Klasteryzacja uczestników eksperymentu z Centrum Nauki Kopernik

Raport z projektu

Michał Pastuszka, Karol Pysiak, Dominik Rafacz

16 czerwca 2019

Spis treści

1 Wstęp

W projekcie zajmowaliśmy się wynikami badań na temat emocji przeprowadzonych w *Centrum Nauki Kopernik*, w których wzięło udział 245 osób. Procedura była komputerowa i składała się z 3 części:

1. Ocena sposobu w jaki uczestnik myśli ogólnie o emocjach; np. Czy „w środku" istnieje jakiś jeden, ukryty wspólny mechanizm, który powoduje, że odczuwamy tą emocję? [ISTOTA]; Czy sposób odczuwania tych emocji zmienił się znacząco na przestrzeni wieków? [STABIL]; Czy jeśli wiemy, że ktoś odczuwa tą emocję, to wiemy dokładnie co ta osoba odczuwa / Mało wiemy na temat tego co ta osoba odczuwa? [INFORM]
2. Trening oceny ekspresji mimicznej - trening przygotowujący do wykonania części 3.
3. Ocena ekspresji mimicznej prezentowanej na zdjęciach. Każdy uczestnik ogląda 36 zdjęć. Przy każdym zdjęciu dokonuje dwóch ocen: w jakim stopniu zdjęcie wyraża smutek/radość oraz ocenia poziom zaufania do przedstawionej osoby.

Celem badania było zbadanie relacji między wartościami parametrów z pierwszej części badania a innymi zmiennymi (np. oceną zaufania zdjęć).

My jako cel postawiliśmy sobie kalsteryzację danych. Chcemy móc pogrupować osoby biorące udział w badaniu na podstawie zmiennych, które je opisują i sprawdzić, czy występują wyraźne podziały. Do dyspozycji posiadamy informacje z pierwszej i trzeciej części badania. Dokonaliśmy analizy powyższych danych, a następnie dwyszkoliliśmy modele grupujące dane w klastry.

2 Eksploracyjna analiza danych

Do dyspozycji mieliśmy dwa zbiory danych – plik *Do analizy czynnikowej esencjalizm.xslx* (nazywany przez nas dalej *esence*), zawierający wyniki badania ankietowego na temat postrzegania emocji przez uczestników oraz plik *Wyniki_CNK_Analiza.xslx* (nazywany dalej *photo*) zawierający odpowiedzi udzielone przez uczestników badania co do emocji widzocznych na zdjęciach.

2.1 Zbiór danych z ankiety

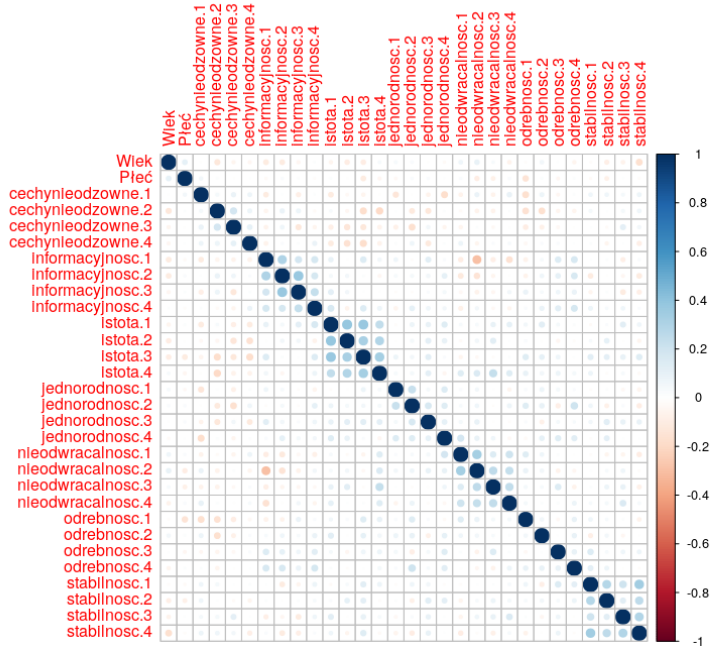
Zbiór danych z ankiety zawierał odpowiedzi na pytania dotyczące postrzegania emocji przez uczestnika. Każdy wiersz zawierał odpowiedzi jednego uczestnika na 28 pytań, po cztery z siedmiu grup:

1. Cechy nieodzowne - czy dana emocja ma swoje konieczne cechy, które ją określają?
2. Informacyjność - jak dużo mówi nam to, że ktoś odczuwa daną emocję?
3. Istota - czy istnieje jeden wspólny mechanizm, który powoduje, że odczuwamy daną emocję?
4. Jednorodność - czy różne przypadki danej emocji mają dużo wspólnych cech?
5. Nieodwracalność - czy dana emocja może się nagle zamienić w inną?
6. Odrębność - czy łatwo jest rozpoznać daną emocję?
7. Stabilność - czy sposób oczuwania tej emocji zmienił się znacząco na przestrzeni wieków?

