P O L I T E C H N I K A Ś W I Ę T O K R Z Y S K A
Wydział Elektrotechniki, Automatyki i Informatyki

Daniel Cieślak

Paweł Marek

Maksymilian Sowula

Jakub Szczur

Implementacja aplikacji do analizy statystycznej kredytów

Raport z przedmiotu Statystyka w Informatyce
na studiach II-go stopnia
na kierunku Informatyka

Prowadzący:

Dr. inż. Damian Frej

Kielce 2025

CEL PRACY

Celem niniejszego projektu było przeprowadzenie kompleksowej analizy statystycznej wybranego zbioru danych oraz zaprezentowanie wyników w formie raportu, aplikacji analitycznej oraz prezentacji multimedialnej. Projekt miał na celu przygotowanie studentów do praktycznego wykorzystania metod statystyki opisowej, wizualizacji danych oraz narzędzi wspomagających analizę, takich jak Excel czy autorska aplikacja zespołu.

W ramach pracy każdy zespół dokonywał samodzielnego pozyskania lub wygenerowania zbioru danych odpowiadającego zdefiniowanemu problemowi badawczemu. W przypadku niniejszego projektu zebrano dane dotyczące aplikacji kredytowych, obejmujące m.in. informacje demograficzne, finansowe oraz wskaźniki oceny zdolności kredytowej. Zbiór ten umożliwiał analizę czynników mogących wpływać na decyzję o przyznaniu bądź odrzuceniu wniosku kredytowego. Minimalna liczba obserwacji wymagana przez prowadzącego została spełniona poprzez pozostawienie w zbiorze dokładnie 30 rekordów.

Głównym celem projektu było:

- opracowanie i uporządkowanie zbioru danych, w tym przygotowanie go do dalszej analizy statystycznej,
- obliczenie podstawowych parametrów statystycznych, takich jak średnia, mediana, dominanta, kwartyle, odchylenie standardowe, skośność i kurtoza,
- interpretacja uzyskanych parametrów z uwzględnieniem charakteru badanej cechy oraz możliwych wniosków dotyczących rozkładu wartości,
- przygotowanie odpowiednich wizualizacji, w tym tabel, wykresów oraz przedstawienie wyników w przejrzystej formie,
- opracowanie trzech sekcji aplikacji: zakładki statystycznej (tabele i wykresy), zakładki prezentującej twarze Chernoffa oraz zakładki z interpretacją wyników,
- generacja oraz interpretacja twarzy Chernoffa dla wybranych obserwacji, co pozwalało na graficzne ukazanie różnic pomiędzy zmiennymi,
- opracowanie kompletnego raportu, zgodnego z zasadami przygotowywania prac dyplomowych, obejmującego opis teoretyczny, metodykę, analizę i wnioski końcowe.

Realizacja powyższych celów pozwoła na pogłębienie umiejętności analizy danych, krytycznej oceny wyników oraz prezentowania informacji w sposób zgodny ze standardami akademickimi. Projekt mial również za zadanie rozwijać kompetencje pracy

zespołowej, dzielenia zadań oraz indywidualnej odpowiedzialności za jakość przygotowanych materiałów. Dzięki temu studenci mogli zdobyć praktyczne doświadczenie w prowadzeniu niewielkiego projektu badawczego, którego struktura odpowiada rzeczywistym wymaganiom stawianym analitykom danych.

SPIS TREŚCI

1. WSTĘP

Statystyka jako dyscyplina naukowa stanowi fundament współczesnej analizy danych. Jej narzędzia pozwalają badać zjawiska ilościowe, opisywać ich strukturę, identyfikować zależności oraz formułować prognozy. W dobie intensywnego rozwoju technologii informatycznych i eksplozji dostępności danych statystyka stała się integralnym elementem pracy analityków, inżynierów danych oraz programistów tworzących rozwiązania wspierające przetwarzanie informacji [1][4]. W praktyce statystykę łączy się z elementami programowania, automatyzacji przetwarzania danych oraz wizualizacji wyników, co pozwala na zastosowanie jej metod w złożonych projektach informatycznych, takich jak modele predykcyjne, systemy scoringowe czy narzędzia wspierające podejmowanie decyzji.

Niniejszy projekt ma na celu praktyczne wykorzystanie metod statystycznych w analizie zbioru danych dotyczącego aplikacji kredytowych. Dane te obejmują zarówno cechy demograficzne, jak i ekonomiczne, co umożliwia wielowymiarową analizę czynników wpływających na decyzję kredytową. W ramach projektu przeprowadzono szereg działań obejmujących przygotowanie danych, obliczenie podstawowych parametrów statystycznych, wizualizację wyników oraz opracowanie narzędzia wspierającego analizę danych. Wstęp teoretyczny przedstawia fundamenty metod wykorzystywanych w dalszych częściach raportu, a także omawia strukturę całego opracowania.

1.1. Podstawowe pojęcia statystyki opisowej

1.1.1. Zmienne i zbiory danych

Zmienną nazywa się cechę, której wartości zmieniają się pomiędzy jednostkami obserwowanymi w badaniu [3]. Może to być np. dochód, wiek, liczba lat zatrudnienia czy wysokość wnioskowanej kwoty kredytu. Zmienna może mieć charakter:

- ilościowy (numeryczny) — np. dochód, liczba lat zatrudnienia, wynik punktowy,
- jakościowy (kategorialny) — np. miasto, decyzja kredytowa (tak/nie).

Zbiór wszystkich zmierzonych wartości zmiennej nazywany jest zbiorem danych, zaś pojedynczy wpis — obserwacją.

1.1.2. Miary tendencji centralnej

Miary tendencji centralnej opisują wartości „typowe” dla zbioru danych. Do analizy wykorzystano:

- średnia – najczęściej stosowaną miarą centralną. Oblicza się ją jako sumę wszystkich wartości podzieloną przez ich liczbę. Jej zaletą jest prostota i podatność na dalsze przetwarzanie matematyczne. Wadą — wrażliwość na wartości odstające.
- mediana – wartość środkowa w uporządkowanym zbiorze danych. Jest odporna na obserwacje ekstremalne, dlatego często lepiej od średniej opisuje zbiór o rozkładzie asymetrycznym [3][4].
- dominanta – najczęściej występująca wartość w zbiorze. Jest szczególnie przydatna w analizie zmiennych jakościowych lub dyskretnych.

1.1.3. Miary rozproszenia

Miary zmienności informują o tym, jak bardzo wartości danej zmiennej różnią się od siebie. Do analizy wykorzystano:

- odchylenie standardowe – jedna z najpowszechniej stosowanych miar zróżnicowania. Informuje o przeciętnym odchyleniu obserwacji od średniej. Wysokie odchylenie wskazuje na znaczne rozproszenie danych.
- suma wartości – informuje o łącznej wartości wszystkich obserwacji w zbiorze; może być przydatna w analizie wielkości zbioru lub agregacji danych.

1.1.4. Kwartyle i podział danych

Kwartyle dzielą uporządkowany zbiór danych na cztery równe części:

- Q1 — pierwszy kwartyl, poniżej którego znajduje się 25% danych,
- Q2 — drugi kwartyl, czyli mediana,
- Q3 — trzeci kwartyl, powyżej którego znajduje się 25% danych.

1.1.5. Miary kształtu rozkładu

Oprócz miar centralnych i miar zmienności istotne znaczenie mają miary opisujące kształt rozkładu wartości zmiennej

Skośność informuje o asymetrii rozkładu.

- Skośność dodatnia świadczy o długim „ogonie” po prawej stronie — większość obserwacji przyjmuje wartości niższe, a pojedyncze wysokie wartości podnoszą średnią.
- Skośność ujemna oznacza długi „ogonie” po lewej stronie — dominują wartości wysokie, a nieliczne niskie ciągną średnią w dół.
- Skośność bliska zeru oznacza rozkład symetryczny.

Skośność odgrywa istotną rolę w interpretacji danych finansowych, np. dochodów, które bardzo często są dodatnio skośne.

Kurtoza określa, czy rozkład jest „bardziej skupiony” lub „bardziej płaski” niż rozkład normalny.

- Kurtoza dodatnia (leptokurtyczność) — większa koncentracja wartości wokół średniej; większa liczba wartości ekstremalnych.
- Kurtoza ujemna (platykurtyczność) — rozkład bardziej płaski, mniejsze zróżnicowanie.
- Kurtoza bliska zeru — rozkład normalny.

Miary skośności i kurtozy są szczególnie cenne w analizie ryzyka kredytowego, gdzie występowanie ekstremów (np. bardzo wysokich lub bardzo niskich dochodów) może wpływać na decyzję kredytową [4][5].

1.2. Wizualizacja danych w statystyce

Wizualizacja danych jest kluczowym etapem analizy statystycznej. Pozwala szybko zauważyć zależności, trendy i anomalie, które mogą nie być widoczne wyłącznie na podstawie tabel liczbowych [1][7].

Do najczęściej stosowanych wykresów należą:

- histogram — przedstawia rozkład wartości zmiennej ilościowej,
- wykres pudełkowy — umożliwia prezentację mediany, kwartylów oraz obserwacji odstających,
- wykres punktowy — pokazuje zależność między dwiema zmiennymi,
- wykres słupkowy — stosowany dla danych kategorialnych,
- wykres liniowy — wykorzystywany najczęściej dla danych czasowych.

W ramach projektu część wizualizacji została zrealizowana przez przygotowaną przez zespół aplikację.

1.3. Twarze Chernoffa

Jednym z bardziej nietypowych sposobów prezentacji danych wielowymiarowych są twarze Chernoffa. Metodę tę zaproponował Hermann Chernoff w 1973 roku [6], zauważając, że ludzie są niezwykle wrażliwi na różnice w wyglądzie twarzy, co można wykorzystać do analizy danych.

1.3.1. Zasada działania

Każdej obserwacji przypisuje się „twarz", której poszczególne elementy (np. kształt oczu, długość nosa, nachylenie brwi, wielkość głowy) odpowiadają wartościom różnych zmiennych. Pozwala to na szybkie porównanie wielu cech jednocześnie — nawet kilkunastu zmiennych w formie jednego rysunku.

1.3.2. Zastosowanie

Twarze Chernoffa znajdują zastosowanie m.in. w:

- analizie porównawczej jednostek (np. klientów, produktów, regionów),
- prezentacji wyników klasyfikacji,
- analizie danych psychologicznych, socjologicznych i finansowych,
- eksploracji danych wielowymiarowych.

W projektach statystycznych metoda ta pełni funkcję wizualizacji wspomagającej analizę – pomaga dostrzec podobieństwa między obserwacjami oraz wyróżnia jednostki nietypowe.

1.4. Struktura raportu i zakres kolejnych rozdziałów

Raport został podzielony na logiczne części, z których każda pełni określoną funkcję w procesie analitycznym i dokumentacyjnym.

1.4.1. Metodyka badawcza (rozdział 2)

W rozdziale „Metodyka badawcza" przedstawiono kompletny proces pozyskania i przygotowania danych. Omówiono źródło zbioru, sposób jego pozyskania, liczbę obserwacji oraz metody zastosowane do czyszczenia danych, identyfikacji wartości odstających, ujednolicania typów danych i łączenia informacji z różnych źródeł. Rozdział obejmuje również szczegółowy opis narzędzi wykorzystanych podczas analizy, takich jak Python (biblioteki Pandas i NumPy), czy Excel.

Ponadto zaprezentowano autorską aplikację stworzoną na potrzeby projektu. Jej funkcjonalności obejmują trzy główne moduły: analizę statystyczną (tabele, wykresy i podstawowe parametry opisowe), generację twarzy Chernoffa oraz prezentację wyników w formie zbiorczych zestawień. W rozdziale zamieszczone zostaną także ilustracje (zrzuty ekranu), które pozwolą przedstawić wygląd interfejsu oraz sposób korzystania z przygotowanego narzędzia.

1.4.2. Analiza wyników (rozdział 3)

Rozdział „Analiza wyników" stanowi centralną część raportu. Zawiera on prezentację wyników uzyskanych w procesie analizy statystycznej wraz z odpowiednimi tabelami, wykresami i opisami. Omówiono w nim podstawowe parametry statystyczne, takie jak suma, średnia arytmetyczna, mediana, dominanta, odchylenie standardowe oraz kwartyle. W rozdziale przedstawiono również interpretację wyników w kontekście badanych cech, zwracając uwagę na strukturę danych, ich zmienność oraz potencjalne zależności między zmiennymi. Zawarte tu wnioski pozwalają zrozumieć, które zmienne wykazują największy wpływ na decyzję kredytową.

1.4.3. Twarze Chernoffa (rozdział 4)

W rozdziale „Twarze Chernoffa" zastosowano antropomorficzną metodę wizualizacji danych wielowymiarowych. Dla kilku wybranych obserwacji wygenerowano twarze Chernoffa, w których poszczególne elementy (kształt oczu, usta, brwi, wielkość głowy itp.) odpowiadają wartościom różnych zmiennych. Poniższa część raportu zawiera zarówno przedstawienia graficzne, jak i ich interpretację – wskazanie, które cechy różnią analizowane obserwacje oraz jak te różnice wpływają na wygląd twarzy. Rozdział ten pełni funkcję pogłębiającą analizę, pozwalając zobaczyć zależności między zmiennymi w mniej konwencjonalny sposób.

1.4.4. Wnioski (rozdział 5)

Rozdział „Wnioski" syntetyzuje wszystkie informacje zebrane podczas projektu. Przedstawiono w nim najważniejsze zależności, które można zaobserwować w danych, dokonano interpretacji parametrów statystycznych, a także odniesiono uzyskane wyniki do celu pracy. Wskazano również potencjalne kierunki dalszych badań, np. rozszerzenie analizy o modele predykcyjne, uwzględnienie dodatkowych zmiennych czy porównanie wyników z rzeczywistymi decyzjami instytucji finansowych.

1.4.5. Bibliografia

Ostatni rozdział raportu zawiera pełny wykaz źródeł naukowych, internetowych oraz dokumentacyjnych, które zostały wykorzystane podczas opracowania raportu. Bibliografia obejmuje zarówno literaturę teoretyczną, jak i praktyczną – w tym podręczniki, dokumentację narzędzi programistycznych i opis zbiorów danych.

W ramach przygotowania niniejszego raportu wykorzystano literaturę dotyczącą zarówno podstaw statystyki, jak i jej zastosowań w analizie danych oraz programowaniu. Jednym z kluczowych źródeł była książka „Statystyka praktyczna w data science. 50 kluczowych zagadnień w językach R i Python. Wydanie II" [1].

Publikacja ta stanowi nowoczesne i praktyczne opracowanie zagadnień statystycznych, skierowane przede wszystkim do analityków danych oraz programistów. Książka podkreśla znaczenie właściwego stosowania metod statystycznych w kontekście data science, a także prezentuje wiele przykładów implementacji w językach R i Python. Autorzy omawiają zarówno klasyczne narzędzia statystyczne, jak i techniki wykorzystywane w uczeniu maszynowym, analizie eksploracyjnej danych oraz modelowaniu.

Zawarte w książce koncepcje, takie jak analiza eksploracyjna, zasady planowania eksperymentów, podstawy regresji czy metody wykrywania anomalii, pozwoliły na lepsze zrozumienie struktury danych kredytowych oraz właściwy dobór parametrów analizy. Publikacja była również pomocna przy interpretacji wyników statystycznych, dzięki klarownemu przedstawieniu praktycznych konsekwencji stosowania poszczególnych metod oraz typowych błędów analitycznych. Książka ta, zgodnie z jej przeznaczeniem, łączy wiedzę statystyczną z podejściem informatycznym, co czyni ją szczególnie przydatną dla studentów kierunków technicznych.

2. METODYKA BADAWCZA

Rozdział „Metodyka badawcza” ma na celu szczegółowe przedstawienie źródeł danych, procesu ich przygotowania do analizy, narzędzi i programów użytych w projekcie, a także opisanie przygotowanej przez zespół aplikacji. W niniejszym opracowaniu opisano również sposób organizacji danych, ich przetwarzanie oraz przygotowanie do dalszej części analizy statystycznej i wizualizacji wyników.

2.1. Źródła danych

Dane wykorzystane w projekcie pochodzą z serwisu Kaggle, z zestawu „Loan Approval Dataset”[2].

Zbiór danych symuluje wnioski kredytowe oraz wyniki decyzji o ich przyznaniu dla 2 000 osób. Zawiera zmienne demograficzne, finansowe i dotyczące zatrudnienia, co umożliwia ocenę ryzyka kredytowego i zastosowanie klasyfikacji w analizie danych.

Dla potrzeb projektu wybrano próbkę 30 rekordów reprezentujących pełny przekrój danych. Wybrane obserwacje obejmują informacje o dochodach, zdolności kredytowej, wnioskowanej kwocie kredytu, długości zatrudnienia, liczbie punktów oceny wniosku oraz decyzji kredytowej.

Dane te umożliwiają praktyczne zastosowanie metod statystycznych, takich jak analiza opisowa, obliczanie miar tendencji centralnej i rozproszenia, wizualizacja oraz generowanie twarzy Chernoffa dla danych wielowymiarowych.

2.2. Przygotowanie zbioru danych

Zbiór danych został poddany procesowi wstępnego przygotowania, który obejmował następujące etapy:

2.2.1. Weryfikacja kompletności danych

Pierwszym krokiem była kontrola liczby rekordów oraz obecności wszystkich zmiennych. Dla każdej obserwacji sprawdzono, czy wszystkie kluczowe zmienne (income, credit_score, loan_amount, years_employed, points, loan_approved) posiadają wartości. W przypadku braków danych w zestawie większym niż 30 obserwacji stosuje się metody uzupełniania, jednak w badanej próbce wszystkie rekordy były kompletne.

2.2.2. Czyszczenie danych

Dane zostały oczyszczone z niepoprawnych i niespójnych wartości. Sprawdzenie obejmowało:

- eliminację wartości ujemnych w kolumnach finansowych (dochód, kwota kredytu),
- kontrolę zakresu punktów oceny wniosku (0–100),
- identyfikację duplikatów i ich usunięcie, jeśli występowały.

2.2.3. Standaryzacja i formatowanie

Dane zostały ujednolicone w formacie CSV z separatorem średnika („;”). Liczby zmiennoprzecinkowe (np. kolumna points) zostały przekształcone do postaci numerycznej, aby umożliwić dalszą analizę statystyczną w aplikacji Python. Nazwy miast i osób zachowano w oryginalnej formie tekstowej, aby zachować identyfikację jednostek.

2.2.4. Wybór zmiennych do analizy

Do dalszej analizy wybrano zmienne, które mają największe znaczenie w kontekście decyzji kredytowych:

- income — miesięczny dochód klienta,
- credit_score — punktacja oceny zdolności kredytowej,
- loan_amount — kwota wnioskowanego kredytu,
- years_employed — liczba lat zatrudnienia,
- points — liczba punktów przyznanych w procesie oceny wniosku,
- loan_approved — decyzja kredytowa.

Zmienna name i city zostały pozostawione do celów identyfikacyjnych i wizualizacji, natomiast w analizie statystycznej uwzględniono jedynie wartości liczbowe.

2.3. Narzędzia i oprogramowanie

Do przygotowania danych oraz przeprowadzenia analizy statystycznej wykorzystano następujące narzędzia:

2.3.1. Excel

Arkusz kalkulacyjny wykorzystano do:

- weryfikacji poprawności danych,
- czyszczenia danych.

2.3.2. Backend aplikacji

Backend aplikacji został opracowany w języku Python 3.12, przy użyciu frameworka Flask 3.0.3, który umożliwia obsługę logiki aplikacji oraz przetwarzanie danych. Do analizy danych wykorzystano biblioteki:

- Pandas 2.2.2 — do wczytywania, filtrowania, agregacji danych oraz obliczania podstawowych parametrów statystycznych, takich jak średnia, mediana, kwartyle, odchylenie standardowe.
- NumPy 1.26.4 — do obliczeń matematycznych i przekształceń danych.
- Flask 3.0.3 — framework backendowy użyty w aplikacji do zarządzania logiką i przetwarzania danych.

2.3.3. Frontend aplikacji

Frontend został opracowany przy użyciu:

- React 18.3.1 — budowa interfejsu użytkownika,
- Vite 5.3.5 — narzędzie do szybkiego uruchamiania środowiska frontendowego,
- TypeScript 5.2.2 — typowanie statyczne kodu,
- Bootstrap 5.3.3 — tworzenie responsywnego interfejsu.

Dzięki połączeniu tych narzędzi możliwe było przygotowanie interaktywnej aplikacji umożliwiającej zarówno wizualizację danych, jak i obliczenie ich podstawowych parametrów statystycznych w czasie rzeczywistym [7][8][9].

2.4. Opis aplikacji

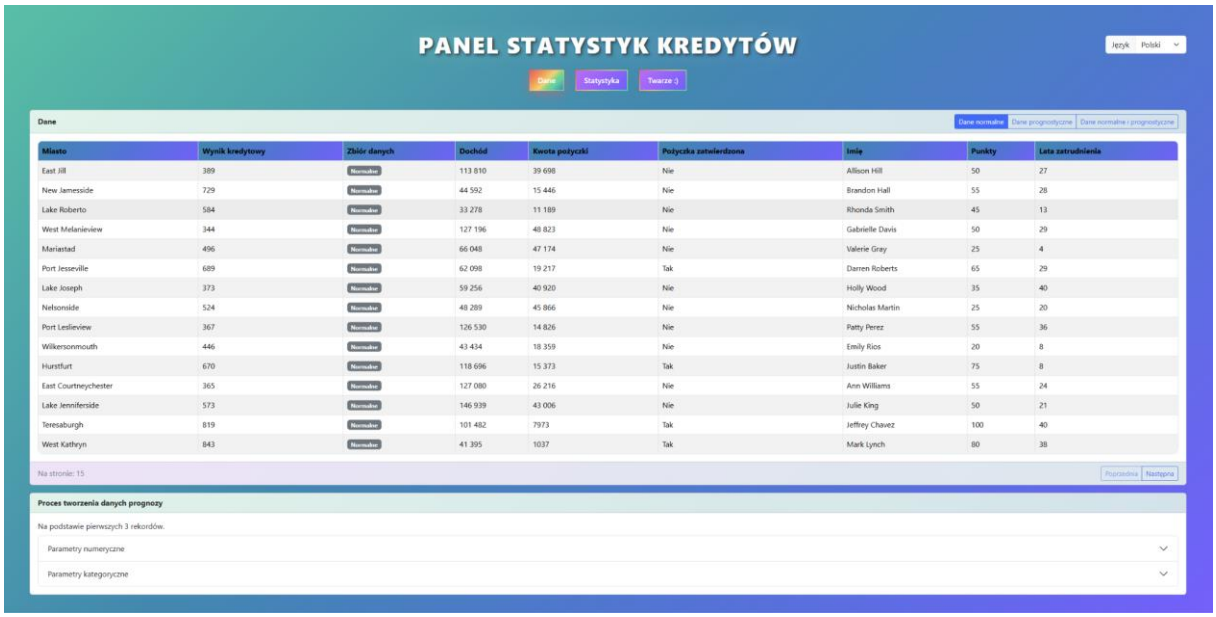
Stworzona aplikacja jest interaktywną platformą analityczną, zrealizowaną w architekturze SPA (Single Page Application), która umożliwia dogłębną analizę i wizualizację danych dotyczących wniosków kredytowych. Głównym celem aplikacji jest przedstawienie złożonych danych w przystępny i zrozumiały sposób, a także demonstracja możliwości prognozowania zdolności kredytowej na podstawie historycznych danych.

Interfejs użytkownika został zaprojektowany z myślą o prostocie i intuicyjnej nawigacji. Główne okno aplikacji podzielone jest na system zakładek, które grupują funkcjonalności w logiczne moduły. W prawym górnym rogu interfejsu umieszczono przełącznik języków, pozwalający na dynamiczną zmianę języka całej aplikacji (dostępne języki: polski, angielski, niemiecki, chiński, koreański), co czyni ją dostępną dla międzynarodowego użytkownika.

Poniżej znajduje się szczegółowy opis poszczególnych modułów (zakładek) aplikacji:

1. Zakładka „Dane” (DataTab)

Jest to ekran startowy aplikacji. Prezentuje on surowy zbiór danych w formie przejrzystej tabeli. Użytkownik może w tym miejscu zapoznać się z poszczególnymi rekordami i atrybutami, takimi jak płeć, status cywilny, liczba posiadanych nieruchomości, dochód, kwota kredytu itp. Zakładka ta stanowi punkt wyjścia do dalszej analizy, dając wgląd w strukturę i zawartość danych.

