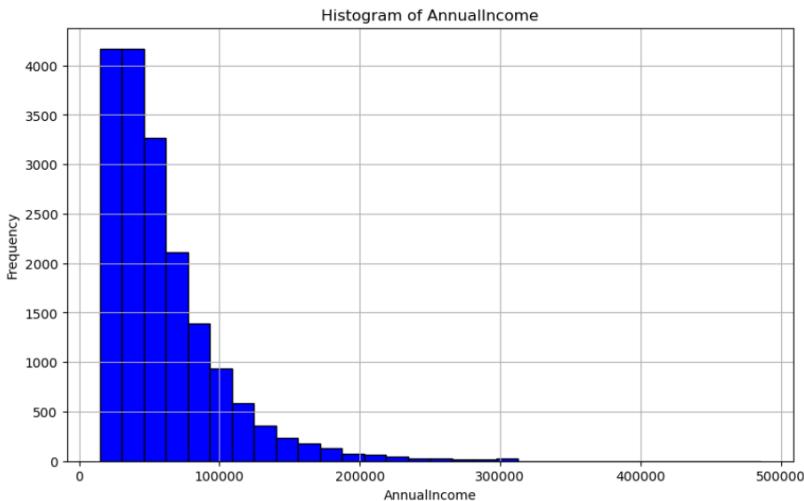
Rysunek 1: Zbiór danych Financial Risk for Loan Approval.

3.2 Histogramy cech



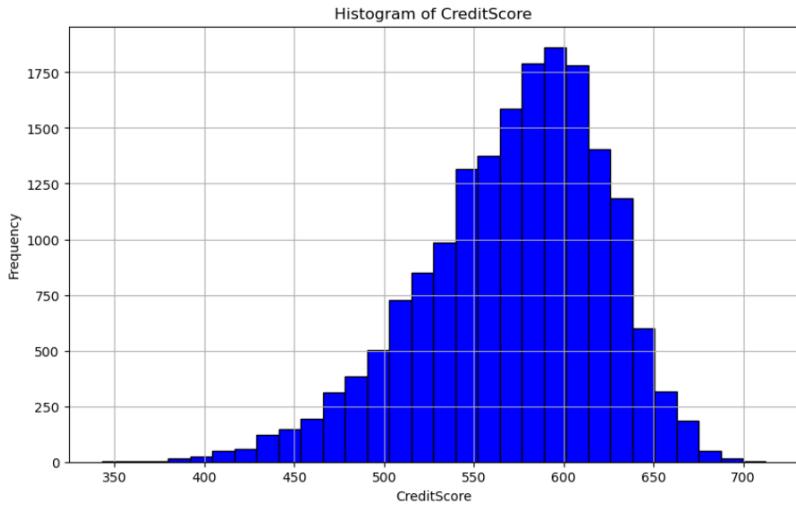
Rysunek 2: Histogram zmiennej Age

Histogram zmiennej Age wskazuje, że większość wnioskodawców ma około 40 lat, a liczba osób maleje wraz z wiekiem, szczególnie powyżej 50 lat. Rozkład jest lekko skośny w prawo, co oznacza, że starsze grupy wiekowe są reprezentowane rzadziej.



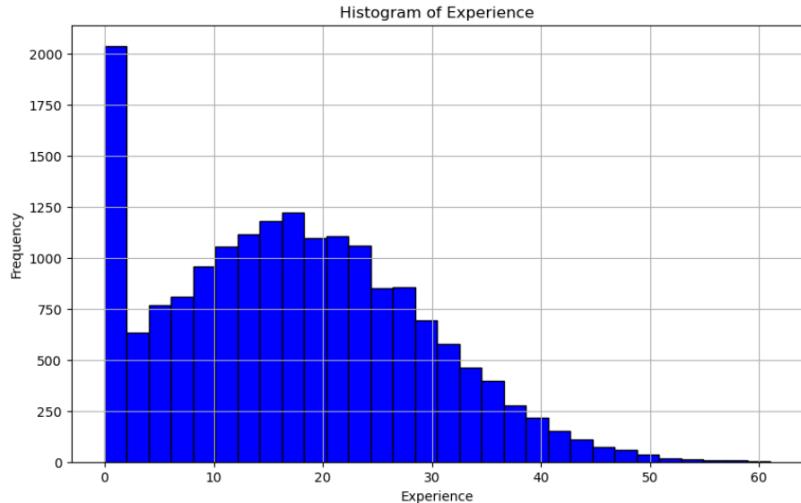
Rysunek 3: Histogram zmiennej AnnualIncome

Histogram zmiennej AnnualIncome pokazuje, że większość wnioskodawców ma roczny dochód poniżej 100,000, a liczba osób maleje wraz ze wzrostem dochodów. Rozkład jest silnie skośny w prawo, co wskazuje, że wyższe dochody są znacznie rzadsze.



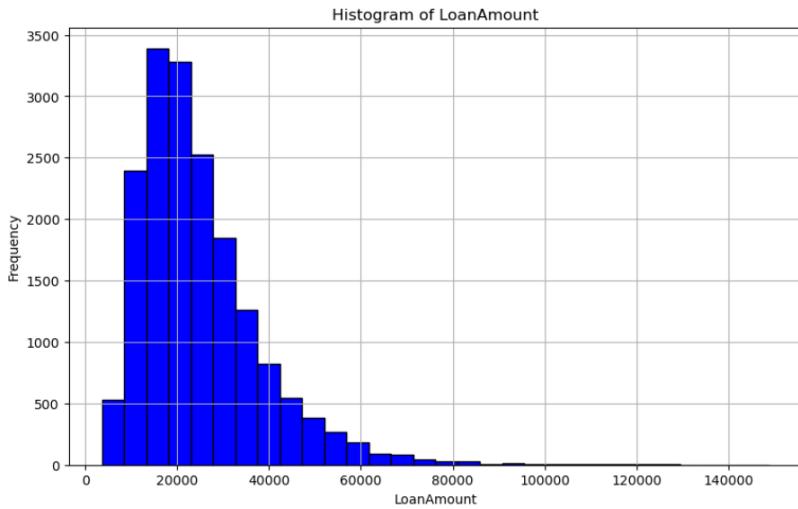
Rysunek 4: Histogram zmiennej CreditScore

Histogram zmiennej CreditScore wskazuje, że większość wnioskodawców ma wynik kredytowy w zakresie od 500 do 650, z największą koncentracją około 600. Rozkład jest zbliżony do normalnego, co oznacza, że niższe i wyższe wartości wyniku kredytowego są rzadsze.



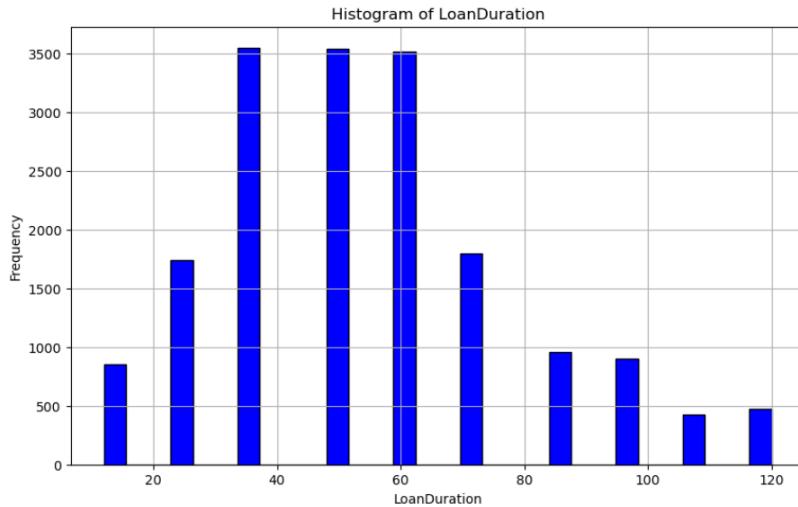
Rysunek 5: Histogram zmiennej Experience

Histogram zmiennej Experience pokazuje, że najwięcej wnioskodawców ma doświadczenie zawodowe około 20 lat, a także że sporo osób ma zerowe doświadczenie. Rozkład jest prawoskóny, z malejącą liczbą osób o dłuższym stażu pracy.



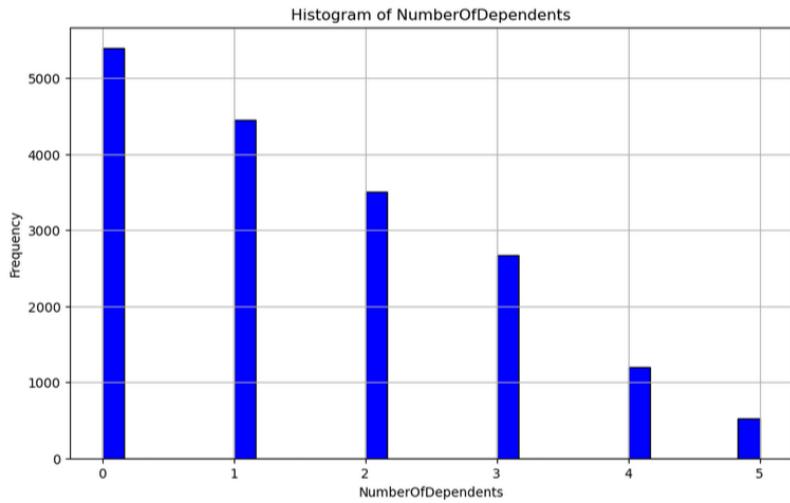
Rysunek 6: Histogram zmiennej LoanAmount

Histogram zmiennej LoanAmount wskazuje, że większość wniosków dotyczy pożyczek o kwotach poniżej 20,000. Rozkład jest skośny w prawo, co oznacza, że większe kwoty pożyczek są rzadsze.



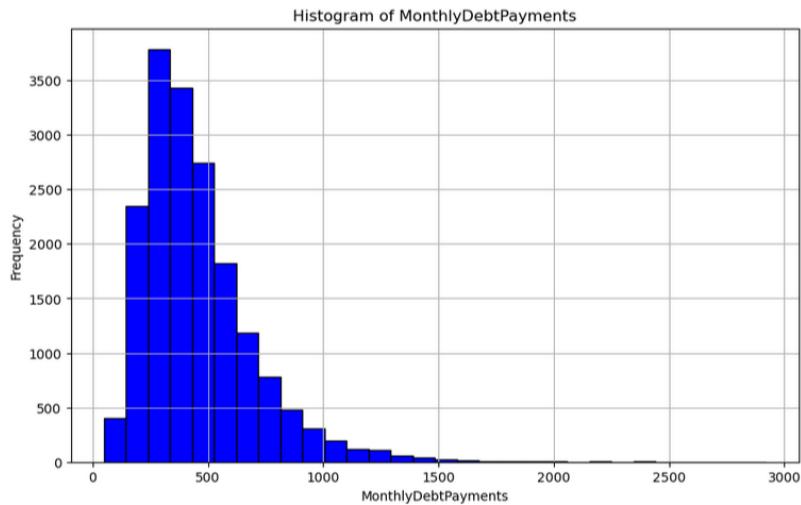
Rysunek 7: Histogram zmiennej LoanDuration

Histogram zmiennej LoanDuration pokazuje, że długości pożyczek są zgrupowane głównie wokół kilku wartości, takich jak 20, 40, 60, 80, co sugeruje, że pożyczki są często oferowane w ustalonych okresach czasu



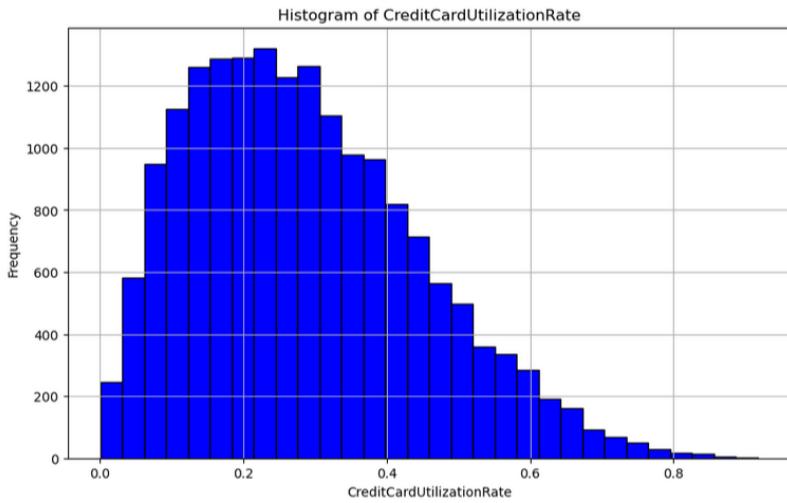
Rysunek 8: Histogram zmiennej NumberOfDependents

Histogram zmiennej NumberOfDependents pokazuje, że większość wnioskodawców ma 0 lub 1 osobę na utrzymaniu. W miarę wzrostu liczby osób na utrzymaniu, liczba wnioskodawców spada, co wskazuje, że mniejsze rodziny są bardziej powszechnie w tym zbiorze danych



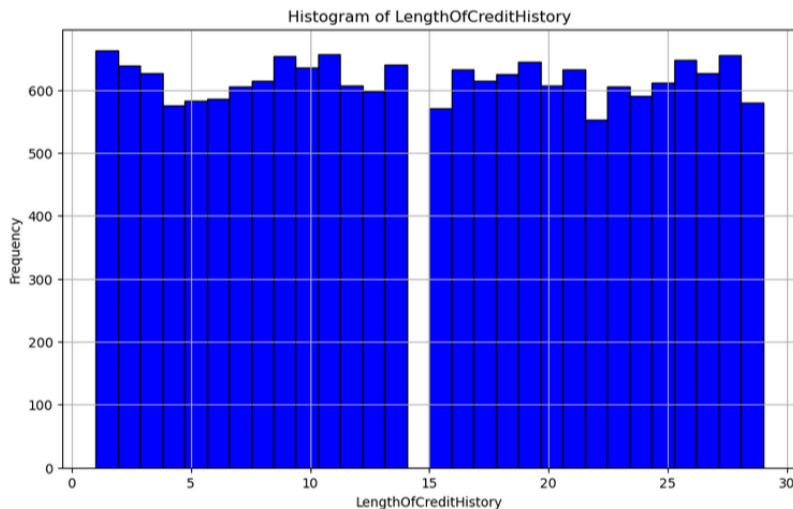
Rysunek 9: Histogram zmiennej MonthlyDebtPayments

Histogram zmiennej MonthlyDebtPayments pokazuje, że większość wnioskodawców ma miesięczne zobowiązania dłużne poniżej 500 jednostek, a liczba osób z wyższymi zobowiązaniami maleje wraz ze wzrostem wysokości płatności. Rozkład jest silnie skośny w prawo, co oznacza, że wyższe wartości miesięcznych zobowiązań występują rzadziej.



