**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

**Praca dyplomowa**
**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. 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 (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.

## 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.

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 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.

## 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.

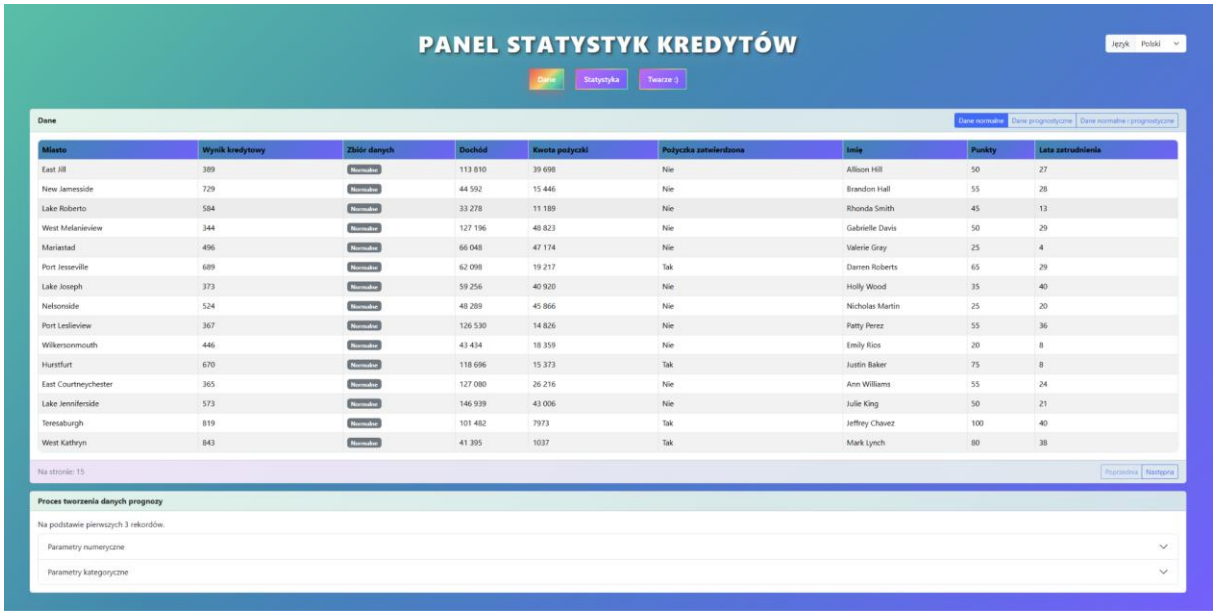
Interfejs użytkownika został zaprojektowany z myślą o prostocie i intuicyjnej nawigacji. Główne okno aplikacji podzielone jest na system zakładek, które grupują

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.



