

Daniel Cieślak

Paweł Marek

Maksymilian Sowula

Jakub Szczur

**Implementacja aplikacji do analizy
statystycznej kredytów**

**Praca dyplomowa
na studiach II-go stopnia
na kierunku Informatyka**

Prowadzący:

Dr. inż. Damian Frej

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 pozwalała na pogłębienie umiejętności analizy danych, krytycznej oceny wyników oraz prezentowania informacji w sposób zgodny ze standardami akademickimi. Projekt miał 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

CEL PRACY **4**

SPIS TREŚCI **6**

1. WSTĘP **8**

<u>1.1. Podstawowe pojęcia statystyki opisowej</u>	9
<u>1.1.1. Podstawowe pojęcia statystyki opisowej</u>	9
<u>1.1.2. Miary tendencji centralnej</u>	9
<u>1.1.3. Miary rozproszenia</u>	10
<u>1.1.4. Kwartyle i podział danych</u>	10
<u>1.1.5. Miary kształtu rozkładu</u>	10
<u>1.2. Wizualizacja danych w statystyce</u>	11
<u>1.3. Twarze Chernoffa</u>	11
<u>1.3.1. Zasada działania</u>	11
<u>1.3.2. Zastosowanie</u>	12
<u>1.4. Struktura raportu i zakres kolejnych rozdziałów</u>	12
<u>1.4.1. Metodyka badawcza (rozdział 2)</u>	12
<u>1.4.2. Analiza wyników (rozdział 3)</u>	12
<u>1.4.3. Twarze Chernoffa (rozdział 4)</u>	13
<u>1.4.4. Wnioski (rozdział 5)</u>	13
<u>1.4.4. Bibliografia</u>	13
<u>1.5. Struktura raportu i zakres kolejnych rozdziałów</u>	14

2. METODYKA BADAWCZA **14**

<u>2.1. Źródła danych</u>	15
<u>2.2. Przygotowanie zbioru danych</u>	15
<u>2.2.1. Weryfikacja kompletności danych</u>	15
<u>2.2.2. Czyszczenie danych</u>	16
<u>2.2.3. Standaryzacja i formatowanie</u>	16
<u>2.2.4. Wybór zmiennych do analizy</u>	16
<u>2.3. Narzędzia i oprogramowanie</u>	16
<u>2.3.1. Excel</u>	16
<u>2.3.2. Backend aplikacji</u>	17
<u>2.3.3. Frontend aplikacji</u>	17
<u>2.4. Opis aplikacji</u>	17

3. ANALIZA WYNIKÓW **21**

<u>3.1. Zakres analizy i zastosowane metody</u>	21
<u>3.2. Prezentacja i interpretacja wyników</u>	22
<u>3.3. Analiza podstawowych parametrów statystycznych - dane normalne</u>	22
<u>3.3. Porównanie różnic w danych progностycznych i danych normalnych</u>	24
<u>3.4. Analiza graficzna wyników</u>	27
<u>3.4.1. Analiza wykresu rozkładu dochodów według decyzji kredytowej</u>	28