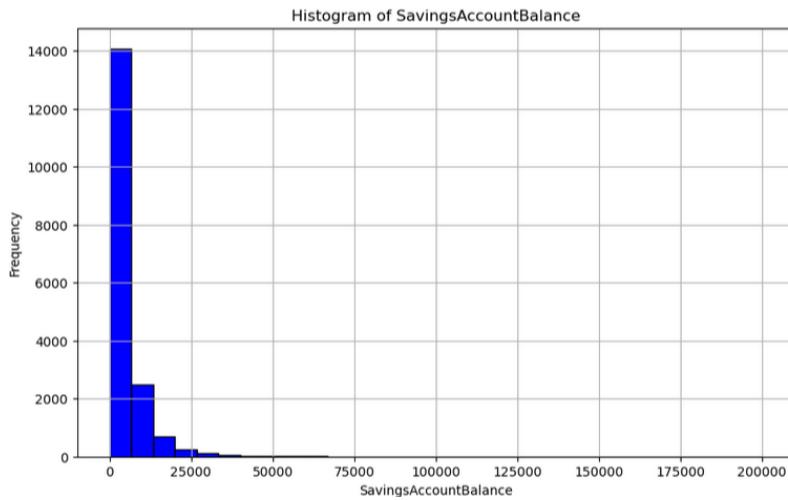
Rysunek 10: Histogram zmiennej CreditCardUtilizationRate

Histogram zmiennej CreditCardUtilizationRate pokazuje, że większość wnioskodawców ma wskaźnik wykorzystania karty kredytowej w okolicach 0.2. Rozkład jest prawoskóny, z malejącą liczbą osób przy wyższych poziomach wykorzystania karty kredytowej.



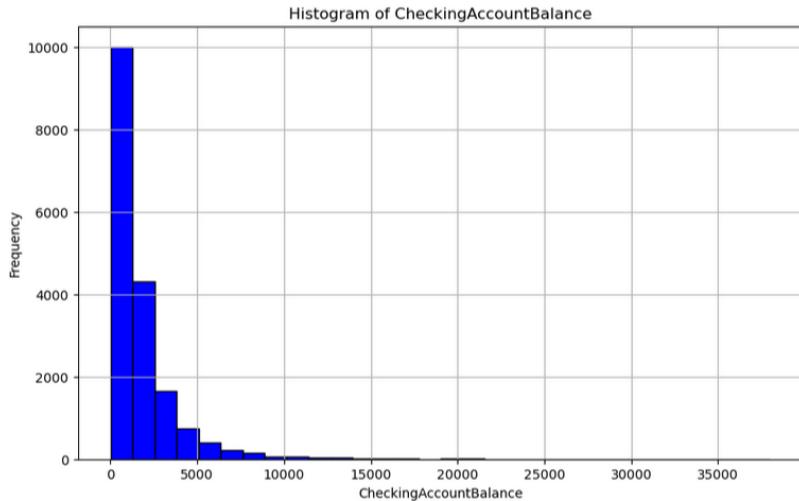
Rysunek 11: Histogram zmiennej LengthOfCreditHistory

Histogram zmiennej LengthOfCreditHistory pokazuje równomierny rozkład w zakresie od 0 do 30 lat historii kredytowej, bez wyraźnego szczytu. Brak dominującej długości historii kredytowej sugeruje, że populacja wnioskodawców ma różnorodne doświadczenia kredytowe.



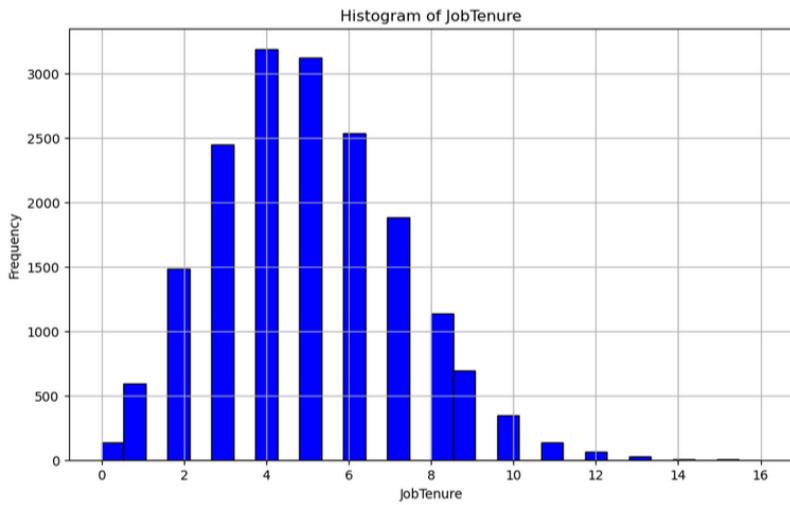
Rysunek 12: Histogram zmiennej SavingsAccountBalance

Histogram zmiennej SavingsAccountBalance pokazuje, że większość wnioskodawców ma niskie salda na koncie oszczędnościowym. Rozkład jest silnie skośny w prawo, co oznacza, że tylko niewielka liczba wnioskodawcy mają wysokie salda oszczędnościowe, a większość posiada znacznie mniejsze oszczędności.



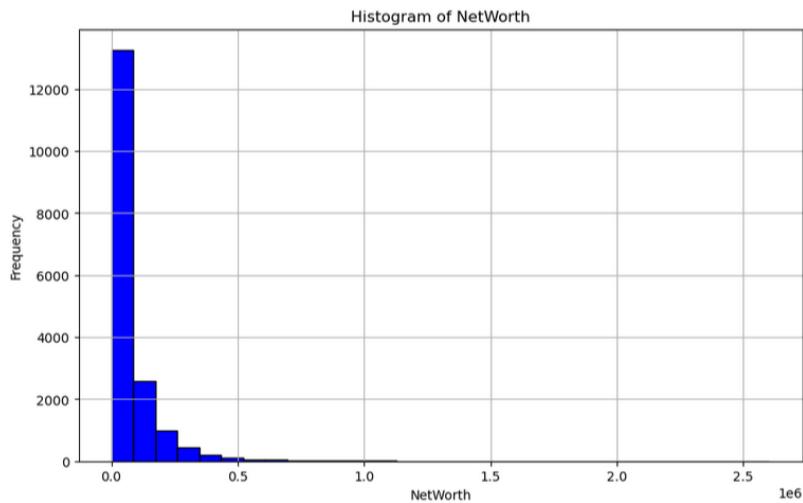
Rysunek 13: Histogram zmiennej CheckingAccountBalance

Histogram zmiennej CheckingAccountBalance pokazuje, że większość wnioskodawców ma niskie salda na koncie bieżącym. Rozkład jest silnie skośny w prawo, co oznacza, że tylko niewielka liczba osób posiada wyższe salda, podczas gdy większość ma znacznie niższe środki na koncie bieżącym.



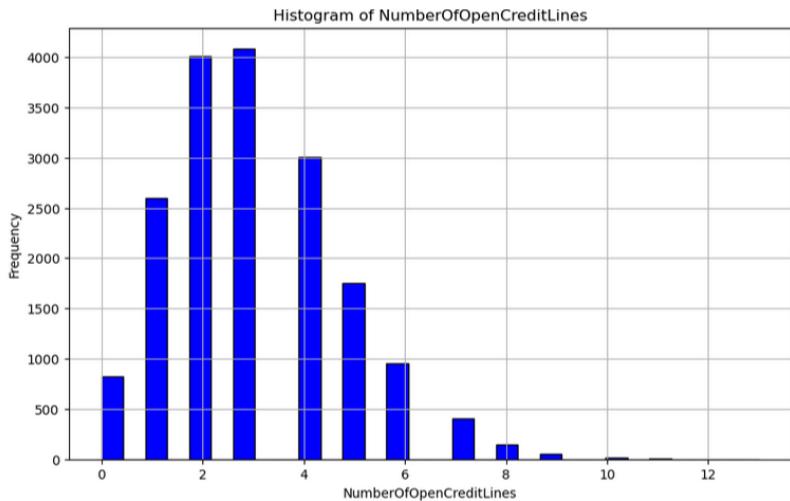
Rysunek 14: Histogram zmiennej JobTenure

Histogram zmiennej JobTenure wskazuje, że większość wnioskodawców ma staż pracy od 3 do 6 lat, z dominacją wartości w przedziale 4-6 lat. Rozkład jest skośny w prawo, co oznacza, że mniej osób ma dłuższy staż pracy, szczególnie powyżej 8 lat.



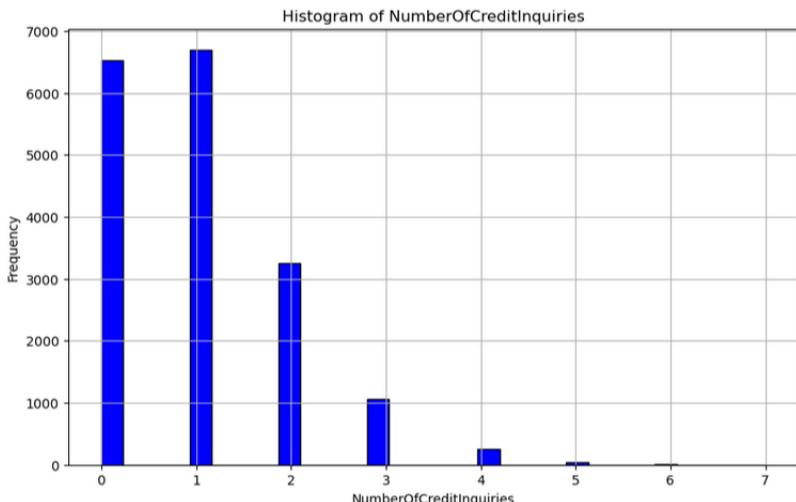
Rysunek 15: Histogram zmiennej NetWorth

Histogram zmiennej NetWorth pokazuje, że większość wnioskodawców ma niską wartość netto majątku, a liczba osób gwałtownie maleje przy wyższych wartościach majątku. Rozkład jest silnie prawoskośny, co oznacza, że wysokie wartości netto są bardzo rzadkie.



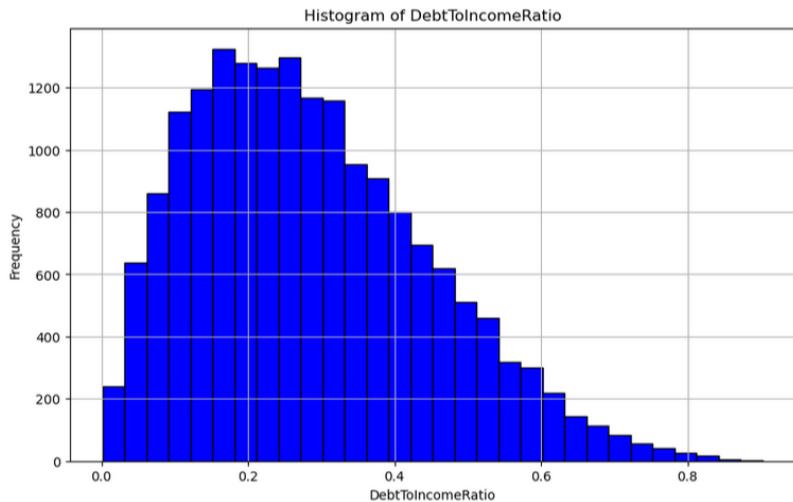
Rysunek 16: Histogram zmiennej NumberOfOpenCreditLines

Histogram zmiennej NumberOfOpenCreditLines pokazuje, że większość wnioskodawców posiada od 2 do 4 otwartej linii kredytowych, co stanowi najczęstsze wartości. Rozkład jest skośny w prawo, z mniejszą liczbą osób mających 6 lub więcej linii kredytowych.



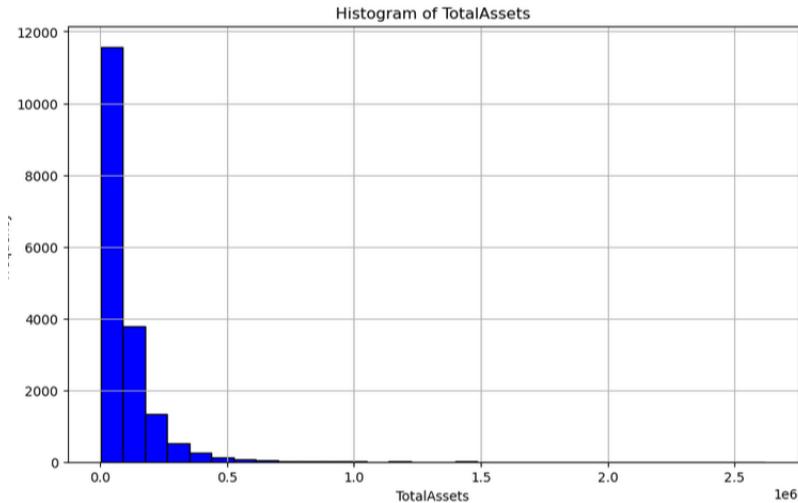
Rysunek 17: Histogram zmiennej NumberOfCreditInquiries

Histogram zmiennej NumberOfCreditInquiries pokazuje, że większość wnioskodawców miała 0 lub 1 zapytanie kredytowe, co stanowi najczęstsze wartości. Liczba zapytań gwałtownie spada wraz ze wzrostem liczby zapytań, a wartości powyżej 3 są rzadko spotykane, co sugeruje, że niewielu wnioskodawców często sprawdza swoją zdolność kredytową.



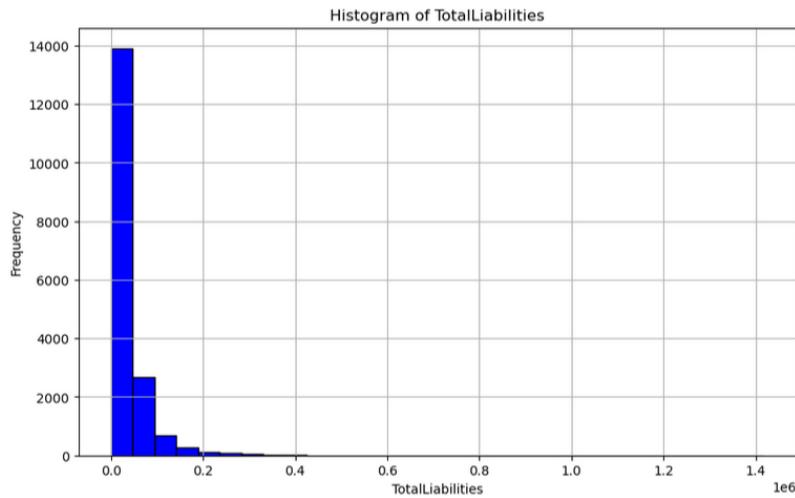
Rysunek 18: Histogram zmiennej DebtToIncomeRatio

Histogram zmiennej DebtToIncomeRatio pokazuje, że większość wnioskodawców ma wskaźnik długu do dochodu w przedziale od 0.1 do 0.4, z najwyższą częstotliwością około 0.2. Rozkład jest silnie skośny w prawo, co oznacza, że wyższe wartości wskaźnika długu do dochodu są rzadsze.



