**2 Metodologia**

**2.1 Eksploracyjna analiza danych (EDA)**

* Definicja EDA, wprowadzenie, dlaczego robimy EDA?
* Metody EDA: wizualizacja danych, czyszczenie danych, walidacja poprawności i niezawodności danych, identyfikacja i traktowania wartości skrajnych oraz brakujących, statystyka opisowa, transformacje danych i przygotowywanie do modelowania (np. normalizacja albo feature engineering)
* Może o pułapkach wizualizacji danych, albo o spurious correlations?
  + Np. oś Y na wykresie kolumnowym, wyciąganie wniosków przyczynowo-skutkowych z analizy eksploracyjnej, korelacji itd.

Eksploracyjna analiza danych (EDA) to podejście do analizy danych zakładające istnienie gotowego zestawu danych wtórnych, który może zostać wykorzystany w projektach naukowych, biznesowych lub personalnych. Celem EDA jest stopniowe pogłębianie zrozumienia zbioru danych poprzez zastosowanie odpowiednich metod analitycznych. Stanowi ona kluczowy, początkowy etap procesu analizy danych. W ramach eksploracyjnej analizy danych stosuje się m.in. wizualizację danych, czyszczenie zbioru (np. identyfikację wartości brakujących, zduplikowanych, skrajnych lub nieprawidłowych), generowanie statystyk opisowych dla zmiennych oraz transformacje danych w celu ich lepszego dostosowania do późniejszych metod modelowania statystycznego. EDA nie tylko dostarcza informacji na temat kolejnych kroków w analizie, lecz także pozwala na identyfikację możliwości oraz ograniczeń związanych z danym zbiorem. Do eksploracyjnej analizy danych można również zaliczyć grupowania obserwacji, takie jak algorytm K-średnich, który klasyfikuje podobne do siebie jednostki w odrębne grupy.

Wizualizacja danych stanowi podstawę eksploracyjnej analizy danych. Pozwala na identyfikowanie kształtów rozkładów zmiennych oraz wstępne zrozumienie związków pomiędzy zmiennymi. Wizualizacja danych jest istotna, ponieważ wykorzystuje ludzką zdolność do identyfikacji skomplikowanych wzorców w formach graficznych. Istotnym zadaniem jest odpowiedni dobór typu wykresu do zmiennej, lub zmiennych które są poddawane wizualizacji. Wykresy powinny być czytelne, zrozumiałe ale jednocześnie przenoszące jak najwięcej informacji. Równie ważny jest dobór wizualizacji do grupy docelowych odbiorców.

Czyszczenie danych jest kolejnym podstawowym zadaniem EDA. Większość zestawów danych generowanych współcześnie zawiera brakujące wartości, błędy, wartości skrajne. Na szczęście w dobie komputerów o wysokich mocach obliczeniowych, sama identyfikacja wartości brakujących nie jest problemem (ciężej jest definitywnie zidentyfikować błędy lub wartości skrajne). Bardziej istotna jest decyzja o traktowaniu tych wartości. Jeżeli jest ich na tyle mało, że usunięcie wierszy zawierających braki nie spowoduje dużej straty informacji, to taka decyzja może mieć sens. Istnieją jednak, lepsze, choć nieperfekcyjne metody na imputację wartości brakujących.

… Opis metod imputacji …

Pomimo tego, że identyfikacja wartości skrajnych jest bardziej zaawansowanym zadaniem istnieje wiele metod na ich identyfikację.

Wizualizacja danych może dawać pierwsze sygnały w kontekście identyfikacji wartości skrajnych. Zaobserwowanie stosunkowo mało licznych przedziałów w ogonach rozkładu zmiennej wizualizowanej przy pomocy histogramu może sygnalizować obecność wartości skrajnych. Wykresy pudełkowe dają szybki obraz identyfikowanych wartości skrajnych. Wykresy rozrzutu pomagają w identyfikacji wartości skrajnych w kontekście dwóch zmiennych.