<u>3.4.2. Analiza wykresu kwota pożyczki vs. ocena kredytowa</u>	<u>29</u>
<u>3.4.3. Analiza wykresu długości zatrudnienia od decyzji o zatwierdzeniu kredytu</u>	<u>31</u>
<u>3.4.4. Analiza macierzy korelacji między zmiennymi</u>	<u>32</u>
<u>3.4.5. Analiza wykresu zależności dochodu od oceny kredytowej</u>	<u>34</u>
<u>3.4.6. Analiza wykresu zależności dochodu od długości zatrudnienia</u>	<u>35</u>
<u>3.4.7. Analiza wykresu rozkładu oceny kredytowej według dochodu i decyzji kredytowej</u>	<u>36</u>
<u>3.4.8. Analiza wykresu średniego dochodu w poszczególnych miastach</u>	<u>38</u>
<u>3.4.9. Analiza wykresu kwoty kredytu w zależności od decyzji o zatwierdzeniu kredytu</u>	<u>39</u>
<u>3.4.10. Analiza rozkładu oceny kredytowej według decyzji kredytowej</u>	<u>41</u>
<u>3.4.11. Analiza histogramu dochodów i rozkład gęstości</u>	<u>42</u>
<u>3.4.12. Analiza wykresu pułkowego dochodów</u>	<u>43</u>
<u>3.4.13. Analiza empirycznej dystrybuanty dochodów</u>	<u>44</u>
<u>3.4.14. Analiza wykresu klientów w przedziałach dochodowych</u>	<u>46</u>
<u>3.4.14. Analiza wykresu klientów w przedziałach dochodowych</u>	<u>47</u>
<u>3.4.15. Analiza wykresu częstości względnej dochodów</u>	<u>48</u>
<u>3.4.16. Analiza wykresu udziału zatwierdzonych i odrzuconych kredytów</u>	<u>49</u>
<u>3.4.17. Analiza porównawcza znormalizowanych średnich cech klienta według decyzji kredytowej</u>	<u>51</u>
<u>3.4.18. Analiza wykresu radarowego średnich znormalizowanych wartości</u>	<u>52</u>
<u>3.4.19. Piramida stażu pracy klientów</u>	<u>54</u>
<u>3.4.20. Analiza wykresu liniowego wartości dochodów</u>	<u>56</u>
<u>3.4.21. Analiza wykresu porównania kurtozy dla wybranych zmiennych numerycznych</u>	<u>57</u>
<u>3.4.22. Analiza rozkładu normalnego dochodów</u>	<u>59</u>
<u>3.4.23. Analiza wykresu rozkładu t-studenta</u>	<u>60</u>
<u>3.4.24. Analiza Wykresu odległości kwartyli od średniej dla zmiennej dochód</u>	<u>61</u>
<u>4. TWARZE CHERNOFFA</u>	<u>63</u>
<u>4.1. Dane podstawowe</u>	<u>64</u>
<u>4.1.1. Ocena Kredytowa</u>	<u>64</u>
<u>4.1.2. Dochód</u>	<u>65</u>
<u>4.1.3. Kwota Pożyczki</u>	<u>66</u>
<u>4.1.4. Punkty</u>	<u>67</u>
<u>4.1.5. Lata Pracy</u>	<u>67</u>
<u>4.1.6. Wszystkie atrybuty</u>	<u>68</u>
<u>4.2. Dane progностyczne</u>	<u>69</u>
<u>4.2.1. Ocena Kredytowa</u>	<u>70</u>
<u>4.2.2. Dochód</u>	<u>71</u>
<u>4.2.3. Kwota Pożyczki</u>	<u>72</u>
<u>4.2.4. Punkty</u>	<u>73</u>
<u>4.2.5. Lata Pracy</u>	<u>74</u>
<u>4.2.6. Wszystkie atrybuty</u>	<u>75</u>
<u>4.3. Dane podstawowe oraz progностyczne</u>	<u>76</u>
<u>4.3.1. Ocena Kredytowa</u>	<u>77</u>

<u>4.3.2. Dochód</u>	<u>78</u>
<u>4.3.3. Kwota Pożyczki</u>	<u>79</u>
<u>4.3.4. Punkty</u>	<u>80</u>
<u>4.3.5. Lata Pracy</u>	<u>80</u>
<u>4.3.6. Wszystkie atrybuty</u>	<u>81</u>

5. WNIOSKI 84

<u>5.1. Podsumowanie wyników i interpretacja zależności</u>	<u>84</u>
<u>5.2. Odniesienie do danych progностycznych</u>	<u>85</u>

<u>PODZIAŁ OBOWIAZKÓW</u>	<u>85</u>
<u>BIBLIOGRAFIA I ŹRÓDŁA</u>	<u>87</u>

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. 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.

1.1.1. Podstawowe pojęcia statystyki opisowej

Zmienną nazywa się cechę, której wartości zmieniają się pomiędzy jednostkami obserwowanymi w badaniu. 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.
- 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 (leptokurytyczność) — większa koncentracja wartości wokół średniej; większa liczba wartości ekstremalnych.
- Kurtoza ujemna (platykurytyczność) — 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.

1.2. Wizualizacja danych w statystyce

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

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, 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.

1.4.4. 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”.

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” dostępnego pod adresem:

<https://www.kaggle.com/datasets/anishdevedward/loan-approval-dataset?resource=download>

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 wstępniego 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.

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 przezroczystej 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.

The screenshot shows the 'PANEL STATYSTYK KREDYTÓW' (Credit Statistics Panel). At the top, there are three tabs: 'Dane' (selected), 'Statystyka', and 'Tworzenie'. A language switcher 'Język: Polski' is also at the top right. The main area contains a table titled 'Dane' with the following columns: Miasto, Wynik kredytowy, Źródło danych, Dochód, Kwota pożyczki, Polityka zatwierdzona, Imię, Punkty, and Lata zatrudnienia. Below the table, there's a note 'Na stronie: 15' and a 'Przedawnia / Następna' button. At the bottom, there's a section titled 'Proces tworzenia danych prognozy' with a note 'Na podstawie pierwszych 3 rekordów.' and two dropdown menus: 'Parametry numeryczne' and 'Parametry kategoryczne'.