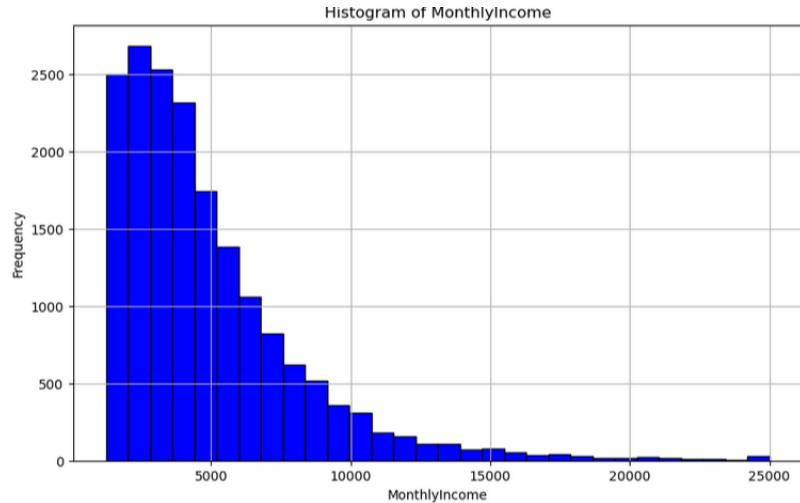
Rysunek 19: Histogram zmiennej TotalAssets

Histogram zmiennej TotalAssets pokazuje, że większość wnioskodawców posiada niską wartość aktywów, głównie poniżej 100,000. Rozkład jest silnie skośny w prawo, co oznacza, że wyższe wartości aktywów są rzadkie i występują tylko w przypadku niewielkiej liczby wnioskodawców.



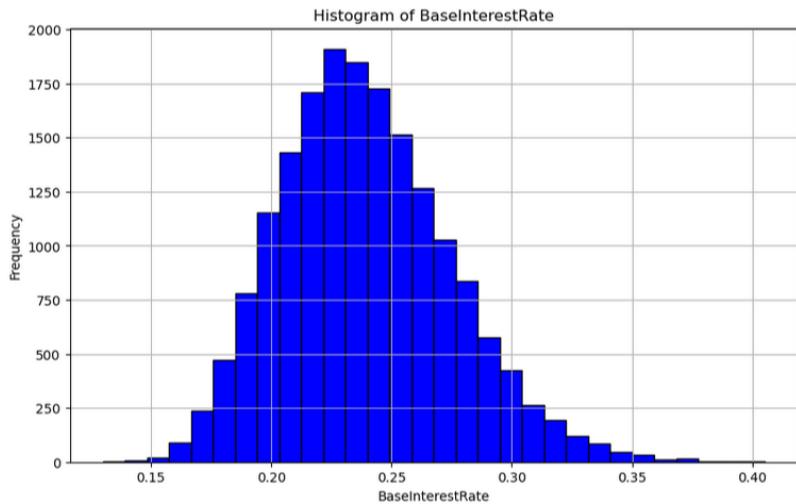
Rysunek 20: Histogram zmiennej TotalLiabilities

Histogram zmiennej TotalLiabilities wskazuje, że większość wnioskodawców ma niskie wartości zobowiązań, głównie poniżej 100,000. Rozkład jest silnie skośny w prawo, co oznacza, że wyższe wartości zobowiązań są rzadkie i dotyczą tylko niewielkiej liczby osób.



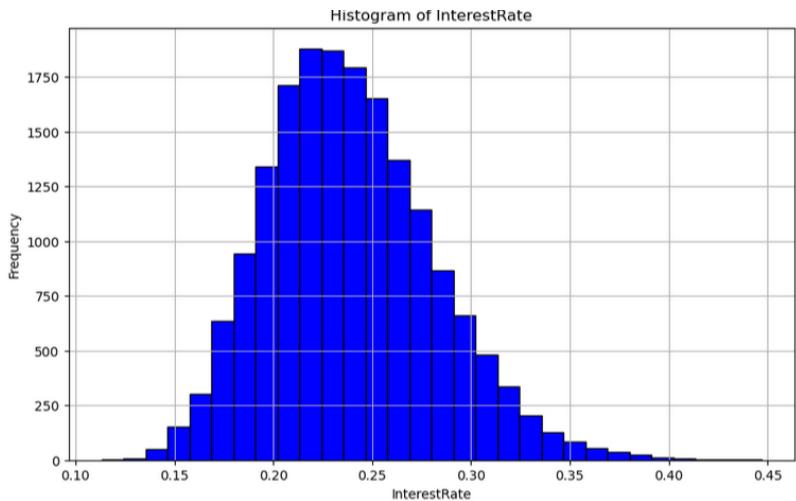
Rysunek 21: Histogram zmiennej MonthlyIncome

Histogram zmiennej BaseInterestRate pokazuje symetryczny rozkład wokół wartości około 0.25, co sugeruje rozkład normalny. Większość wartości skupia się w przedziale od 0.2 do 0.3, co wskazuje na dominację tych poziomów stopy bazowej w zbiorze danych.



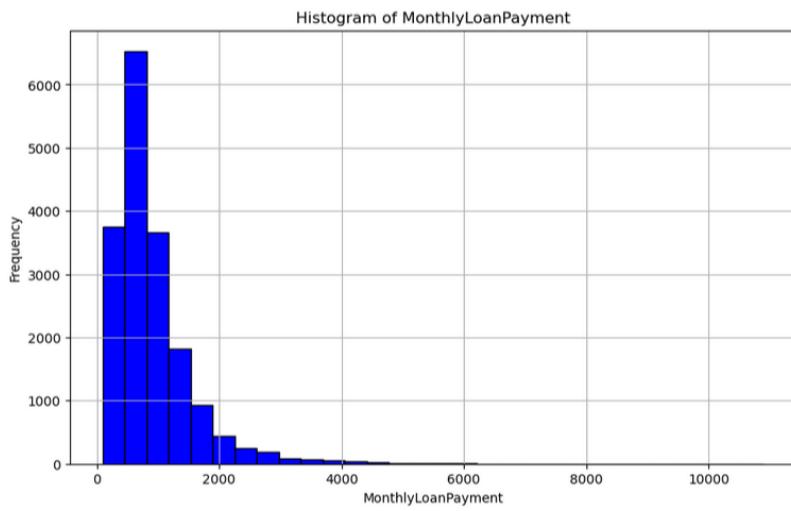
Rysunek 22: Histogram zmiennej BaseInterestRate

Histogram zmiennej BaseInterestRate pokazuje, że większość wartości jest skupiona wokół 0.25, co stanowi szczyt rozkładu. Rozkład jest symetryczny, co sugeruje, że wartości stóp procentowych są równomiernie rozmieszczone wokół średniej.



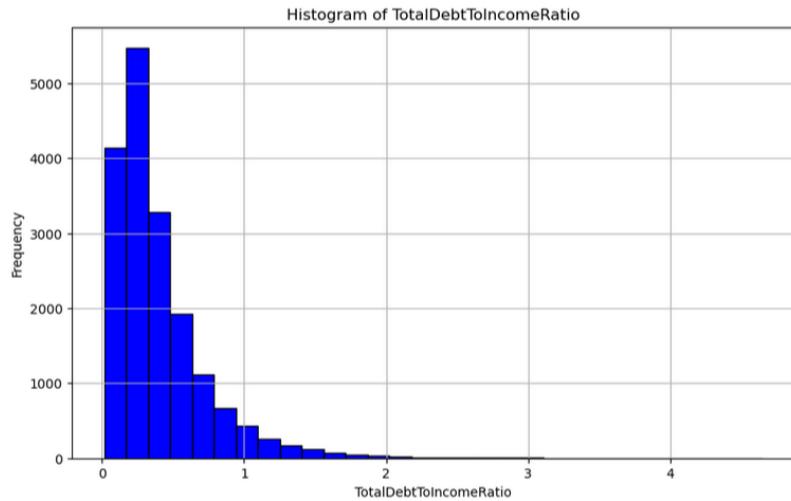
Rysunek 23: Histogram zmiennej InterestRate

Histogram zmiennej InterestRate wskazuje na symetryczny rozkład wokół wartości około 0.25, podobny do rozkładu normalnego. Większość wartości znajduje się w zakresie od 0.2 do 0.3, co oznacza, że te poziomy stóp procentowych są najczęstsze w analizowanym zbiorze danych.



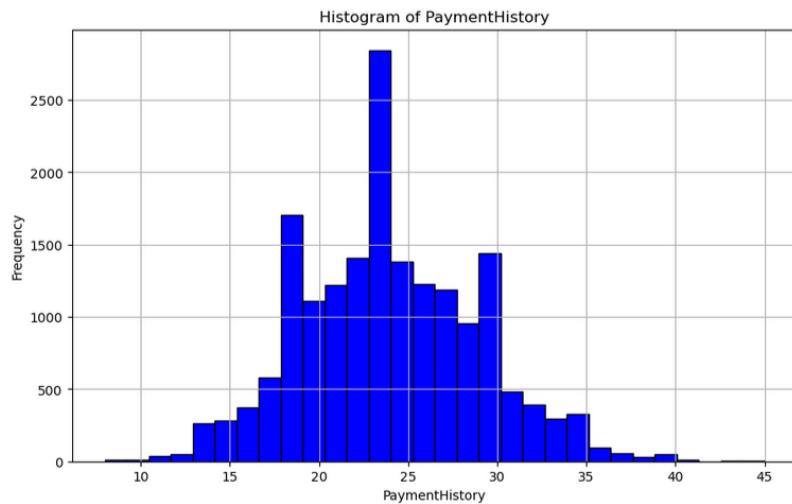
Rysunek 24: Histogram zmiennej MonthlyLoanPayment

Histogram zmiennej MonthlyLoanPayment pokazuje silnie skośny rozkład w prawo. Większość miesięcznych płatności kredytowych jest niska, z przewagą wartości poniżej 2000.



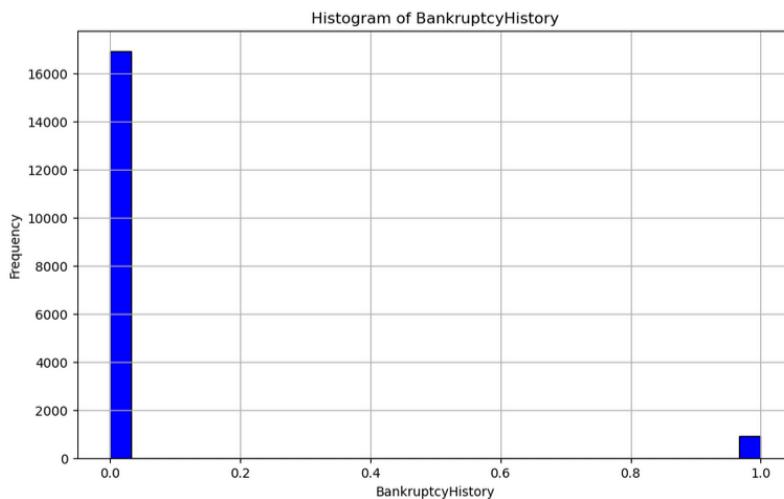
Rysunek 25: Histogram zmiennej TotalDebtToIncomeRatio

Histogram zmiennej TotalDebtToIncomeRatio pokazuje silnie skośny rozkład w prawo. Większość osób ma stosunek całkowitego zadłużenia do dochodu poniżej 1, z przewagą wartości bliskich 0, co wskazuje na niski poziom zadłużenia w stosunku do dochodu w tej grupie.



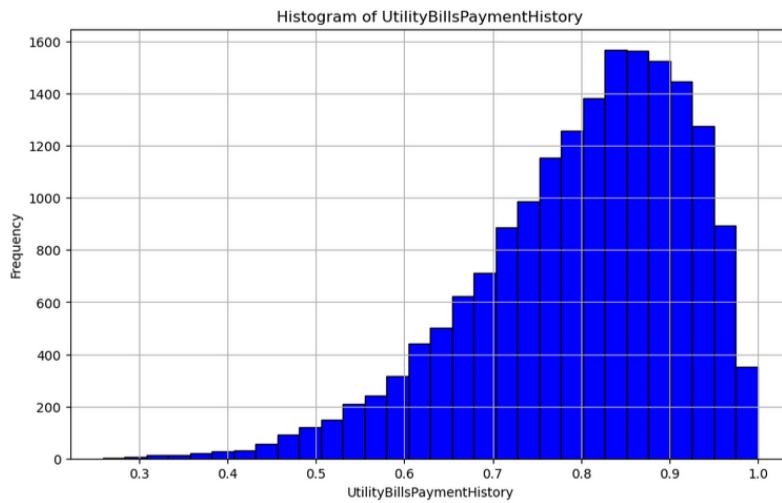
Rysunek 26: Histogram zmiennej PaymentHistory

Histogram zmiennej PaymentHistory ma rozkład zbliżony do normalnego, z wyraźnym szczytem wokół wartości 25. Wartości w tej zmiennej są skoncentrowane w zakresie od 20 do 30, co może wskazywać na typowe wartości historii płatności dla większości osób. Rozkład ma również kilka mniejszych szczytów, co może świadczyć o pewnych grupach lub okresach, które wyróżniają się częstotliwością występowania.



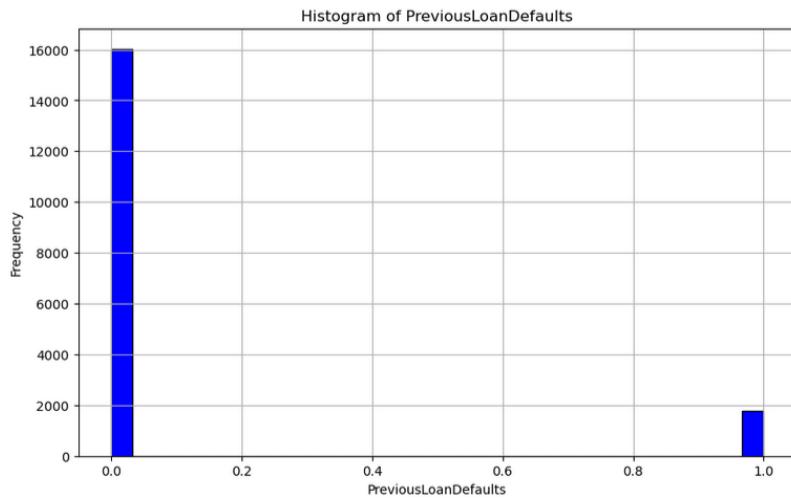
Rysunek 27: Histogram zmiennej BankruptcyHistory

Histogram zmiennej BankruptcyHistory pokazuje, że zdecydowana większość wnioskodawców nie ma historii bankructwa (wartość 0), natomiast niewielki odsetek posiada taką historię (wartość 1). Wskazuje to na rzadkość występowania bankructwa wśród badanych.



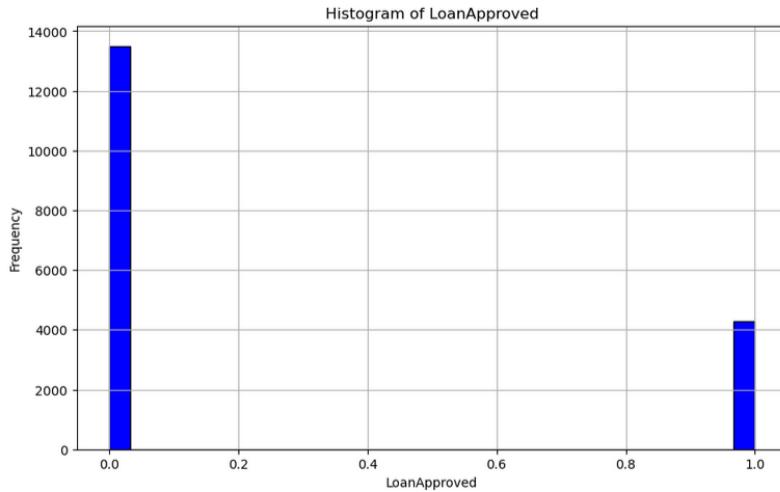
Rysunek 28: Histogram zmiennej UtilityBillsPaymentHistory

Histogram zmiennej UtilityBillsPaymentHistory wskazuje, że większość osób ma wysoki wskaźnik płatności rachunków za media (wartości w okolicach 0.8-1.0). Rozkład jest skośny w lewo, co sugeruje, że osoby o niższym wskaźniku płatności są mniej liczne.