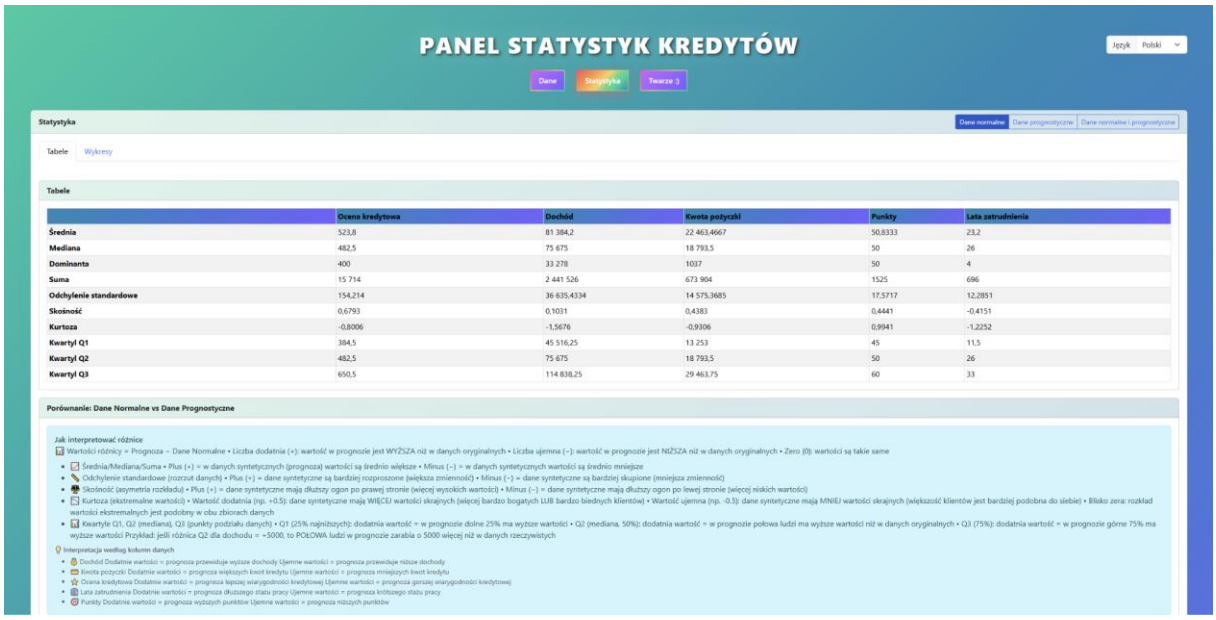


Rysunek 2.1. Zakładka „Dane”

2. Zakładka „Statystyki” (StatisticsTab)

Ten moduł jest sercem analitycznym aplikacji. Umożliwia on generowanie i przeglądanie kluczowych statystyk opisowych dla całego zbioru danych. Aplikacja prezentuje dwie główne tabele:

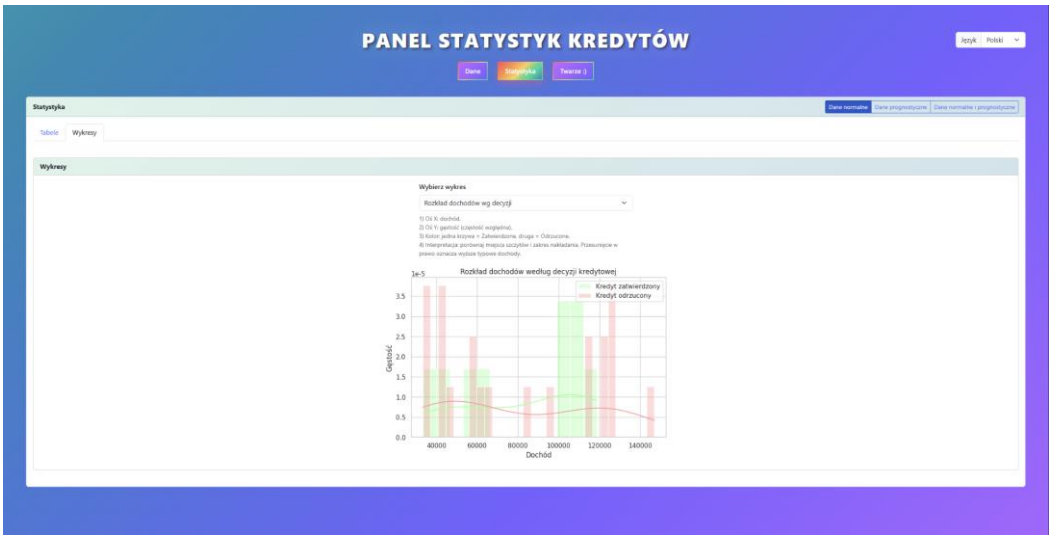
- Podsumowanie statystyczne: Zawiera podstawowe miary statystyczne (np. średnia, mediana, odchylenie standardowe, wartości minimalne i maksymalne) dla poszczególnych kolumn numerycznych.
- Porównanie statystyk: Pozwala na porównanie statystyk pomiędzy dwiema grupami – na przykład osobami, które otrzymały kredyt, i tymi, którym go odmówiono. Ułatwia to identyfikację kluczowych różnic między grupami.



Rysunek 2.2. Zakładka „Statystyki”

3. Zakładka „Wykresy” (ChartsTab)

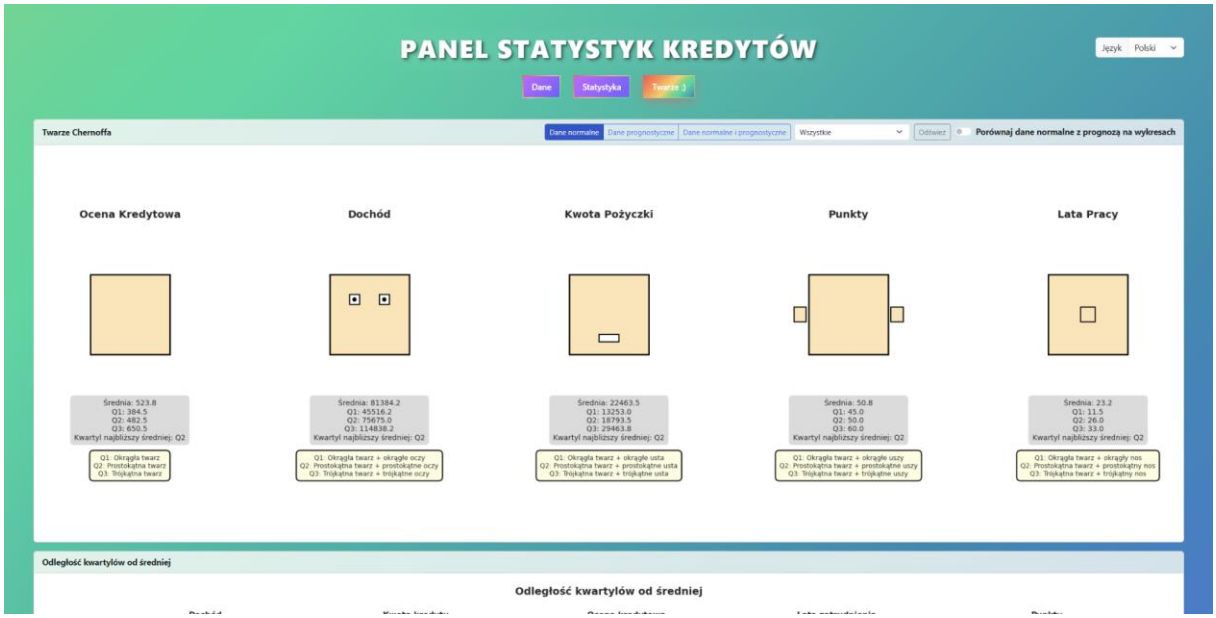
W celu ułatwienia interpretacji danych liczbowych, aplikacja oferuje moduł do wizualizacji. Zakładka "Wykresy" dynamicznie generuje graficzne reprezentacje statystyk, takie jak wykresy słupkowe czy kołowe. Użytkownik może w sposób wizualny porównać rozkłady poszczególnych cech, np. procentowy udział kobiet i mężczyzn w zbiorze danych czy rozkład statusu edukacyjnego kredytobiorców.



Rysunek 2.3. Zakładka „Wykresy”

4. Zakładka "Twarze Chernoffa" (ChernoffFacesTab)

Jest to zaawansowana i unikalna metoda wizualizacji danych wielowymiarowych [1][4][5]. Aplikacja implementuje algorytm Twarzy Chernoffa, który mapuje poszczególne zmienne (atrybuty kredytobiorcy) na cechy ludzkiej twarzy, takie jak kształt głowy, wielkość oczu, krzywizna ust czy długość nosa. Dzięki tej metodzie analityk jest w stanie w sposób intuicyjny i niemal natychmiastowy identyfikować wzorce, klastry oraz wartości odstające w danych, które byłyby trudne do zauważenia w tradycyjnej tabeli czy na standardowych wykresach.



Rysunek 2.4. Zakładka „Twarze Chernoffa”

3. ANALIZA WYNIKÓW

Niniejszy rozdział poświęcony jest szczegółowej analizie statystycznej danych kredytowych użytkowników, stanowiącej kluczowy element oceny wiarygodności kredytowej i prognozowania przyszłych wyników. Celem tej analizy jest zrozumienie charakterystyki danych rzeczywistych (Normalne Dane) oraz porównanie ich z danymi syntetycznymi (Dane Prognostyczne) w celu oceny jakości i przydatności modelu prognostycznego.

3.1. Zakres analizy i zastosowane metody

Analiza została przeprowadzona z wykorzystaniem podstawowych i zaawansowanych parametrów statystycznych oraz technik wizualizacji danych. Zastosowane metody obejmują:

- Analizę podstawowych parametrów statystycznych: Obliczono i zinterpretowano kluczowe miary tendencji centralnej (średnia, mediana, dominanta) oraz miary rozrzutu i kształtu rozkładu (odchylenie standardowe, skośność, kurtoza),
- Analizę pozycyjną: Wykorzystano kwartyle (Q1, Q2, Q3) do określenia punktów podziału danych i zrozumienia ich rozkładu,
- Porównanie danych normalnych i prognostycznych: Przedstawiono i zinterpretowano różnice między danymi rzeczywistymi a prognozowanymi dla wszystkich kluczowych zmiennych,
- Wizualizację danych: Wyniki analizy zostały przedstawione za pomocą tabel statystycznych, wykresów oraz Wizualizacji Twarzy Chernoffa (prezentującej dane normalne i prognozę na wykresach).

3.2. Prezentacja i interpretacja wyników

W kolejnych sekcjach rozdziału przedstawiono:

- Dane statystyczne: Szczegółowe tabele zawierające zestawienie parametrów statystycznych dla zmiennych numerycznych takich jak Ocena kredytowa, Dochód, Kwota pożyczki, Punkty i Lata zatrudnienia.
- Analiza różnic (prognoza - dane normalne): Tabela zawierająca wartości różnic w kluczowych parametrach statystycznych, co pozwala ocenić, czy dane syntetyczne przewidują wartości wyższe (+) czy niższe (-) niż dane oryginalne, a także czy są

bardziej rozproszone (większa zmienność) czy bardziej skupione (mniejsza zmienność).

- Obserwowane zależności i tendencje: Interpretacja wyników statystycznych, w tym opis zaobserwowanych asymetrii rozkładu (skośność) i występowania wartości skrajnych (kurtoza), co jest kluczowe dla oceny ryzyka kredytowego i profilu klientów.
- Interpretacja wyników ma na celu wyciągnięcie wniosków na temat dynamiki i trendów w danych, co stanowi podstawę do podejmowania decyzji biznesowych związanych z przyznawaniem kredytów

3.3. Analiza podstawowych parametrów statystycznych - dane normalne

Poniższa tabela zawiera kluczowe parametry statystyczne obliczone przez stworzone w ramach projektu oprogramowanie, dla zbioru Danych Normalnych (rzeczywistych). Dostarcza ona obraz charakterystyki populacji klientów ubiegających się o kredyt.

Tabela 3.1. Parametry statystyczne zbioru danych normalnych

Parametr	Ocena kredytowa	Dochód	Kwota pożyczki	Punkty	Lata zatrudnienia
Średnia	523,8	81 384,2	22 463,47	50,83	23,2
Mediana (Q2)	482,5	75 675	18 793,5	50	26
Dominanta	400	33 278	1037	50	4
Odchylenie Standardowe	154,214	36 635,43	14 575,37	17,57	12,285
Skośność	0,6793	0,1031	0,4383	0,4441	-0,4151
Kurtoza	-0,8006	-1,5676	-0,9306	0,9941	-1,2252
Kwartyl Q1	384,5	45 516,25	13 253	45	11,5

Kwartyl Q3	650,5	114 838,25	29 463,75	60	33
Suma	15 714	2 441 526	673 904	1525	696

W przypadku analizy tendencji centralnej (średnia i mediana) zaobserwowano następujące zależności:

- Ocena kredytowa: Prognozowane dane wskazują na wyraźnie wyższą ocenę kredytową klientów w porównaniu z danymi rzeczywistymi. Zarówno średnia (+77,42), jak i mediana (+130,35) mają wartości dodatnie, co oznacza, że przeciętny klient w prognozie uzyskuje znacząco lepszy wynik kredytowy. Mediana sugeruje, że połowa klientów ma ocenę wyższą o około 130 punktów względem danych normalnych, co wskazuje na ogólne podniesienie wiarygodności kredytowej w danych syntetycznych.
- Dochód: W przypadku dochodu prognoza przewiduje niższe wartości w porównaniu z danymi rzeczywistymi. Zarówno średnia (-31 038,12), jak i mediana (-15 888,61) mają wartości ujemne, co oznacza, że przeciętny klient w danych syntetycznych zarabia mniej. Połowa klientów ma dochód niższy o około 15 889 zł, co wskazuje na przesunięcie całego rozkładu dochodów w dół.
- Lata zatrudnienia: Średnia różnica lat zatrudnienia jest dodatnia (+2,3), co oznacza, że prognoza zakłada nieco dłuższy przeciętny staż pracy. Jednocześnie mediana wynosi 0, co sugeruje, że dla połowy klientów staż pracy pozostaje taki sam jak w danych rzeczywistych. Wskazuje to, że wzrost przeciętnej wartości wynika głównie z klientów o bardzo długim stażu.
- Kwota pożyczki: W prognozie klienci wnioskują przeciętnie o wyższe kwoty pożyczek. Zarówno średnia (+494,54), jak i mediana (+1 745,36) są dodatnie, co oznacza, że większość klientów zgłasza zapotrzebowanie na wyższe kwoty niż w danych normalnych.

Analizując zmienność danych na podstawie wartości odchylenia standardowego, można sformułować następujące wnioski dotyczące stopnia rozproszenia zmiennych oraz różnic pomiędzy danymi prognozowanymi a rzeczywistymi:

- Dochód: Odchylenie standardowe dochodu jest dodatnie (+8 397,21), co oznacza, że prognozowane dane wykazują większą zmienność niż dane rzeczywiste. Wskazuje to, że model generuje bardziej zróżnicowaną populację pod względem wysokości zarobków.

- Dla pozostałych zmiennych - Oceny kredytowej, Kwoty pożyczki, Punktów oraz Lat zatrudnienia odchylenia standardowe mają wartości ujemne, co oznacza mniejszą zmienność danych syntetycznych w stosunku do danych rzeczywistych. Innymi słowy, model generuje bardziej „jednorodnych" klientów w tych kategoriach.

Dokonano również analizy kształtu rozkładu, obejmującej ocenę skośności i kurtozy. Na tej podstawie odnotowano kierunek asymetrii poszczególnych zmiennych oraz stopień spiczastości lub spłaszczenia rozkładów, co pozwoliło określić obecność wartości skrajnych oraz ogólną charakterystykę rozkładu danych syntetycznych względem danych rzeczywistych.

- Skośność:
  - Ocena kredytowa i Dochód: Ujemne wartości skośności wskazują, że prognozowane rozkłady mają dłuższy ogon po lewej stronie, czyli więcej niskich wartości. Oznacza to, że w danych syntetycznych częściej pojawiają się klienci z niższymi wynikami kredytowymi lub niższymi dochodami.
  - Kwota pożyczki i Punkty: Wartości skośności bliskie zeru sugerują, że ich rozkłady nie odbiegają znacząco od symetrii.
  - Lata zatrudnienia: Dodatnia skośność (+1,9837) wskazuje na dłuższy ogon po prawej stronie, czyli obecność osób o znacząco dłuższym stażu pracy.
- Kurtoza
  - Lata zatrudnienia: Bardzo wysoka dodatnia kurtoza (+4,48) świadczy o dużej liczbie wartości skrajnych — w prognozie występują zarówno bardzo krótkie, jak i wyjątkowo długie staże pracy.
  - Punkty: Znacznie ujemna kurtoza (-2,62) wskazuje na bardziej spłaszczony rozkład z mniejszą liczbą wartości skrajnych niż w danych rzeczywistych.
  - Pozostałe zmienne: Wartości kurtozy dla Oceny kredytowej, Dochodu i Kwoty pożyczki są bliskie zeru, co sugeruje, że ich kształt jest zbliżony do rozkładu normalnego, z umiarkowaną liczbą obserwacji skrajnych.

Przeprowadzona analiza porównawcza danych syntetycznych z danymi rzeczywistymi pozwala stwierdzić, że model predykcyjny generuje populację klientów o wyraźnie zmodyfikowanych parametrach. Prognoza zakłada znacznie wyższą ocenę kredytową oraz większe kwoty pożyczek, co wskazuje na bardziej optymistyczną charakterystykę zdolności kredytowej klientów. Jednocześnie przewiduje niższe dochody

oraz nieznacznie dłuższy przeciętny staż pracy, przy czym wartości ekstremalnie długiego stażu pojawiają się w modelu częściej niż w danych rzeczywistych.

Analiza zmienności pokazuje, że syntetyczne dane dochodowe są bardziej zróżnicowane, natomiast pozostałe zmienne stają się bardziej jednorodne. Ocena skośności i kurtozy potwierdza zmiany w kształcie rozkładów. Model generuje więcej niskich wartości w zakresie dochodów i ocen, a jednocześnie wytwarza skrajne przypadki w stażu pracy.

Przyjęta w projekcie predykcja danych przesuwa parametry klientów w kierunku bardziej korzystnego profilu kredytowego, ale jednocześnie wprowadza pewne odchylenia strukturalne, szczególnie widoczne w stażu pracy i wysokości dochodów. Wyniki te wskazują, że dane syntetyczne różnią się od rzeczywistych zarówno pod względem poziomu wartości, jak i kształtu rozkładu.

3.4. Analiza graficzna wyników

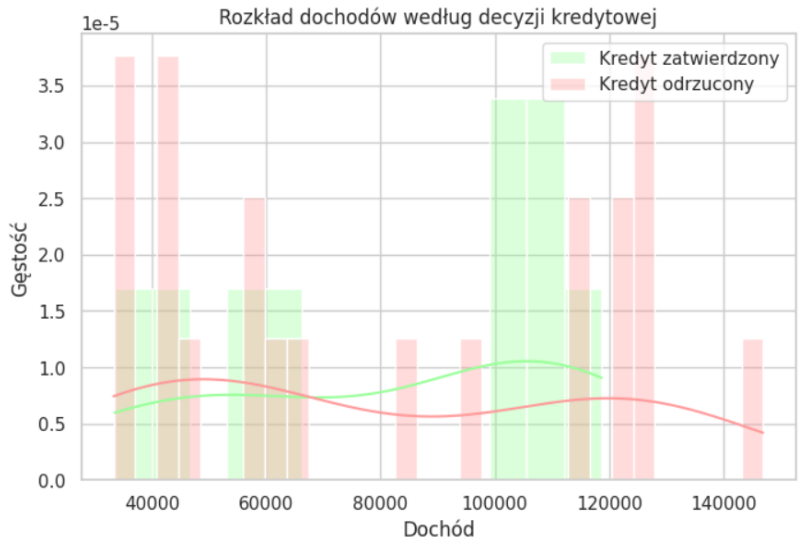
W celu pełniejszego zrozumienia struktury danych oraz różnic pomiędzy danymi rzeczywistymi a danymi syntetycznymi, wygenerowano zestaw wykresów ilustrujących najważniejsze zależności, rozkłady i porównania. Wizualizacja danych stanowi kluczowy etap analizy, ponieważ umożliwia szybkie wychwycenie trendów, odchyleń, asymetrii czy wartości odstających, które mogą być mniej czytelne w tabelach statystycznych.

W dalszej części rozdziału przedstawiono interpretację poszczególnych wykresów, koncentrując się zarówno na ich ogólnym kształcie, jak i szczegółowych różnicach pomiędzy danymi normalnymi a prognozowanymi. Każdy podpunkt omawia oddzielny typ wykresu, opisuje zaobserwowane zależności oraz wskazuje możliwe przyczyny i konsekwencje przedstawionych wyników.

3.4.1. Analiza wykresu rozkładu dochodów według decyzji kredytowej

Pierwszym elementem analizy graficznej jest ocena rozkładu dochodów klientów w kontekście podjętej decyzji kredytowej. Wykres gęstości pozwala zobrazować, jak kształtują się dochody osób, którym kredyt przyznano, oraz tych, których wnioski zostały odrzucone. Taka wizualizacja umożliwia bezpośrednie porównanie obu grup oraz identyfikację różnic i podobieństw, które mogą wpływać na proces decyzyjny. Poniżej przedstawiono interpretację zauważonych zależności.

1) Oś X: dochód.
2) Oś Y: gęstość (częstość względna).
3) Kolor: jedna krzywa = Zatwierdzone, druga = Odrzucone.
4) Interpretacja: porównaj miejsca szczytów i zakres nakładania. Przesunięcie w prawo oznacza wyższe typowe dochody.



Rysunek 3.1. Wykres rozkładu dochodów według decyzji kredytowej

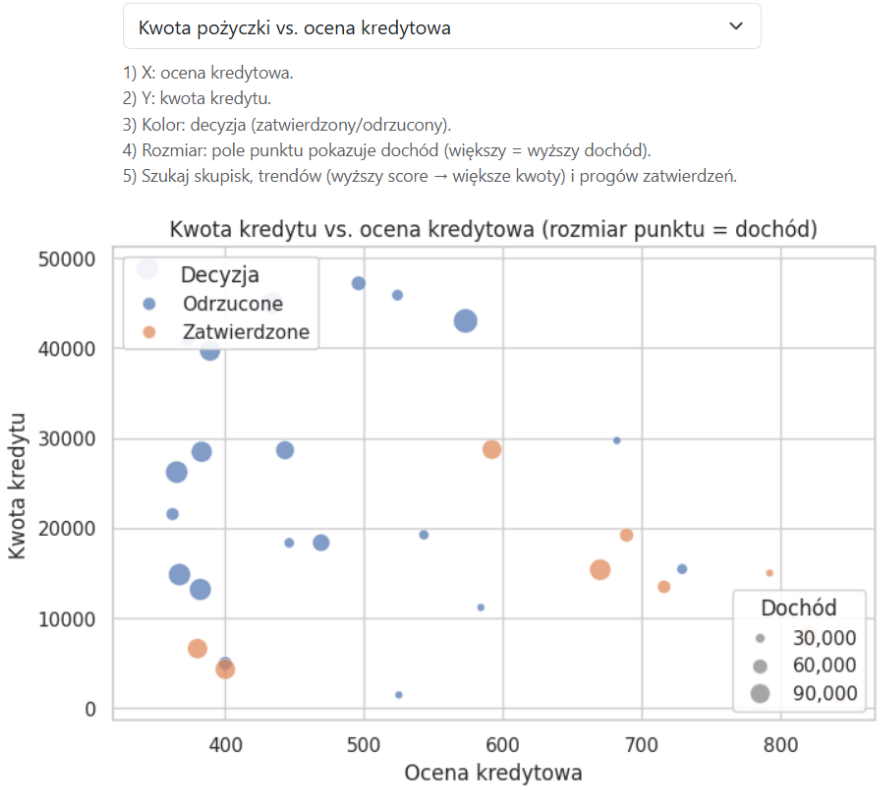
Prezentowany wykres gęstości przedstawia, jak rozkładają się dochody klientów w zależności od tego, czy ich wniosek kredytowy został zatwierdzony (zielony), czy odrzucony (czerwony). Można z niego odczytać, że klienci, którym zatwierdzono kredyt (zielony), mają tendencję do koncentrowania się w wyższych przedziałach dochodowych (szczyt zielonej krzywej jest przesunięty w prawo, zwłaszcza w okolicach 100 000 zł i więcej), podczas gdy klienci, których wnioski odrzucono (czerwony), dominują w niższych i średnich przedziałach dochodowych. Mimo to, występuje znaczne nakładanie się rozkładów, co oznacza, że duża grupa klientów o zbliżonych dochodach otrzymuje sprzeczne decyzje, sugerując, że dochód jest ważnym, ale nie jedynym czynnikiem decyzyjnym.

Taki wykres jest niezbędny i użyteczny, ponieważ pozwala wizualnie zrozumieć politykę ryzyka i jej konsekwencje. Wskazuje on, jaki typ klienta pod względem dochodowym jest akceptowany, a jaki odrzucany, a także identyfikuje obszary niejednoznaczne (nakładanie się), w których dochód nie przesądza o wyniku, co zachęca do dalszej, bardziej szczegółowej analizy innych zmiennych.