Jedną z najprostszych formalnych metod identyfikacji wartości skrajnych jest metoda rozstępu międzykwartylowego. Wartości skrajne oznaczane na wykresach pudełkowych są oznaczane przy pomocy tej metody. Obserwacja zostanie uznana za skrajną, jeżeli wartość badanej cechy znajduje się w odległości 1,5 razy rozstęp międzykwartylowy (IQR) poniżej pierwszego kwartyla (Q1) lub powyżej trzeciego kwartyla (Q3). Matematycznie granice wykrywania wartości skrajnych można wyrazić jako:

Istotnym jest zdeterminowanie czy zidentyfikowana wartość skrajna jest wynikiem błędu (np. błędnego wprowadzenia przez człowieka lub wadliwego odczytu sensora), czy też stanowi rzeczywistą wartość, opisującą obserwację znacząco odbiegającą od pozostałych w zbiorze danych. Zdrowy rozsądek oraz wiedza ekspertów danej dziedziny pomaga w identyfikacji wartości skrajnych powstałych w wyniku błędu. Prostym przykładem może być zbiór danych dotyczący wzrostu ludzi, w którym znajduje się obserwacja o wartości -150 centymetrów. Wzrost jest zmienną ograniczoną do wartości nieujemnych, więc logicznym wnioskiem jest uznanie tej obserwacji za błędną. W przypadku zidentyfikowania wartości błędnej rozsądnym jest usunięcie całego wiersza obserwacji, lub usunięcie wartości jedynie błędnej zmiennej i użycie odpowiedniej metody imputacji. Z kolei w sytuacji, gdy wartość skrajna nie jest wynikiem błędu, lecz rzeczywistą cechą danej obserwacji, jej usunięcie może prowadzić do utraty istotnej informacji. Atrakcyjnością usunięcia takiej wartości jest potencjalne lepsze przygotowanie danych do danego modelu, który został wybrany jako następny krok analizy. Jednakże, celem eksploracji danych jest między innymi właśnie poprawienie świadomości dotyczącej charakterystyki danego zbioru danych oraz wybór następnych kroków w analizie. Dlatego zamiast ignorować wartości skrajne, bardziej zasadnym podejściem jest transparentne uwzględnienie ich w analizie oraz jawne zakodowanie założenia występowania wartości skrajnych poprzez odpowiedni wybór narzędzi statystycznych, które uwzględniają ich wpływ na estymację i niepewność predykcji.

Jeżeli wcześniejsza eksploracja danych wykazała, że w zbiorze mogą występować podgrupy, to przydatne mogą się okazać metody nienadzorowanej klasyfikacji, których celem jest przypisanie każdej obserwacji do grupy, w której znajdują się obserwacje jak najbardziej podobne do siebie. Każda grupa powinna być jak najbardziej jednorodna, a jednocześnie jak najbardziej różna od pozostałych. W zależności od zastosowanej metody, liczba grup może być determinowana w trakcie procesu uczenia modelu lub ustalana przez badacza. Liczba grup, którą ustala badacz, powinna być informowana rezultatami wcześniejszej analizy eksploracyjnej. Atrakcyjność tych metod polega na możliwości tworzenia profili grup, w ramach których badacz analizuje każdą grupę osobno i opisuje jej cechy. Dodatkowo, grupowanie otwiera nowe możliwości dalszej analizy, na przykład umożliwiając tworzenie osobnych modeli dla każdej grupy lub zastosowanie modeli hierarchicznych.

Ważnym aspektem eksploracyjnych metod analizy danych jest uwzględnienie ich ograniczeń. Podczas EDA poszukujemy hipotez, jednak wnioski wyciągane na tym etapie są mocno ograniczone, a łatwość wyciągania błędnych wniosków stanowi jedną z głównych pułapek tego procesu. Jednym z takich błędów jest nadinterpretacja korelacji i fałszywe przypisywanie przyczynowości. Kolejnym zagrożeniem jest niewłaściwy dobór narzędzi statystyki opisowej do danej zmiennej. Innym przykładem błędu może być niewłaściwy dobór wykresu lub niepoprawne ustalenie skali na osiach wykresu, co może prowadzić do poważnych nadużyć w interpretacji danych. Wnioski o wyższej wadze wyciągane są dopiero na późniejszych etapach procesu analizy danych. EDA pełni rolę informacyjną, wskazując potencjalne hipotezy lub obszary, które wymagają dalszej analizy i weryfikacji. Na etapie EDA badacz wykonuje prace detektywistyczną na danych, poszukując wskazówek, które pozwolą nadać sens danym, zrozumieć je oraz sformułować kolejne hipotezy.