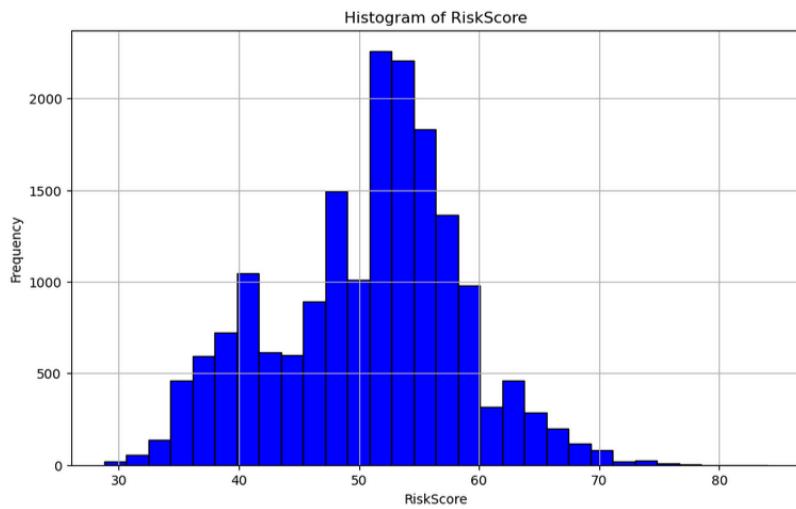
Rysunek 29: Histogram zmiennej PreviousLoanDefaults

Histogram zmiennej PreviousLoanDefaults pokazuje, że zdecydowana większość osób nie ma wcześniejszych zaległości w spłacie pożyczek (wartość 0), a jedynie niewielka liczba przypadków wykazuje wcześniejsze problemy ze spłatą (wartość 1).



Rysunek 30: Histogram zmiennej LoanApproved

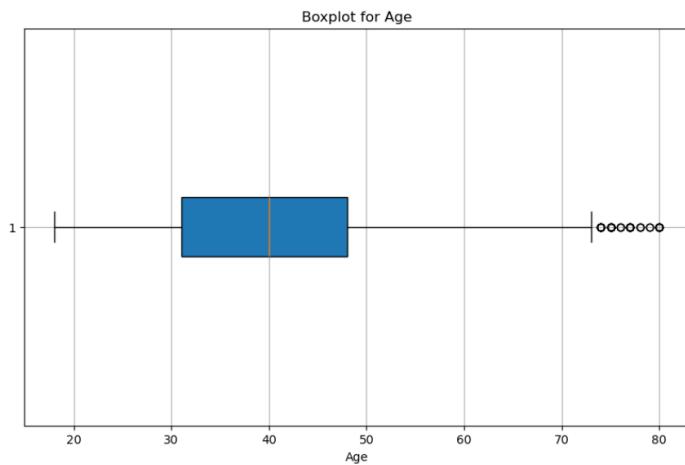
Histogram zmiennej LoanApproved pokazuje, że większość wniosków o pożyczkę została odrzucona (0), podczas gdy mniejsza liczba wniosków została zatwierdzona (1).



Rysunek 31: Histogram zmiennej RiskScore

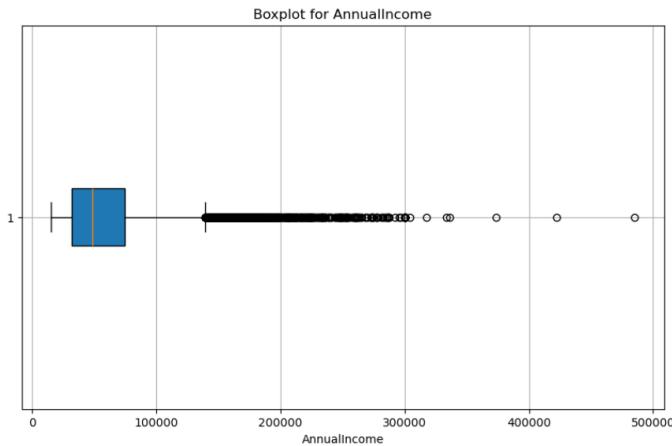
Histogram zmiennej RiskScore pokazuje, że większość wnioskodawców ma wynik ryzyka w okolicach 50, co stanowi szczyt rozkładu. Rozkład jest symetryczny, co sugeruje, że wartości wyniku ryzyka są równomiernie rozmiieszczone wokół średniej.

3.3 Boxploty cech



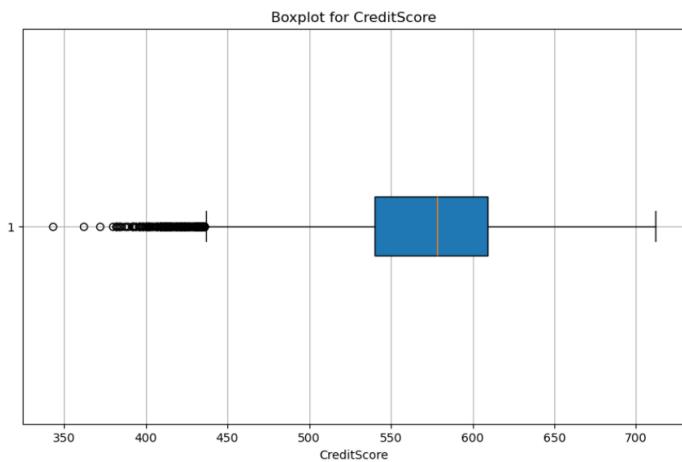
Rysunek 32: Boxplot zmiennej Age

Boxplot dla zmiennej Age pokazuje rozkład wieku. Większość wartości znajduje się w przedziale od około 30 do 50 lat. Mediana wieku wynosi około 40 lat. Są także wartości odstające po prawej stronie, które reprezentują osoby w wieku powyżej 70 lat.



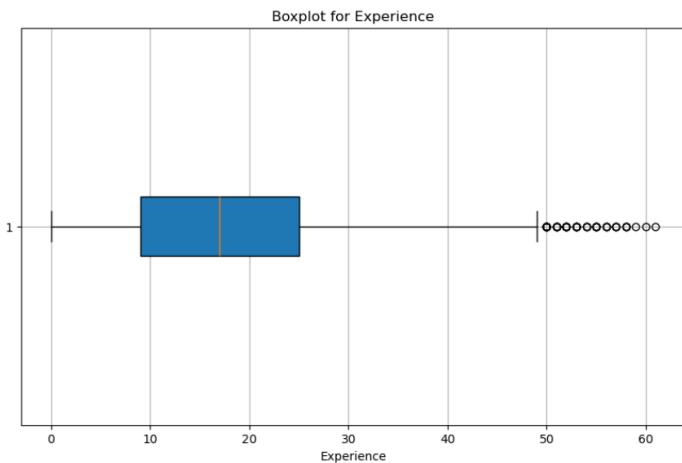
Rysunek 33: Boxplot zmiennej AnnualIncome

Boxplot dla zmiennej AnnualIncome pokazuje, że większość wartości jest skoncentrowana blisko dolnej granicy rozkładu, z medianą poniżej 100,000. Występuje wiele wartości odstających po prawej stronie, które reprezentują wyższe dochody, rozciągające się aż do około 500,000.



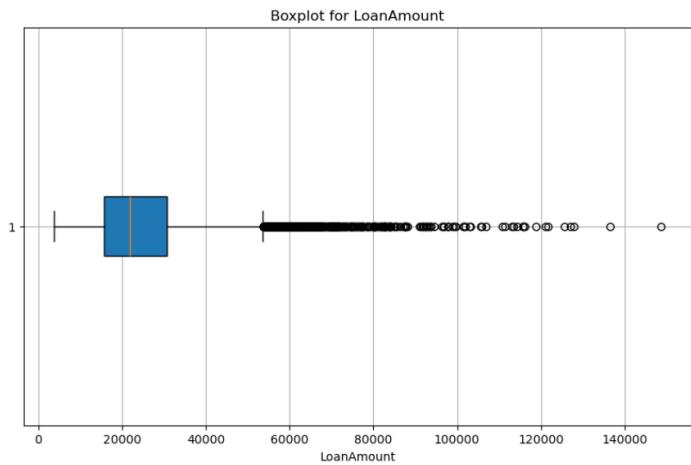
Rysunek 34: Boxplot zmiennej CreditScore

Boxplot dla zmiennej CreditScore pokazuje, że większość wartości mieści się w zakresie od około 500 do 650, z medianą około 600. Istnieje jednak wiele wartości odstających po lewej stronie, poniżej 500, co oznacza niższe wartości punktacji kredytowej dla niektórych osób



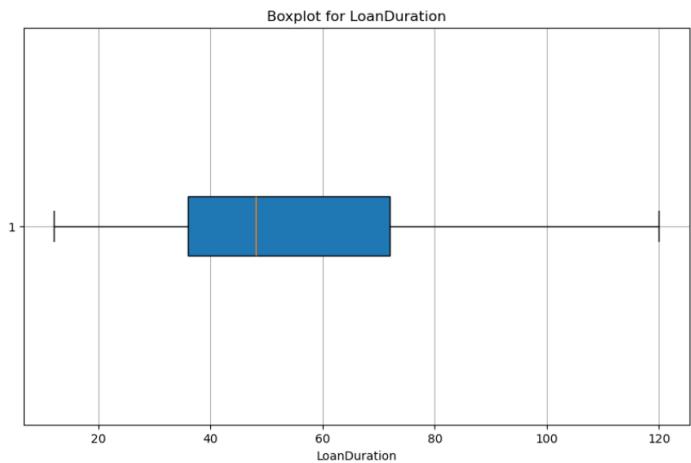
Rysunek 35: Boxplot zmiennej Experience

Boxplot dla zmiennej Experience wskazuje, że większość wartości mieści się w zakresie od około 10 do 30 lat doświadczenia, z medianą w okolicach 20 lat. Istnieje kilka wartości odstających powyżej 40 lat, co reprezentuje osoby z wyjątkowo długim stażem pracy.



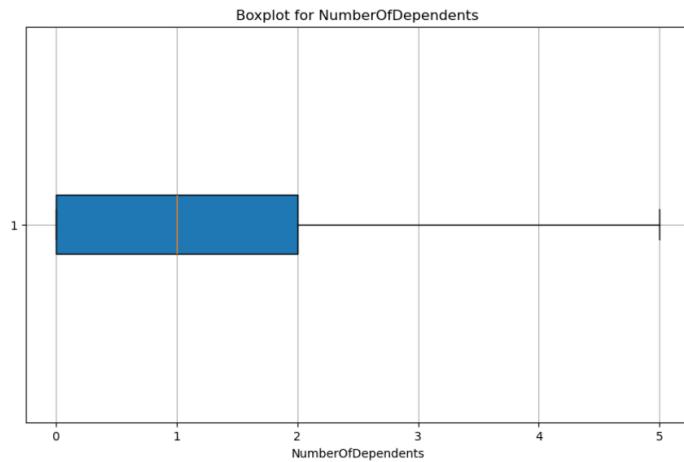
Rysunek 36: Boxplot zmiennej LoanAmount

Boxplot dla zmiennej LoanAmount pokazuje, że większość wartości pożyczek jest skoncentrowana poniżej 20,000, z medianą około tej wartości. Występuje wiele wartości odstających powyżej 30,000, co wskazuje na obecność kilku dużych pożyczek.



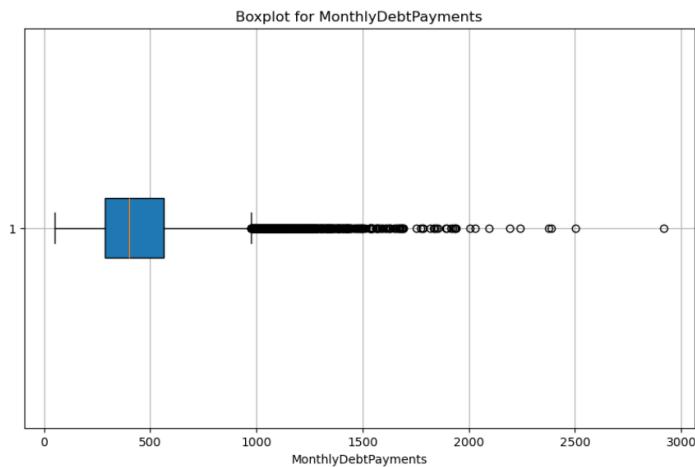
Rysunek 37: Boxplot zmiennej LoanDuration

Boxplot dla zmiennej LoanDuration pokazuje, że czas trwania pożyczek jest rozłożony od około 10 do 120 miesięcy, z medianą wynoszącą około 50 miesięcy. Zakres wartości jest szeroki, ale nie obserwuje się wyraźnych wartości odstających.



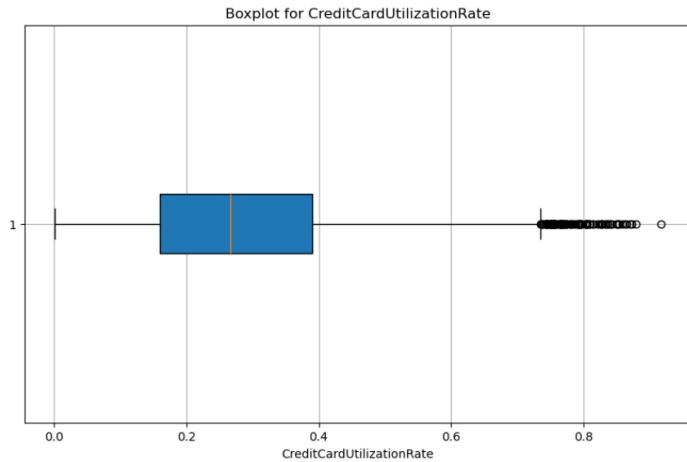
Rysunek 38: Boxplot zmiennej NumberOfDependents

Boxplot dla zmiennej NumberOfDependents wskazuje, że liczba osób na utrzymaniu waha się od 0 do 5, z medianą wynoszącą około 1. Rozkład jest równomierny w zakresie 0-2, a wartości powyżej 3 są mniej powszechnne.