3.4.2. Analiza wykresu kwota pożyczki vs. ocena kredytowa

Kolejnym elementem analizy jest wykres rozrzutu przedstawiający zależności pomiędzy kwotą pożyczki, oceną kredytową oraz dochodem klientów, z uwzględnieniem

decyzji kredytowej. Taka forma wizualizacji umożliwia jednoczesne uchwycenie wpływu wielu czynników na proces podejmowania decyzji oraz pozwala obserwować, jak poszczególne zmienne oddziałują na siebie w praktyce.



Rysunek 3.2. Wykres kwoty kredytu w odniesieniu do oceny kredytowej klienta

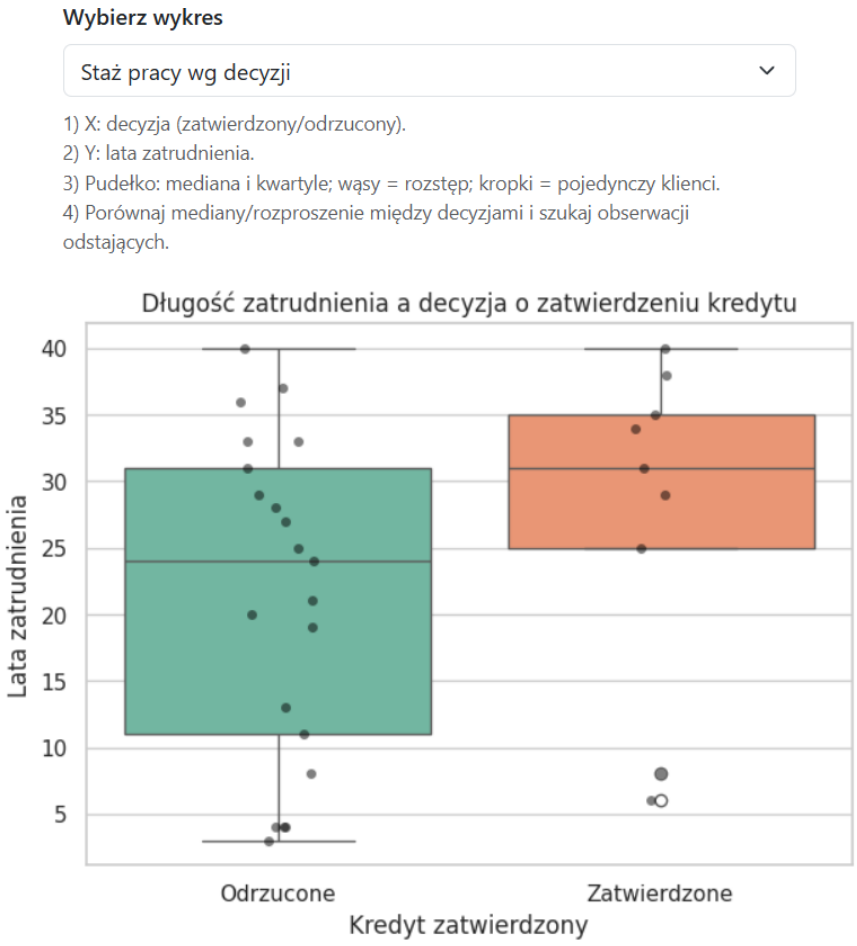
Prezentowany wykres jest diagramem rozrzutu, który ukazuje kluczowe zależności trójwymiarowe między Oceną kredytową (Oś X), Kwotą kredytu (Oś Y) a Decyzją (kolor: czerwony/niebieski), przy czym dodatkowo wielkość punktu reprezentuje dochód klienta. Wizualna analiza potwierdza silny trend: wnioski o kredyty zatwierdzone (pomarańczowe) koncentrują się w obszarze wyższych ocen kredytowych (głównie powyżej 650) i to zarówno dla małych, jak i dużych kwot. Co ciekawe, w grupie klientów odrzuconych (niebieskie) dominuje słaba lub przeciętna ocena kredytowa (poniżej 600), a punkty te są liczne i rozproszone, w tym również te o dużym rozmiarze (wysokim dochodzie) i wnioskujące o wysokie kwoty pożyczki. Wykres ten uwidacznia kluczowy próg decyzyjny - mianowicie, poniżej pewnego poziomu (okolice na poziomie oceny kredytowej 600), ocena kredytowa jest głównym czynnikiem odrzucającym, niezależnie od wysokiego dochodu klienta.

Taki wykres został wybrany, ponieważ w sposób przejrzysty identyfikuje progi ryzyka i segmentuje klientów na podstawie wielu kryteriów jednocześnie. Pozwala on na

szybkie dostrzeżenie, że sama Ocena kredytowa jest najsilniejszym predyktorem akceptacji – klienci z bardzo dobrymi ocenami dostają kredyt bez względu na wysokość dochodu i wnioskowanej kwoty, co ma znaczenie dla walidacji strategii kredytowej i dalszego kalibrowania modelu prognostycznego.

3.4.3. Analiza wykresu długości zatrudnienia od decyzji o zatwierdzeniu kredytu

Kolejnym elementem analizy graficznej jest wykres pudełkowy przedstawiający rozkład długości zatrudnienia w zależności od decyzji kredytowej.



Rysunek 3.3. Wykres długości zatrudnienia od decyzji o zatwierdzeniu kredytu

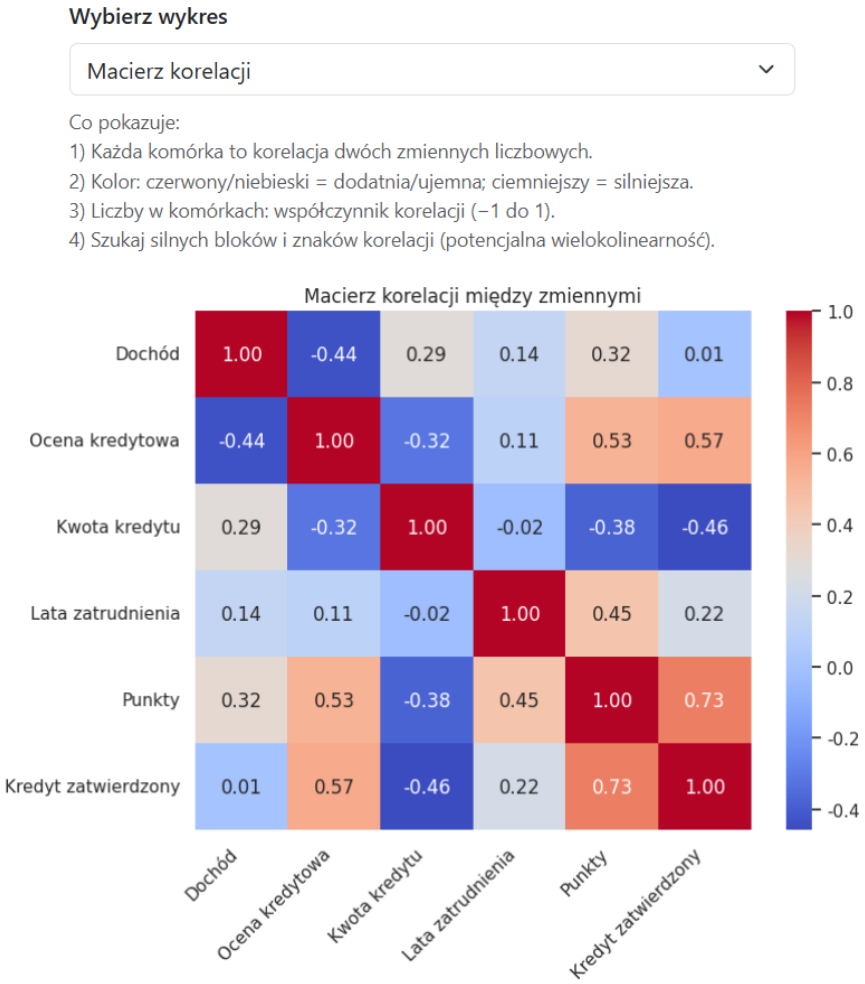
Prezentowany wykres jest wykresem pudełkowym (box plot). Porównuje on rozkład Lata zatrudnienia (Oś Y) dla klientów, których kredyt został odrzucony (zielony) i zatwierdzony (pomarańczowy). Analiza wizualna wskazuje na kluczową różnicę: klienci z zatwierdzonym kredytem mają wyraźnie wyższą medianę (linia wewnątrz pudełka) lat zatrudnienia, która znajduje się w okolicach 30 lat, w porównaniu do mediany dla klientów

odrzuconych, która wynosi około 24 lata. Wniosek ten jest wsparty przez fakt, że całe "pudełko" (zawierające 50% środkowych danych) dla zatwierdzonych jest przesunięte w górę. Mimo to, grupa odrzuconych cechuje się znacznie większym rozrzutem (dłuższe pudełko i wąsy), co oznacza większą niejednorodność i obecność klientów zarówno z bardzo krótkim, jak i bardzo długim stażem pracy (do 40 lat).

Taki typ wykresu został zastosowany, gdyż kwantyfikuje znaczenie stabilności zawodowej w procesie decyzyjnym, pokazując, że dłuższy i bardziej stabilny staż pracy (reprezentowany przez wyższą medianę) jest czynnikiem sprzyjającym akceptacji kredytu. Pomaga on też zidentyfikować, że sama długość zatrudnienia nie jest jedynym kryterium, gdyż w grupie odrzuconych są obserwacje skrajne (kropki) z bardzo długim stażem (np. 35+ lat), co wskazuje, że inna negatywna zmienna (np. niska Ocena kredytowa) musiała przeważyć

3.4.4. Analiza macierzy korelacji między zmiennymi

Kolejnym narzędziem wizualnym w analizie jest macierz korelacji, która pozwala ocenić siłę i kierunek zależności liniowych między wszystkimi zmiennymi numerycznymi. Analiza kolorystyczna i liczbowa korelacji ułatwia zauważenie zarówno pozytywnych, jak i negatywnych zależności, co jest istotne przy interpretacji wyników i dalszym budowaniu modelu prognostycznego.



Rysunek 3.4. Macierz korelacji między zmiennymi

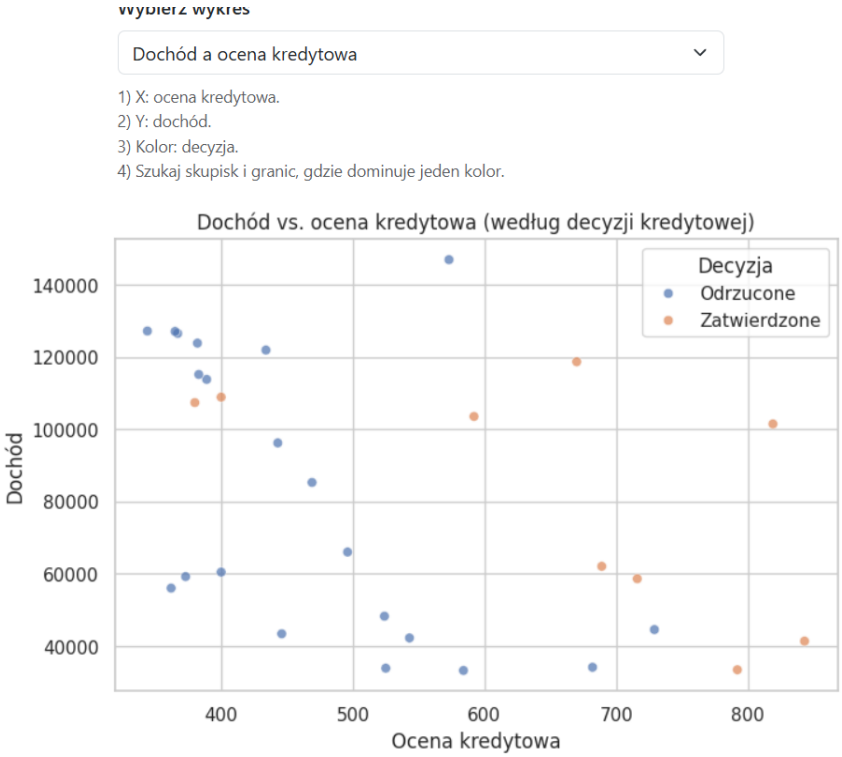
Ten wykres to Macierz korelacji, która w formie kolorystycznej i liczbowej prezentuje siłę i kierunek liniowych zależności pomiędzy wszystkimi analizowanymi zmiennymi numerycznymi. Kolor czerwony oznacza dodatnią korelację (wzrost jednej zmiennej towarzyszy wzrostowi drugiej), a niebieski oznacza ujemną korelację (wzrost jednej zmiennej towarzyszy spadkowi drugiej). Najsilniejsze korelacje dodatnie (ciemny czerwony) występują między Kredyt zatwierdzony a Punkty (0,73) oraz Kredyt zatwierdzony a Ocena kredytowa (0,57). Z kolei Ocena kredytowa wykazuje wyraźną ujemną korelację z Dochodem (-0,44) i Kwotą kredytu (-0,32), co jest interesujące i sugeruje, że nie zawsze wyższy dochód lub wyższa kwota kredytu idą w parze z lepszą oceną kredytową.

Takie wykresy są niezwykle użyteczne, ponieważ można za ich pomocą zidentyfikować ważne predyktory decyzyjne i potencjalną wieloliniowość. Jasno można na nim odczytać, że Punkty oraz Ocena kredytowa są najsilniejszymi czynnikami

wpływającymi na ostateczną decyzję kredytową. Ujemna korelacja między Oceną kredytową a Dochód/Kwota kredytu stanowi ważną informację dla modelu prognostycznego, wskazując, że te zmienne są w pewnym stopniu niezależne, co pozwala na budowanie bardziej odpornych i trafnych modeli oceny ryzyka.

3.4.5. Analiza wykresu zależności dochodu od oceny kredytowej

Kolejnym wykresem w analizie jest diagram rozrzutu przedstawiający zależność między dochodem a oceną kredytową klientów, z podziałem na decyzję kredytową (zatwierdzony vs. odrzucony). Tego typu wizualizacja pozwala jednocześnie ocenić wpływ dwóch kluczowych zmiennych na wynik decyzji oraz dostrzec interakcje między nimi.



Rysunek 3.5. Wykres zależności dochodu od oceny kredytowej

Ten wykres rozrzutu bada bezpośrednią relację pomiędzy Oceną kredytową (Oś X) a Dochodem (Oś Y), jednocześnie oznaczając decyzję kredytową (kolor: niebieski dla odrzuconych, pomarańczowy dla zatwierdzonych). Wykres uwydatnia dwie kluczowe grupy i brak silnej korelacji dodatniej między dochodem a oceną kredytową:

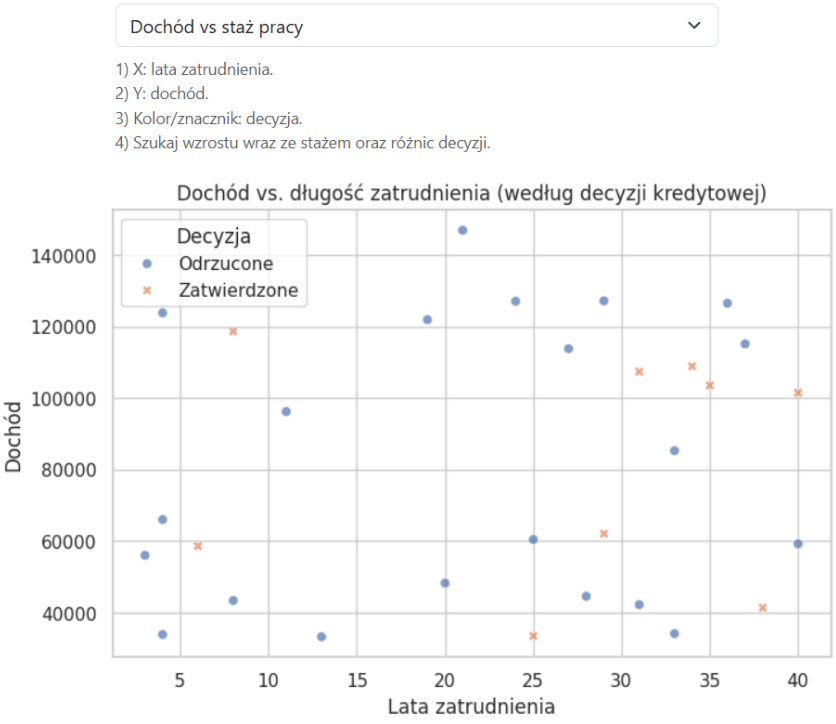
- Wnioski odrzucone (niebieskie) koncentrują się w zakresie niskich i średnich Ocen kredytowych (poniżej 600), ale rozciągają się przez cały zakres dochodów, włączając w to klientów o dochodach powyżej 120 000 zł. Potwierdza to, że niska Ocena kredytowa jest główną przeszkodą, nawet dla klientów bogatych.

- Wnioski zatwierdzone (pomarańczowe) koncentrują się niemal wyłącznie w obszarze wysokich Ocen kredytowych (powyżej 650) i wykazują duży rozrzut dochodów. Widoczne jest, że klienci o wysokiej Ocenie kredytowej są akceptowani bez względu na to, czy zarabiają 40 000 zł, czy 120 000 zł.

Taki wykres jest użyteczny, gdyż pokazuje on, że Ocena kredytowa jest czynnikiem dominującym, silniejszym niż dochód, w przeciwieństwie do intuicyjnych oczekiwań, że wyższy dochód zawsze gwarantuje akceptację. Pozwala to na ukierunkowanie dalszej analizy statystycznej i prognostycznej na zmienne jakościowe i scoringowe, a nie tylko na sam dochód klienta.

3.4.6. Analiza wykresu zależności dochodu od długości zatrudnienia

Kolejnym elementem analizy jest wykres rozrzutu przedstawiający zależność między dochodem a długością zatrudnienia klientów, z podziałem na decyzję kredytową (zatwierdzony vs. odrzucony). Tego typu wizualizacja umożliwia jednoczesną ocenę wpływu stabilności zawodowej i sytuacji finansowej klienta na decyzję kredytową.



Rysunek 3.6. Wykres zależności dochodu od długości zatrudnienia

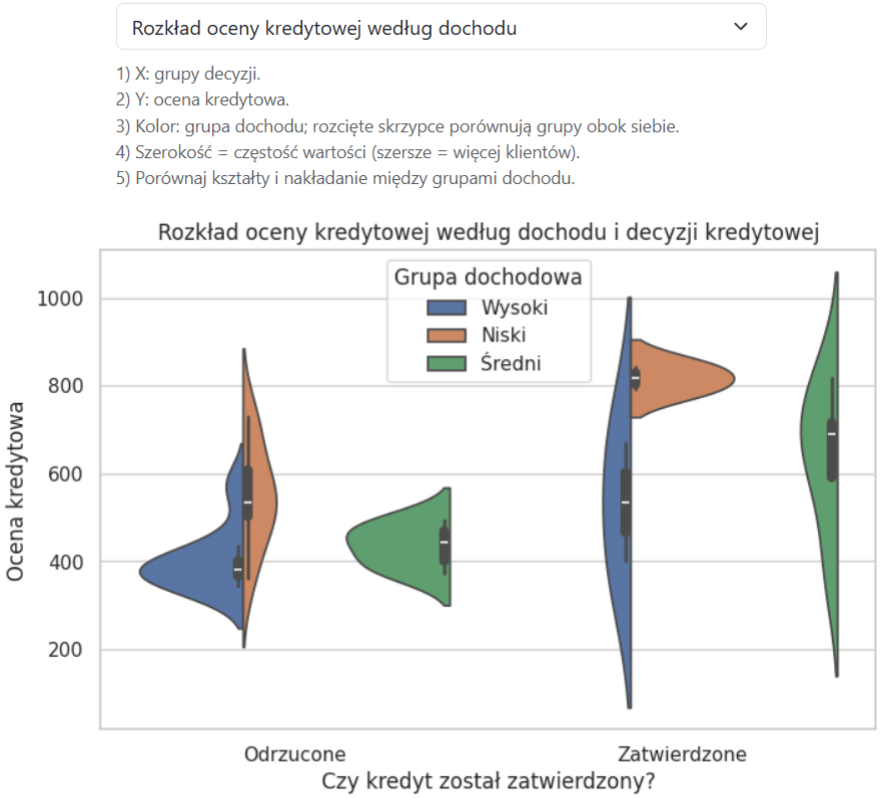
Wykres ten ilustruje relację pomiędzy latami zatrudnienia (Oś X) a dochodem (Oś Y), z podziałem na decyzję kredytową (pomarańczowy dla zatwierdzonych, niebieski dla odrzuconych). Wizualna analiza ukazuje brak wyraźnej, rosnącej zależności między

stażem pracy a dochodem, co jest kluczowe: punkty nie układają się w linii rosnącej. Wnioski zatwierdzone (pomarańczowe) są mocno rozrzucone na osi Y (Dochód) i występują przy różnych latach zatrudnienia (np. 5, 29, 36 lat), ale ich mediana lat zatrudnienia jest wyższa (co widzieliśmy na poprzednim wykresie pudełkowym). Natomiast odrzucone (niebieskie) wnioski dominują w całym spektrum, zwłaszcza w obszarze bardzo wysokich dochodów (powyżej 120 000 zł) z niskim i średnim stażem pracy (5-30 lat), oraz w obszarze niskich dochodów (poniżej 60 000 zł) przy różnym stażu. Sugeruje to, że ani sam dochód, ani sam staż nie są wystarczające do akceptacji bez dobrej Ocena kredytowej (co potwierdziła Macierz Korelacji).

Wykres ten jest użyteczny z punktu widzenia instytucji udzielającej kredyt, z uwagi na to, że negatywnie waliduje prostą hipotezę, że długi staż pracy lub wysoki dochód bezpośrednio prowadzą do akceptacji. Ujawnia, że decyzja jest rezultatem złożonej kombinacji tych czynników z innymi zmiennymi (głównie Oceną kredytową). Pokazuje on wyraźnie, że wysoki dochód nie jest gwarancją sukcesu, jeśli towarzyszą mu inne czynniki ryzyka (np. niska Ocena kredytowa), a co za tym idzie, konieczność polegania na wieloczynnikowym modelu prognostycznym.

3.4.7. Analiza wykresu rozkładu oceny kredytowej według dochodu i decyzji kredytowej

Kolejnym wykresem w analizie jest wizualizacja przedstawiająca rozkład oceny kredytowej w podziale na poziom dochodu i decyzję kredytową. Tego typu wykres pozwala zobaczyć, jak poziom dochodu wpływa na rozkład ocen kredytowych w grupach klientów zatwierdzonych i odrzuconych.



Rysunek 3.7. Wykres rozkładu oceny kredytowej według dochodu i decyzji kredytowej

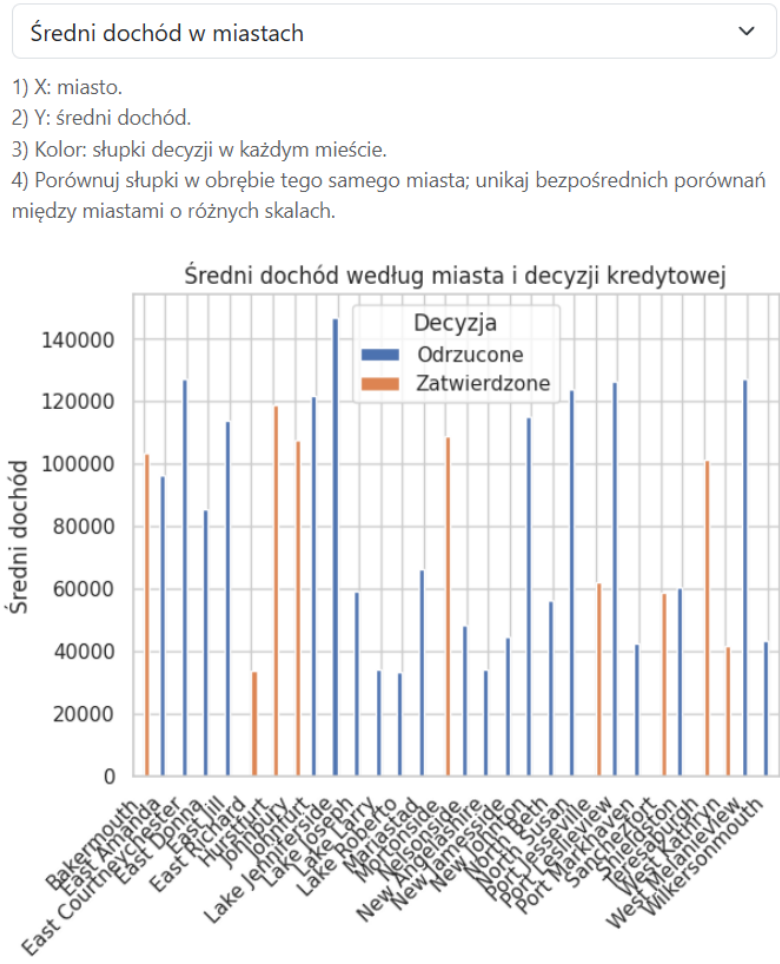
Aby zobrazować tę zależność został zastosowany wykres skrzypcowy (violin plot) ukazuje rozkład Oceny kredytowej (Oś Y) w zależności od Decyzji kredytowej (Oś X: odrzucone/zatwierdzone) oraz dzieli te rozkłady na grupy dochodowe (kolory: Wysoki, Niski, Średni). Wnioski są następujące: po pierwsze, dla wniosków odrzuconych rozkłady grup dochodowych są silnie przesunięte w kierunku niższych Ocen kredytowych (poniżej 600), z wyjątkiem Grupy Niskich Dochodu, której rozkład jest bardziej skoncentrowany w dolnym zakresie. Po drugie, dla wniosków zatwierdzonych rozkłady wszystkich grup dochodowych są skoncentrowane w wyższych ocenach kredytowych (powyżej 650). Najważniejsza obserwacja dotyczy grupy niskiego dochodu (pomarańczowy) w obszarze zatwierdzone - ich ocena kredytowa ma bardzo wysoką medianę (linia środkowa), co potwierdza, że wysoka ocena kredytowa jest głównym i najskuteczniejszym czynnikiem akceptującym, silniejszym niż dochód, umożliwiając akceptację nawet klientom najmniej zamożnym.

