**Instytut Informatyki  
Wydział Nauk Ścisłych i Technicznych  
Uniwersytet Rzeszowski**

**Przedmiot:**

**Hurtownie danych**

**Dokumentacja projektu:**

**Analiza danych statystyk postów na Facebooku**

**Wykonał: Mateusz Florian**

**Prowadzący: mgr inż. Adam Szczur**

**Rzeszów 2025**

Spis treści

[1. Temat i cel projektu 3](#_Toc195194591)

[2. Techniczne aspekty projektu 3](#_Toc195194592)

[2.1. Funkcjonalności aplikacji 3](#_Toc195194593)

[2.2. Wykorzystane technologie 3](#_Toc195194594)

[2.3. Projekt GUI 3](#_Toc195194595)

[3. Wygląd i użytkowanie aplikacji 3](#_Toc195194596)

[3.1. Wymagania do uruchomienia aplikacji 3](#_Toc195194597)

[3.2. Obsługa aplikacji 3](#_Toc195194598)

[3.2.1. Wczytanie zbioru danych 4](#_Toc195194599)

[3.2.2. Funkcjonalność 2 4](#_Toc195194600)

[3.2.3. Funkcjonalność 3 4](#_Toc195194601)

[4. Eksperymenty na danych 4](#_Toc195194602)

[4.1. Wykorzystane zbiory danych 4](#_Toc195194603)

[4.2. Przebieg eksperymentu i wyniki 4](#_Toc195194604)

[4.3. Analiza uzyskanych wyników i wnioski 4](#_Toc195194605)

[5. Literatura 4](#_Toc195194606)

# Temat i cel projektu

Celem projektu jest zaprojektowanie oraz wykonanie prostej aplikacji w Pythonie, umożliwiającej analizę zbioru danych dotyczącego postów opublikowanych na stronach firmowych na Facebooku. Projekt obejmuje wczytanie danych, wykonanie podstawowych analiz statystycznych (średnia, mediana, odchylenie standardowe) oraz wizualizację korelacji między wybranymi cechami. Wyniki analizy pozwolą na ocenę, jakie cechy postów (np. typ, godzina publikacji, użycie płatnej promocji) wpływają na zaangażowanie użytkowników (liczba polubień, komentarzy, udostępnień).

# Techniczne aspekty projektu

## Funkcjonalności aplikacji

Aplikacja będzie posiadać następujące funkcjonalności:

* Wczytanie zbioru danych w formacie CSV.
* Obliczanie podstawowych miar statystycznych dla wybranych cech:
  + średnia,
  + mediana,
  + minimum i maksimum,
  + odchylenie standardowe.
* Wyznaczanie macierzy korelacji (Pearsona) pomiędzy cechami dotyczącymi zaangażowania.
* Wizualizacja danych:
  + wykresy rozrzutu (np. komentarze vs. lajki),
  + mapa korelacji (heatmapa),
  + histogramy wybranych cech.
* Generowanie prostego raportu tekstowego lub graficznego z wynikami.

## Wykorzystane technologie

Projekt zostanie wykonany przy użyciu:

* **Python 3.12** — to nowoczesna wersja popularnego języka programowania, znana z czytelnej składni i bogatego ekosystemu narzędzi, co czyni go idealnym do analizy danych i tworzenia aplikacji.

**Biblioteki Python:**

* **Pandas** — umożliwia łatwe wczytywanie, przetwarzanie i filtrowanie danych w formie tabelarycznej. Umożliwia manipulacje takimi strukturami jak DataFrame, co jest kluczowe przy analizach statystycznych i eksploracyjnych.
* **Numpy** — zapewnia wydajne operacje na macierzach i wektorach, co przyspiesza obliczenia numeryczne — idealne przy obróbce dużych zbiorów danych i analizie korelacji.
* **Matplotlib** i **Seaborn** — to biblioteki graficzne. Matplotlib umożliwia tworzenie podstawowych wykresów (słupkowych, liniowych, kołowych), a Seaborn — bardziej zaawansowanych wizualizacji z lepszą estetyką i integracją ze strukturami Pandas.
* **Scikit-learn** — to zestaw narzędzi do uczenia maszynowego, który można wykorzystać m.in. . do obliczania korelacji między zmiennymi, redukcji wymiarowości czy skalowania danych — co pomaga lepiej zrozumieć zależności i przygotować dane do dalszych analiz.

**GUI:**

* **Tkinter** — pozwala zbudować prosty, ale funkcjonalny graficzny interfejs użytkownika. Dzięki temu użytkownik może np. wybierać pliki do analizy, wyświetlać wyniki w oknach czy generować wykresy bez potrzeby korzystania z terminala.