Miasto	Wynik kredytowy	Źródło danych	Dochód	Kwota pożyczki	Polityka zatwierdzona	Imię	Punkty	Lata zatrudnienia
East Jill	389	Niemieckie	113 810	39 698	Nie	Allison Hill	50	27
New Jamesside	729	Niemieckie	44 592	15 446	Nie	Brandon Hall	55	28
Lake Roberto	584	Niemieckie	33 278	11 189	Nie	Rhonda Smith	45	13
West Malariewiew	344	Niemieckie	127 196	48 823	Nie	Gabrielle Davis	50	29
Mariastad	496	Niemieckie	66 048	47 174	Nie	Valerie Gray	25	4
Port Jesseville	689	Niemieckie	62 098	19 217	Tak	Darren Roberts	65	29
Lake Joseph	373	Niemieckie	59 256	40 920	Nie	Holly Wood	35	40
Nebonside	524	Niemieckie	48 289	45 866	Nie	Nicholas Martin	25	20
Port Leslieview	367	Niemieckie	126 530	14 826	Nie	Patty Perez	55	36
Wilkensoomouth	446	Niemieckie	43 434	18 359	Nie	Emily Ross	20	8
Hurstflut	670	Niemieckie	118 696	15 373	Tak	Justin Baker	75	8
East Courtneychester	365	Niemieckie	127 080	26 216	Nie	Ann Williams	55	24
Lake Jenniferseite	573	Niemieckie	146 939	43 006	Nie	Julie King	50	21
Teresaburgh	819	Niemieckie	101 482	7973	Tak	Jeffrey Chavez	100	40
West Kathryn	843	Niemieckie	41 395	1037	Tak	Mark Lynch	80	38

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:

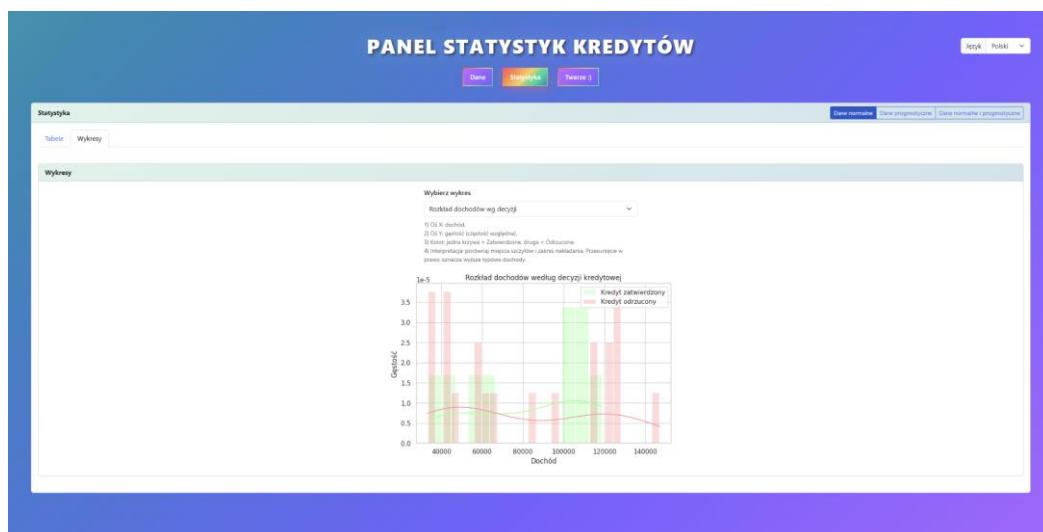
- Podsumowanie statystyczne: Zawiera podstawowe miary statystyczne (np. średnia, mediana, odchylenie standardowe, wartości minimalne i maksymalne) dla poszczególnych kolumn numerycznych.