Taki wykres jest niezwykle użyteczny i niezbędny, ponieważ kwantyfikuje znaczenie scoringu ponad kryteriami finansowymi. Pozwala on na natychmiastowe zrozumienie hierarchii ważności zmiennych w modelu decyzyjnym: jeśli klient ma niskie dochody, musi mieć bardzo dobrą ocenę kredytową (co widać po skoncentrowaniu się "skrzypiec"

z niskim dochodem w wysokich wartościach na osi Y w sekcji "Zatwierdzone"). To spostrzeżenie jest kluczowe dla kalibracji modelu prognostycznego i ustalenia, które zmienne mają największą wagę decyzyjną.

3.4.8. Analiza wykresu średniego dochodu w poszczególnych miastach

Kolejnym wykresem w analizie jest wizualizacja średniego dochodu klientów w poszczególnych miastach w podziale na decyzję kredytową. Tego typu wykres słupkowy pozwala ocenić, jak sytuacja finansowa klientów w różnych lokalizacjach wpływa na przyznanie lub odrzucanie kredytów.



Rysunek 3.8. Wykres średniego dochodu w poszczególnych miastach

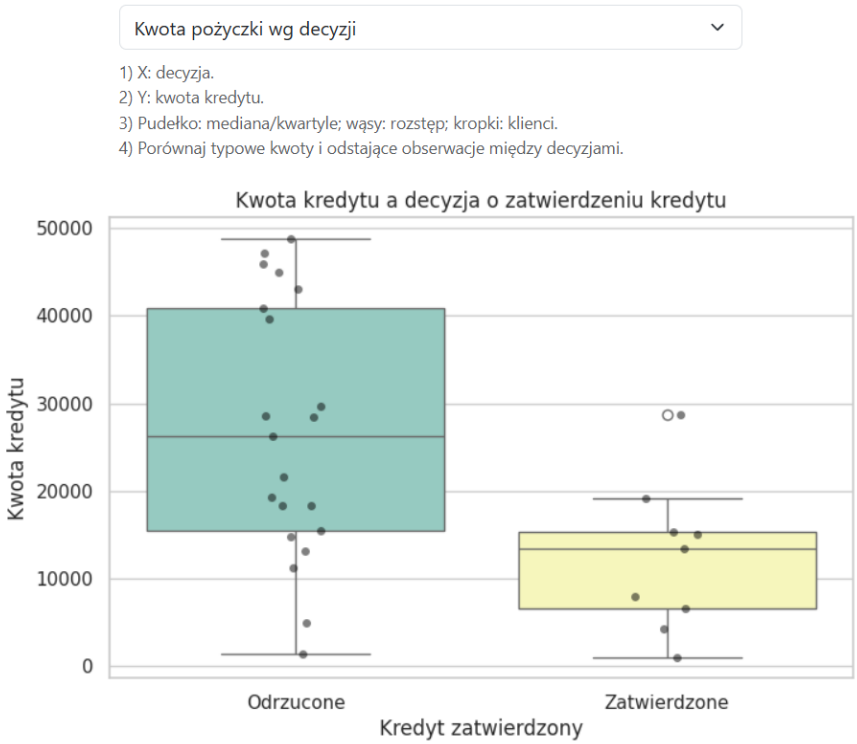
Ukazany wykres słupkowy ilustruje Średni dochód (Oś Y) w różnych miastach (Oś X), dzieląc te dane na wnioski Odrzucone (niebieskie) i Zatwierdzone (pomarańczowe) w obrębie każdej lokalizacji. Wnioski z tego wykresu są lokalnie zróżnicowane, ale podsumowują się następująco: w większości miast, w których występują obie decyzje (np. East Courtneychester, Lake Joseph), średni dochód klientów

zatwierdzonych (pomarańczowy) jest niższy niż średni dochód klientów odrzuconych (niebieski). To jest kluczowa anomalia, która jest zgodna z poprzednimi obserwacjami: sam dochód nie jest wystarczającym warunkiem akceptacji. W miastach takich jak Lake Jennifer, Port Jesseville czy West Kathryn, zatwierdzane są wnioski klientów zarabiających mniej niż ci, którzy zostali odrzuceni. Dodatkowo, w niektórych miastach, np. Lake Joseph, średni dochód klientów odrzuconych jest ekstremalnie wysoki, co znowu sugeruje, że ich Ocena kredytowa musiała być na tyle niska, że wykluczyła ich z akceptacji.

Taki wykres jest niezbędny, ponieważ kwestionuje założenie o dochodzie jako głównym predyktorze decydującym o przyznaniu kredytu i identyfikuje anomalie geograficzne. Waliduje on hipotezę, że niskie wyniki scoringowe (np. ocena kredytowa) są znacznie silniejszym czynnikiem ryzyka niż wysoki dochód, a decyzje są podejmowane niezależnie od średniej zarobkowej w danym regionie, co jest kluczowe dla ustalenia priorytetów w algorytmach oceny ryzyka.

3.4.9. Analiza wykresu kwoty kredytu w zależności od decyzji o zatwierdzeniu kredytu

Kolejnym wykresem w analizie jest wizualizacja przedstawiająca zależność między kwotą wnioskowanej pożyczki a decyzją kredytową. Tego typu wykres umożliwia ocenę, jak wysokość wnioskowanej kwoty wpływa na prawdopodobieństwo zatwierdzenia lub odrzucenia wniosku.



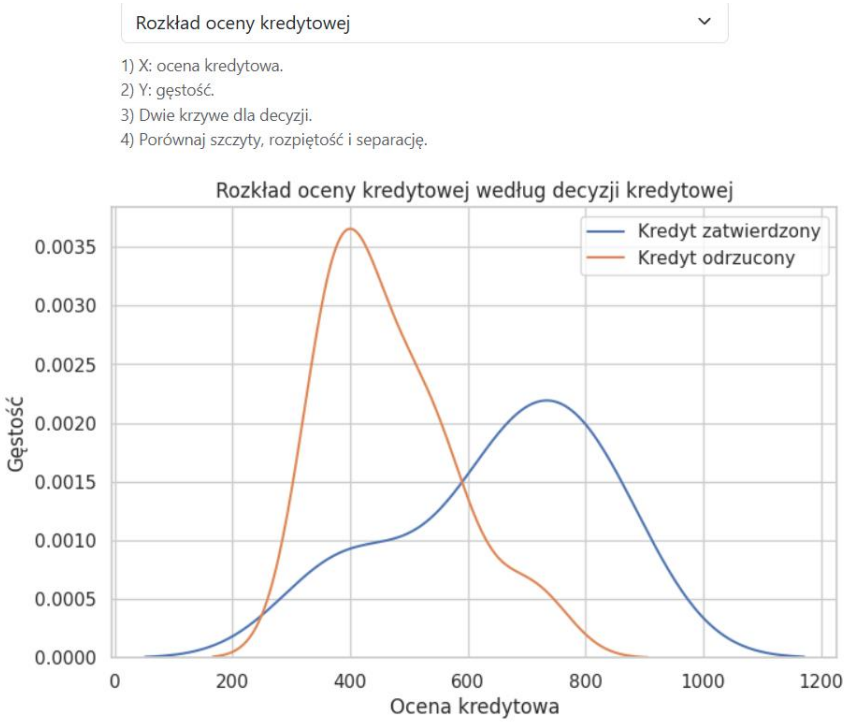
Rysunek 3.9. Wykres kwoty kredytu w zależności od decyzji o zatwierdzeniu kredytu

Prezentowany wykres pudełkowy (box plot) porównuje rozkład Kwoty kredytu (Oś Y) dla klientów, których kredyt został odrzucony (zielony/turkusowy) i zatwierdzony (żółty). Analiza wizualna ujawnia kluczową, kontrintuicyjną zależność: klienci, których wnioski odrzucono, typowo ubiegali się o znacznie wyższe kwoty kredytów; całe "pudełko" (zakres 50% środkowych danych) dla odrzuconych jest znacznie wyższe (mediana w okolicach 27 000 zł) niż dla klientów zatwierdzonych (mediana poniżej 15 000 zł). Rozrzut kwot (rozmiar pudełka) jest również znacznie większy w grupie odrzuconej, co oznacza, że bank jest bardziej skłonny akceptować wnioski o niższe i bardziej jednorodne kwoty, nawet jeśli z wcześniejszej analizy wiemy, że są to klienci z dobrą Oceną kredytową. Potwierdza to strategiczne ostrożności banku.

Taki wykres jest niezwykle użyteczny, ponieważ kwantyfikuje awersję do ryzyka w kontekście kwoty. Pokazuje, że pomimo silnej korelacji między dobrą Oceną kredytową a akceptacją, bank utrzymuje ostre ograniczenia w ekspozycji na ryzyko, preferując akceptację mniejszych pożyczek. Stanowi to podstawę do wyciągnięcia wniosków na temat maksymalnych kwot udzielanych klientom o danym profilu ryzyka, co jest krytyczne dla zarządzania płynnością i ryzykiem portfela.

3.4.10. Analiza rozkładu oceny kredytowej według decyzji kredytowej

Kolejnym wykresem w analizie jest diagram rozrzutu przedstawiający zależność między dochodem a oceną kredytową klientów, z podziałem na decyzję kredytową (zatwierdzony vs. odrzucony). Tego typu wizualizacja pozwala jednocześnie ocenić wpływ dwóch kluczowych zmiennych na wynik decyzji oraz dostrzec interakcje między nimi.



Rysunek 3.10. Rozkład oceny kredytowej według decyzji kredytowej

Wykres gęstości (density plot) przedstawia rozkład Oceny kredytowej (Oś X) dla klientów, których kredyt zatwierdzono (niebieska krzywa) i odrzucono (pomarańczowa krzywa). Jest to najczystszy i najbardziej kluczowy wykres podsumowujący politykę ryzyka:

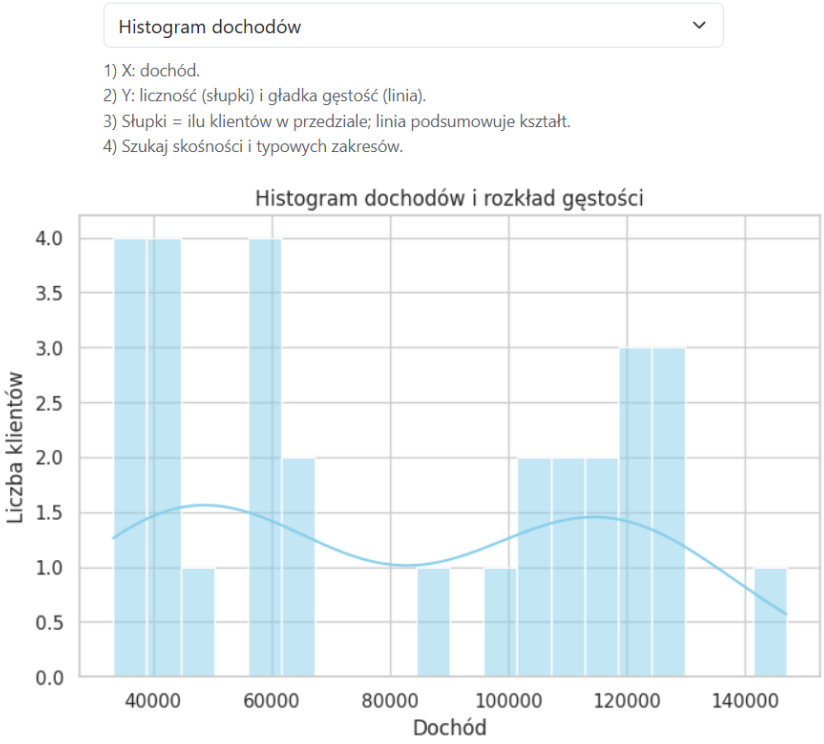
- Kredyt Odrzucony (pomarańczowy): Krzywa jest wąska, wysoka i silnie skupiona wokół niskich wartości Ocena kredytowa, ze szczytem w okolicach 400-500. Oznacza to, że większość klientów odrzuconych ma niską lub przeciętną ocenę kredytową, a ich odrzucenie jest zdominowane przez ten jeden czynnik.
- Kredyt Zatwierdzony (niebieski): Krzywa jest szersza, niższa i przesunięta w prawo, ze szczytem w okolicach 750-850. Oznacza to, że klienci zatwierdzeni mają typowo wysokie oceny kredytowe, a ich rozkład jest bardziej rozciągnięty, co sugeruje, że akceptowane są również osoby z nieco niższymi wynikami, o ile inne czynniki są pozytywne.

- Separacja: Rozkłady te są wyraźnie rozdzielone z minimalnym nakładaniem się, co potwierdza, że Ocena kredytowa jest najsilniejszym predyktorem decyzyjnym i działa jak wyraźny próg akceptacji/odrzucenia (potwierdzając wcześniejsze wnioski z analizy rozrzutu).

Wykres wizualnie potwierdza siłę dyskryminacyjną zmiennej Oceny kredytowej. Służy jako główny dowód na to, że niezależnie od Dochodu, Kwoty Kredytu czy Stażu Pracy, Ocena kredytowa jest dominującym czynnikiem ryzyka. Jest to kluczowy element do omówienia w rozdziale analizy wyników, gdyż potwierdza priorytet tej zmiennej w modelu prognostycznym.

3.4.11. Analiza histogramu dochodów i rozkład gęstości

Kolejnym wykresem w analizie jest histogram dochodów połączony z rozkładem gęstości, który ilustruje częstotliwość występowania poszczególnych przedziałów dochodów oraz kształt całego rozkładu w badanej populacji klientów.



Rysunek 3.11. Histogram dochodów i rozkład gęstości

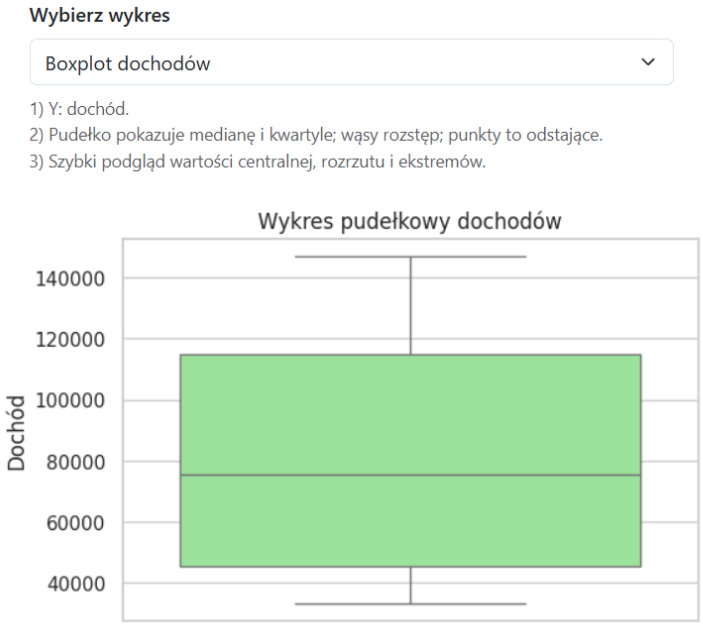
Histogram przedstawiony powyżej ilustruje liczbę klientów (Oś Y) w poszczególnych przedziałach Dochodu (Oś X) z gładką krzywą gęstości (niebieska linia), która sumarycznie oddaje kształt rozkładu. Wizualna analiza potwierdza, że rozkład

dochodów jest złożony i wielomodalny (wieloszczytowy). Największe koncentracje klientów występują w dwóch głównych obszarach: niskich i średnich dochodów (okolice 40 000 zł i 60 000 zł) oraz wysokich dochodów (okolice 120 000 zł). Pomiędzy tymi szczytami (szczególnie w zakresie 80 000 zł - 100 000 zł) widoczna jest znacząca "dolina" (mniej klientów). Taka bimodalność lub złożoność rozkładu sugeruje, że portfel kredytowy obsługuje co najmniej dwie wyraźne grupy klientów: liczną grupę klientów mniej zamożnych i drugą, również liczną, grupę klientów zamożnych.

Taki wykres jest ważny, ponieważ identyfikuje segmenty populacji kredytobiorców i wyjaśnia parametry statystyczne (np. kurtozę i skośność). Złożony kształt rozkładu (wielomodalność) ma znaczenie dla modelowania ryzyka, ponieważ proste modele zakładające rozkład normalny mogą nie oddawać wiernie charakterystyki populacji.

3.4.12. Analiza wykresu pudełkowego dochodów

Kolejnym wykresem w analizie jest wykres pudełkowy (box plot) przedstawiający skondensowaną dystrybucję zmiennej Dochód.



Rysunek 3.12. Wykres pudełkowy dochodów

Na podstawie analizy tego wykresu można zauważyć, że:

- Mediana (Q2) - linia wewnątrz pudełka znajduje się w okolicach 75 675 zł, co oznacza, że połowa klientów zarabia mniej, a połowa więcej niż ta kwota. Mediana jest

nieznacznie niższa od średniej (81 384,2 zł), co sugeruje lekką asymetrię rozkładu dochodów.

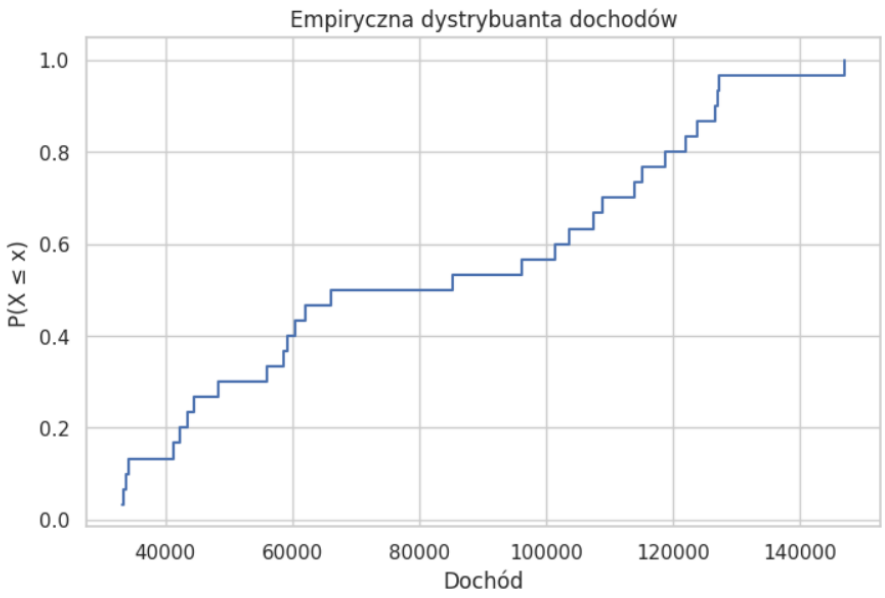
- Pudełko (Rozrzut danych (IQR)) rozciąga się od dolnego kwartyla Q1 (~45 516 zł) do górnego kwartyla Q3 (~114 838 zł). Szeroki zakres potwierdza znaczną zmienność dochodów w populacji, zgodnie z wcześniejszą obserwacją wysokiego odchylenia standardowego (36 635,43 zł). Wskazuje to, że populacja obejmuje zarówno klientów niskodochodowych, jak i wysokodochodowych.
- Pudełko jest stosunkowo symetryczne względem mediany, jednak górny „wąs" wydaje się nieco dłuższy, co wskazuje na lekką dodatnią skośność rozkładu (0,1031). Oznacza to obecność niewielkiej liczby bardzo wysokich dochodów, które „ciągną" średnią w górę.
- Chociaż na wykresie nie widać wielu punktów oznaczających wartości odstające, rozciągnięcie wąsów informuje o istnieniu ekstremów dochodowych. Jednocześnie ujemna kurtoza (-1,5676) wskazuje, że liczba tych ekstremów nie jest masowa, a większość klientów mieści się w typowym zakresie rozkładu.

Wykres ten jest użyteczny, ponieważ potwierdza ogromną wewnętrzną zmienność bazy klientów w zakresie dochodów. Umożliwia natychmiastowe zidentyfikowanie, że główna masa (50%) klientów ma dochody w bardzo szerokim przedziale, co czyni dochód mniej efektywnym wskaźnikiem decyzyjnym niż, na przykład, Ocena kredytowa. Jest to kluczowe do zrozumienia, dlaczego model decyzyjny musi polegać na innych, bardziej dyskryminujących zmiennych.

3.4.13. Analiza empirycznej dystrybuanty dochodów

Przedstawiony wykres ilustruje empiryczną dystrybuantę dochodów klientów, czyli funkcję $P(X \leq x)$, która pokazuje odsetek klientów z dochodem nieprzekraczającym danej wartości x. Dzięki temu można szybko ocenić, jaki procent populacji znajduje się poniżej określonego progu dochodowego.

1) X: dochód.
2) Y: odsetek klientów z dochodem ≤ x.
3) Percentyle: wejdź do 0,5 (50%) i odczytaj medianę; 0,9 to 90. percentyl.



Rysunek 3.13. Wykres empirycznej dystrybuanty dochodów

Analizując przedstawiony wykres zaobserwować można następujące zależności:

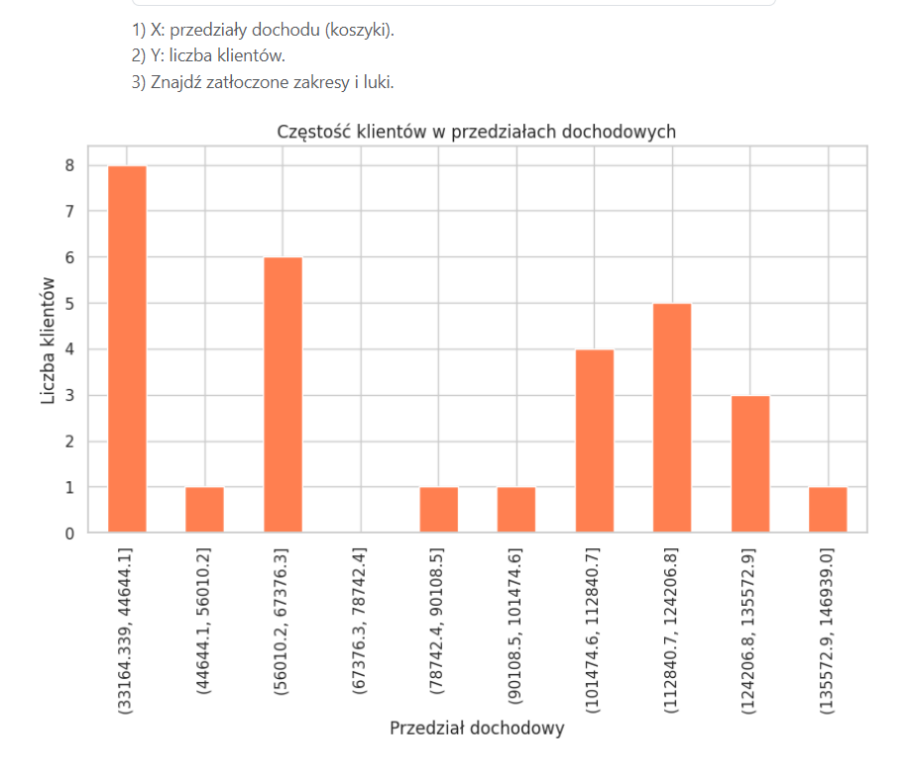
- Mediana dochodów, odpowiadająca wartości $P(X \leq x)=0,5$, znajduje się w okolicach 75 675 zł. Oznacza to, że połowa klientów zarabia mniej, a połowa więcej niż ta kwota, co jest zgodne z obserwacją z wykresu pudełkowego.
- Dystrybuanta pokazuje, że około 25% klientów zarabia poniżej 45 000 zł (dolny kwartyl), natomiast 75% nie przekracza 115 000 zł (górny kwartyl). Szeroki przedział między 25. a 75. percentylem potwierdza dużą zmienność dochodów w populacji, co jest spójne z wysokim odchyleniem standardowym obserwowanym wcześniej.
- Wykres wskazuje lekką dodatnią asymetrię – funkcja rośnie nieco wolniej w środkowym przedziale, a szybciej w górnym zakresie dochodów, co sugeruje obecność niewielkiej liczby bardzo wysokich dochodów. Rozkład jest stosunkowo równomierny w przedziale środkowym, co potwierdza wcześniejsze obserwacje o umiarkowanej koncentracji dochodów w środku rozkładu.