Pytania w każdej grupie odnosiły się do czterech, losowo wybranych emocji z grupy: gniew, odraza, strach, szczęście, zazdrość, miłość, duma, smutek, wstyd, zaskoczenie. Były one różne dla różnych uczestników, i nie mamy informacji o tym, o jakich konkretnie dotyczyły odpowiedzi w zbiorze. Odpowiedzią na każde z pytań była liczba od 0 do 10, określająca, jak bardzo uczestnik zgadzał się z jednym z dwóch przeciwstawnych stwierdzeń na temat tego pytania.

2.1.1 Korelacja zmiennych

Ze względu na podział pytań, które miały zbadać postrzeganie emocji przez uczestnika na grupy, możemy się spodziewać korelacji zmiennych w posczególnych grupach. Z wykresu 1 widać dużą korelację w pytaniach o istotę, stabilność, informacyjność i nieodwracalność. Nie widać jej natomiast w pytaniach o odrębność i cechy nieodzowne, co sugeruje, że odpowiedzi na te pytania zależały od konkretnych emocji, o które pytano.



Rysunek 1: Wykres korelacji odpowiedzi na pytania z ankiety

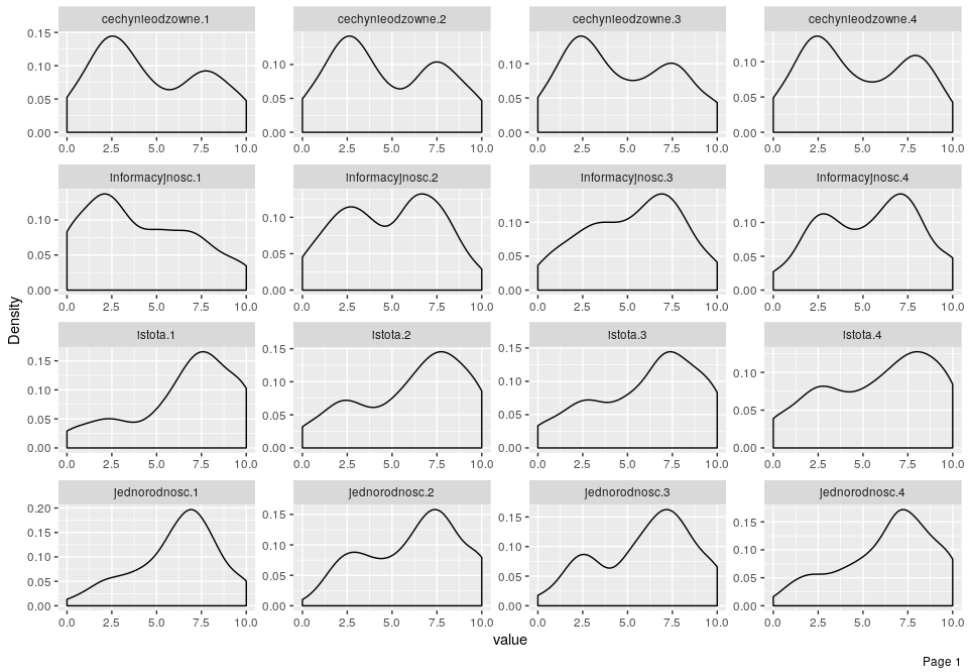
2.1.2 Rozkład odpowiedzi

Patrząc na wykresy rozkładów odpowiedzi na poszczególne pytania 2, 3, możemy zauważyć kilka ciekawych zależności. Po pierwsze dla wszystkich pytań widać mało odpowiedzi blisko wartośći 5. Po drugie na odpowiedzi na część pytań, jak na przykład stabilność, rozkładają się blisko obu końców. W pozostałych wyraźnie dominuje jeden z nich, jak to ma miejsce w przypadku pytania o odrębność.