**2.2 Analiza przyczynowa**

* Czym jest i dlaczego analiza przyczynowa jest istotna?
* Podejście statystyki bayesowskiej w analizie przyczynowej, programowanie probabilistyczne
  + Zalety, ukazywanie pełnej niepewności w estymatach i parametrach
  + Mozliwosc transpartnego enkodowania założeń do modelu (Przykład heteroskedastycznosci i sigmy jako funkcji liniowej zmiennej x)
* Istotność transparentności przyjmowanych założeń oraz konkretności pytań badawczych
* DAGi, Backdoor Criterion, identyfikacja confounding variables i wybór adjustment set
* Generatywne symulacje danych syntetycznych w celu odzyskania ustalonych parametrów symulacji w modelu statystycznym.
* Backdoo

**2.3 Analiza predykcyjna i uczenie maszynowe**

* Zalety metod uczenia maszynowego (szybkość generowania predykcji, trenowania modeli, wysoka dokładność, umożliwienie przez szybkie i dostępne komputery)
* Klasyfikacja multiklasowa, metody

**3 Analiza Danych**

Analiza własna została przeprowadzona w trzech odrębnych, ale powiązanych i równie istotnych etapach, które odpowiadają strukturze rozdziału drugiego. W pierwszej sekcji wykorzystane zostaną eksploracyjne metody analizy danych w celu dogłębnego zapoznania się z zestawem danych i jego lepszego zrozumienia. Druga sekcja będzie poświęcona analizie przyczynowo-skutkowej i spróbuje odpowiedzieć na pytanie dotyczące relacji przyczynowych pomiędzy różnymi zmiennymi. Zostanie zaprezentowany graf przyczynowy, ilustrujący założenia dotyczące relacji przyczynowych. Następnie przeprowadzona zostanie generatywna symulacja, której rezultatem będzie syntetyczny zbiór danych, wykorzystany do oceny gotowości modelu przyczynowego do pracy na rzeczywistych danych. Trzecia sekcja będzie próbowała odpowiedzieć na pytanie dotyczące gotowości wykorzystywania urządzeń ubieralnych do dokładnego przewidywania poziomu stresu użytkowników w czasie rzeczywistym. Zostaną opracowane modele uczenia maszynowego, które będą przewidywać odczuwany poziom stresu na podstawie innych zmiennych.