## Projekt GUI

Obraz zawierający tekst, zrzut ekranu, numer, wyświetlacz

Zawartość wygenerowana przez AI może być niepoprawna.

Główny wygląd.

Obraz zawierający zrzut ekranu, tekst, oprogramowanie, Oprogramowanie multimedialne

Zawartość wygenerowana przez AI może być niepoprawna.

Import danych.

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, wyświetlacz

Zawartość wygenerowana przez AI może być niepoprawna.

Tryb ciemny.

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Zawartość wygenerowana przez AI może być niepoprawna.

Eksport do CSV.

Obraz zawierający tekst, zrzut ekranu, wyświetlacz, oprogramowanie

Zawartość wygenerowana przez AI może być niepoprawna.Przykładowy wygląd wyniku eksperymentu.

Obraz zawierający tekst, zrzut ekranu, Wielobarwność, wzór

Zawartość wygenerowana przez AI może być niepoprawna.

Wykres jednego wyników.

# Wygląd i użytkowanie aplikacji

## Wymagania do uruchomienia aplikacji

Aby poprawnie uruchomić i korzystać z aplikacji, niezbędne są określone wymagania sprzętowe i programowe.

**Wymagania sprzętowe:**

* Komputer z systemem operacyjnym Windows, Linux lub macOS,
* Procesor: minimum 2-rdzeniowy (rekomendowany Intel i5 lub odpowiednik AMD),
* Pamięć RAM: minimum **4 GB** (rekomendowane **8 GB**),
* Miejsce na dysku: co najmniej **200 MB** wolnego miejsca,
* Rozdzielczość ekranu: minimum **1280x720** dla poprawnego działania interfejsu GUI.

**Wymagania programowe:**

* **Python 3.12** (lub nowszy),
* Zainstalowane biblioteki:
  + pandas – do wczytywania i manipulacji danymi,
  + numpy – do operacji numerycznych,
  + matplotlib, seaborn – do wizualizacji danych,
  + scikit-learn – do analizy statystycznej, modelowania, skalowania i obliczania korelacji,
  + tkinter – do budowy graficznego interfejsu użytkownika (część standardowej biblioteki Pythona).

**Środowisko uruchomieniowe (opcjonalne, ale zalecane):**

* Python IDE, np. **PyCharm**, **Visual Studio Code** lub **Jupyter Notebook**,
* System zarządzania pakietami: pip, venv lub conda,
* Możliwość uruchamiania skryptów .py.

## Obsługa aplikacji

Aplikacja została zaprojektowana z myślą o intuicyjnej obsłudze, umożliwiającej użytkownikowi szybkie wczytanie danych, wykonanie analiz oraz wizualizację wyników za pomocą graficznego interfejsu użytkownika (GUI). Obsługa aplikacji przebiega zgodnie z poniższymi krokami:

**Krok 1: Uruchomienie aplikacji**

* Po uruchomieniu pliku main.py, otworzy się główne okno interfejsu z czytelnymi przyciskami i tabelą danych.
* Nie są wymagane żadne parametry wejściowe – aplikacja jest gotowa do pracy od razu po starcie.

**Krok 2: Wczytanie pliku z danymi**

* Kliknij przycisk **„Wczytaj dane (CSV)”** i wybierz plik z danymi (np. facebook\_metrics.csv).
* Dane zostaną załadowane i wyświetlone w tabeli na ekranie.

**Krok 3: Wstępna analiza danych**

* Użytkownik może przeprowadzić:
  + **Sprawdzenie braków danych**
  + **Wykrywanie wartości odstających**
  + **Wyświetlanie statystyk opisowych**
* Każda z tych opcji dostępna jest jako osobny przycisk w interfejsie.

**Krok 4: Filtrowanie i analiza**

* Użytkownik może filtrować dane, np. według typu posta lub zakresu dat.
* Istnieje możliwość generowania podstawowych wykresów (słupkowych, pudełkowych, korelacji) oraz uruchomienia klasyfikatora skuteczności posta.

**Krok 5: Eksport wyników**

* Wyniki analiz oraz zmodyfikowany zbiór danych można zapisać klikając **„Zapisz jako CSV”**.
* Dodatkowo, możliwe jest wygenerowanie raportów lub tabeli podsumowującej wyniki eksperymentów.

### Wczytanie zbioru danych

Obraz zawierający oprogramowanie, tekst, Oprogramowanie multimedialne, Ikona komputerowa

Zawartość wygenerowana przez AI może być niepoprawna.