Dystrybuanta ta jest szczególnie istotnym wykresem, gdyż pozwala wizualnie ocenić, jakie odsetki klientów znajdują się w określonych przedziałach dochodowych. W kontekście oceny zdolności kredytowej pokazuje, że dochód sam w sobie nie jest wystarczającym predyktorem akceptacji kredytu – wiele wniosków pochodzi od klientów

z dochodem w środkowym, zróżnicowanym przedziale, co wymaga uwzględnienia innych czynników decyzyjnych, takich jak Ocena kredytowa czy długość zatrudnienia.

3.4.14. Analiza wykresu klientów w przedziałach dochodowych

W tej analizie wyznaczono wykres częstości klientów w przedziałach dochodowych i przedstawiono go w celu oceny liczebności i gęstości klientów w różnych grupach dochodowych, co pozwala na dokładniejsze zrozumienie struktury finansowej portfela klienta.



Rysunek 3.14. Wykres klientów w przedziałach dochodowych

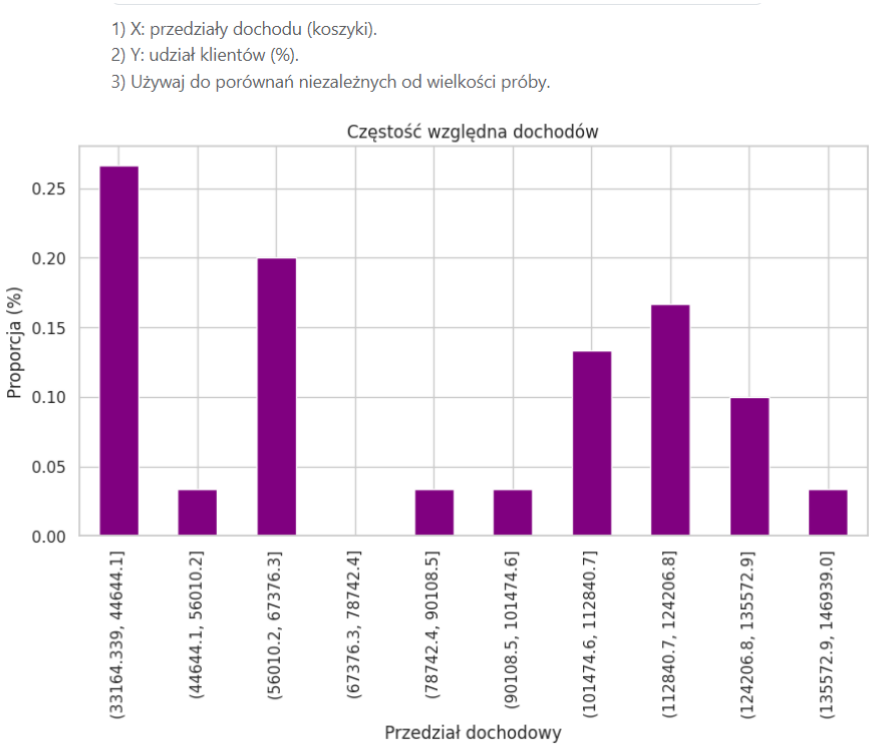
Wykres ten to histogram częstości (wykres słupkowy), który ilustruje liczbę klientów (Oś Y) wpadających do poszczególnych przedziałów dochodowych (koszyków, Oś X). Zauważyć można, że wyraźne ukazanie nierównomiernego, i wieloszczytowego rozkładu dochodów. Najwyższe koncentracje klientów występują w najniższym przedziale (ok. 33 000 zł do 44 000 zł, 8 klientów) oraz w przedziale średnio-niskim (ok. 56 000 zł do 67 000 zł, 6 klientów). Druga, mniejsza koncentracja występuje w obszarze wysokich dochodów (ok. 112 000 zł do 124 000 zł, 5 klientów). Dodatkowo pomiędzy tymi szczytami widoczne są wyraźne "doliny" lub luki w częstościach (np. w przedziałach 67 000 zł – 78 000 zł oraz 90 000 zł – 101 000 zł), co sugeruje, że nasz portfel kredytowy obsługuje wyraźnie dwie odrębne grupy klientów (niskodochodowych

i wysokodochodowych), a mniej jest osób zarabiających kwoty przejściowe. Wobec czego można stwierdzić, że ten histogram potwierdza obserwacje z Wykresu Pudełkowego Dochodów i Rozkładu Gęstości, które również wskazywały na złożoną i niejednorodną strukturę populacji.

Taki wykres okazuje się bardzo użyteczny, ponieważ dokładnie kwantyfikuje segmenty dochodowe i potwierdza strukturalną niejednorodność portfela, co ma kluczowe znaczenie dla zarządzania ryzykiem. Umożliwia bankowi precyzyjne targetowanie i różnicowanie polityki kredytowej dla dominujących koszyków dochodowych, a także uwzględnienie, że decyzja kredytowa musi być silnie uzależniona od Oceny kredytowej, która skuteczniej dzieli klientów na zaakceptowanych i odrzuconych, niezależnie od faktu, w którym koszyku dochodowym się znajdują.

3.4.15. Analiza wykresu częstości względnej dochodów

W tej analizie wyznaczono wykres częstość względna dochodów i przedstawiono go w celu oceny proporcjonalnego udziału klientów w poszczególnych koszykach dochodowych, co jest fundamentalnym krokiem w statystycznym opisie populacji. Jest to kluczowe narzędzie, ponieważ normalizuje liczebność poszczególnych segmentów, umożliwiając porównania niezależne od całkowitej wielkości próby



Rysunek 3.15. Wykres częstości względnej dochodów

Z wykresu odczytać można następujące zależności:

- Największy udział procentowy klientów (ponad 25%) znajduje się w najniższym przedziale dochodowym (ok. 33 164 zł do 44 644 zł). Drugi kluczowy segment (około 20%) to klienci średnio-niskich dochodów (ok. 56 010 zł do 67 376 zł). Fakt, że te dwa segmenty stanowią niemal połowę naszej bazy, ma krytyczne znaczenie dla ustalenia polityki ryzyka i jej wrażliwości na najmniej zamożne grupy.
- Wykres wyraźnie pokazuje znaczące spadki proporcji (nawet do niemal zerowego udziału) w środkowych przedziałach dochodowych (np. w koszykach od ok. 67 000 zł do 78 000 zł oraz 90 000 zł do 101 000 zł). Ta nieciągłość, z lukami o znikomej proporcji, jest mocnym dowodem na to, że rozkład dochodów jest bimodalny lub złożony, co oznacza, że operujemy na dwóch oddzielnych populacjach klientów, a mniejsza jest liczba osób zarabiających kwoty przejściowe.
- Klienci o wyższych dochodach (powyżej 101 000 zł) stanowią drugą, rozproszoną grupę, z największym pojedynczym udziałem w przedziale ok. 112 840 zł do 124 206 zł (ok. 17%).

Wykres ten precyzyjnie kwantyfikuje wagę każdego segmentu dochodowego w stosunku do całej populacji, co czyni go szczególnie pomocnym w podejmowaniu decyzji dotyczących zarządzania ryzykiem. Umożliwia on bankowi precyzyjne targetowanie i różnicowanie polityki kredytowej dla dominujących koszyków dochodowych, potwierdzając jednocześnie, że decyzja kredytowa musi być silnie uzależniona od czynników ryzyka (takich jak Ocena kredytowa), które skuteczniej dzielą klientów na zaakceptowanych i odrzuconych, niezależnie od ich przynależności do danego koszyka dochodowego. Znormalizowana częstość względna pozwala na wiarygodne porównania tego rozkładu z innymi, zewnętrznymi zbiorami danych

3.4.16. Analiza wykresu udziału zatwierdzonych i odrzuconych kredytów

W tej analizie wyznaczono wykres udział zatwierdzonych i odrzuconych kredytów i przedstawiono go w celu szybkiego, proporcjonalnego podziału całej puli wniosków kredytowych według ostatecznej decyzji (akceptacja/odrzucenie).

1) Udziały zatwierdzonych vs odrzuconych.
2) Dokładne wartości w etykietach/procentach.
3) Dobre do szybkiego podziału (nie do precyzyjnych porównań).

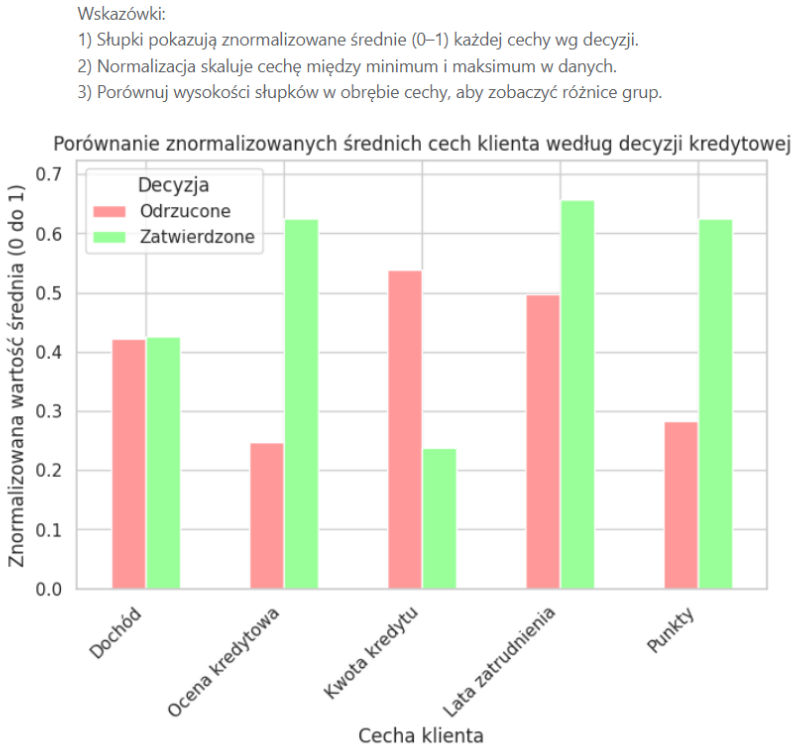
Rysunek 3.16. Wykres udziału zatwierdzonych i odrzuconych kredytów

Ten wykres to diagram kołowy (pie chart) , który przedstawia częstość względną (proporcję) dwóch kategorii decyzyjnych: Odrzucone (czerwony) i Zatwierdzone (zielony). Analiza wykresu wskazuje na wyraźną dominację odrzuceń kredytów - kategoria „Odrzucone kredyty" stanowi 70% wszystkich wniosków, podczas gdy kredyty zatwierdzone to jedynie 30% całkowitej puli. Taki podział potwierdza restrykcyjną politykę kredytową banku i jego awersję do ryzyka – odrzucenie aż 70% wniosków świadczy o bardzo selektywnym procesie akceptacji. Wynik ten jest zgodny z analizą rozkładu Oceny kredytowej, która wykazała, że większość odrzuconych klientów posiadała niskie wyniki (około 400–500), a grupy zatwierdzonych i odrzuconych wniosków były wyraźnie oddzielone. Wysoki wymóg Oceny kredytowej jest głównym czynnikiem ograniczającym liczbę pozytywnie zweryfikowanych wniosków.

Przedstawiony wykres pełni funkcję natychmiastowego podsumowania skuteczności i restrykcyjności procesu decyzyjnego. Stanowi podstawowy wskaźnik, pokazując podział decyzji, który jest punktem wyjścia do dalszej analizy i optymalizacji - w celu zwiększenia akceptacji przy jednoczesnym utrzymaniu akceptowalnego poziomu ryzyka.

3.4.17. Analiza porównawcza znormalizowanych średnich cech klienta według decyzji kredytowej

W tej analizie wyznaczono wykres Porównanie znormalizowanych średnich cech klienta według decyzji kredytowej i przedstawiono go w celu skondensowanego zestawienia różnic między dwiema grupami klientów – Odrzuconymi i Zatwierdzonymi.



Rysunek 3.17. Porównanie znormalizowanych średnich cech klienta według decyzji kredytowej

Analiza grupowanego wykresu słupkowego przedstawiającego znormalizowane średnie pięciu cech klientów (Dochód, Ocena kredytowa, Kwota kredytu, Lata zatrudnienia, Punkty) dla grup Odrzucone (czerwone) i Zatwierdzone (zielone) wskazuje na następujące obserwacje:

- Najsilniejszym dyskryminatorem decyzji kredytowej jest Ocena kredytowa. Średnia znormalizowana dla klientów Zatwierdzonych wynosi ponad 0,6, podczas gdy dla Odrzuconych spada poniżej 0,3. Różnica ta potwierdza, że Ocena kredytowa jest dominującym predyktorem akceptacji kredytu, zgodnie z wcześniejszą analizą rozkładu oceny kredytowej według decyzji.
- Interesującym wnioskiem jest to, że klienci Odrzuceni mają wyższą znormalizowaną średnią kwotę kredytu (około 0,55) niż klienci Zatwierdzeni (około 0,25). Wskazuje to, że bank jest bardziej skłonny odrzucać wnioski o wyższe kwoty, nawet przy niskim

ryzyku w innych obszarach, co odzwierciedla silną awersję do dużych ekspozycji finansowych.

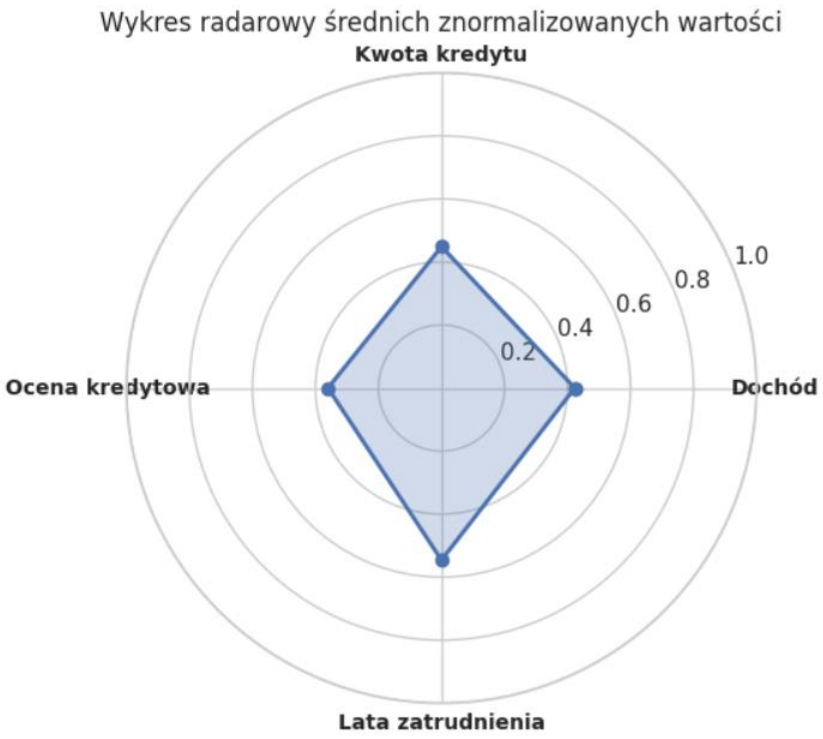
- Obie cechy mają wyższą średnią w grupie Zatwierdzonej (Lata zatrudnienia ≈ 0,5, Punkty ≈ 0,3), co sugeruje, że długi staż pracy i wyższa liczba punktów są pozytywnymi, choć wtórnymi wskaźnikami akceptacji kredytu, wspierającymi decyzję opartą głównie na Oceny kredytowej.
- Dochód wykazuje niemal identyczne średnie znormalizowane w obu grupach (około 0,42), co wskazuje, że nie jest istotnym dyskryminatorem decyzji kredytowej. Jest to zgodne z wcześniejszą analizą wykresu Dochód vs. Ocena kredytowa, gdzie punkty odrzucone były rozrzucone w całym spektrum dochodowym.

Wykres stanowi użyteczne narzędzie w procesie eksploracji danych, identyfikując hierarchię siły predykcyjnej zmiennych. Ocena kredytowa jest w jego przypadku kluczowa, następnie istotna jest Kwota kredytu jako czynnik ryzyka, a Lata zatrudnienia i Punkty pełnią rolę wspierającą. Dochód natomiast jest najmniej przydatny w modelowaniu decyzji kredytowej. Wnioski te stanowią istotny punkt wyjścia do dalszej analizy i modelowania.

3.4.18. Analiza wykresu radarowego średnich znormalizowanych wartości

W tej analizie wyznaczono Wykres radarowy średnich znormalizowanych wartości i przedstawiono go w celu wizualnej oceny i porównania względnej wielkości średnich wartości kluczowych cech klienta (Kwota kredytu, Dochód, Lata zatrudnienia, Ocena kredytowa) w całym portfelu (dla wszystkich klientów, niezależnie od decyzji).

1) Ramiona: wybrane cechy.
2) Promień: średnia znormalizowana wartość (0–1).
3) Dalej od środka = większa średnia.
4) Służy do porównania względnych wielkości cech.



Rysunek 3.18. Wykres radarowy średnich znormalizowanych wartości

Analiza wykresu radarowego, przedstawiającego średnie znormalizowane wartości czterech kluczowych zmiennych numerycznych (Dochód, Kwota kredytu, Ocena kredytowa, Lata zatrudnienia), wskazuje na następujące obserwacje:

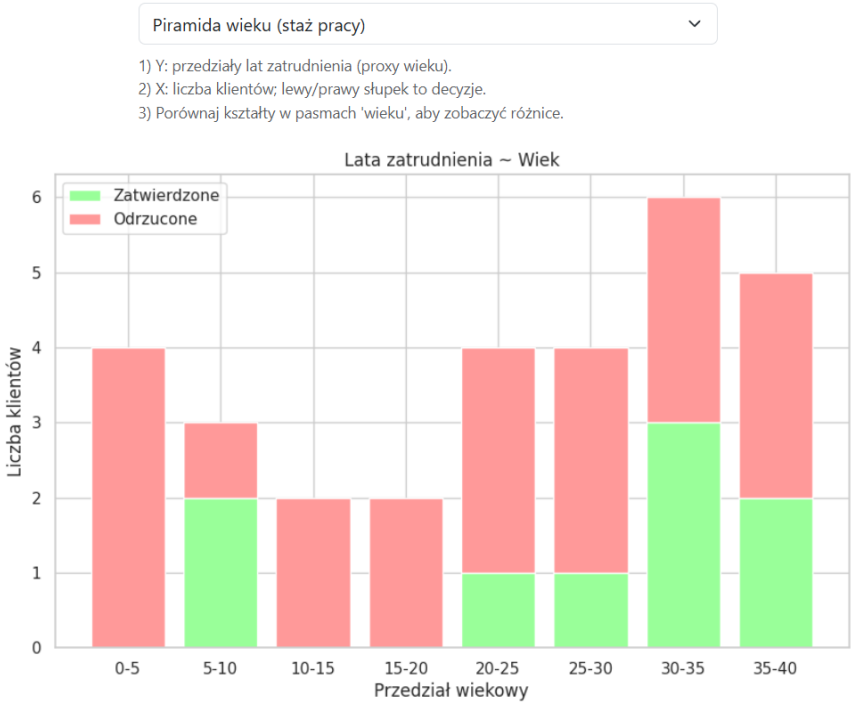
- Średnie znormalizowane wartości Dochodu i Kwoty kredytu wynoszą około 0,45, co oznacza, że typowe wartości tych cech są bliskie połowie ich maksymalnego zakresu. Wskazuje to na stosunkowo wysokie oczekiwania finansowe klientów względem ich możliwości oraz wnioskowanych kwot kredytowych.
- Średnia znormalizowana Ocena kredytowa jest najniższa (około 0,3), co sugeruje, że większość klientów ma wyniki bliższe dolnemu zakresowi możliwych ocen. Jest to spójne z wcześniejszymi statystykami (średnia 523,8) i potwierdza, że niski poziom Oceny kredytowej jest głównym czynnikiem ograniczającym akceptację wniosków.
- Średnia znormalizowana długość stażu pracy wynosi około 0,4, czyli znajduje się między poziomem Dochodu/Kwoty kredytu a Oceny kredytowej. Wskazuje to na neutralny wpływ tej cechy na ryzyko przyznania kredytu – typowy klient ma średni staż pracy.

- Czworokąt wykresu jest wydłużony w kierunku Dochodu i Kwoty kredytu, a ściśnięty w kierunku Oceny kredytowej. Wizualnie potwierdza to asymetrię ryzyka w przyznawaniu kredytu: klienci mają stosunkowo wysokie oczekiwania finansowe, ale stosunkowo niską wiarygodność kredytową.

Wykres radarowy dostarcza skoncentrowanego obrazu średniego profilu klientów i natychmiast widoczna jest kluczowa słabość – niska Ocena kredytowa. Jest to istotny wniosek popierający hipotezę, że Ocena kredytowa jest głównym czynnikiem odrzucania wniosków. Skoncentrowanie się na jej podniesieniu lub zastosowanie bardziej restrykcyjnego filtrowania w tym zakresie może znacząco poprawić jakość usług prowadzonych przez instytucję udzielającą kredytów swoim klientom.

3.4.19. Piramida stażu pracy klientów

W tej analizie wyznaczono wykres Piramidy wieku (dla stażu pracy) i przedstawiono go w celu oceny liczebności klientów w różnych przedziałach stażu pracy, jednocześnie różnicując te grupy ze względu na decyzję kredytową (Zatwierdzone vs. Odrzucone). Jest to kluczowy element do zrozumienia, w jakim wieku zawodowym bank najczęściej akceptuje, a w jakim odrzuca wnioski.



Rysunek 3.19. Wykres piramidy stażu pracy klientów

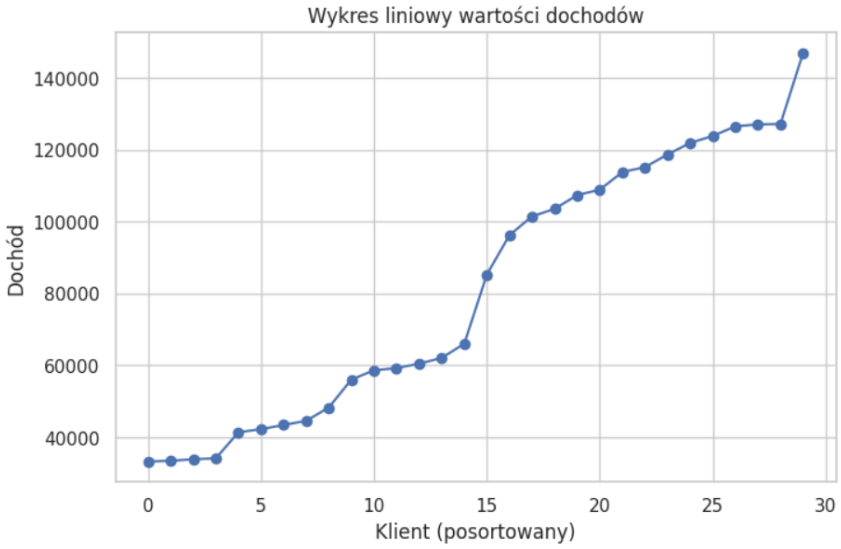
Analiza skumulowanego wykresu słupkowego - Piramida Wieku (w kategorii lata zatrudnienia), przedstawiającego liczbę klientów w poszczególnych przedziałach stażu pracy z podziałem na Zatwierdzonych (zielone) i Odrzuconych (czerwone), wskazuje na to, że w większości przedziałów stażu pracy liczba odrzuconych klientów przewyższa liczbę zatwierdzonych, co jest zgodne z ogólnym wskaźnikiem akceptacji na poprzednim wykresie wynoszącym 30%. Dodatkowo w przedziale 0–5 lat liczba odrzuconych wynosi 4, przy braku zatwierdzonych wniosków. W przedziałach 5–10 i 10–15 lat odrzuconych jest odpowiednio 1 i 2, przy 2 i 0 zatwierdzonych. Wniosek: Bank wyraźnie preferuje klientów z większą stabilnością zawodową i unika akceptacji osób z krótkim stażem pracy. Również przedziały 20–25 i 25–30 lat mają po 4 klientów, przy czym odrzuconych jest znacznie więcej. Najwyższa równowaga występuje w przedziale 30–35 lat (3 zatwierdzonych vs. 3 odrzuconych). W przedziale 35–40 lat odrzuconych jest 3, a zatwierdzonych 2.

Wykres jest użyteczny do podziału ryzyka według długości stażu pracy. Pokazuje, że najbezpieczniejszą grupą pod względem proporcji akceptacji są osoby z ponad 30-letnim stażem. Choć dłuższy staż pracy generalnie zwiększa szanse na akceptację, nie gwarantuje pozytywnej decyzji kredytowej – nawet wśród najbardziej stabilnych zawodowo klientów część wniosków została odrzucona. Utrzymująca się ogólna restrykcyjność polityki kredytowej (70% odrzuceń) podkreśla, że decydującym kryterium pozostaje Ocena kredytowa, która jest dominującym czynnikiem w procesie decyzyjnym.