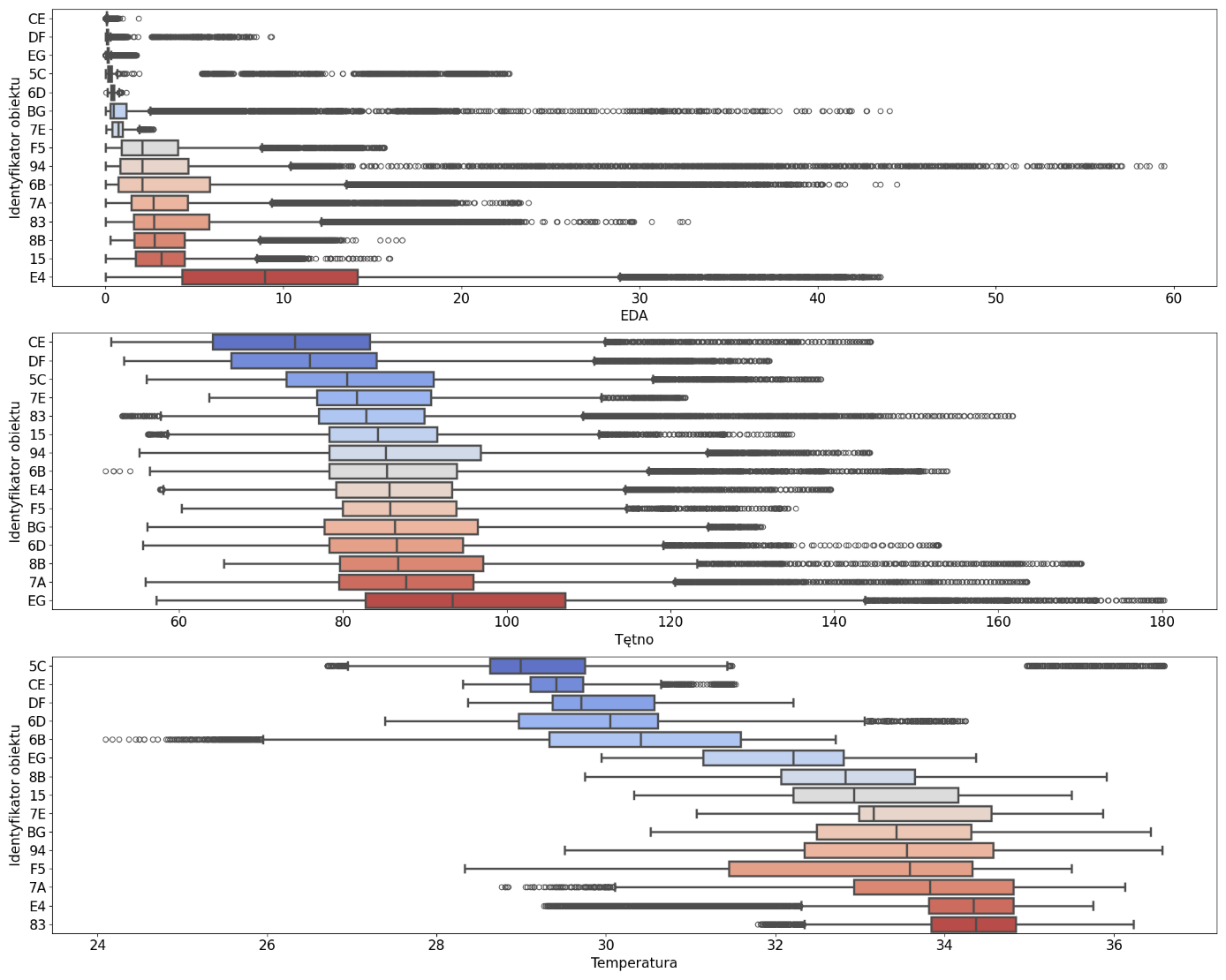
Zbiór danych obejmuje dane z urządzeń ubieralnych noszonych przez piętnaście pielęgniarek, które monitorowały m.in. tętno, reakcję skórno-galwaniczną (*en. Electrodermal activity, EDA*) oraz temperaturę skóry.

**3.1 Eksploracyjna analiza danych (EDA)**

Podczas eksploracji danych duży nacisk położono na procesowanie danych przygotowujących je do dalszej analizy. Dokonano resamplingu danych do interwałów czasowych wynoszących 1 sekundę, co pozwoliło na znaczną redukcję rozmiaru zbioru danych z pierwotnych 11,5 miliona wierszy do 356 561 wierszy.

W kolejnym kroku zbiór danych został sprawdzony pod kątem brakujących wartości. Stwierdzono, że w zbiorze danych nie występują brakujące wartości.

Następnie skupiono się na wizualizacji danych. Stworzono wykresy pudełkowe, macierz korelacji oraz wykresy rozrzutu dla cech ciągłych zbioru danych.



Wykres obrazuje znaczną zmienność wartości pomiarowych pomiędzy poszczególnymi uczestnikami badania. W szczególności obiekt E4 charakteryzuje się wysokimi wartościami miar położenia dla zmiennych związanych z aktywnością elektrodermalną (EDA). Ponadto zaobserwowano znaczną różnorodność wartości EDA na poziomie indywidualnych uczestników.