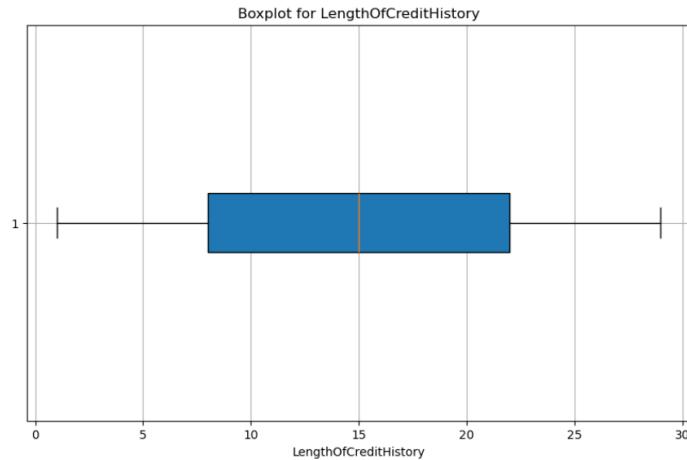
Rysunek 39: Boxplot zmiennej MonthlyDebtPayments

Boxplot dla zmiennej MonthlyDebtPayments pokazuje, że większość miesięcznych spłat długów wynosi mniej niż 500. Rozkład zawiera wiele wartości odstających, które przekraczają 1000, wskazując na pewną liczbę klientów z wyższymi miesięcznymi spłatami.



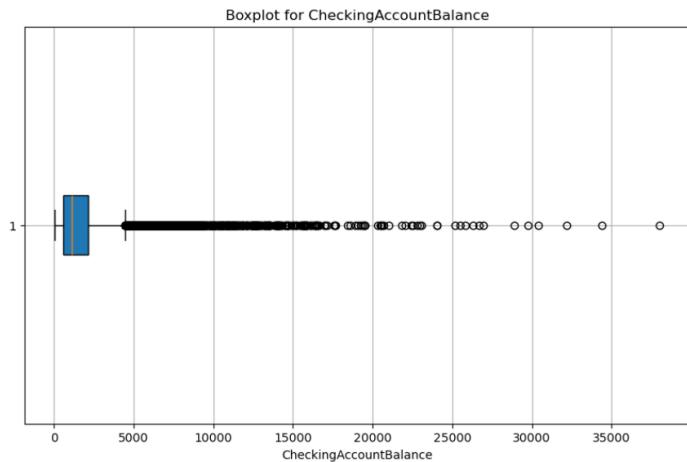
Rysunek 40: Boxplot zmiennej CreditCardUtilizationRate

Boxplot dla zmiennej CreditCardUtilizationRate pokazuje, że większość wartości współczynnika wykorzystania karty kredytowej jest skoncentrowana poniżej 0.4, co sugeruje umiarkowany poziom wykorzystania przez większość klientów. Istnieje jednak znaczna liczba wartości odstających powyżej tej wartości, wskazując na klientów o wyższym współczynniku wykorzystania, którzy mogą wykazywać większe ryzyko zadłużenia.



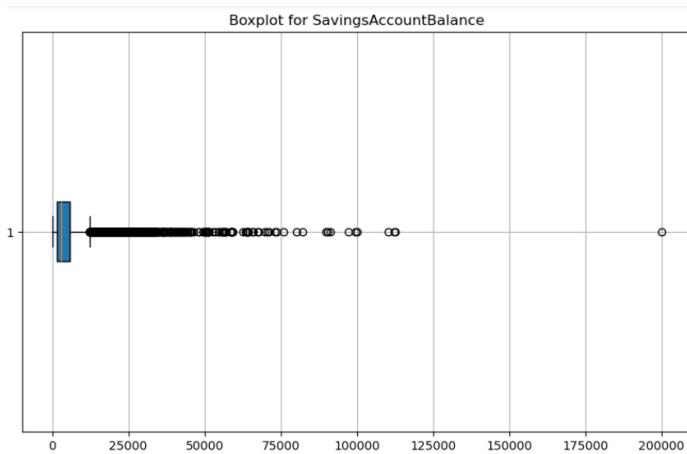
Rysunek 41: Boxplot zmiennej LengthOfCreditHistory

Boxplot dla zmiennej LengthOfCreditHistory wskazuje, że długość historii kredytowej większości klientów wynosi od 10 do 20 lat. Nie ma widocznych wartości odstających



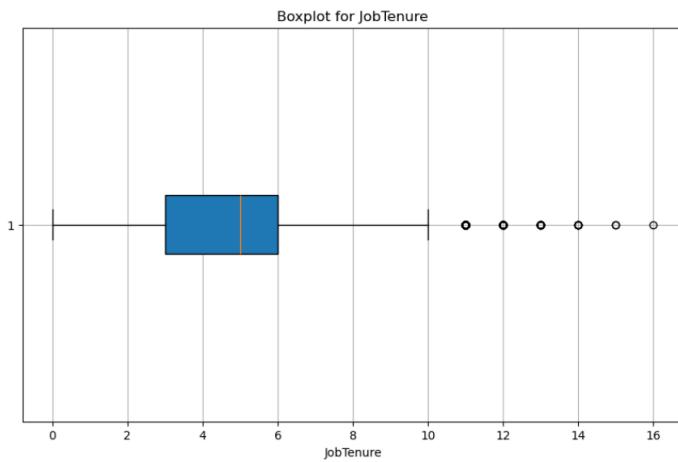
Rysunek 42: Boxplot zmiennej CheckingAccountBalance

Boxplot dla zmiennej CheckingAccountBalance pokazuje, że większość sald na rachunkach bieżących jest skoncentrowana w niskim zakresie wartości, co sugeruje, że klienci mają stosunkowo niewielkie kwoty na tych kontach. Występuje wiele wartości odstających, które wskazują na znacznie wyższe salda u niektórych klientów, co pokazuje dużą różnorodność w posiadanych środkach na kontach bieżących..



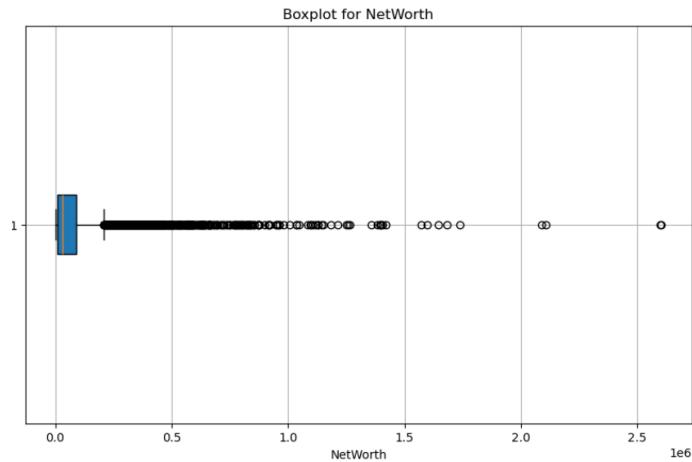
Rysunek 43: Boxplot zmiennej SavingsAccountBalance

Boxplot dla zmiennej SavingsAccountBalance pokazuje, że większość wartości znajduje się na niskim poziomie, co oznacza, że saldo kont oszczędnościowych u większości klientów jest relatywnie niskie. Istnieje duża liczba wartości odstających, wskazujących na nielicznych klientów z wysokimi saldami na kontach oszczędnościowych, co świadczy o znaczącej różnorodności w oszczędnościach klientów.



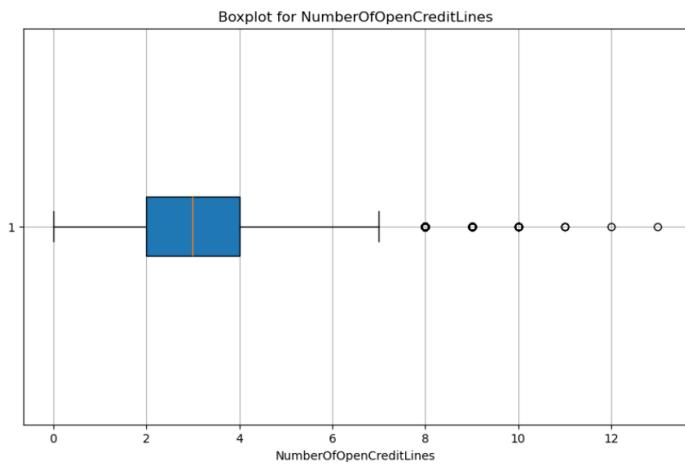
Rysunek 44: Boxplot zmiennej JobTenure

Boxplot dla zmiennej JobTenure wskazuje, że większość klientów ma staż pracy między 3 a 6 lat, z medianą około 5 lat. Istnieje kilka wartości odstających, które reprezentują klientów o znacznie dłuższym stażu pracy (powyżej 10 lat).



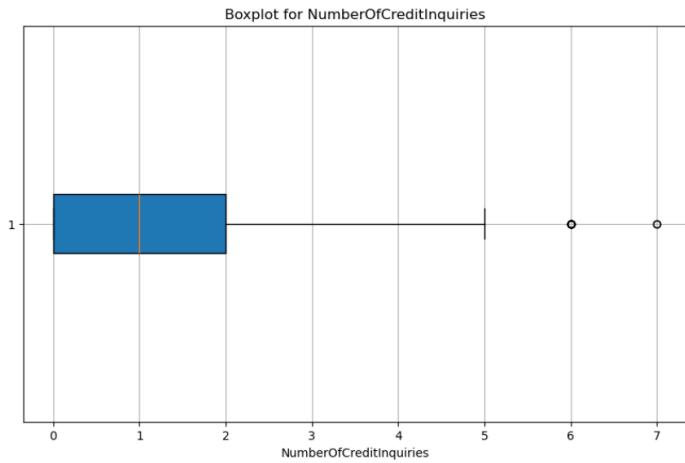
Rysunek 45: Boxplot zmiennej NetWorth

Boxplot dla zmiennej NetWorth pokazuje, że większość klientów ma wartość netto bliską zeru, co sugeruje, że znacząca część danych jest skupiona przy dolnej granicy. Są wiele wartości odstających, które reprezentują klientów o znacznie wyższej wartości, sięgającej nawet do 2,5 miliona.



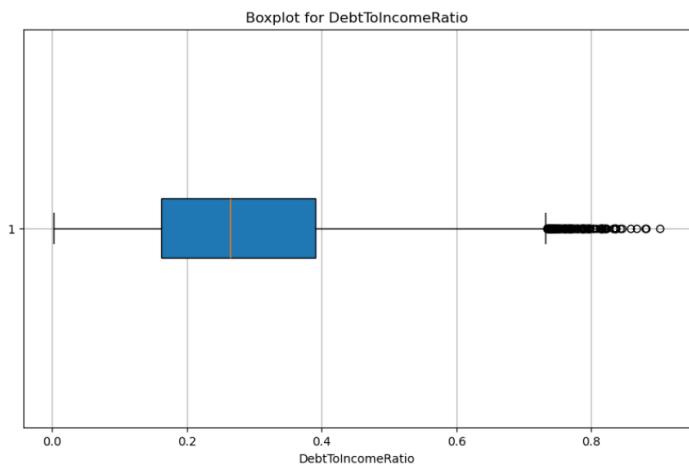
Rysunek 46: Boxplot zmiennej NumberOfOpenCreditLines

Boxplot dla zmiennej NumberOfOpenCreditLines pokazuje, że większość klientów posiada od 2 do 4 otwartych linii kredytowych. Wartości odstające znajdują się po prawej stronie, przy liczbach powyżej 8, co wskazuje na nieliczne przypadki klientów z wyższą liczbą otwartych linii kredytowych.



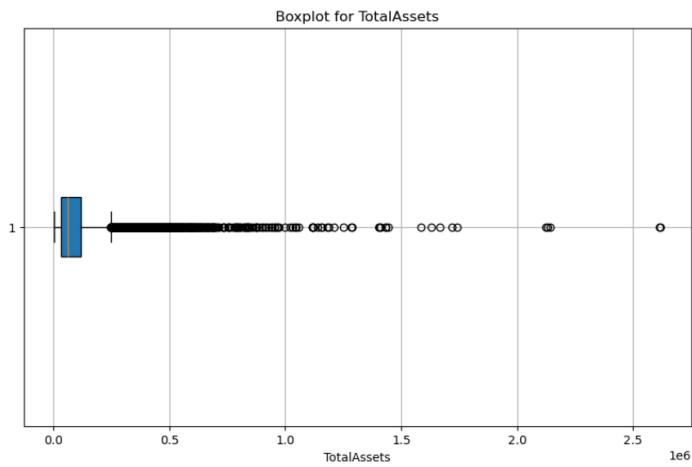
Rysunek 47: Boxplot zmiennej NumberOfCreditInquiries

Boxplot dla zmiennej NumberOfCreditInquiries pokazuje, że większość klientów ma od 0 do 2 zapytań kredytowych. Wartości odstające znajdują się powyżej 5 zapytań, co sugeruje, że tylko nieliczni klienci mają więcej niż 5 zapytań kredytowych.



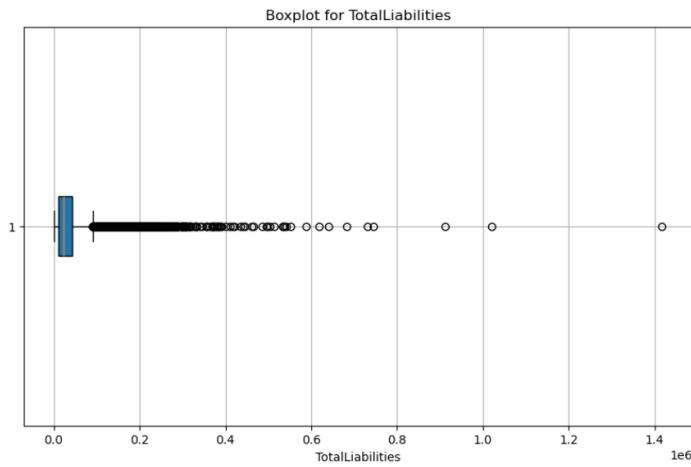
Rysunek 48: Boxplot zmiennej DebtToIncomeRatio

Boxplot dla zmiennej DebtToIncomeRatio pokazuje, że większość wartości wskaźnika długu do dochodu klientów znajduje się pomiędzy 0.1 a 0.4. Występuje kilka wartości odstających powyżej 0.6, co oznacza, że niektórzy klienci mają wyższy stosunek długu do dochodu, ale są to raczej rzadkie przypadki.



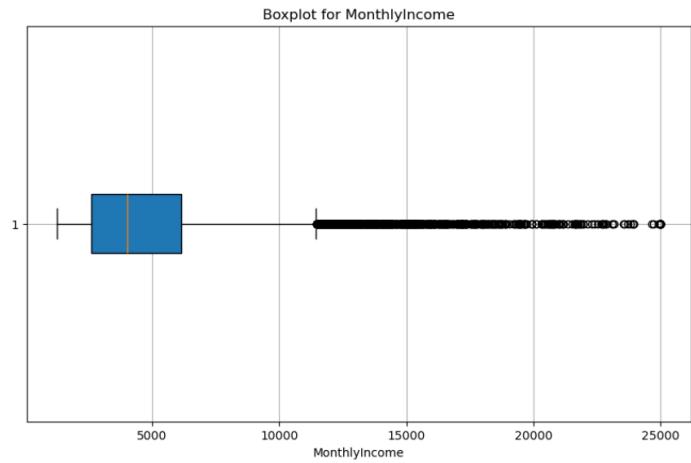
Rysunek 49: Boxplot zmiennej TotalAssets

Boxplot dla zmiennej TotalAssets wskazuje, że większość wartości całkowitych aktywów klientów koncentruje się w dolnym zakresie, blisko zera. Jednak występuje znaczna liczba wartości odstających, sięgających nawet 2,5 miliona, co pokazuje, że niektórzy klienci posiadają bardzo wysokie aktywa w porównaniu do ogólna.