3.4.20. Analiza wykresu liniowego wartości dochodów

W tej analizie wyznaczono wykres Wykres liniowy wartości dochodów i przedstawiono go w celu wizualnej oceny ciągłości i rozkładu dochodów po posortowaniu klientów od najmniej do najbardziej zamożnych. Służy do identyfikacji luk i ekstremalnych wartości w danych.

1) X: klienci posortowani wg dochodu.
2) Y: dochód.
3) Skoki oznaczają luki; płaskie odcinki = podobne dochody; ogon pokazuje ekstrema.



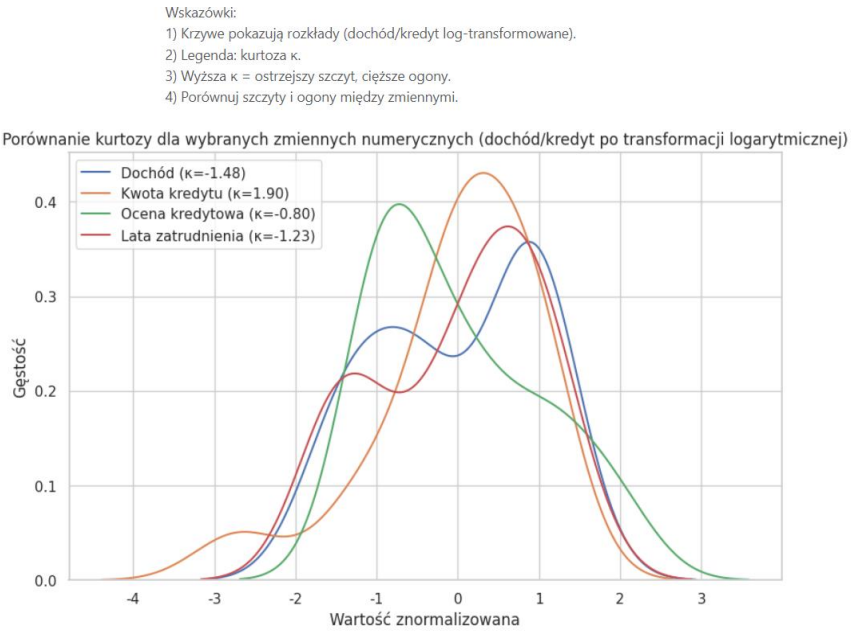
Rysunek 3.20. Wykres liniowy dochodów klientów

Wykres ten przedstawia liniowy rozkład dochodów klientów, posortowanych według rosnących wartości, i na jego podstawie można zaobserwować, że w początkowej części wykresu (klienci 0–8) linia jest stosunkowo płaska, co wskazuje na grupę klientów o bardzo podobnych, niskich dochodach (około 33 000–50 000 zł). Obserwacja ta jest zgodna z dominantą dochodu wynoszącą 33 278 zł oraz z histogramem częstości względnej, gdzie najwyższy słupek również przypada w tym zakresie. Co więcej wzrost dochodów nie jest ciągły – pojawiają się wyraźne skoki (np. między klientami 13 a 15, dochód rośnie z około 65 000 zł do 95 000 zł) oraz płaskie odcinki. Potwierdza to bimodalny charakter rozkładu dochodów, który był wcześniej widoczny na histogramach, oraz brak klientów w niektórych środkowych przedziałach dochodowych. W końcowej fazie (klienci 25–30) wykres staje się bardziej stromy, odzwierciedlając większe różnice między najwyższymi dochodami. Maksymalny dochód wynosi 146 939 zł i nie jest pojedynczym odstającym, lecz częścią narastającego trendu.

Użyteczność tego wykresu polega na tym, że dostarcza jasnego obrazu każdej obserwacji (w kontekście posortowania), ilustrując jednocześnie wyraźne skoki i luki widoczne na wykresie są kluczowym dowodem na to, że zmienna dochód jest niejednorodna i będzie wymagała ostrożnego traktowania w modelowaniu, prawdopodobnie poprzez kategoryzację (podział na koszyki) zamiast użycia jej jako zmiennej ciągłej.

3.4.21. Analiza wykresu porównania kurtozy dla wybranych zmiennych numerycznych

W tej analizie wyznaczono porównanie kurtozy dla wybranych zmiennych numerycznych (dochód/kredyt po transformacji logarytmicznej) i przedstawiono go w celu oceny kształtu rozkładu (kurtozy i skośności) kluczowych zmiennych po ich log-transformacji i normalizacji. Transformacja logarytmiczna jest często używana w danych finansowych, aby zbliżyć rozkłady do normalnego i zminimalizować wpływ ekstremalnych wartości.



Rysunek 3.21. Wykres porównania kurtozy dla wybranych zmiennych numerycznych (dochód/kredyt po transformacji logarytmicznej)

W celu oceny kształtu rozkładu kluczowych zmiennych numerycznych (Dochód, Kwota kredytu, Ocena kredytowa, Lata zatrudnienia) przeprowadzono log-transformację i normalizację danych. Transformacja logarytmiczna jest często stosowana w danych finansowych, aby zbliżyć rozkłady do normalnego i zmniejszyć wpływ ekstremalnych wartości.

Kurtoza mierzy „ostrość" szczytu rozkładu w porównaniu do rozkładu normalnego. Dodatnia kurtoza powoduje, że szczyt jest wyższy i węższy oraz widoczne są cięższe ogony. Ujemna kurtoza powoduje, że szczyt jest niższy i szerszy, lżejsze ogony. Analiza kurtozy dla badanych zmiennych:

- Ocena kredytowa - kurtoza wynosząca -0,80 oznacza to że wykres tej zmiennej ma najwyższy i najwęższy szczyt spośród wszystkich zmiennych. Dane są najbardziej

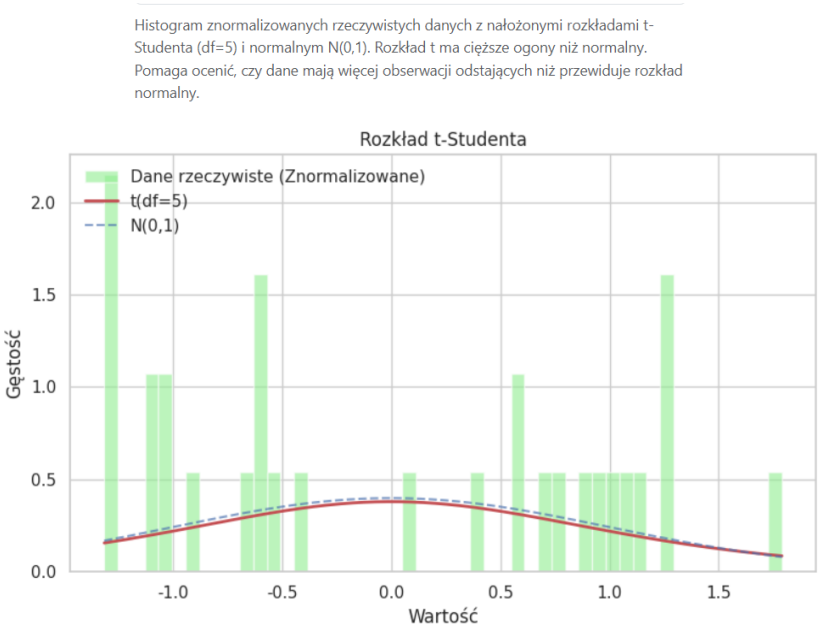
skupione wokół średniej i najbardziej przypominają normalny dzwon, czyli rozkład normalny. Większość klientów ma podobną ocenę kredytową, niewiele jest ekstremalnie niskich lub wysokich wartości.

- Zmienna Lata zatrudnienia - kurtoza wynosząca -1,23 sugeruje, że szczyt wykresu jest niższy i szerszy niż w przypadku Oceny kredytowej, co oznacza, że wartości lat zatrudnienia są bardziej rozproszone, mniej klientów ma identyczny staż pracy, a różnice między nimi są większe.
- Dochód - kurtoza wynosząca -1,48 sugeruje, że szczyt wykresu jest jeszcze niższy i bardzo szeroki, oraz że rozkład jest płaski. Mamy zarówno osoby z niskimi, średnimi, jak i wysokimi dochodami. Dane są rozproszone i trudniej przewidzieć typowego klienta po dochodzie.
- Kwota kredytu - kurtoza wynosząca -1,90 - wykres jest najbardziej spłaszczony i rozlany spośród wszystkich zmiennych, co za tym idzie kwoty kredytów bardzo się różnią, nie ma jednej typowej wartości. Dane są rozproszone, a koncentracja wokół średniej jest najmniejsza – klienci składają zarówno małe, jak i bardzo duże wnioski

Wykres gęstości po log-transformacji jest więc niezbędnym narzędziem diagnostycznym, pozwalającym ocenić efektywność transformacji i identyfikować zmienne wymagające specjalnego traktowania w modelowaniu (np. Dochód – poprzez segmentację lub podział na koszyki). Niska kurtoza dla wszystkich zmiennych sugeruje, że zbiór danych zawiera dużą różnorodność klientów, z mniejszą liczbą skrajnie odstających wartości.

3.4.22. Analiza wykresu rozkładu t-studenta

W tej analizie wyznaczono wykres rozkładu t-Studenta i przedstawiono go w celu oceny, który rozkład teoretyczny (normalny czy t-Studenta) lepiej opisuje kształt znormalizowanych danych rzeczywistych. Jest to element diagnostyczny służący do zrozumienia stabilności danych i potencjalnego ryzyka wynikającego z wartości ekstremalnych.



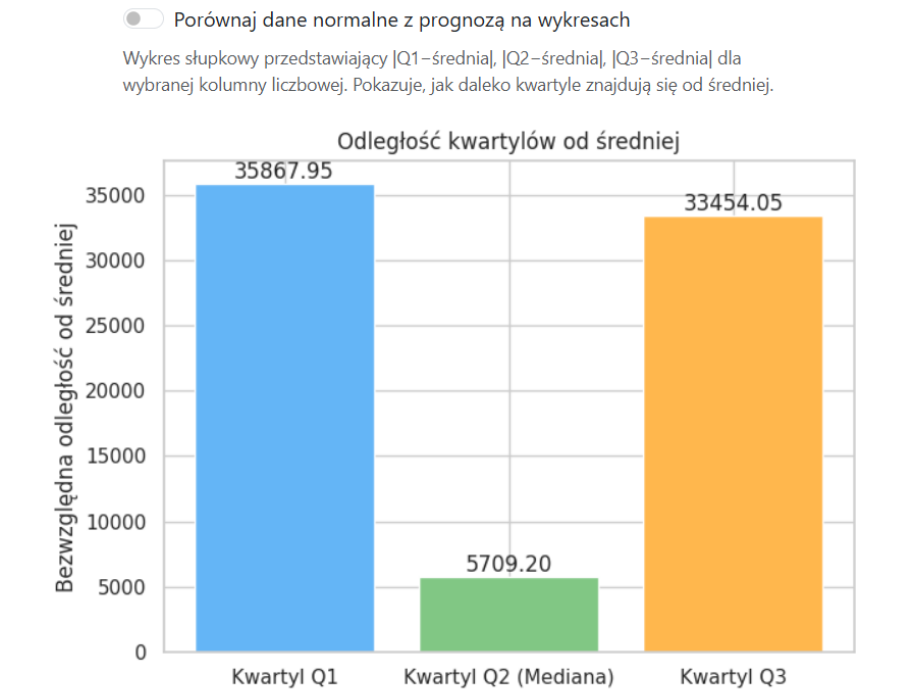
Rysunek 3.22. Wykres rozkładu t-studenta

Histogram danych pokazuje wyraźnie, że ich rzeczywiste rozłożenie znacząco odbiega od idealnej krzywej dzwonowej rozkładu normalnego. W normalnym rozkładzie większość obserwacji powinna być skupiona w centrum (w okolicach wartości 0 po standaryzacji), a ogony powinny opadać szybko i symetrycznie po obu stronach. Tymczasem znormalizowane dane prezentują nieregularny, „poszarpany" kształt, z wyraźnymi skupiskami obserwacji w obszarach skrajnych. Oznacza to, że w zbiorze znajduje się więcej wartości ekstremalnych, niż przewidywał rozkład normalny. Wyższe słupki histogramu w ogonach – zarówno w lewym, jak i prawym – potwierdzają, że dane są obarczone cięższymi ogonami, co jest typowym sygnałem niestabilności rozkładu i zwiększonego ryzyka występowania odstających klientów. Dlatego rozkład t-studenta o 5 stopniach swobody lepiej odwzorowuje ten charakter danych. Krzywa t-Studenta ma naturalnie cięższe ogony niż rozkład normalny, co pozwala jej „łapać" więcej obserwacji w obszarach dalekich od średniej. Na wykresie widać, że krzywa t(df=5) przebiega bliżej słupków histogramu, szczególnie w regionach skrajnych, gdzie rozkład normalny wyraźnie niedoszacowuje liczby obserwacji. Oznacza to, że t-Student jest znacznie adekwatniejszym modelem teoretycznym, jeśli chcemy opisywać zmienne o zwiększonej zmienności i wysokim udziale wartości odstających. W dodatku wynik tej analizy jest spójny z wcześniejszymi pomiarami kurtozy, które wskazywały na spłaszczone, niestandardowe rozkłady (np. dla dochodu $K \approx -1{,}57$, dla kwoty kredytu $K \approx -0{,}93$). Takie wartości kurtozy potwierdzają niską koncentrację w centrum rozkładu oraz zwiększone rozproszenie danych, co bezpośrednio przekłada się na cięższe ogony i większą zmienność.

Wykres ten potwierdza diagnozę z poprzednich analiz, zwłaszcza niską kurtozę dla większości zmiennych. Wskazuje, że użycie modeli zakładających normalność może prowadzić do niedoszacowania tego ryzyka, szczególnie w przypadku klientów reprezentujących wartości skrajne. Dlatego bardziej odpowiednią metodą jest modelowanie oparte na rozkładzie t-Studenta lub zastosowanie metod nieparametrycznych, które nie wymagają założenia normalności i są bardziej odporne na nieregularne, rzeczywiste struktury danych.

3.4.23. Analiza Wykresu odległości kwartyli od średniej dla zmiennej dochód

Jako ostatni wykres przeanalizowano odległość kwartyli od średniej , który przedstawia bezwzględną różnicę między wartościami kwartyl (Q1, mediana Q2, Q3) a wartością średnią dla zmiennej dochód.



Rysunek 3.23. Wykres odległości kwartyli od średniej dla zmiennej dochód

Wykres ten przedstawiaja odległość kwartylów od średniej dla zmiennej Dochód. Pokazuje, jak poszczególne części rozkładu dochodów klientów różnią się od wartości przeciętnej. Analiza ta polega na porównaniu średniego dochodu z wartościami charakterystycznymi dla dolnej ćwiartki danych (Q1), mediany (Q2) oraz górnej ćwiartki (Q3), przy czym odległości te zostały wyrażone jako wartości bezwzględne, co pozwoliło

na obiektywne porównanie kierunków rozkładu bez uwzględniania tego, czy dany kwartyl znajduje się powyżej czy poniżej średniej.

Średnia dochodów klientów wynosi 81 384,2 zł i jest zbliżona do mediany, która osiąga poziom 75 675 zł. Bardzo niewielka odległość mediany od średniej (wynosząca 5 709,20 zł) świadczy o tym, że centralna część rozkładu jest stosunkowo stabilna i spójna z wartością średnią. Taka zależność potwierdza niewielką skośność rozkładu dochodów (0,1031), co oznacza, że dochody są rozłożone w sposób niemal symetryczny, bez znaczącego przechylenia w stronę wartości niskich lub wysokich.

Znacznie większe różnice obserwujemy jednak w przypadku dolnego i górnego kwartyla. Największą odległość od średniej wykazuje kwartyl Q1, wynoszący 35 867,95 zł. Oznacza to, że grupa klientów o najniższych dochodach (dolne 25% próby) to klienci o najniższych zarobkach - to oni najmocniej odbiegają od typowego klienta. Ta część populacji charakteryzuje się dużą rozbieżnością względem średniej, co wskazuje na jej mniejszą stabilność finansową oraz większą niejednorodność. Z kolei kwartyl trzeci (Q3) znajduje się bliżej średniej niż Q1, ale jego odległość wciąż jest znacząca i wynosi 33 454,05 zł. Różnica między Q1 i Q3 wskazuje, że rozkład dochodów jest bardziej rozciągnięty po stronie niższych wartości niż po stronie wyższych, co oznacza, że dochody klientów o niskich zarobkach są bardziej rozproszone i mniej przewidywalne niż w przypadku osób najlepiej zarabiających. Zjawisko to jest spójne z wartością dominanty wynoszącą zaledwie 33 278 zł, która potwierdza, że najczęściej występujące dochody należą do segmentu o najniższych zarobkach.

W analizie ryzyka kredytowego wykres ten dostarcza istotnych informacji. Wyraźne oddalenie dolnego kwartyla od średniej oznacza, że to właśnie klienci o najniższych dochodach stanowią najbardziej zróżnicowaną i najmniej przewidywalną grupę osób którym można udzielić kredytu. W praktyce oznacza to konieczność ostrożniejszego podejścia do oceny zdolności kredytowej w tym segmencie osób, ponieważ odbiegają one najbardziej od typowego profilu klienta i generują największą niepewność z punktu widzenia instytucji finansowej.

4. TWARZE CHERNOFFA

Poniższy rozdział zawiera prezentację twarzy Chernoffa dla trzech zestawów danych - podstawowego składającego się z trzydziestu rekordów, prognostycznego zawierającego osiem rekordów syntetycznych utworzonych na podstawie rekordów z podstawowego zestawu danych oraz z zestawu danych podstawowych oraz prognostycznych, który składa się łącznie z trzydziestu ośmiu rekordów. Twarze te wizualizują odległość najbliższego kwartyla od wartości średniej dla wskazanego atrybutu.

Dla atrybutu “Ocena Kredytowa” okrągła twarz wizualizuje to, że kwartyl Q1 jest najbliższy wartości średniej, prostokątna twarz to, że wartość kwartyla Q2 jest najbliższa wartości średniej, a trójkątna twarz to, że wartość kwartyla Q3 jest najbliższa.

W przypadku atrybutu “Dochód” okrągła twarz z okrągłymi oczami wizualizuje to, że kwartyl Q1 jest najbliższy wartości średniej, prostokątna twarz z prostokątnymi oczyma to, że wartość kwartyla Q2 jest najbliższa wartości średniej, a trójkątna twarz z trójkątnymi oczami to, że wartość kwartyla Q3 jest najbliższa.

Dla wartości atrybutu “Kwota Pożyczki” okrągła twarz z okrągłymi ustami wizualizuje to, że kwartyl Q1 jest najbliższy wartości średniej, prostokątna twarz z prostokątnymi ustami to, że wartość kwartyla Q2 jest najbliższa wartości średniej, a trójkątna twarz z trójkątnymi ustami to, że wartość kwartyla Q3 jest najbliższa.

W przypadku wartości atrybutu “Punkty” okrągła twarz z okrągłymi uszami wizualizuje to, że kwartyl Q1 jest najbliższy wartości średniej, prostokątna twarz z prostokątnymi uszami to, że wartość kwartyla Q2 jest najbliższa wartości średniej, a trójkątna twarz z trójkątnymi uszami to, że wartość kwartyla Q3 jest najbliższa.

Dla atrybutu “Lata Pracy” okrągła twarz z okrągłymi uszami wizualizuje to, że kwartyl Q1 jest najbliższy wartości średniej, prostokątna twarz z prostokątnymi uszami to, że wartość kwartyla Q2 jest najbliższa wartości średniej, a trójkątna twarz z trójkątnymi uszami to, że wartość kwartyla Q3 jest najbliższa. Poniższy rysunek przedstawia legendę rysowania twarzy Chernoffa.

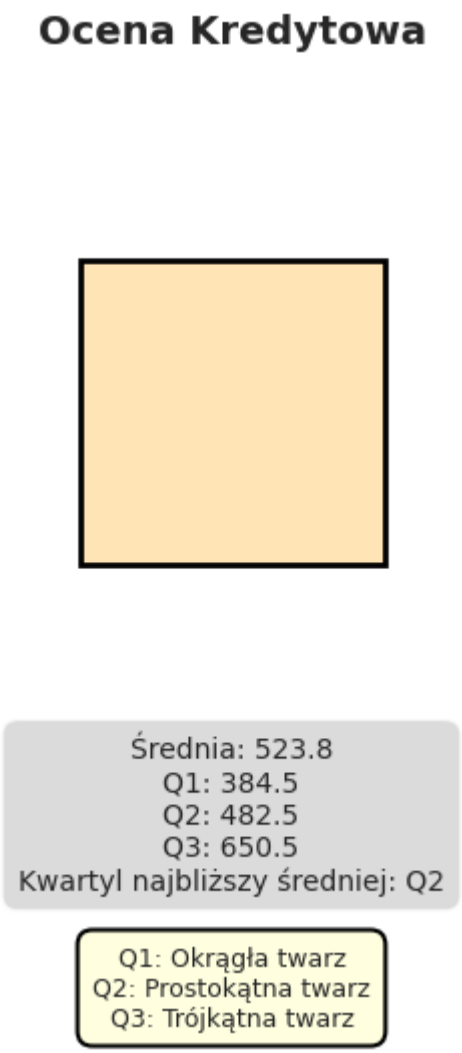
Rysunek 4.1. Wykres odległości kwartyli od średniej dla zmiennej dochód

4.1. Dane podstawowe

Poniższy podrozdział prezentuje twarze Chernoffa dla podstawowego zbioru danych wraz z wyjaśnieniem znaczenia kształtu twarzy i różnic w nich zawartych.

4.1.1. Ocena Kredytowa

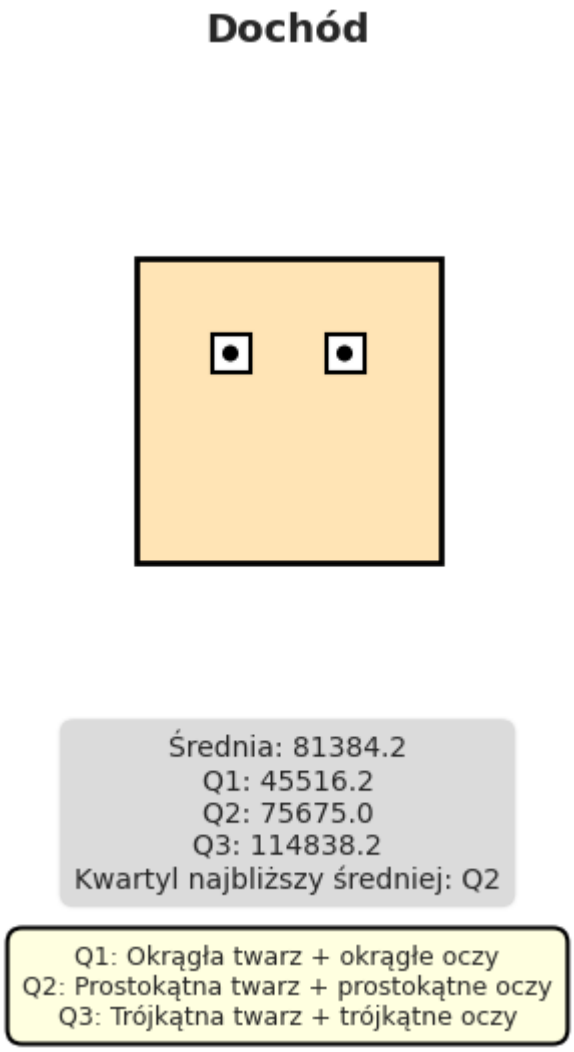
Poniższy rysunek przedstawia wizualizację odległości najbliższego kwartyla od wartości średniej dla oceny kredytowej wszystkich rekordów w zbiorze. Na rysunku 4.2 widoczny jest kwadratowy kształt twarzy Chernoffa, który przedstawia fakt, że wartość środkowa (kwartyl Q2 - mediana) znajduje się bardzo blisko średniej oceny kredytowej. Oznacza to, że typowa ocena kredytowa osób w zbiorze niemal w całości pokrywa się z wartością średnią co wskazuje na niewielki rozrzut danych w centralnej części rozkładu.