- 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)

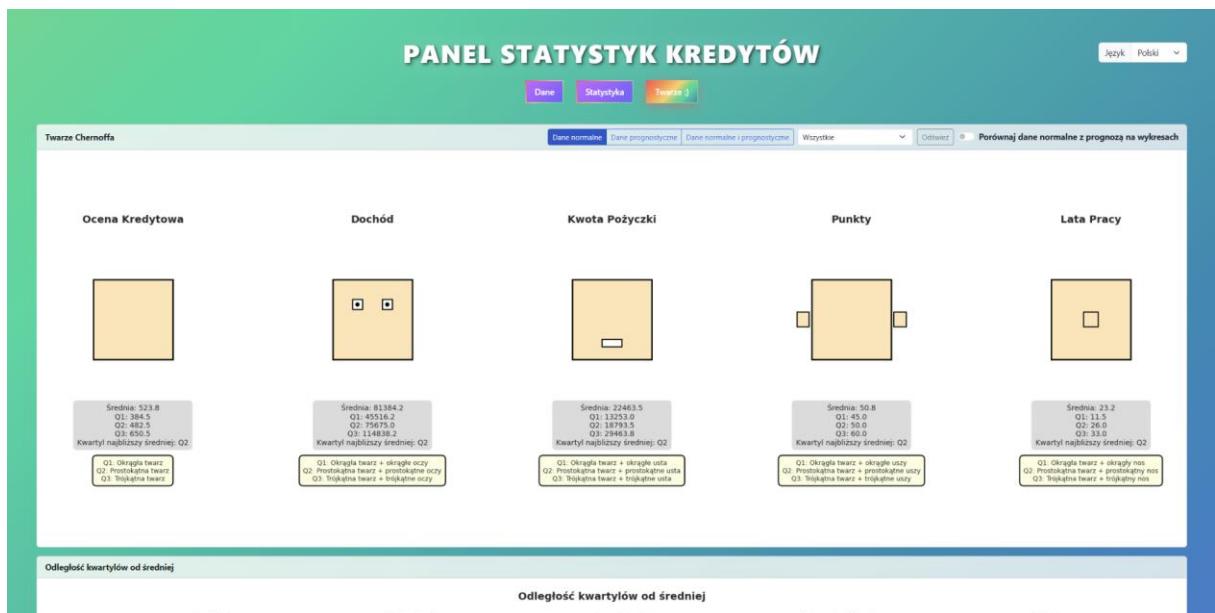
W celu ułatwienia interpretacji danych liczbowych, aplikacja oferuje moduł do wizualizacji. Zakładka "Wykresy" dynamicznie generuje graficzne reprezentacje statystyk, takie jak wykresy słupkowe czy kołowe. Użytkownik może w sposób wizualny porównać rozkłady poszczególnych cech, np. procentowy udział 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. 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 2.3. Zakładka „Wykresy”

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 znaczco 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 spiczastosci 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 tworzy skrajne przypadki w stażu pracy.

Przyjęta w projekcie predykcja danych przesywa 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

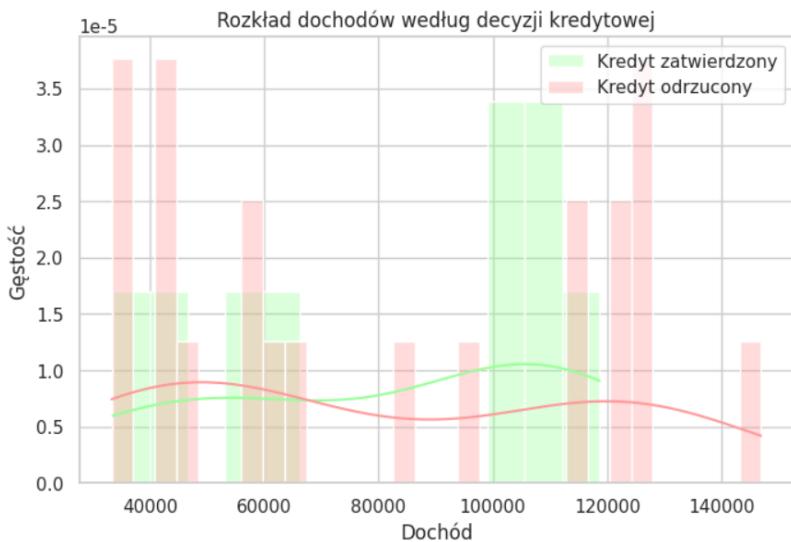
W celu pełniejszego zrozumienia struktury danych oraz różnic pomiędzy danymi rzeczywistymi a danymi syntetycznymi, wygenerowano zestaw wykresów ilustrujących najważniejsze zależności, rozkłady i porównania. Wizualizacja danych stanowi kluczowy etap analizy, ponieważ umożliwia szybkie wyciągnięcie trendów, 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 wzgledna).
- 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