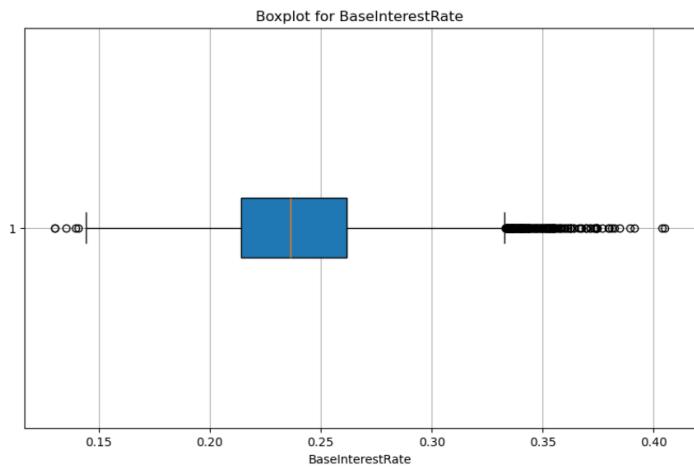
Rysunek 50: Boxplot zmiennej TotalLiabilities

Boxplot dla zmiennej TotalLiabilities wskazuje, że większość zobowiązań finansowych klientów skupia się w niskim zakresie, blisko zera. Niemniej jednak widoczna jest znacząca liczba wartości odstających, które sięgają nawet 1,4 miliona. To sugeruje, że niektórzy klienci posiadają bardzo wysokie zobowiązania w porównaniu do reszty populacji.



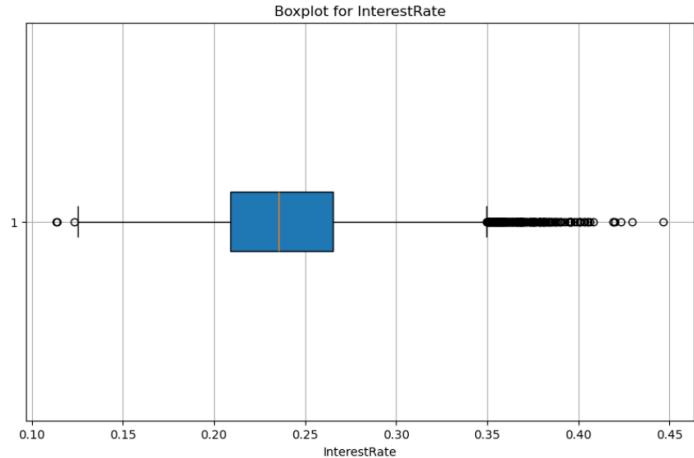
Rysunek 51: Boxplot zmiennej MonthlyIncome

Boxplot dla zmiennej MonthlyIncome wskazuje, że większość miesięcznych dochodów klientów mieści się w zakresie od około 2000 do 6000. Wartości odstające są rozproszone aż do 25000, co sugeruje, że niektórzy klienci posiadają znacznie wyższe dochody niż większość. Środkowy kwartyl (medianą) jest wyraźnie bliżej dolnej granicy, co może oznaczać asymetryczny rozkład dochodów.



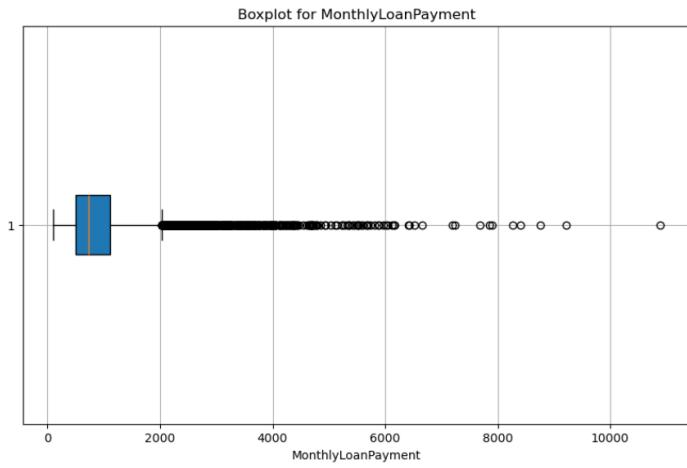
Rysunek 52: Boxplot zmiennej BaseInterestRate

Boxplot dla zmiennej BaseInterestRate pokazuje, że większość wartości stopy bazowej mieści się w przedziale od około 0,20 do 0,30, z medianą blisko środka tego zakresu. Istnieje kilka wartości odstających, zarówno po stronie niższej, jak i wyższej, dochodzących do około 0,40. To sugeruje, że dla niektórych klientów ustalana jest wyższa stopa bazowa, jednak większość danych skupia się w środkowej części rozkładu.



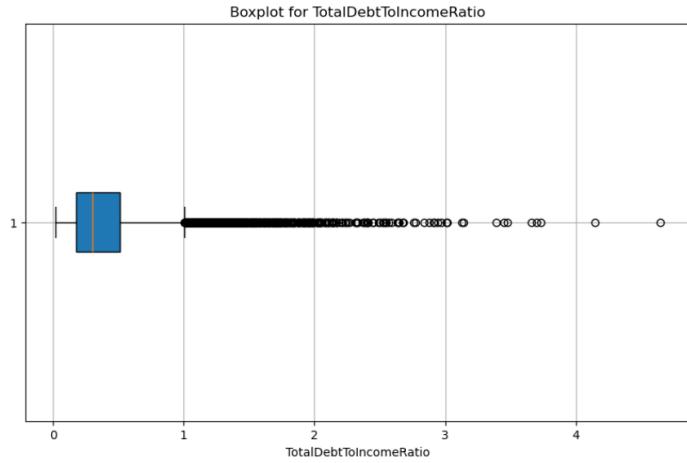
Rysunek 53: Boxplot zmiennej InterestRate

Boxplot dla zmiennej InterestRate wskazuje, że większość wartości stóp procentowych oscyluje w przedziale od około 0,20 do 0,30, z medianą w okolicach 0,25. Występuje kilka wartości odstających, zwłaszcza po stronie wyższych stóp procentowych, które sięgają do około 0,45. Sugeruje to, że chociaż większość klientów ma stopę procentową blisko mediany, niektórzy mogą mieć znacznie wyższe oprocentowanie.



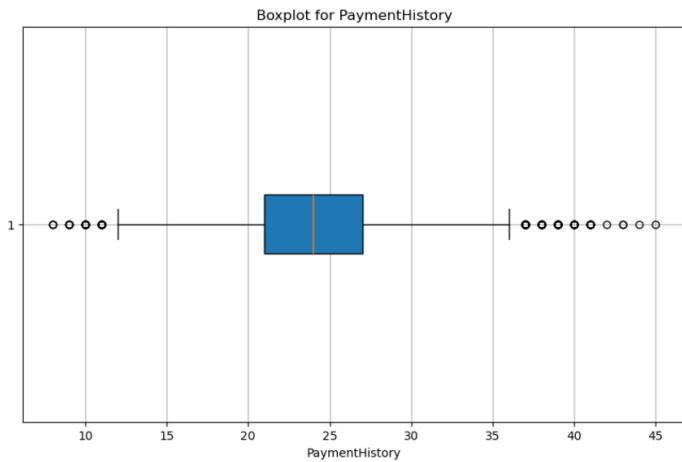
Rysunek 54: Boxplot zmiennej MonthlyLoanPayment

Boxplot dla zmiennej MonthlyLoanPayment wskazuje, że większość wartości miesięcznych płatności pożyczkowych koncentruje się w dolnym zakresie, w pobliżu 1000 jednostek, z medianą na poziomie około 500–600. Widoczne są liczne wartości odstające, które sięgają nawet do 10,000, co sugeruje, że niektórzy klienci mają znacznie wyższe płatności miesięczne w porównaniu do reszty.



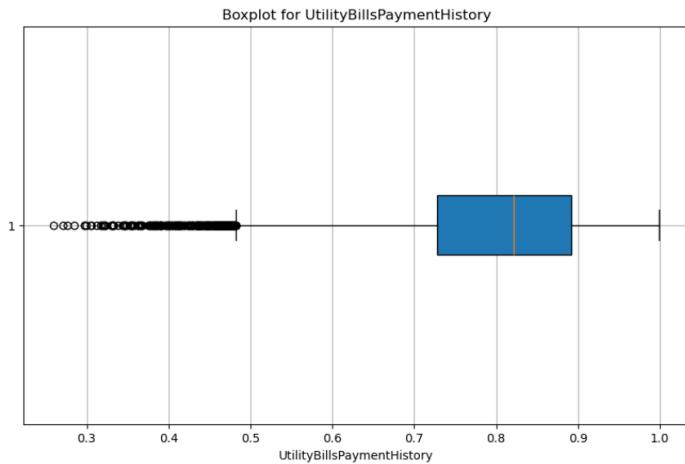
Rysunek 55: Boxplot zmiennej TotalDebtToIncomeRatio

Boxplot dla zmiennej TotalDebtToIncomeRatio pokazuje, że większość wartości współczynnika dłużu do dochodu mieści się poniżej 1, co oznacza, że dług większości klientów nie przekracza ich rocznego dochodu. Medianą znajduje się blisko 0.3. Widoczne są liczne wartości odstające, które sięgają nawet powyżej 4, wskazując, że niektórzy klienci mają dług znacznie przewyższający ich dochód, co może sugerować wyższe ryzyko kredytowe.



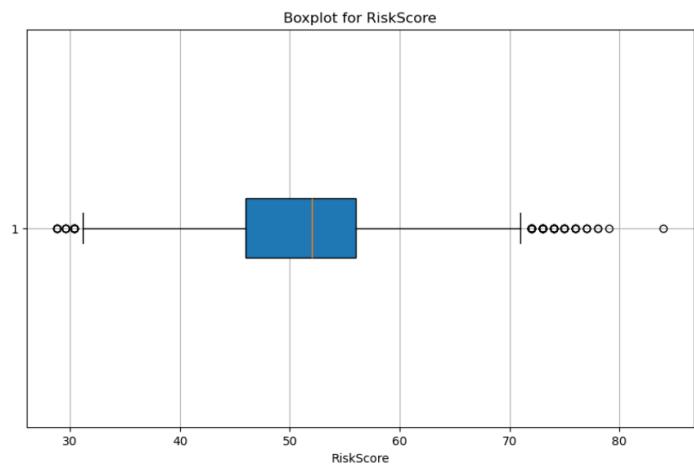
Rysunek 56: Boxplot zmiennej PaymentHistory

Boxplot dla zmiennej PaymentHistory pokazuje, że większość wartości mieści się w przedziale od około 20 do 30, z medianą blisko 25. Sugeruje to, że większość klientów ma umiarkowaną historię płatności. Wykres wskazuje także na istnienie wartości odstających zarówno po niższej stronie (około 10), jak i po wyższej stronie (powyżej 35), co może oznaczać klientów z wyjątkowo dobrą lub złą historią płatności.



Rysunek 57: Boxplot zmiennej UtilityBillsPaymentHistory

Boxplot dla zmiennej UtilityBillsPaymentHistory wskazuje, że większość wartości znajduje się w zakresie od około 0.7 do 1.0, co sugeruje, że większość klientów ma pozytywną historię płatności za rachunki za media. Występuje jednak kilka wartości odstających poniżej 0.5, co może wskazywać na klientów z historią problemów z płatnościami za media.

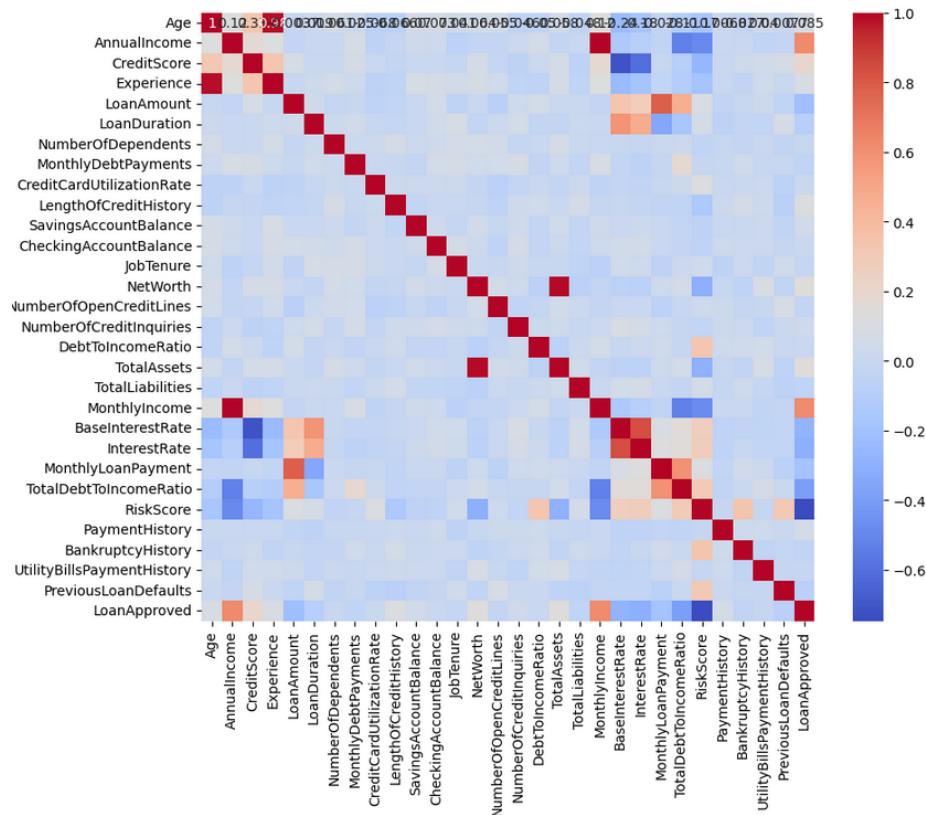


Rysunek 58: Boxplot zmiennej RiskScore

Boxplot dla zmiennej RiskScore pokazuje, że większość wartości mieści się w zakresie od około 45 do 60, z medianą blisko 52. Wyniki te wskazują, że większość klientów ma przeciętny poziom ryzyka. Widoczne są również wartości odstające zarówno po niższej stronie (około 30), jak i po wyższej (powyżej 70), co sugeruje, że niewielka liczba klientów ma wyjątkowo niskie lub wysokie ryzyko kredytowe.