Rysunek 4.2. T warz Chernoffa dla atrybutu „Ocena Kredytowa" z podstawowego zbioru danych

4.1.2. Dochód

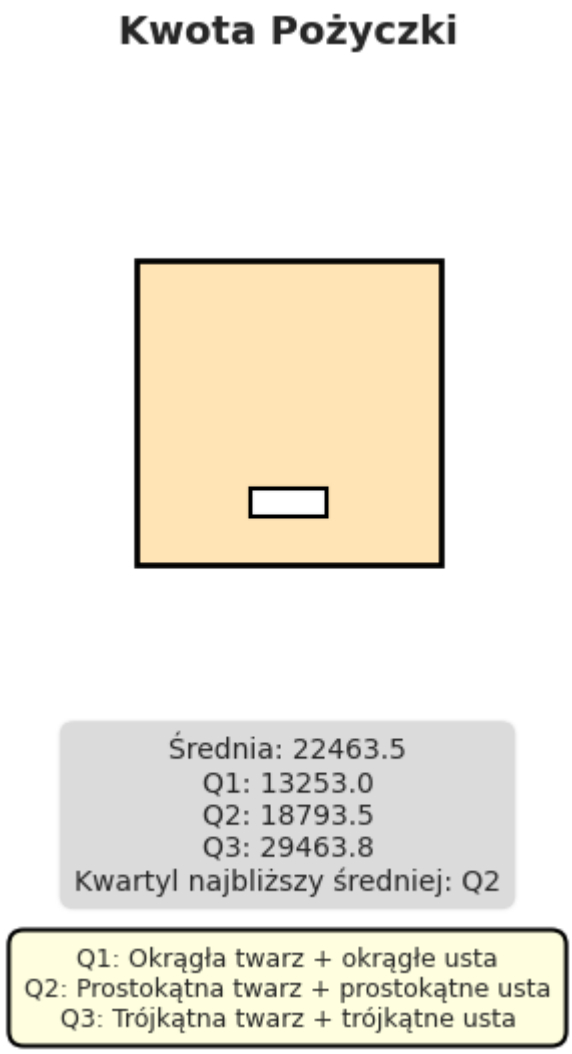
Poniższy rysunek przedstawia wizualizację odległości najbliższego kwartyla od wartości średniej dla dochodu wszystkich rekordów w zbiorze. Na rysunku 4.3 widoczna jest kwadratowa twarz Chernoffa z kwadratowymi oczami, która odzwierciedla fakt, że wartość środkowa (kwartyl Q2 – mediana) dochodu znajduje się bardzo blisko średniej wartości dochodu.



Rysunek 4.3. Twarz Chernoffa dla atrybutu „Dochód" z podstawowego zbioru danych

4.1.3. Kwota Pożyczki

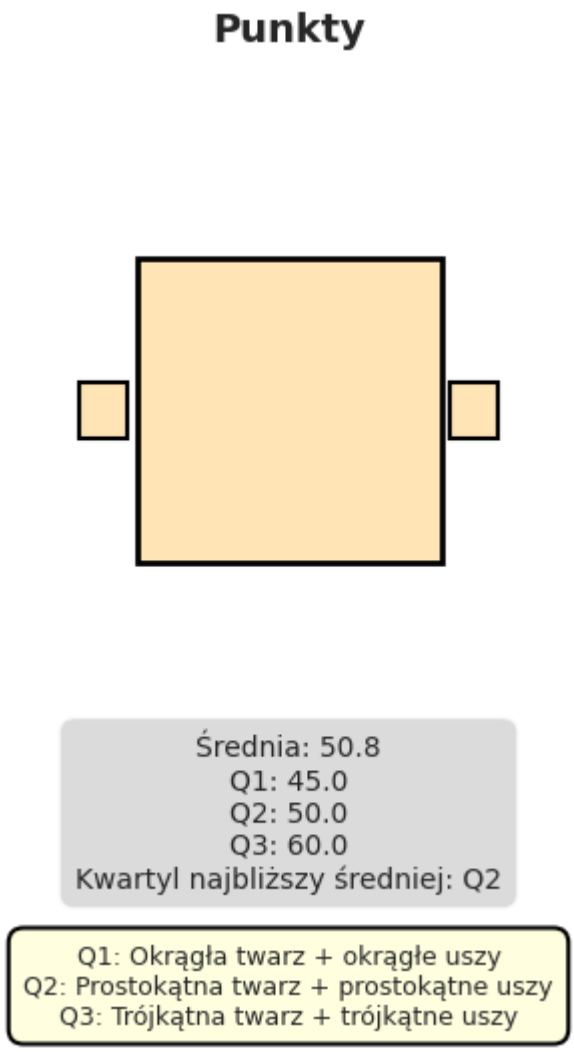
Poniższy rysunek przedstawia wizualizację odległości najbliższego kwartyla od wartości średniej dla kwoty pożyczki wszystkich rekordów w zbiorze. Na rysunku 4.4 widoczna jest kwadratowa twarz Chernoffa z kwadratowymi uszami, co wskazuje, że wartość środkowa (kwartyl Q2 – mediana) kwoty pożyczki znajduje się bardzo blisko średniej wartości udzielonych pożyczek. W praktyce oznacza to, że najczęściej wybierana kwota pożyczki niemal pokrywa się z wartością średnią, co sugeruje niewielką zmienność danych w centrum rozkładu.



Rysunek 4.4. Twarz Chernoffa dla atrybutu „Kwota Pożyczki" z podstawowego zbioru danych

4.1.4. Punkty

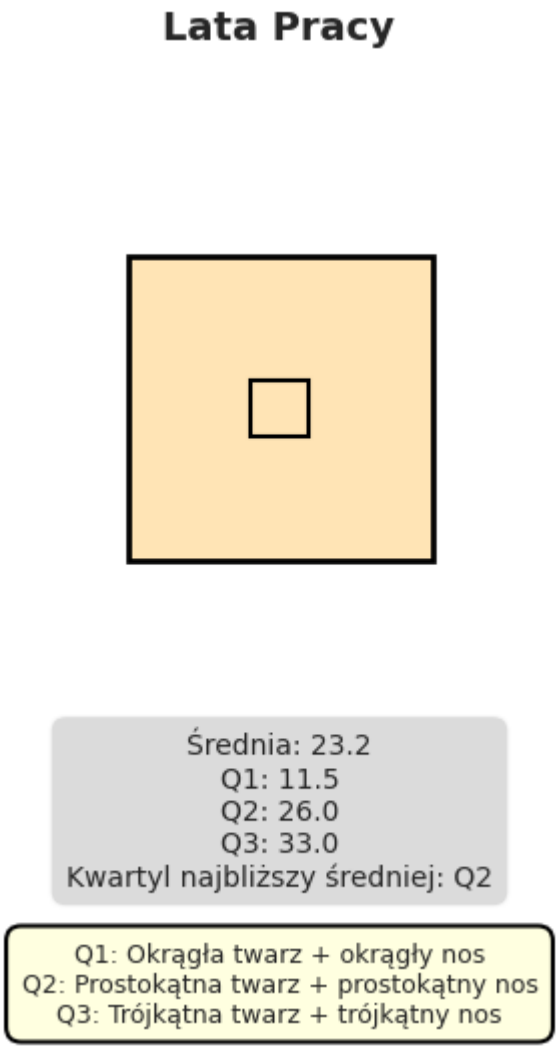
Poniższy rysunek przedstawia wizualizację odległości najbliższego kwartyla od wartości średniej dla liczby punktów przypisanych wszystkim rekordom w zbiorze. Na rysunku 4.5 widoczna jest kwadratowa twarz Chernoffa z prostokątnymi uszami, reprezentująca fakt, że wartość środkowa (kwartyl Q2 – mediana) znajduje się bardzo blisko średniej liczby punktów. Oznacza to, że typowa liczba punktów nie odbiega znacząco od średniej, co potwierdza stabilny i mało rozproszony rozkład wyników.



Rysunek 4.5. Twarz Chernoffa dla atrybutu „Punkty” z podstawowego zbioru danych

4.1.5. Lata Pracy

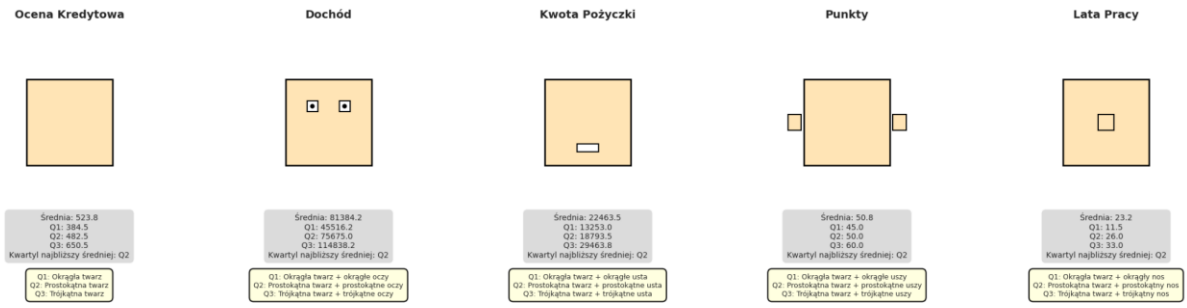
Poniższy rysunek przedstawia wizualizację odległości najbliższego kwartyla od wartości średniej dla liczby lat pracy wszystkich osób w zbiorze. Na rysunku 4.6 widoczna jest kwadratowa twarz Chernoffa z kwadratowym nosem, co oznacza, że wartość środkowa (kwartyl Q2 – mediana) lat pracy znajduje się bardzo blisko średniej liczby lat pracy. W praktyce wskazuje to, że większość osób ma doświadczenie zawodowe zbliżone do wartości typowej, a dane w centralnej części rozkładu są słabo zróżnicowane.



Rysunek 4.6. Twarz Chernoffa dla atrybutu „Lata Pracy" z podstawowego zbioru danych

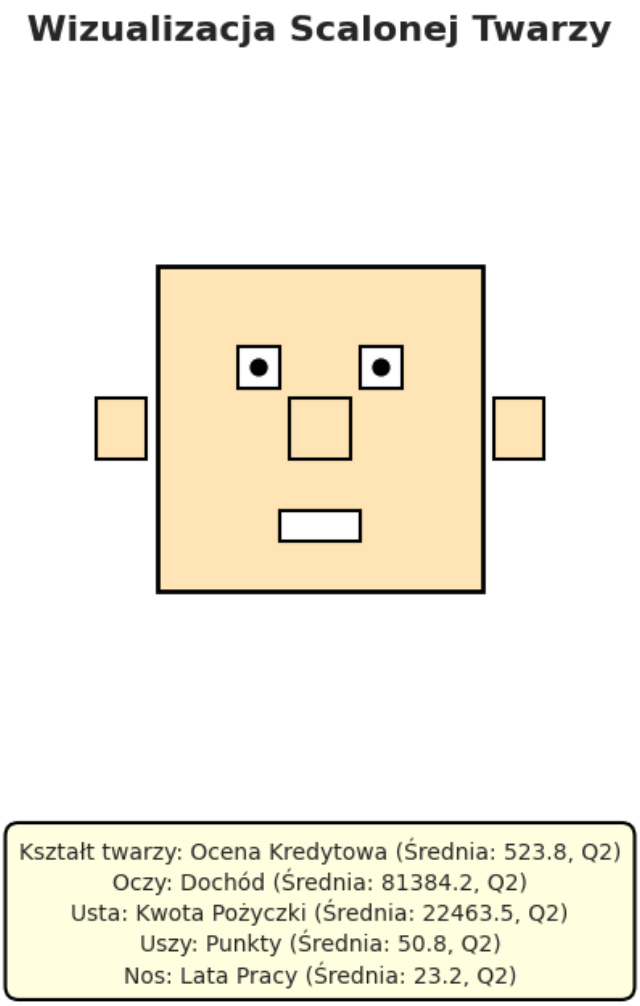
4.1.6. Wszystkie atrybuty

Na poniższym rysunku zaprezentowano Twarze Chernoffa reprezentujące odległość najbliższego kwartyla do wartości średniej dla wszystkich atrybutów.



Rysunek 4.7. Wszystkie twarze Chernoffa dla wszystkich atrybutów z podstawowego zbioru danych

Poniższy rysunek przedstawia widok złączonych twarzy Chernoffa, prezentujący, iż dla wszystkich atrybutów odległość drugiego kwartyla od średniej była najbliższa.



Rysunek 4.8. Połączona twarz Chernoffa wizualizująca zbiorczo dla wszystkich atrybutów, który kwartyl miał najmniejszą odległość od średniej

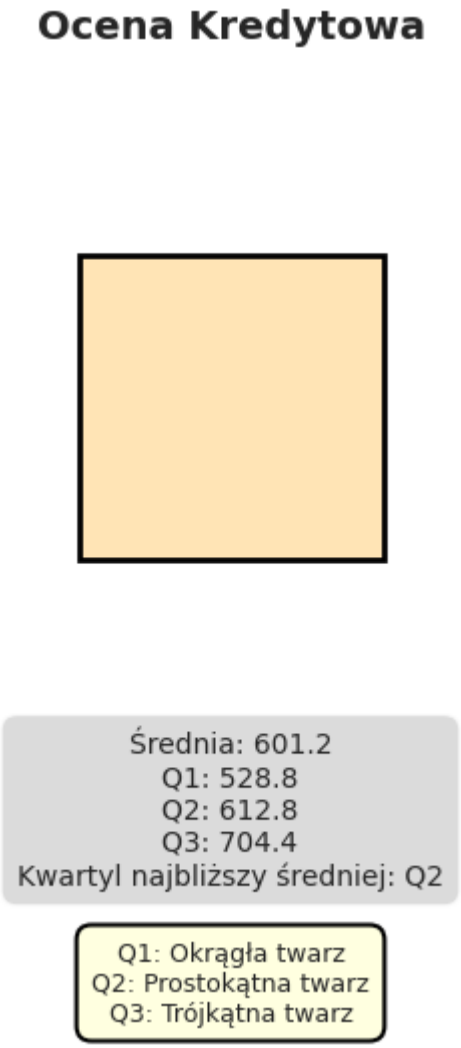
4.2. Dane prognostyczne

W niniejszym podrozdziale przedstawiono twarze Chernoffa wygenerowane na podstawie danych prognostycznych. Podobnie jak w przypadku zbioru podstawowego, kształty poszczególnych elementów twarzy pozostają takie same, ponieważ wynikają z relacji najbliższego kwartyla do wartości średniej. Różnice dotyczą jedynie wartości liczbowych w kwartylach i średnich, co wpływa na stopień regularności wizualizacji, ale nie na sam typ prezentowanego kształtu.

4.2.1. Ocena Kredytowa

W danych prognostycznych średnia ocena kredytowa wynosi 601.2, natomiast wartości kwartylowe to Q1 = 528.8, Q2 = 612.8, Q3 = 704.4. Ponieważ mediana

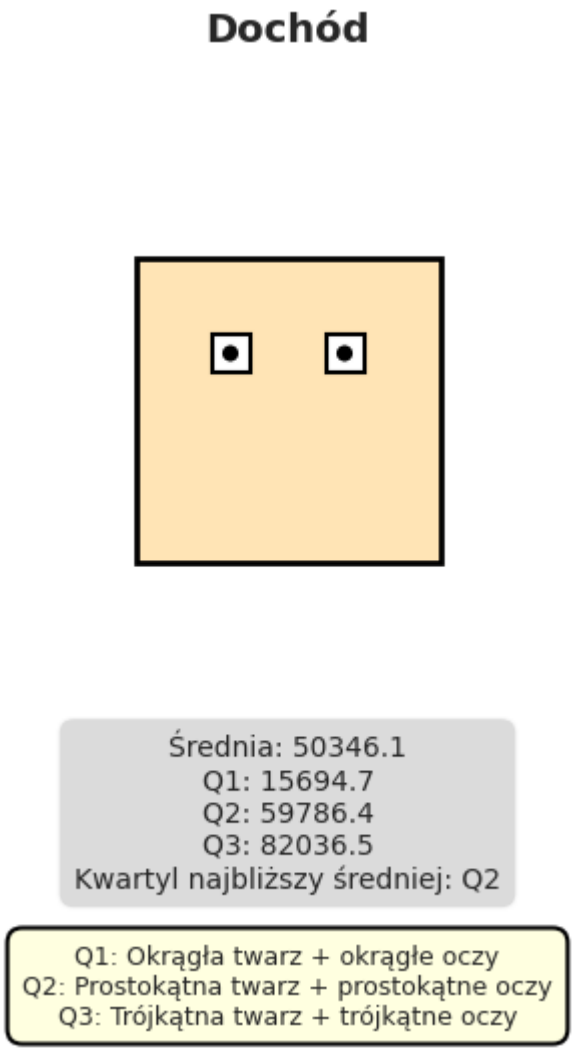
(Q2) ponownie znajduje się bardzo blisko średniej, twarz Chernoffa przyjmuje kwadratowy kształt, podobnie jak w zbiorze podstawowym. Świadczy to o tym, że również w prognozach środkowa część rozkładu ocen kredytowych jest stabilna i mało rozproszona.



Rysunek 4.9. Twarz Chernoffa dla atrybutu „Ocena Kredytowa" z prognostycznego zbioru danych

4.2.2. Dochód

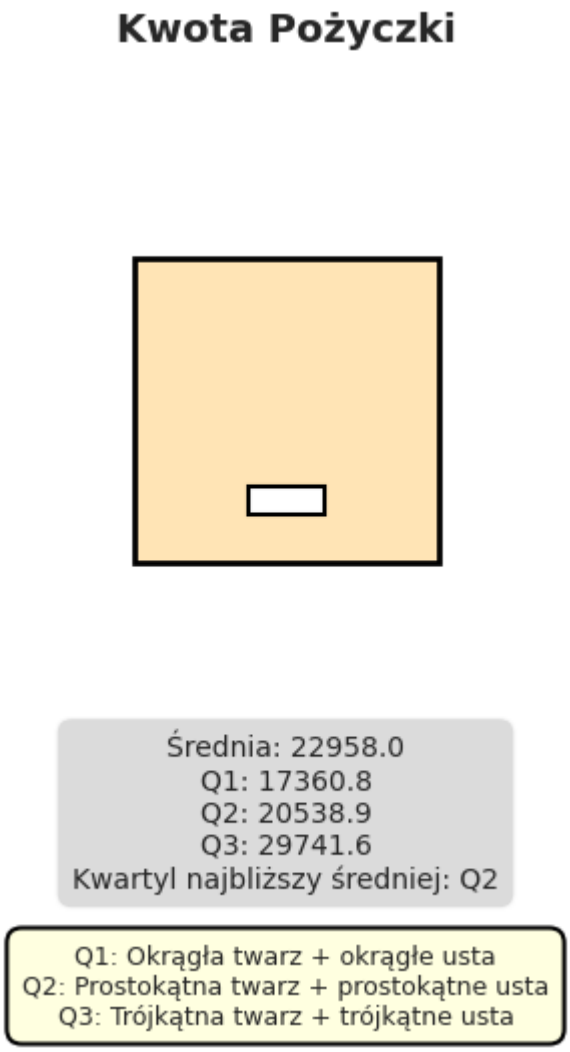
Dla danych prognostycznych średnia wartość dochodu to 50346.1, natomiast kwartyle wynoszą Q1 = 15694.7, Q2 = 59786.4, Q3 = 82036.5. Pomimo różnic liczbowych względem danych podstawowych, mediana ponownie leży blisko wartości średniej, dlatego twarz Chernoffa zachowuje kwadratowy kształt z kwadratowymi oczami. Sugeruje to, że prognozowane dochody również tworzą rozklad o relatywnie małej zmienności w części centralnej.



Rysunek 4.10. Twarz Chernoffa dla atrybutu „Dochód” z prognostycznego zbioru danych

4.2.3. Kwota Pożyczki

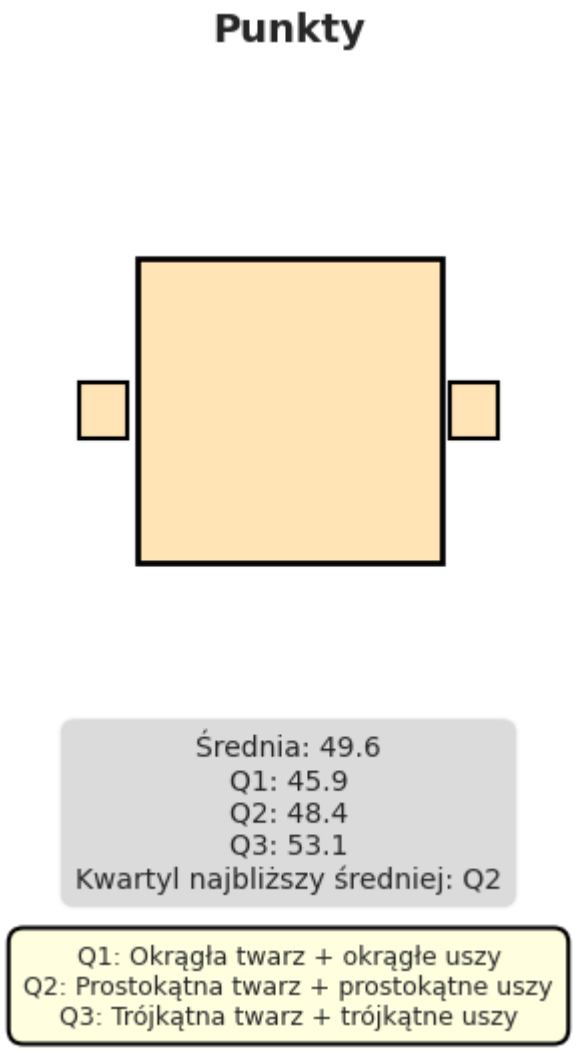
W prognozach średnia kwota pożyczki wynosi 22958, a wartości kwartylowe to Q1 = 17360.8, Q2 = 20538.9, Q3 = 29741.6. Różnice względem danych podstawowych są niewielkie, a mediana pozostaje najbliżej średniej, co powoduje, że twarz Chernoffa zachowuje kwadratowy kształt z kwadratowymi uszami. Oznacza to, że prognozowane kwoty pożyczek utrzymują porównywalny do danych podstawowych poziom centralnej stabilności.



Rysunek 4.11. Twarz Chernoffa dla atrybutu „Kwota pożyczki” z prognostycznego zbioru danych

4.2.4. Punkty

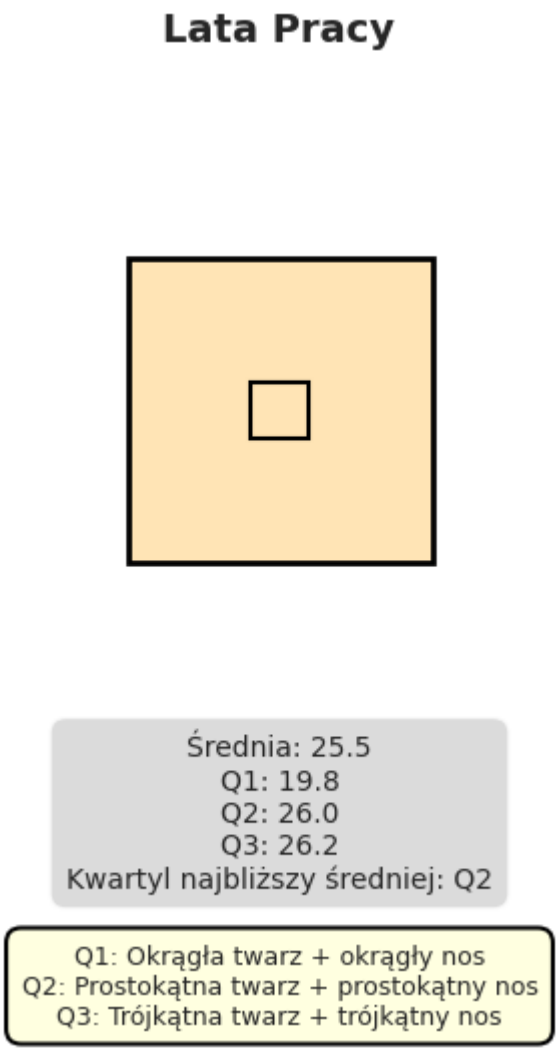
Dla danych prognostycznych średnia liczba punktów wynosi 49.6, natomiast kwartyle to Q1 = 45.9, Q2 = 48.4, Q3 = 53.1. Mediana ponownie jest wartością najbliższą średniej, dlatego wizualizacja przyjmuje kwadratową twarz z prostokątnymi uszami. Oznacza to, że prognozowane wartości punktów również układają się w zbliżony, stabilny sposób jak w danych podstawowych.



Rysunek 4.12. Twarz Chernoffa dla atrybutu „Punkty" z prognostycznego zbioru danych

4.2.5. Lata Pracy

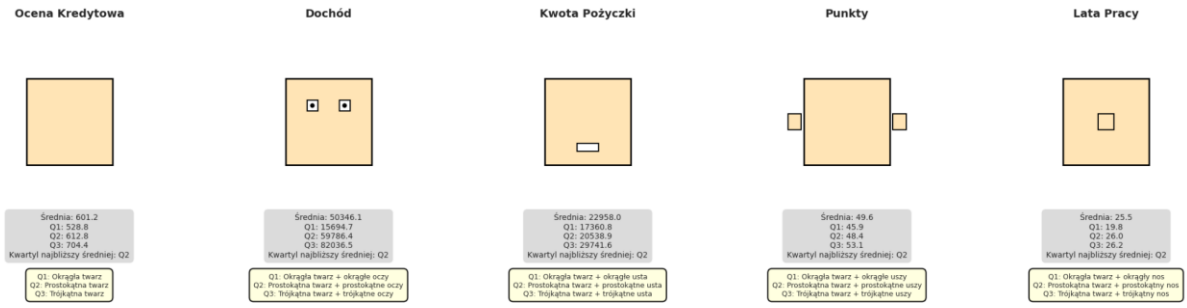
W przypadku lat pracy średnia prognoza wynosi 25.5, a kwartyle to Q1 = 19.8, Q2 = 26.0, Q3 = 26.2. Ponieważ mediana ponownie niemal pokrywa się ze średnią, twarz Chernoffa zachowuje kwadratowy kształt z kwadratowym nosem. Wskazuje to, że także w prognozach doświadczenie zawodowe osób skupia się wokół typowej wartości, z niewielką zmiennością w centrum rozkładu.