### Zmiana widoku

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

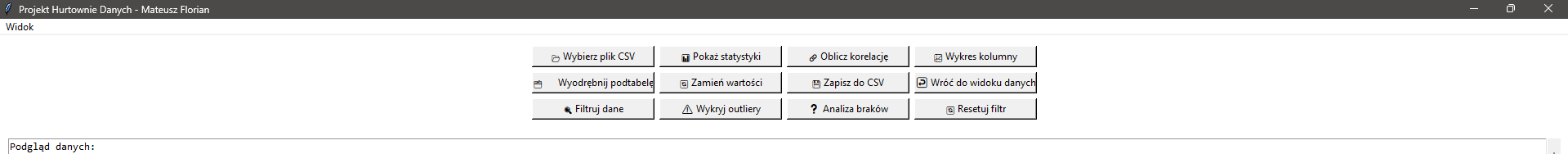
Zawartość wygenerowana przez AI może być niepoprawna.

### Zapis do pliku CSV

Obraz zawierający tekst, oprogramowanie, Oprogramowanie multimedialne, Ikona komputerowa

Zawartość wygenerowana przez AI może być niepoprawna.

### Przyciski do eksperymentów



# Eksperymenty na danych

W projekcie wykorzystano zbiór danych **"Facebook metrics"**.

**Charakterystyka zbioru:**

* **Źródło:** Zbiór danych **Facebook Metrics** pochodzi z UCI Machine Learning Repository

https://archive.ics.uci.edu/dataset/368/facebook+metrics

* **Liczba instancji:** 500 postów.
* **Liczba cech:** 19 zmiennych (ilościowe i jakościowe). 7 cech dotyczy informacji znanych przed publikacją posta (np. dzień tygodnia, godzina, typ posta), 12 cech opisuje efektywność posta (np. zasięg, liczba interakcji, kliknięcia, komentarze, udostępnienia).
* **Typy danych:** zmienne liczbowe (np. liczba lajków) oraz kategoryczne (np. typ posta, dzień tygodnia).

## Wykorzystane zbiory danych

**Pochodzenie danych:** Został opracowany przez Sérgio Moro, Paulo Ritę i Bernardo Valę i opublikowany jako część badań w *Journal of Business Research* w 2016 roku. Dane pochodzą z oficjalnej strony dużej marki kosmetycznej na Facebooku i obejmują aktywność w 2014 roku.

**Tematyka i problematyka:** Zbiór dotyczy skuteczności postów na Facebooku. Pozwala analizować, jakie czynniki wpływają na zasięg i zaangażowanie użytkowników, co jest istotne dla marketingu w mediach społecznościowych.

## Przebieg eksperymentu i wyniki

W ramach projektu przeprowadzono cztery uzupełniające się eksperymenty mające na celu zrozumienie struktury danych, wykrycie anomalii oraz identyfikację zależności między cechami opisującymi skuteczność postów na Facebooku.

**1. Analiza brakujących danych**

Sprawdzono kompletność zbioru. Spośród 9500 komórek tylko **6 zawierało braki**, co stanowi **zaledwie 0,06%**. Dane brakowały w kolumnach: *Paid*, *like* oraz *share*. Ponieważ liczba braków była znikoma, zdecydowano się je usunąć bez wpływu na jakość dalszej analizy.

**2. Wykrywanie wartości odstających (outlierów)**

Zidentyfikowano wartości odstające w 17 kolumnach, m.in. . w *Lifetime Post Total Reach*, *Total Interactions*, *like*, *share* i *comment*. Najwięcej odstających wartości wykryto w kolumnie **Lifetime Post Consumptions** (49 przypadków) i **Lifetime Post Total Reach** (69 przypadków). Część outlierów przekraczała próg z-score > 3, co może wskazywać na kampanie reklamowe lub wiralowe posty. Wyniki te są istotne przy dalszym modelowaniu — warto je uwzględnić lub odfiltrować w zależności od celu analizy.

**3. Statystyka opisowa zbioru danych**

Dla każdej cechy obliczono podstawowe statystyki (średnia, mediana, kwartyle, min/max, odchylenie standardowe). Przykładowo:

* Średnia liczba **polubień (like)**: 177.95,
* Maksymalna liczba interakcji łącznie: **6334**,
* Większość postów była publikowana w godzinach porannych (mediana = 9),
* Typ „Photo” dominował wśród postów (426 z 500 przypadków).

Statystyki te ukazują silną asymetrię niektórych zmiennych, co może mieć znaczenie przy dalszej analizie.