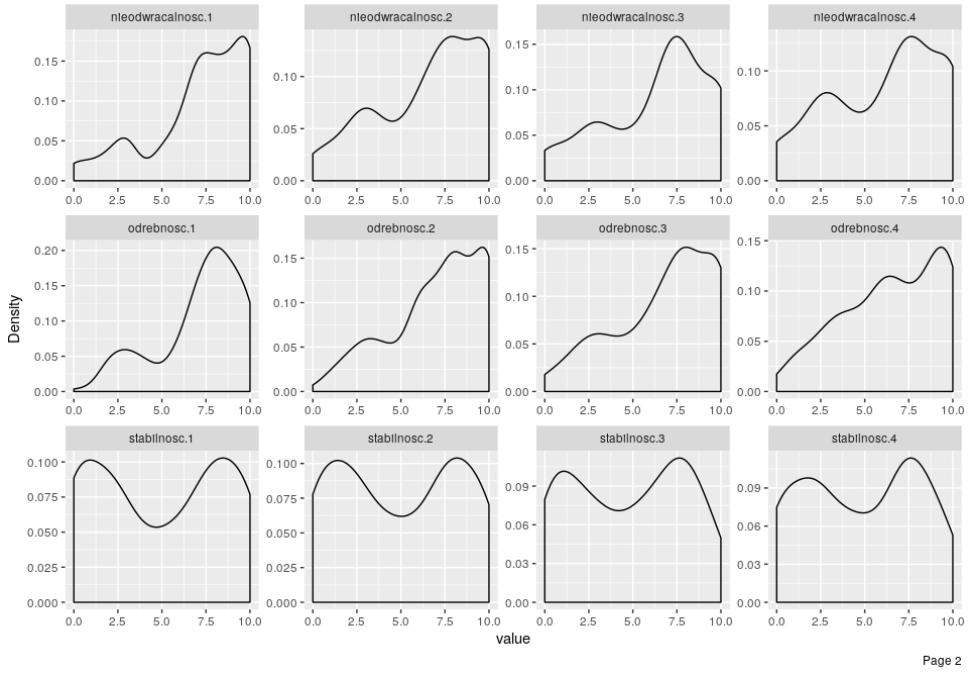
2.1.3 Redukcja wymiarów

Ze względu na powiązanie pytań ze sobą, można się spodziewać, że zbiór jest podatny na redukcję wymiarów. Z wykresu 4 widać, że dużą część wariancji przedstawiają cztery składowe główne.

Próbowaliśmy wykorzystać algorytm t-SNE, aby wykryć, czy w zbiorze istnieją naturalne klastry. Z rysunku 5 widać jednak, że dane rozkładają się równomiernie.