Rysunek 4.13. Twarz Chernoffa dla atrybutu „Lata Pracy" z prognostycznego zbioru danych

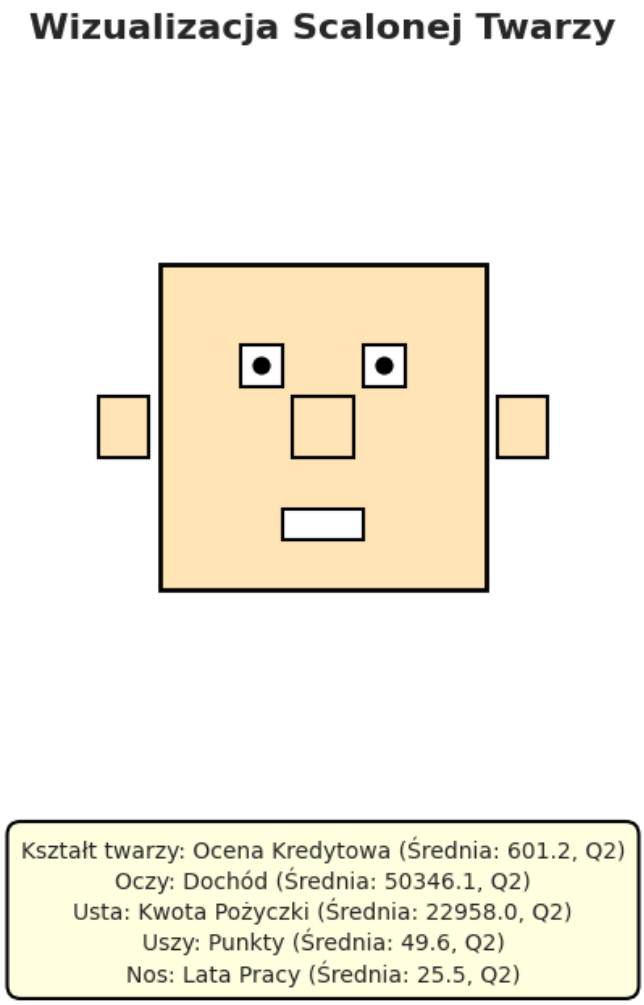
4.2.6. Wszystkie atrybuty

Na poniższym rysunku przedstawiono Twarze Chernoffa ilustrujące odległość najbliższego kwartyla od wartości średniej dla wszystkich atrybutów w danych prognostycznych.



Rysunek 4. 14. Wszystkie twarze Chernoffa dla wszystkich atrybutów z prognostycznego zbioru danych

Poniższy rysunek przedstawia widok złączonych twarzy Chernoffa, prezentujący, iż dla wszystkich atrybutów odległość drugiego kwartyla od średniej była najbliższa.



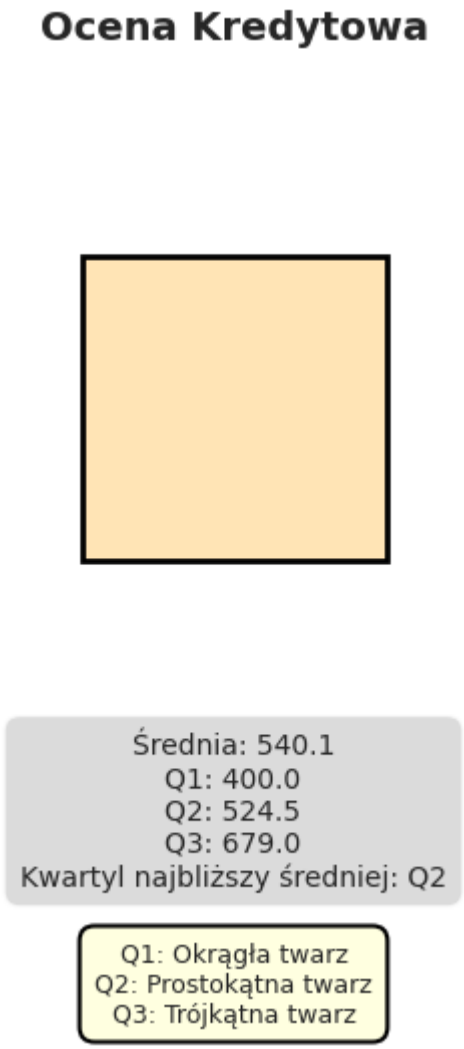
Rysunek 4.15. Połączona twarz Chernoffa wizualizująca zbiorczo dla wszystkich atrybutów, który kwartyl miał najmniejszą odległość od średniej

4.3. Dane podstawowe oraz prognostyczne

W niniejszym podrozdziale przedstawiono twarze Chernoffa wygenerowane na podstawie połączonego zbioru danych, zawierającego zarówno dane podstawowe, jak i prognostyczne. Połączenie dwóch zbiorów skutkuje zmianą wartości średnich i kwartylowych, jednak relacja najbliższego kwartyla względem średniej pozostaje podobna jak w poprzednich przypadkach. Z tego powodu również w tym zestawieniu kształty twarzy Chernoffa pozostają niezmienione, natomiast ich regularność odzwierciedla stopień rozproszenia danych w części centralnej.

4.3.1. Ocena Kredytowa

W połączonym zbiorze danych średnia ocena kredytowa wynosi 540.1, natomiast wartości kwartylowe to Q1 = 400, Q2 = 524.5, Q3 = 679. Ponieważ mediana (Q2) ponownie znajduje się najbliżej wartości średniej, twarz Chernoffa zachowuje kwadratowy kształt. Oznacza to, że rozkład ocen kredytowych dla połączonych danych charakteryzuje się stosunkowo małą zmiennością w centrum, podobnie jak w danych podstawowych i prognostycznych.

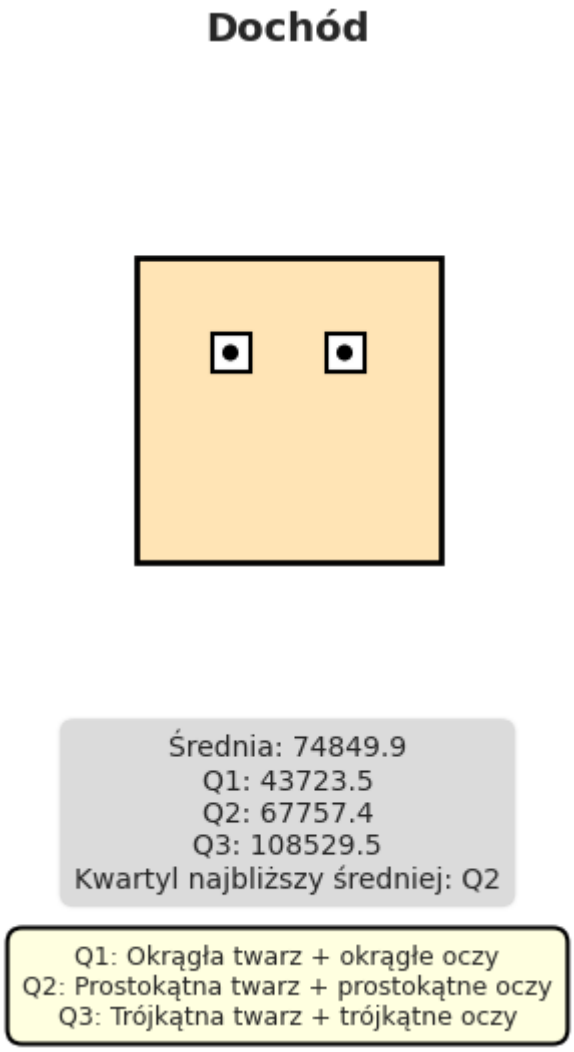


Rysunek 4.16. Twarz Chernoffa dla atrybutu „Ocena Kredytowa" z połączonego zbioru danych

4.3.2. Dochód

W przypadku dochodu, średnia dla połączonego zbioru wynosi 74849.9, natomiast wartości kwartylowe to Q1 = 43723.5, Q2 = 67757.4, Q3 = 108529.5. Najbliższym wartością średniej ponownie pozostaje kwartyl środkowy (Q2), dlatego twarz Chernoffa

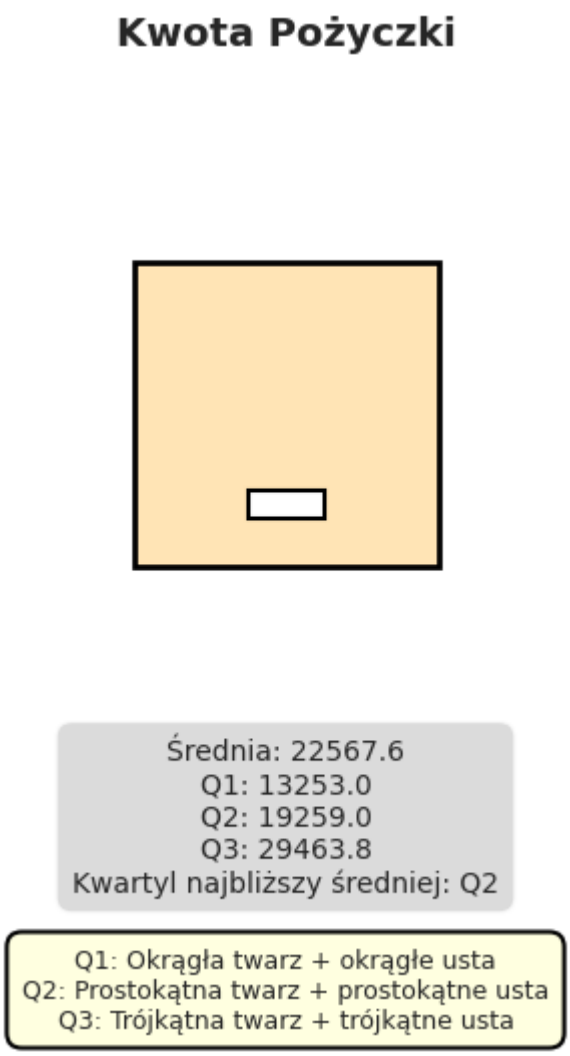
przyjmuje kwadratowy kształt z kwadratowymi oczami. Wskazuje to, że prognozy oraz dane rzeczywiste układają się w spójny rozkład dochodów, o ograniczonej zmienności w centralnej części.



Rysunek 4.17. Twarz Chernoffa dla atrybutu „Dochód" z połączonego zbioru danych

4.3.3. Kwota Pożyczki

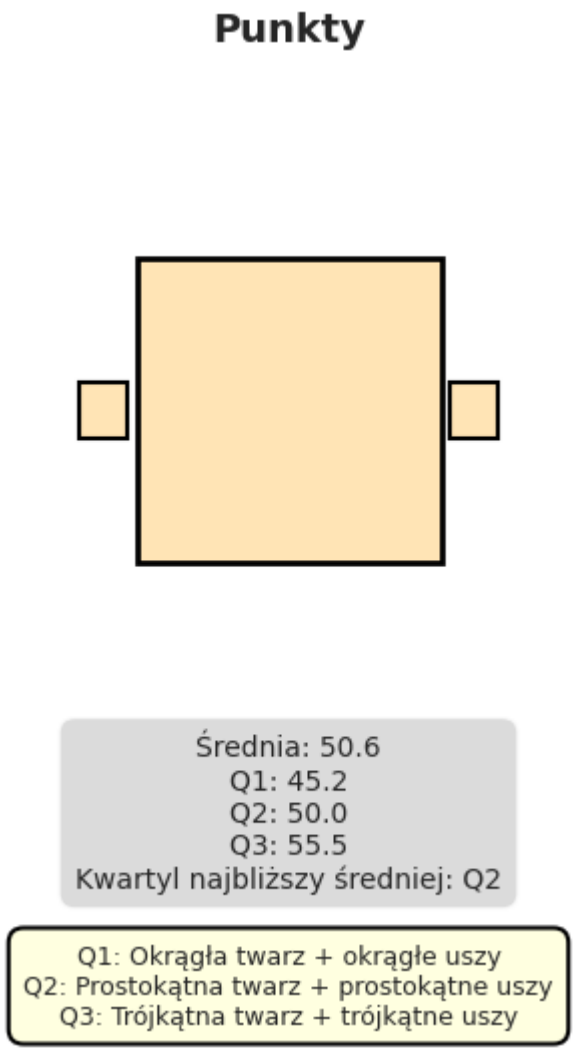
Dla połączonego zbioru średnia kwota pożyczki wynosi 22567.6, natomiast kwartyle to Q1 = 13253, Q2 = 19259, Q3 = 29463.8. Mediana jest najbliższa średniej, co powoduje, że twarz Chernoffa przybiera kwadratowy kształt z kwadratowymi uszami. W praktyce oznacza to, że również po połączeniu danych kwoty pożyczek pozostają stabilne i mało rozproszone w środkowej części rozkładu.



Rysunek 4.18. Twarz Chernoffa dla atrybutu „Kwota Pożyczki" z połączonego zbioru danych

4.3.4. Punkty

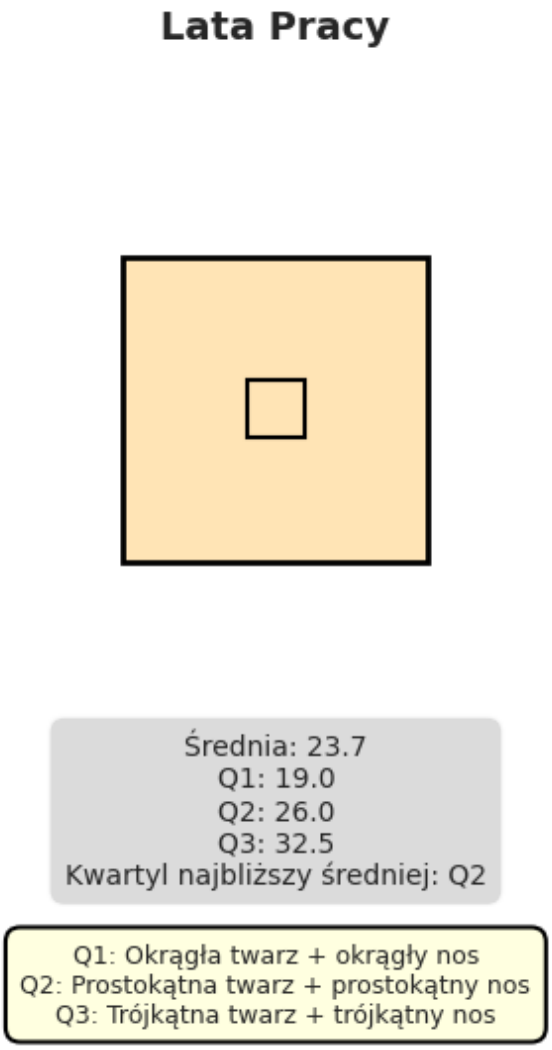
W połączonym zbiorze średnia liczba punktów wynosi 50.6, natomiast wartości kwartylowe to Q1 = 45.2, Q2 = 50, Q3 = 55.5. Najbliższą średniej wartością pozostaje Q2, dzięki czemu wizualizacja zachowuje kwadratową twarz z prostokątnymi uszami. Oznacza to, że wartości punktowe w połączonym zbiorze są bardzo spójne, a środkowa część rozkładu pozostaje dobrze skupiona.



Rysunek 4.19. Twarz Chernoffa dla atrybutu „Punkty" z połączonego zbioru danych

4.3.5. Lata Pracy

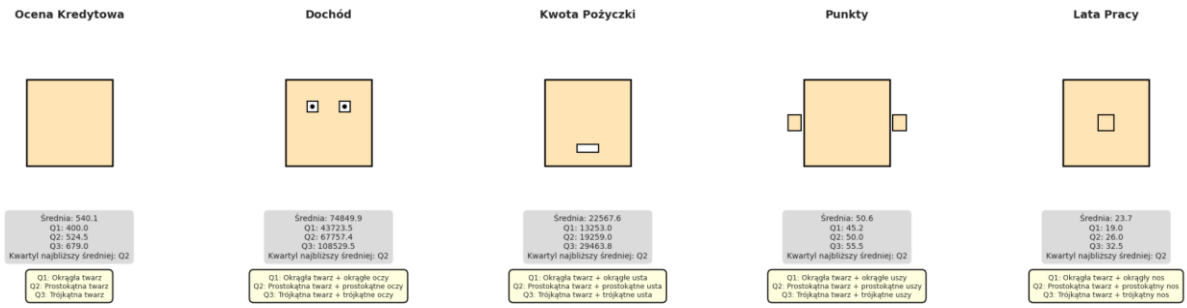
W danych połączonych średnia liczba lat pracy wynosi 23.7, natomiast kwartyle to Q1 = 19, Q2 = 26, Q3 = 32.5. Podobnie jak wcześniej, mediana (Q2) znajduje się najbliżej wartości średniej, dlatego twarz Chernoffa zachowuje kwadratowy kształt z kwadratowym nosem. Świadczy to o tym, że również w połączonym zbiorze wartości dotyczące doświadczenia zawodowego są skupione wokół typowej wartości, z niewielkim rozproszeniem.



Rysunek 4.20. Twarz Chernoffa dla atrybutu „Lata Pracy" z połączonego zbioru danych

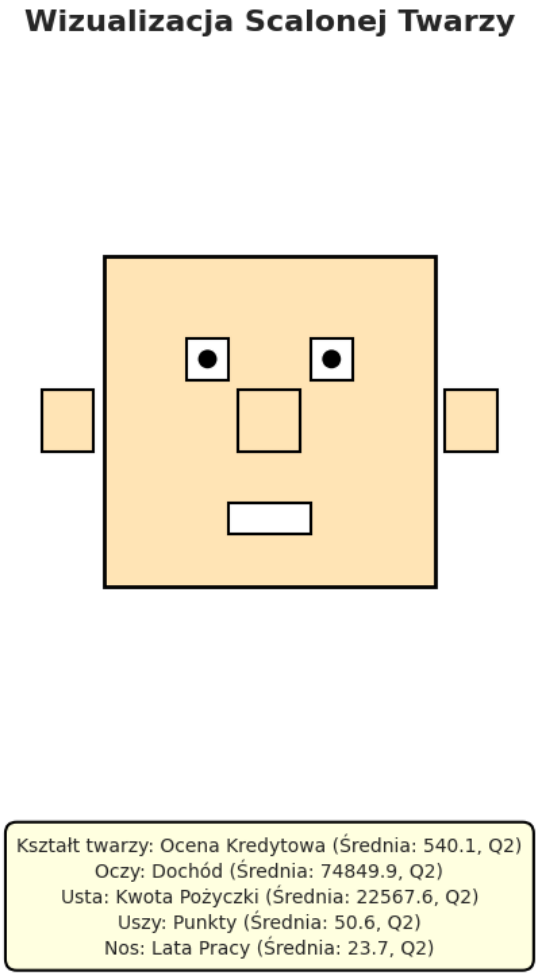
4.3.6. Wszystkie atrybuty

Na poniższym rysunku zaprezentowano Twarze Chernoffa obrazujące odległość najbliższego kwartyla od średniej dla wszystkich atrybutów w połączonym zbiorze danych.



Rysunek 4.21. Wszystkie twarze Chernoffa dla wszystkich atrybutów z podstawowego zbioru danych

Poniższy rysunek przedstawia widok złączonych twarzy Chernoffa, prezentujący, iż dla wszystkich atrybutów odległość drugiego kwartyla od średniej była najbliższa.



Rysunek 4.22. Połączona twarz Chernoffa wizualizująca zbiorczo dla wszystkich atrybutów, który kwartyl miał najmniejszą odległość od średniej

5. WNIOSKI

Celem niniejszej pracy było zaprojektowanie i implementacja aplikacji do analizy statystycznej oraz przeprowadzenie kompleksowego badania zbioru danych kredytowych. Cel ten został w pełni zrealizowany. Stworzone narzędzie pozwoliło na efektywne przetworzenie danych, wygenerowanie statystyk opisowych oraz wizualizację wielowymiarowych zależności przy użyciu zarówno standardowych wykresów, jak i zaawansowanej metody Twarzy Chernoffa.

5.1. Podsumowanie wyników i interpretacja zależności

Przeprowadzona analiza, obejmująca m.in. wykresy gęstości i korelacji, jednoznacznie wykazała, że ocena kredytowa (credit score) jest najsilniejszym czynnikiem determinującym decyzję o przyznaniu pożyczki. Wnioski zatwierdzone, stanowiące około 30% zbioru, charakteryzowały się wysoką punktacją z medianą powyżej 700, podczas gdy wnioski odrzucone (około 70%) koncentrowały się w przedziale niskim (400–500), niezależnie od innych parametrów. Wbrew intuicyjnym założeniom, wysoki dochód nie stanowi gwarancji otrzymania kredytu. Wykazano słabą korelację ujemną między dochodem a oceną kredytową. Rozkład dochodów okazał się bimodalny, złożony z grupy mniej i bardziej zamożnej, a liczne odrzucenia wniosków występowały również w grupie o najwyższych zarobkach, co potwierdza, że historia kredytowa jest ważniejsza niż bieżąca płynność finansowa.

Analiza kwot pożyczek ujawniła ponadto, że bank stosuje restrykcyjną politykę kredytową, co wskazuje na strategię minimalizacji ekspozycji na ryzyko. Wnioski o wysokie kwoty są częściej odrzucane, a mediana kwot zatwierdzonych jest wyraźnie niższa niż kwot wnioskowanych przez grupę odrzuconą. W aspekcie statystycznym zauważono, że większość analizowanych zmiennych (z wyjątkiem średnich wartości punktowych) nie podlega idealnemu rozkładowi normalnemu. Wykryto zjawiska leptokurtyczności (ciężkie ogony) oraz asymetrii, co sugeruje, że do modelowania ryzyka w tym zbiorze lepiej sprawdzają się rozkłady typu t-Studenta niż klasyczny rozkład Gaussa.

Istotnym elementem pracy było również zastosowanie metody Twarzy Chernoffa, która pozwoliła na szybką, wzrokową ocenę wielowymiarową. Analiza wykazała, że twarze dla danych rzeczywistych i prognostycznych zachowują zbliżone kształty,

z dominacją wartości typowych bliskich medianie. Świadczy to o stabilności struktury danych mimo występowania wartości odstających.

5.2. Odniesienie do danych prognostycznych

Porównanie danych rzeczywistych z danymi syntetycznymi pozwoliło zidentyfikować istotne różnice w strukturze obu zbiorów. Zauważono, że model prognostyczny przejawia tendencję do optymistycznego szacowania zdolności kredytowej (wyższe oceny i kwoty wnioskowane) przy konserwatywnym podejściu do dochodów. Obserwacja ta stanowi kluczowy wniosek dla dalszego rozwoju narzędzia, wskazując na precyzyjne obszary wymagające kalibracji, aby model wierniej odwzorowywał rygorystyczne mechanizmy scoringowe banku.

BIBLIOGRAFIA I ŹRÓDŁA

1. BRUCE, Peter; BRUCE, Andrew; GEDECK, Peter. *Statystyka praktyczna w data science. 50 kluczowych zagadnień w językach R i Python*. Wyd. II. Gliwice: Helion SA, 2021.
2. EDWARD, Anish Dev. *Loan Approval Dataset* [zbiór danych online]. Kaggle, 2023. Dostęp: https://www.kaggle.com/datasets/anishdevedward/loan-approval-dataset [16.12.2025].
3. EVERITT, Brian S.; SKRONDAL, Anders. *The Cambridge Dictionary of Statistics*. 4th ed. Cambridge: Cambridge University Press, 2010.
4. MONTGOMERY, Douglas C.; RUNGER, George C. *Applied Statistics and Probability for Engineers*. 7th ed. Hoboken: Wiley, 2018.
5. HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning*. 2nd ed. New York: Springer, 2009.
6. CHERNOFF, Herman. *The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association*, 1973, vol. 68(342), s. 361–368.
7. McKINNEY, Wes. *Python for Data Analysis*. 3rd ed. Sebastopol: O'Reilly Media, 2022.
8. Dokumentacja biblioteki Pandas [online]. Dostęp: https://pandas.pydata.org/docs/ [16.12.2025].
9. Dokumentacja biblioteki NumPy [online]. Dostęp: https://numpy.org/doc/ [16.12.2025].