Podobne zróżnicowanie dotyczyło zmiennej opisującej tętno. Średnia arytmetyczna tętna dla uczestnika z najniższymi odczytami wynosiła 75,47 uderzeń na minutę, natomiast dla osoby z najwyższą średnią – 84,68 uderzeń na minutę.

Temperatura skóry również różnicowała badanych. Można zaobserwować dwie główne grupy obiektów, w kontekście miar położenia temperatury. Pierwsza z nich zawiera niższe wartości, w okolicach 29 stopni Celsjusza, a druga wyższe, około 33 stopni. Szczególną uwagę zwracają wartości też wartości zmiennej oscylujące wokół 25°C. Należy jednak pamiętać, że odczyt sensorów różni się od tradycyjnych metod pomiaru temperatury ciała. Co więcej, urządzenia ubieralne mierzyły temperaturę skóry na nadgarstku, który jest bardziej podatny na wychłodzenie niż wnętrze organizmu.

Uwagę zwraca również duża liczba wartości skrajnych, zidentyfikowanych na podstawie rozstępu międzykwartylowego, we wszystkich badanych cechach. Na tym etapie analizy nie można jednak jednoznacznie stwierdzić, że są to błędne odczyty - wszystkie wartości mieszczą się w przewidywalnych zakresach zmienności.

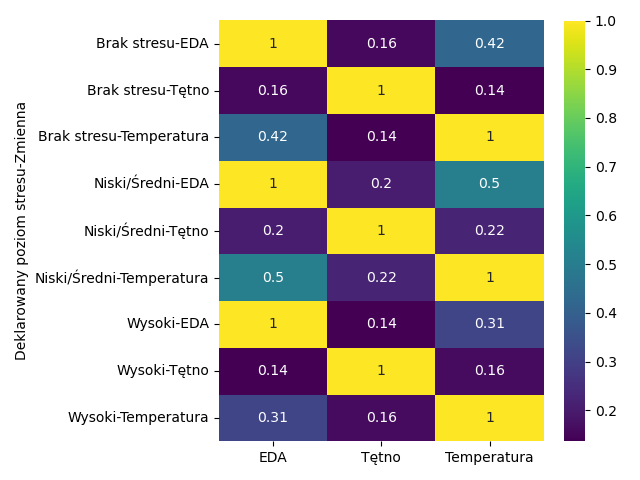
Następnie stworzono macierz korelacji, która dostarcza informacji o współzależności zmiennych.

A yellow and purple squares with white text

AI-generated content may be incorrect.