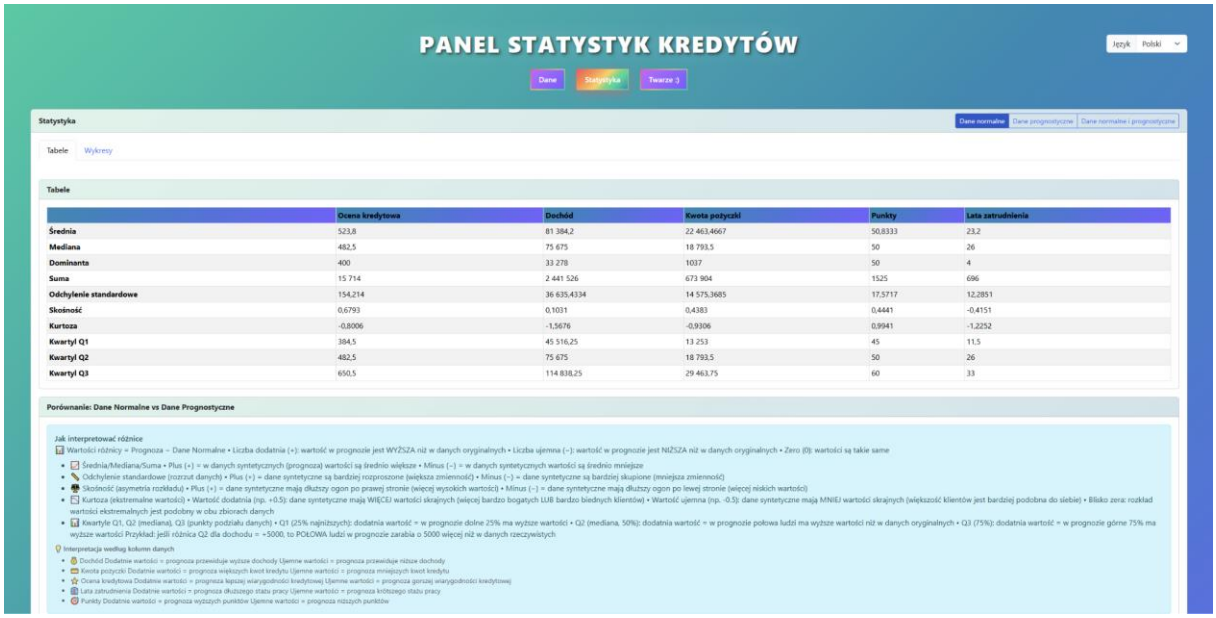

*Rysunek 2.1. Zakładka „Dane”*

### 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.

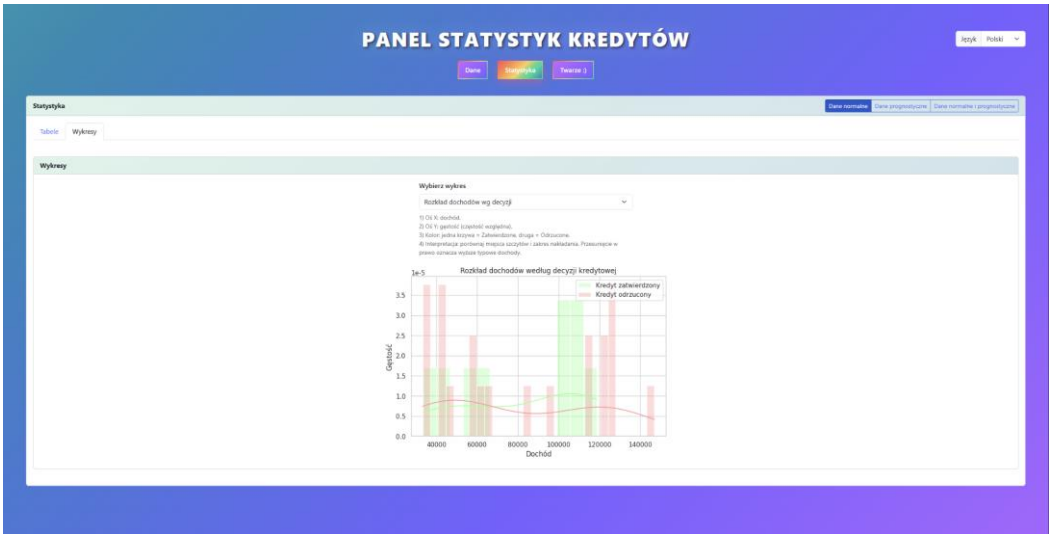
- Porównanie statystyk: Pozwala na porównanie statystyk pomiędzy dwiema grupami – na przykład osobami, które otrzymały kredyt, i tymi, którym go odmówiono. Ułatwia to identyfikację kluczowych różnic między grupami.




*Rysunek 2.2. Zakładka „Statystyki”*

## 3. Zakładka "Wykresy" (ChartsTab)

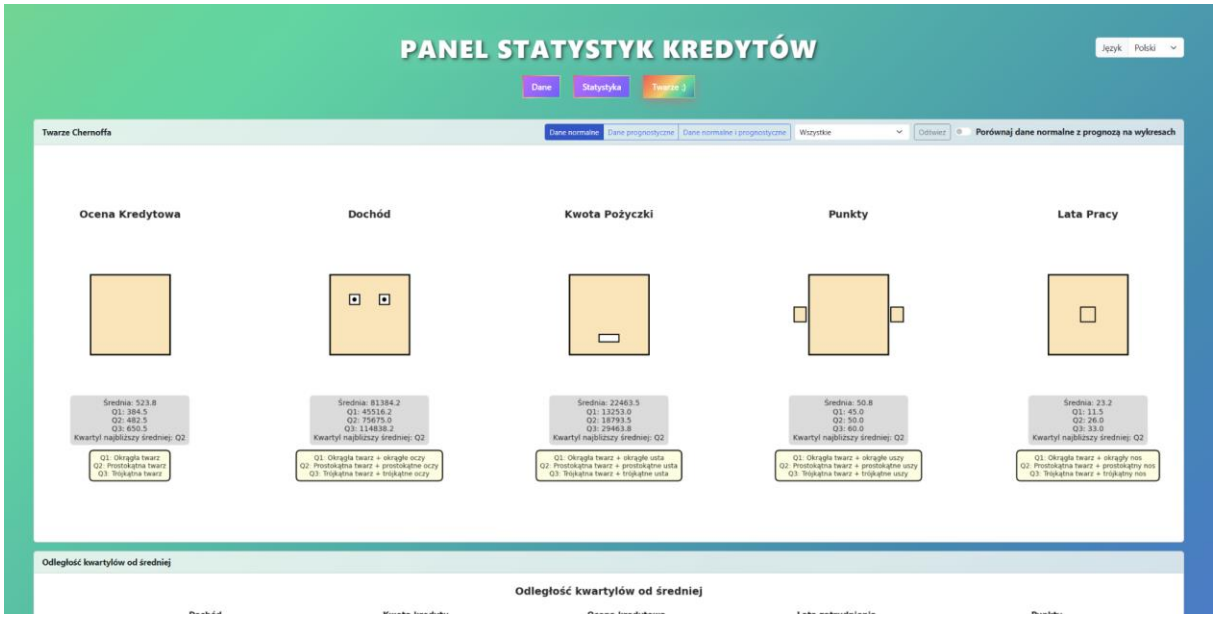
W celu ułatwienia interpretacji danych liczbowych, aplikacja oferuje moduł do wizualizacji. Zakładka "Wykresy" dynamicznie generuje graficzne reprezentacje statystyk, takie jak wykresy słupkowe czy kołowe. Użytkownik może w sposób wizualny porównać rozkłady poszczególnych cech, np. procentowy udział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

## LITERATURA

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