**4. Analiza korelacji między cechami**

Zbudowano macierz korelacji, aby zidentyfikować silne zależności między zmiennymi. Najciekawsze wnioski:

* **Total Interactions** jest silnie skorelowane z:
  + *like* (0.9977),
  + *share* (0.9286),
  + *comment* (0.8649).
* **Lifetime Engaged Users** silnie koreluje z:
  + *Lifetime Post Consumers* (0.9682),
  + *Lifetime People who liked Page and engaged* (0.8393).
* Większość cech związanych z czasem publikacji (godzina, miesiąc, dzień tygodnia) ma niewielką korelację z wynikami postów (|r| < 0.1), co sugeruje, że czynniki treściowe i budżetowe mają większy wpływ.

**Wnioski ogólne:** Eksperymenty pozwoliły dokładnie poznać strukturę danych, zidentyfikować anomalie i zależności. Wskazano na kluczowe zmienne wpływające na zaangażowanie (reakcje, udostępnienia, komentarze) i potwierdzono, że dane nadają się do dalszego modelowania predykcyjnego lub segmentacji postów.

## Analiza uzyskanych wyników i wnioski

Analiza wyników przeprowadzonych eksperymentów pozwoliła uzyskać szersze zrozumienie czynników wpływających na skuteczność postów publikowanych na Facebooku oraz jakość samego zbioru danych.

**Jakość danych**

Zbiór okazał się niemal kompletny — brakujące wartości stanowiły zaledwie **0,06%** danych, a ich rozkład nie wykazywał strukturalnych zależności. Dzięki temu możliwe było przeprowadzenie analizy bez skomplikowanego imputowania danych. Dane były czytelnie uporządkowane, a liczby cech i obserwacji — wystarczające do wyciągania sensownych wniosków statystycznych.

**Wartości odstające i rozkład danych**

Występowanie wartości odstających w wielu zmiennych (np. *Lifetime Post Total Reach*, *like*, *share*) wskazuje na obecność postów o niezwykle wysokim zasięgu i zaangażowaniu. Z jednej strony mogą to być kampanie sponsorowane, z drugiej — posty wiralowe, które skutecznie przyciągnęły uwagę odbiorców. Obecność outlierów sugeruje, że rozkład wielu cech jest silnie skośny, co potwierdzają też statystyki opisowe (np. średnie znacznie przewyższające mediany).

**Zależności między cechami**

Z analizy korelacji wyłoniły się silne zależności między wskaźnikami zaangażowania:

* *like*, *comment* i *share* są silnie dodatnio skorelowane z **Total Interactions**, co potwierdza ich wspólną naturę jako miar zaangażowania.
* *Lifetime Engaged Users* silnie koreluje z *Post Consumers* i *People who liked the Page and engaged*, co wskazuje na wspólne źródło interakcji.

Jednocześnie cechy związane z czasem publikacji (dzień tygodnia, godzina, miesiąc) miały niewielki wpływ na efektywność posta (|r| < 0.1), co sugeruje, że **kluczowe znaczenie mają czynniki treściowe** (typ posta) oraz zasięg promocyjny.

**Praktyczne wnioski**

* Posty graficzne (np. zdjęcia) pojawiały się najczęściej, co warto było uwzględnić w analizach jako dominującą kategorię.
* Płatne kampanie cechowały się wyższym zasięgiem, co potwierdzają korelacje zmiennej *Paid* z zasięgami i interakcjami.
* Duże rozproszenie danych sugeruje, że warto rozważyć **grupowanie postów** (np. skuteczność niska/średnia/wysoka) lub dalsze modelowanie za pomocą metod klasyfikacyjnych.

# Literatura

 Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, 69(9), 3341–3351. — **Publikacja źródłowa dotycząca zbioru Facebook Metrics.**

 Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. — **Opis biblioteki Scikit-learn, wykorzystywanej do budowy modelu klasyfikacyjnego.**

 McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 56–61. — **Źródło dokumentujące bibliotekę pandas.**

 Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. — **Dotyczące wizualizacji danych.**

 Seaborn documentation: https://seaborn.pydata.org/ — **Oficjalna dokumentacja biblioteki Seaborn.**

 Tkinter documentation (Python standard library): https://docs.python.org/3/library/tkinter.html — **Źródło opisujące narzędzia używane do stworzenia GUI.**

 UCI Machine Learning Repository — Facebook Metrics Dataset: https://archive.ics.uci.edu/ml/datasets/Facebook+metrics — **Link do zbioru danych.**