3.4 Pair ploty

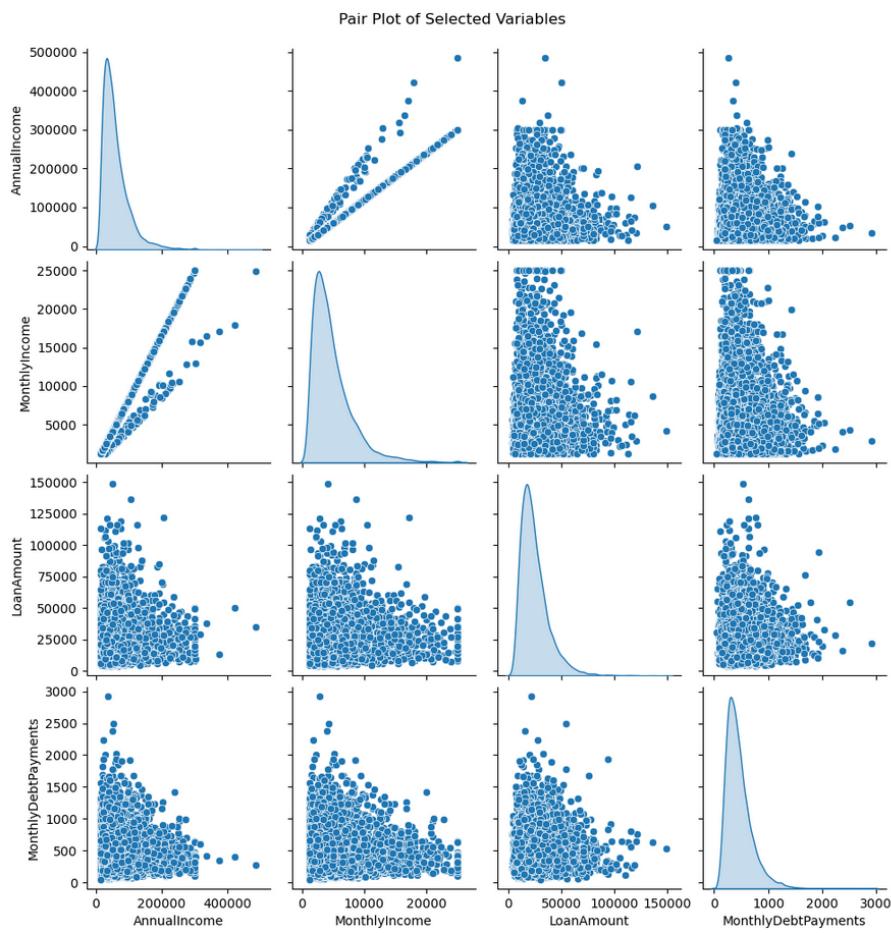
Zbiór danych zawiera dużą liczbę cech, stworzyliśmy macierz korelacji, aby przeanalizować zależności między zmiennymi i wybrać istotne cechy do wizualizacji za pomocą pairplotu.



Rysunek 59: Macierz korelacji

Analiza macierzy korelacji pokazuje, że silne korelacje, jak między AnnualIncome a MonthlyIncome oraz TotalAssets i NetWorth, są logiczne, gdyż wyższe roczne dochody oznaczają wyższe miesięczne, a aktywa wpływają na wartość netto. Umiarkowane korelacje, np. między LoanAmount a MonthlyLoanPayment, wskazują, że większa kwota pożyczki wiąże się z wyższymi ratami. Słabe korelacje, takie jak między CreditScore a LoanAmount oraz wiekiem a innymi zmiennymi, sugerują brak bezpośredniego związku. Brak silnych ujemnych korelacji potwierdza, że zmienne nie są odwrotnie skorelowane.

Na podstawie tych wyników wybrano do pair plotu zmienne finansowe i zobowiązania



Rysunek 60: Pair plot Cechy finansowe i zobowiązania

Pair plot pokazuje silną dodatnią korelację między AnnualIncome a MonthlyIncome, co jest logiczne, ponieważ dochód miesięczny stanowi część rocznego. LoanAmount jest umiarkowanie skorelowany z MonthlyDebtPayments, co sugeruje, że większe pożyczki wiążą się z wyższymi miesięcznymi zobowiązaniami. Zależność między AnnualIncome a LoanAmount jest słabsza, wskazując, że dochód roczny nie jest bezpośrednim predyktorem kwoty pożyczki. Inne pary zmiennych nie wykazują wyraźnych zależności.

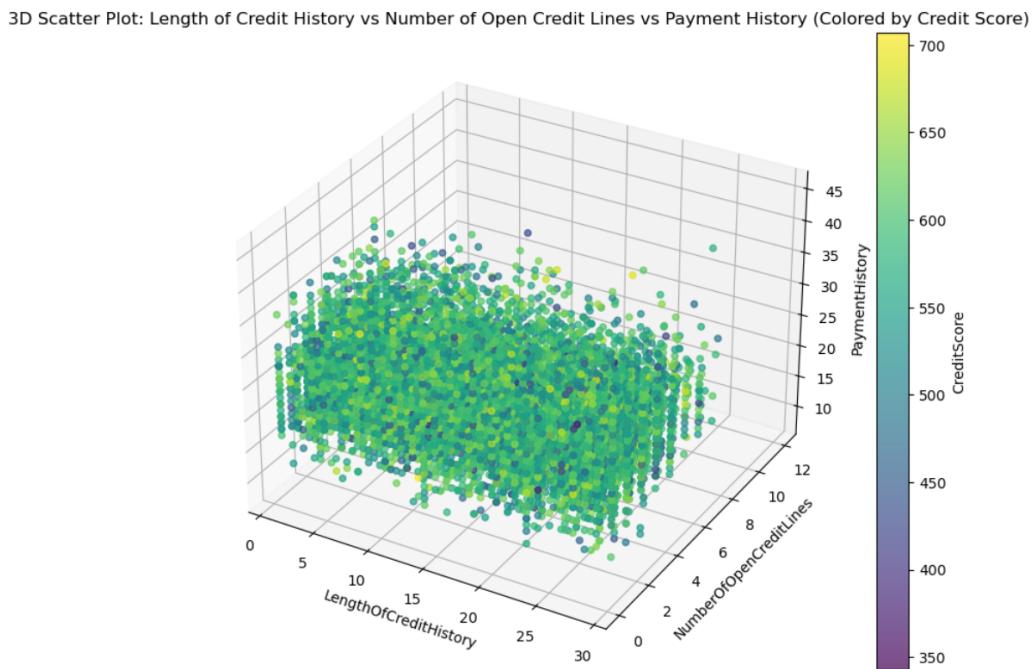
3.5 Wizualizacja danych



Rysunek 61: Scatter Plot. Zależność między kwotą pożyczki a oceną kredytową z uwzględnieniem wyniku ryzyka i miesięcznego dochodu

Wykres przedstawia zależność między kwotą pożyczki a oceną kredytową. Punkty są kolorowane na podstawie wyniku ryzyka, a ich wielkość odpowiada miesięcznym dochodom.

- Nie widać wyraźnej korelacji między kwotą pożyczki a oceną kredytową, co sugeruje, że wysokość pożyczki nie zależy bezpośrednio od oceny kredytowej.
- Większe kwoty pożyczek występują zarówno przy niskich, jak i wysokich wynikach oceny kredytowej.
- Wynik ryzyka jest zróżnicowany, ale najwyższe wartości są bardziej widoczne przy wyższych ocenach kredytowych.
- Punkty reprezentujące wyższe miesięczne dochody są większe, co jest zgodne z oczekiwaniem, że osoby o wyższych dochodach mogą ubiegać się o większe pożyczki.



Rysunek 62: 3D Scatter Plot. Zależność między długością historii kredytowej a liczbą otwartych kredytów z uwzględnieniem historii płatności i wyniku kredytowego.

Wykres przedstawia długość historii kredytowej, liczbę otwartych linii kredytowych oraz historię płatności. Kolory punktów reprezentują ocenę kredytową, gdzie jaśniejsze punkty oznaczają wyższą ocenę.

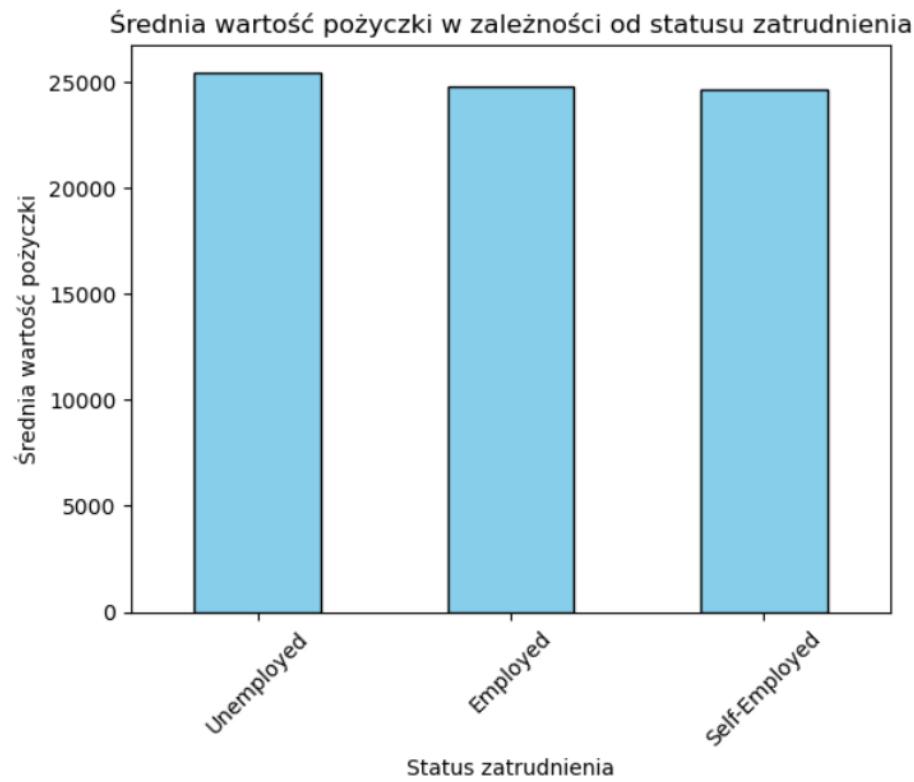
- Wykres nie pokazuje jednoznacznej zależności między zmiennymi. Dane są rozproszone równomiernie, co sugeruje brak silnych korelacji między długością historii kredytowej, liczbą otwartych kredytów i historią płatności.
- Wynik kredytowy zmienia się losowo w całym zakresie danych, co sugeruje, że ani długość historii kredytowej, ani liczba otwartych kredytów, ani historia płatności nie mają wyraźnego wpływu na ocenę kredytową.
- Na wykresie nie widać wyraźnych trendów ani wzorców wskazujących na zależność między zmiennymi. Brak wyraźnego nachylenia czy skupisk może sugerować, że te zmienne są niezależne lub wpływają na siebie w niewielkim stopniu.

3.6 Analizę danych bazując na pandas

- Średnia wartość pożyczki w zależności od statusu zatrudnienia

```
selected_columns = ['LoanAmount', 'EmploymentStatus']
clean_data = data[selected_columns].dropna()
average_loan_amount = clean_data.groupby('EmploymentStatus')['LoanAmount'].mean()
    .sort_values(ascending=False)
print(average_loan_amount)
average_loan_amount.plot(kind='bar', color='skyblue', edgecolor='black')
plt.title('Średnia wartość pożyczki w zależności od statusu zatrudnienia')
plt.xlabel('Status zatrudnienia')
plt.ylabel('Średnia wartość pożyczki')
plt.xticks(rotation=45)
plt.show()
```

```
EmploymentStatus
Unemployed      25484.028244
Employed        24775.911861
Self-Employed   24660.179299
Name: LoanAmount, dtype: float64
```



Rysunek 63: Średnia wartość pożyczki w zależności od statusu zatrudnienia

- Średni wynik kredytowy w zależności od wieku

```
selected_columns = ['Age', 'CreditScore']
clean_data = data[selected_columns].dropna()
average_credit_score = clean_data.groupby('Age')['CreditScore'].mean().
    sort_index()
print(average_credit_score)
plt.figure(figsize=(10, 6))
plt.plot(average_credit_score.index, average_credit_score.values, marker='o',
         color='b', linestyle='--')
plt.title('Średni wynik kredytowy w zależności od wieku')
plt.xlabel('Wiek')
plt.ylabel('Średni wynik kredytowy')
plt.grid()
plt.show()
```

```
Age
18.0      544.742947
19.0      540.733945
20.0      544.963964
21.0      544.011299
22.0      546.292818
...
76.0      658.333333
77.0      642.333333
78.0      627.500000
79.0      674.000000
80.0      646.857143
Name: CreditScore, Length: 63, dtype: float64
```



Rysunek 64: Średni wynik kredytowy w zależności od wieku

- Częstotliwość występowania pożyczek niezatwierdzonych w zależności od historii bankructw

```

selected_columns = ['LoanApproved', 'BankruptcyHistory']
clean_data = data[selected_columns].dropna()
filtered_data = clean_data[clean_data['LoanApproved'] == 0]
bankruptcy_counts = filtered_data['BankruptcyHistory'].value_counts().
    sort_index()
print(bankruptcy_counts)
plt.figure(figsize=(8, 5))
plt.bar(bankruptcy_counts.index, bankruptcy_counts.values, color='skyblue')
plt.xlabel('Historia Bankructwa (0 - brak, 1 - wystąpiło)')
plt.ylabel('Częstotliwość Niezatwierdzonych Pożyczek')
plt.title('Częstotliwość występowania pożyczek niezatwierdzonych w zależności od historii bankructw')
plt.xticks([0, 1], ['Brak Bankructwa', 'Bankructwo'])
plt.show()

```

```

BankruptcyHistory
0.0      11321
1.0       726
Name: count, dtype: int64

```



Rysunek 65: Częstotliwość występowania pożyczek niezatwierdzonych w zależności od historii bankructw

- Porównanie miesięcznego dochodu pomiędzy osobami z różnym wynikiem ryzyka

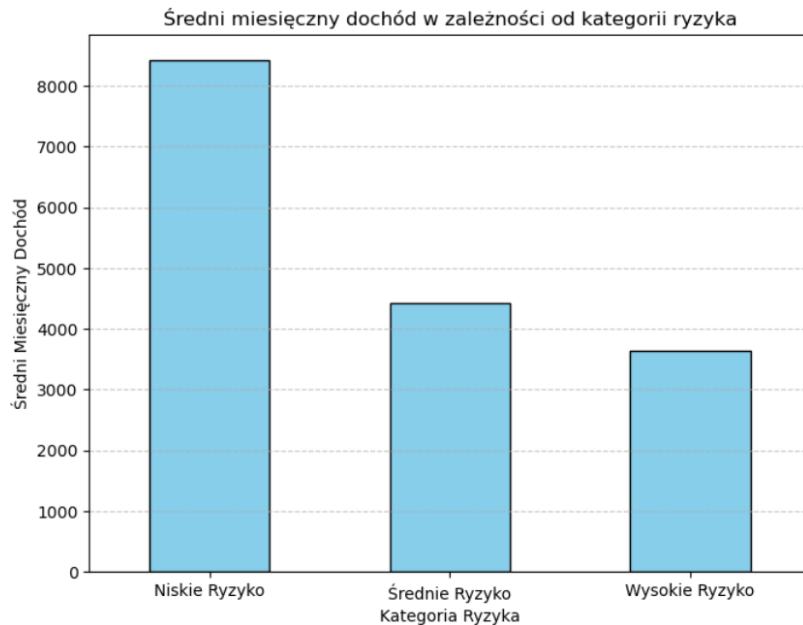
```

selected_columns = ['MonthlyIncome', 'RiskScore']
clean_data = data[selected_columns].dropna()
bins = [0, 40, 60, 100]
labels = ['Niskie Ryzyko', 'Średnie Ryzyko', 'Wysokie Ryzyko']
clean_data['RiskCategory'] = pd.cut(clean_data['RiskScore'], bins=bins, labels=labels)
mean_income = clean_data.groupby('RiskCategory')['MonthlyIncome'].mean()
print(mean_income)
plt.figure(figsize=(8, 6))
mean_income.plot(kind='bar', color='skyblue', edgecolor='black')
plt.title('Średni miesięczny dochód w zależności od kategorii ryzyka')
plt.xlabel('Kategoria Ryzyka')
plt.ylabel('Średni Miesięczny Dochód')
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