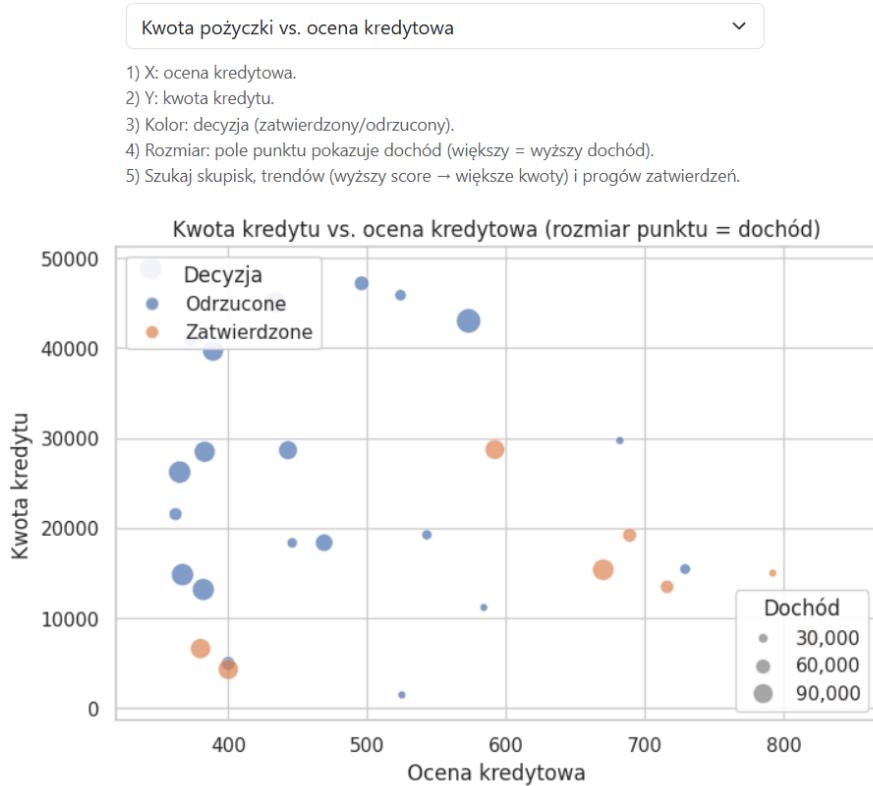
Prezentowany wykres gęstości przedstawia, jak rozkładają się dochody klientów w zależności od tego, czy ich wniosek kredytowy został zatwierdzony (zielony), czy odrzucony (czerwony). Można z niego odczytać, że klienci, którym zatwierdzono kredyt (zielony), mają tendencję do koncentrowania się w wyższych 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 kwoty 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

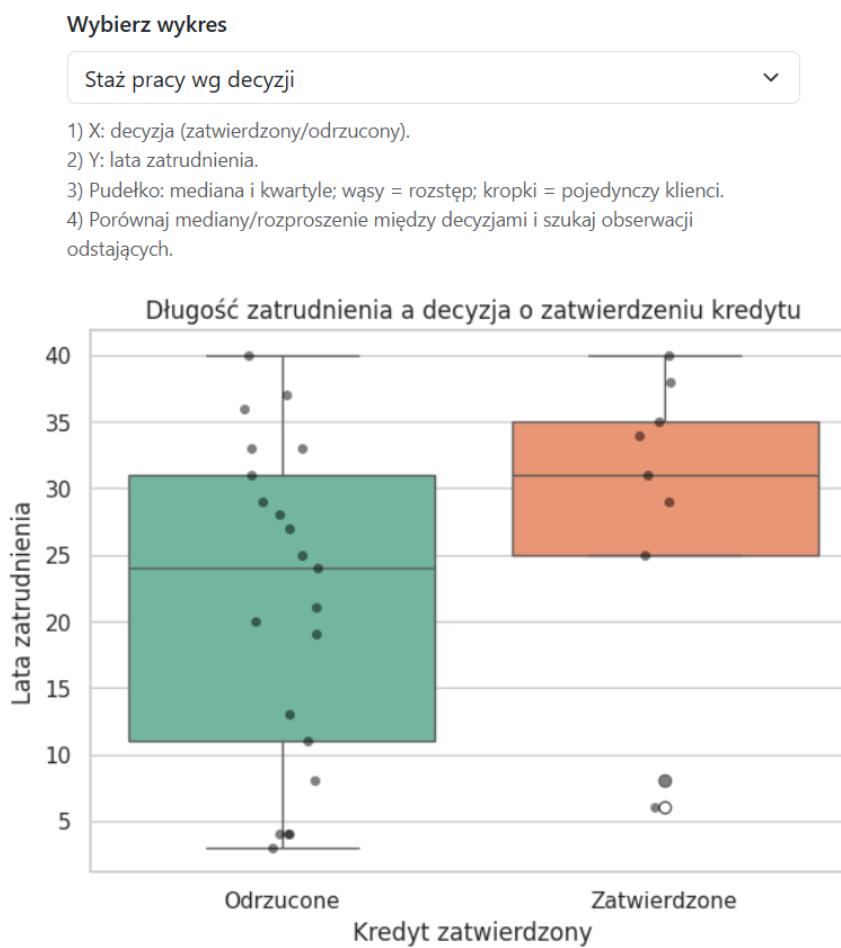
Prezentowany wykres jest diagramem rozrzutu, który ukazuje kluczowe zależności trójwymiarowe między Oceną kredytową (Oś X), Kwotą kredytu (Oś Y) a Decyzją (kolor: czerwony/niebieski), przy czym dodatkowo wielkość punktu reprezentuje dochód klienta. Wizualna analiza potwierdza silny trend: wnioski o kredyty 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