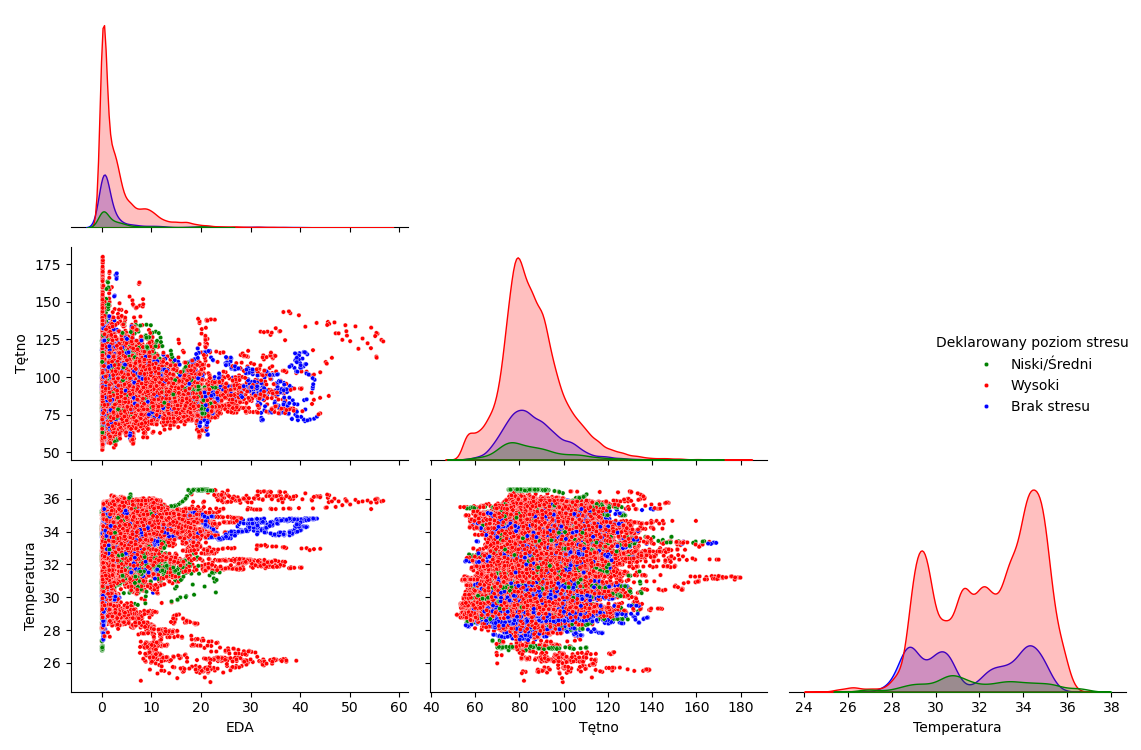
Analiza korelacji wskazuje, że najsilniejszy związek występuje pomiędzy temperaturą skóry a aktywnością elektrodermalną (r = 0,35). Oznacza to umiarkowaną dodatnią korelację – wyższe wartości EDA są związane z wyższą temperaturą skóry. Z kolei korelacja pomiędzy tętnem a temperaturą skóry jest słabsza (r = 0,16), ale nadal dodatnia, co sugeruje, że wyższa temperatura skóry może być nieznacznie związana z wyższym tętnem. Najmniejsza współzależność dodatnia występuje pomiędzy EDA a tętnem (r = 0,14).

Siła związku zmienia się po stratyfikacji ze względu na deklarowany poziom stresu.



Przykładowo, wśród badanych deklarujących niski/średni poziom stresu korelacja pomiędzy temperaturą skóry a EDA jest silniejsza (r = 0.504).

W kolejnym kroku skonstruowano wykresy rozrzutu dla każdej pary zmiennych ilościowych. Dodatkowo obserwowany był rozkład każdej cechy przy użyciu wykresu jądrowego estymatora gęstości, czyli gładkiego histogramu. Cała wizualizacja jest stratyfikowana ze względu na poziom stresu



Pierwszym wnioskiem z wykresu, jest zauważalna nierównowaga w liczbie obserwacji w poszczególnych kategoriach stresu. Partycypanci najwięcej czasu spędzali odczuwając wysoki poziomu stresu.

Widoczna jest duża asymetria prawostronna w rozkładzie zmiennej opisującej aktywność elektrodermalną. Rozkład tętna jest najbardziej zbliżony do rozkładu normalnego. W przypadku temperatury skóry, ponownie można zaobserwować dwie grupy obiektów.

Eksploracyjna analiza danych dostarczyła cennych informacji o związkach pomiędzy zmiennymi, kształtach rozkładów zmiennych, różnic fizjologicznych pomiędzy poszczególnymi obiektami oraz dostarczyła wskazówki dotyczące kierunku dalszej analizy.

**3.2 Wpływ deklarowanego poziomu stresu na zmienne fizjologiczne**

W pierwszej kolejności uwagę skupiono na zmianach i różnicach w zmierzonych zmiennych fizjogicznych wśród partycypantów. Konkretnie, usiłowano odpowiedzieć na pytanie badawcze: Czy istnieje, i jaka jest siła efektu deklarowanego stresu na zmienne fizjologiczne. Zmienne fizjologiczne istotnie różnią się w zależności od obiektu, więc skupiono się na hierarchicznych model liniowych, dla trzech zmiennych objaśnianych: Temperatura skóry, tętno oraz aktywność elektrodermalna.