```

RiskCategory	Mean Income
Niskie Ryzyko	8420.368417
Średnie Ryzyko	4422.561493
Wysokie Ryzyko	3634.831680

Name: MonthlyIncome, dtype: float64



Rysunek 66: Porównanie miesięcznego dochodu pomiędzy osobami z różnym wynikiem ryzyka

- Średnie zadłużenie w zależności od liczby posiadanych linii kredytowych

```
selected_columns = ['TotalLiabilities', 'NumberOfOpenCreditLines']
clean_data = data[selected_columns].dropna()
average_liabilities = clean_data.groupby('NumberOfOpenCreditLines')[
    'TotalLiabilities'].mean()
print(average_liabilities)
plt.figure(figsize=(10, 6))
plt.plot(average_liabilities.index, average_liabilities.values, marker='o',
         color='b', linestyle='--')
plt.xlabel('Liczba posiadanych linii kredytowych')
plt.ylabel('Średnie zadłużenie')
plt.title('Średnie zadłużenie w zależności od liczby posiadanych linii kredytowych')
plt.grid()
plt.show()
```

```
NumberOfOpenCreditLines
0.0      36493.828017
1.0      37285.737826
2.0      35718.287629
3.0      37479.382458
4.0      36577.768556
5.0      33864.321958
6.0      35656.279481
7.0      33255.352113
8.0      42167.984127
9.0      35018.612245
10.0     48649.230769
11.0     58853.666667
12.0     16484.000000
13.0     98345.000000
Name: TotalLiabilities, dtype: float64
```

3.7 Analiza brakujących danych

Analizując ilość brakujących danych, mamy około 10% - 12% brakujących danych:

ApplicationDate	2207
Age	2239
AnnualIncome	2181
CreditScore	2146
EmploymentStatus	2198
EducationLevel	2149
Experience	2139
LoanAmount	2258
LoanDuration	2253
MaritalStatus	2137
NumberOfDependents	2241
HomeOwnershipStatus	2227
MonthlyDebtPayments	2123
CreditCardUtilizationRate	2194
NumberOfOpenCreditLines	2143
NumberOfCreditInquiries	2126
DebtToIncomeRatio	2183
BankruptcyHistory	2157
LoanPurpose	2220
PreviousLoanDefaults	2194
PaymentHistory	2169
LengthOfCreditHistory	2120
SavingsAccountBalance	2219
CheckingAccountBalance	2207
TotalAssets	2218
TotalLiabilities	2162
MonthlyIncome	2159
UtilityBillsPaymentHistory	2132
JobTenure	2187
NetWorth	2210
BaseInterestRate	2180
InterestRate	2247
MonthlyLoanPayment	2110
TotalDebtToIncomeRatio	2171
LoanApproved	2224
RiskScore	2193
dtype:	int64

Dane zostały określone na różne typy kolumn, aby poprawnie przygotować dane do wypełniania braków i dalszej analizy. Kolumny zostały podzielone na cztery główne typy:

- Kolumny binarne - zawierają tylko dwie unikalne wartości 0 i 1
- Kolumny dyskretne - zawierają wartości całkowite

- Kolumny ciągłe - zawierają wartości zmiennoprzecinkowe
- kolumny kategoryczne - zawierają wartości opisowe

Podział na te typy umożliwił zastosowanie właściwych metod wypełniania braków oraz przekształceń dla każdego typu kolumn, co z kolei wpłynęło na poprawę jakości danych. Braki danych zostały wypełnione trzema różnymi metodami:

- Wypełnianie medianą i mode.

W pierwszym podejściu, braki w zmiennych numerycznych zostały uzupełnione medianą, a w zmiennych kategorycznych mode.

- Kolumny binarne – braki uzupełniane mode, czyli najczęściej występującą wartością.
- Kolumny dyskretne – braki uzupełniane medianą, a następnie zaokrąglane do liczb całkowitych.
- Kolumny ciągłe – braki wypełniane medianą.
- Kolumny kategoryczne – braki uzupełniane mode, a kolumny konwertowane do typu category.

```
ds1 = pd.read_csv('data/filled_data/dataset_filled_median.csv')

missing_values = ds1.isnull().sum()
if missing_values.sum() == 0:
    print("Brak brakujących wartości w zbiorze danych.")
else:
    print("Są jakieś braki w następujących kolumnach:")
    print(missing_values[missing_values > 0])
ds1.head()
```

Brak brakujących wartości w zbiorze danych.

	ApplicationDate	Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	MaritalStatus	...	MonthlyIncome
0	2018-01-01	45	39948	617	Employed	Master	22	13152	48	Married	...	3329.000000
1	2018-01-02	38	39709	628	Employed	Associate	15	21916	48	Single	...	3309.083333
2	2018-01-03	47	40724	578	Employed	Bachelor	26	17627	48	Married	...	3393.666667
3	2018-01-04	58	69084	545	Employed	High School	34	37898	96	Single	...	4023.500000
4	2018-01-05	37	103264	594	Employed	Associate	17	9184	36	Married	...	8605.333333

Rysunek 67: Dataset 1: Wypełnienie medianą/modą

- Wypełnianie średnią i mode.

W drugim podejściu, braki w zmiennych numerycznych zostały uzupełnione średnią, a w zmiennych kategorycznych mode.

- Kolumny binarne – braki uzupełniane mode, czyli najczęściej występującą wartością.
- Kolumny dyskretne – braki wypełniane są średnią, zaokrągloną do liczb całkowitych.
- Kolumny ciągłe – braki wypełniane są średnią.
- Kolumny kategoryczne – braki uzupełniane mode, a kolumny konwertowane do typu category.

	ApplicationDate	Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	MaritalStatus	...	MonthlyIncome
0	2018-01-01	45	39948	617	Employed	Master	22	13152	54	Married	...	3329.000000
1	2018-01-02	38	39709	628	Employed	Associate	15	24825	48	Single	...	3309.083333
2	2018-01-03	47	40724	572	Employed	Bachelor	26	17627	54	Married	...	3393.666667
3	2018-01-04	58	69084	545	Employed	High School	34	37898	96	Single	...	4881.070582
4	2018-01-05	37	103264	594	Employed	Associate	17	9184	36	Married	...	8605.333333

Rysunek 68: Dataset 2: Wypełnienie średnią/modą

- Wypełnianie za pomocą wytrenowanych modeli regresji.

W trzecim sposobie uzupełniania braków zastosowano połączenie metod regresji wielomianowej z uzupełnianiem wartości za pomocą mody, mediany i średniej.

- Dla cech docelowych AnnualIncome, LoanAmount, LoanDuration i MonthlyLoanPayment stworzono modele regresji wielomianowej na danych bez braków w wybranych cechach wejściowych.
- Jakość modeli oceniono metrykami R^2 i MSE (trening/test).
- Modele użyto do przewidywania brakujących wartości w kolumnach docelowych.
- Braki w kolumnach binarnych uzupełniono najczęstszą wartością.
- Braki w kolumnach kategorycznych uzupełniono modą i przekształcono na typ category.
- Braki w kolumnach dyskretnych i ciągłych uzupełniono medianą dla połowy braków i średnią dla pozostałych.

	ApplicationDate	Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	MaritalStatus	...	MonthlyIncome	U
0	2018-01-01	45	39948	617	Employed	Master	22	13152	50	Married	...	3329.000000	
1	2018-01-02	38	39709	628	Employed	Associate	15	23964	48	Single	...	3309.083333	
2	2018-01-03	47	40724	578	Employed	Bachelor	26	17627	-151	Married	...	3393.666667	
3	2018-01-04	58	69084	545	Employed	High School	34	37898	96	Single	...	4023.500000	
4	2018-01-05	37	103264	594	Employed	Associate	17	9184	36	Married	...	8605.333333	

Rysunek 69: Dataset 3: Wypełnianie braków za pomocą regresji

3.8 Skalowanie cech

Po uzupełnieniu brakujących danych przy użyciu trzech różnych metod (wypełnienie medianą, średnią oraz za pomocą regresji wielomianowej), stworzono trzy nowe zbiorów danych. Kolejnym krokiem było przekształcenie i przeskalarowanie cech, aby przygotować je do dalszej analizy oraz modelowania.

Każdy z tych trzech zbiorów danych został przeskalarowany dwoma metodami:

- Min-Max Scaling
 - Usunięcie kolumny ApplicationDate. Cechy czasowe nie są istotne w modelach analizy numerycznej, więc kolumna została usunięta.
 - Przetwarzanie cech kategorycznych. Cechy: HomeOwnershipStatus, MaritalStatus, LoanPurpose przetworzono za pomocą One-Hot Encoding. Cechy: EmploymentStatus, EducationLevel przetworzono za pomocą Label Encoding.
 - Skalowanie dotyczyło wszystkich cech numerycznych, z wyjątkiem kolumn binarnych.
 - Metoda ta przekształca wartości cech na zakres [0, 1], co pomaga zredukować wpływ różnej skali danych i poprawia efektywność algorytmów uczących się.
 - Przeskalowane dane zapisano do nowego pliku csv.

```
: min_max = pd.read_csv('data/scaled_data/min_max/dataset_filled_mean_min_max.csv')
min_max.head()
```

	Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	NumberOfDependents	MonthlyDebtPayments	...
0	0.435484	0.053042	0.742547	0	4	0.360656	0.065320	0.388889	0.4	0.046358	...
1	0.322581	0.052534	0.772358	0	0	0.245902	0.145766	0.333333	0.2	0.155455	...
2	0.467742	0.054692	0.620596	0	1	0.426230	0.096160	0.388889	0.4	0.296968	...
3	0.645161	0.114989	0.547425	0	3	0.557377	0.235862	0.777778	0.2	0.245730	...
4	0.306452	0.187660	0.680217	0	0	0.278689	0.037973	0.222222	0.2	0.078076	...

5 rows × 45 columns

Rysunek 70: Min-Max Scaling

- Standaryzacja
 - Usunięcie kolumny ApplicationDate. Cechy czasowe nie są istotne w modelach analizy numerycznej, więc kolumna została usunięta.
 - Przetwarzanie cech kategorycznych. Cechy: HomeOwnershipStatus, MaritalStatus, LoanPurpose przetworzono za pomocą One-Hot Encoding. Cechy: EmploymentStatus, EducationLevel przetworzono za pomocą Label Encoding.
 - Dla kolumn numerycznych zastosowano standaryzację przy użyciu StandardScaler.
 - Standaryzacja przekształca dane tak, aby każda cecha miała średnią równą 0 i odchylenie standardowe równe 1.
 - Przeskalowane dane zapisano do nowego pliku csv.

	Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	NumberOfDependents	MonthlyDebtPayments	...
0	0.480780	-0.506366	0.943372	0	4	0.412483	-0.933505	-0.003341	0.327243	-1.188795	...
1	-0.158677	-0.512590	1.172094	0	0	-0.242712	-0.000022	-0.261327	-0.433522	0.185701	...
2	0.663482	-0.486157	0.007691	0	1	0.786880	-0.575642	-0.003341	0.327243	1.968594	...
3	1.668341	0.252418	-0.553717	0	3	1.535674	1.045418	1.802560	-0.433522	1.323064	...
4	-0.250027	1.142561	0.465135	0	0	-0.055514	-1.250824	-0.7777299	-0.433522	-0.789181	...

5 rows × 45 columns

Rysunek 71: Standaryzacja

3.9 Usunięcie wartości odstających

Wartości odstające to obserwacje znacznie odbiegające od pozostałych danych w zbiorze, które mogą mieć istotny wpływ na wyniki analizy oraz jakość budowanych modeli predykcyjnych. Obecność wartości odstających może prowadzić do błędnych wniosków oraz obniżenia dokładności modeli regresji i klasyfikacji. Dlatego ważnym etapem przetwarzania danych jest ich identyfikacja oraz odpowiednie przetworzenie.

W niniejszym rozdziale przeprowadzono analizę wartości odstających w przetworzonych zbiorach danych, które zostały wcześniej wypełnione brakującymi wartościami oraz przeskalowane. Jako kołumnę do usunięcia wartości odstających wybrano kołumnę TotalAssets oraz MonthlyIncome, ponieważ te kołumnę mogą mieć duży wpływ w przypadku użycia regresji.

Poniżej przedstawiono wyniki usunięcia wartości odstających iteracyjnie, dopóki ich nie będzie zero. Widoczne jest że zostało usunięto 3206 rekordów.

Liczba wierszy przed usunięciem wartości odstających: 20000

Liczba wierszy po usunięciu wartości odstających: 16794

3.10 Modele klasyfikacyjne/regresyjne

W niniejszej pracy wykorzystano dwa podejścia do analizy regresji: regresję liniową oraz regresję k-Nearest Neighbors (kNN). Oba te modele zostały zastosowane na zbiorze danych zarówno przed, jak i po usunięciu wartości odstających, aby zbadać ich wpływ na jakość predykcji.

Wartości odstające mogą znacząco zaburzać wyniki analizy, szczególnie w przypadku regresji liniowej, która jest podatna na ich obecność. Dlatego często przeprowadza się proces detekcji i eliminacji obserwacji odstających w celu poprawy dokładności modelu. Regresja kNN, która bazuje na sąsiadach, może być mniej wrażliwa na wartości odstające, jednak również jej wyniki mogą ulec poprawie po usunięciu ekstremalnych obserwacji.

W niniejszym opracowaniu przedstawiono proces budowy i oceny modeli regresji liniowej oraz kNN, a także analizę porównawczą wyników uzyskanych na danych z wartościami odstającymi oraz po ich usunięciu.

```
Wyniki dla zbioru `df`:  
Mean Squared Error (MSE): 0.07  
R2 Score: 0.62
```

```
Wyniki dla zbioru `filtered_df`:  
Mean Squared Error (MSE): 0.05  
R2 Score: 0.60
```

Porównanie wyników:

Różnica MSE: 0.01

Różnica R²: 0.03

Rysunek 72: Wyniki metryki regresji dla danych z usunięciem wartości odstających oraz bez

- Wnioski

- Usunięcie wartości odstających poprawiło MSE, co sugeruje, że model lepiej radzi sobie z przewidywaniem bez wpływu skrajnych wartości.
- Niewielki spadek R² Score może być wynikiem zmniejszenia rozmiaru zbioru danych, co wpłynęło na dopasowanie modelu.
- Warto zastosować usunięcie wartości odstających (lub winsoryzację) dla poprawy jakości modelu, zwłaszcza gdy mamy do czynienia z dużą liczbą wartości odstających.

- 4 Przegląd literatury**
- 5 Motivation**
- 6 Evaluation**
- 7 Zasoby**
- 8 Zastosowane metody**