szystkie 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 wewnętrz 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 "pułapko" (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

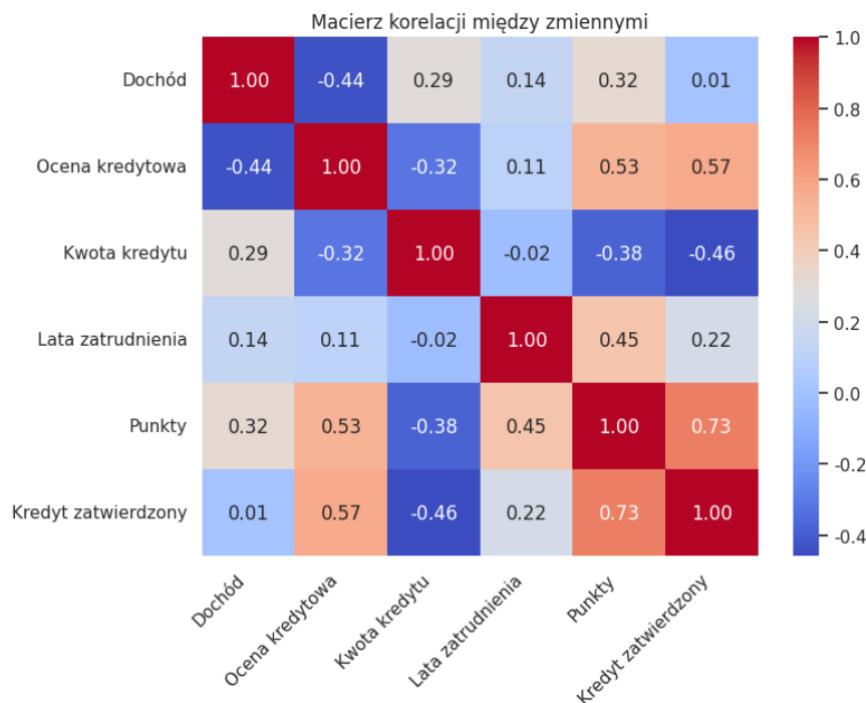
Kolejnym narzędziem wizualnym w analizie jest macierz korelacji, która pozwala ocenić siłę i kierunek zależności liniowych między wszystkimi zmiennymi numerycznymi. Analiza kolorystyczna i liczbowa korelacji ułatwia zauważenie zarówno pozytywnych, jak i negatywnych zależności, co jest istotne przy interpretacji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

Wybierz wykres

Macierz korelacji

Co pokazuje:

- 1) Każda komórka to korelacja dwóch zmiennych liczbowych.
- 2) Kolor: czerwony/niebieski = dodatnia/ujemna; ciemniejszy = silniejsza.
- 3) Liczby w komórkach: współczynnik korelacji (-1 do 1).
- 4) Szukaj silnych bloków i znaków korelacji (potencjalna wielokolinearność).



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 prediktory 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.

LITERATURA

1. Chion, M. (2019). *Audio-Vision: Sound on Screen*. Stany Zjednoczone: Columbia University Press.