Rysunek 2: Rozkłady odpowiedzi na pytania o istotę emocji 1/2

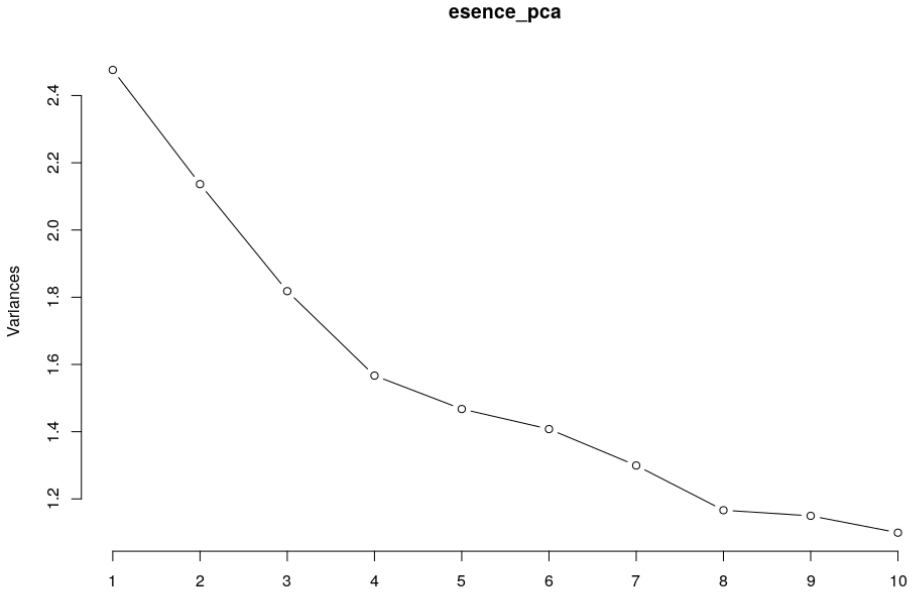


Rysunek 3: Rozkłady odpowiedzi na pytania o istotę emocji 2/2

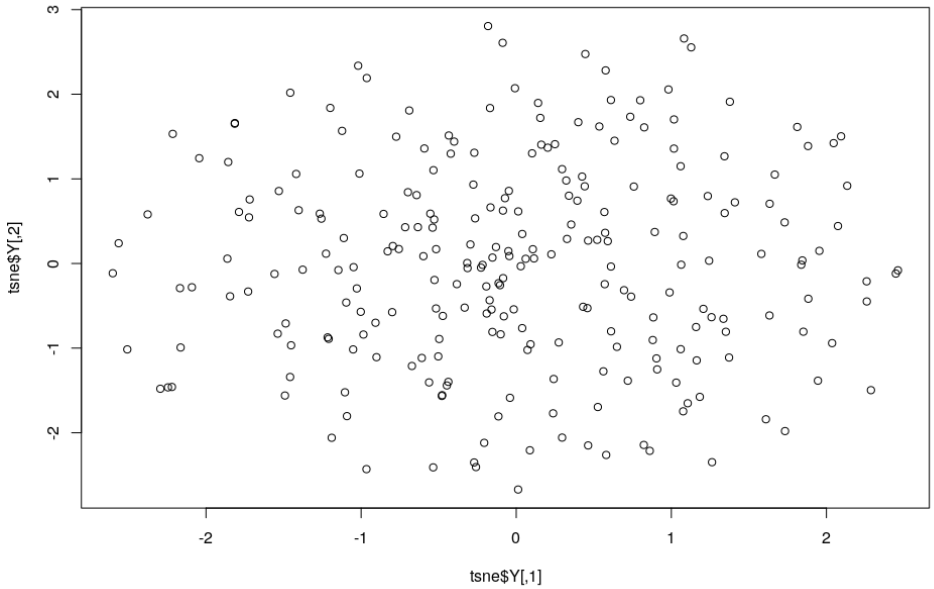
2.2 Zbiór danych z eksperymentu

Nasz zbiór danych z eksperymentu zawierał odpowiedzi na pytania do zdjęć. Uczestnicy testu mieli odpowiedzieć na pytania:

1. Jaką emocję wyraża przedstawiana twarz - należało ocenić w skali od 0 do 10 jaką emocję wyraża osoba na zdjęciu (0 – najsmutniejsza, 10 – najszczęśliwsza)



Rysunek 4: Wariancja wyjaśniana przez składowe główne



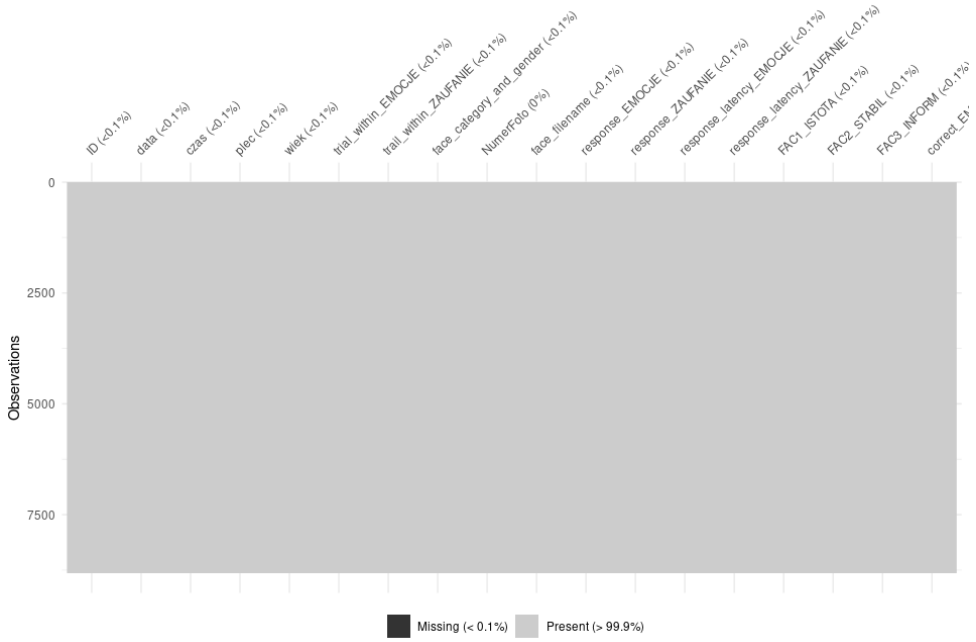
Rysunek 5: Wizualizacja algorytmem t-SNE

2. Jak dużym zaufaniem darzysz osobę przedstawiającą tą emocję - należało ocenić w sali od 0 do 100 jak dużym zaufaniem darzymy osobę, która widzimy na zdjęciu, oczywiście biorąc pod uwagę przedstawianą emocję (0 – brak zaufania, 100 – pełne zaufanie)

Rejestrowany był także czas odpowiedzi na każde pytanie. W tym zbiorze mieliśmy także zawarte poprawne odpowiedzi na przedstawiane emocje oraz podstawowe dane o osobie udzielającej odpowiedzi tj. płeć, wiek oraz zagregowany wynik wyżej wymienionej ankiety.

2.2.1 Brakujące wartości

W tym przypadku mieliśmy szczęście, gdyż wszystkie brakujące wartości zlokalizowane były w jednym zdeformowanym wierszu.



Rysunek 6: Wykres brakujących wartości w zbiorze danych z eksperymentu

2.2.2 Korelacja zmiennych

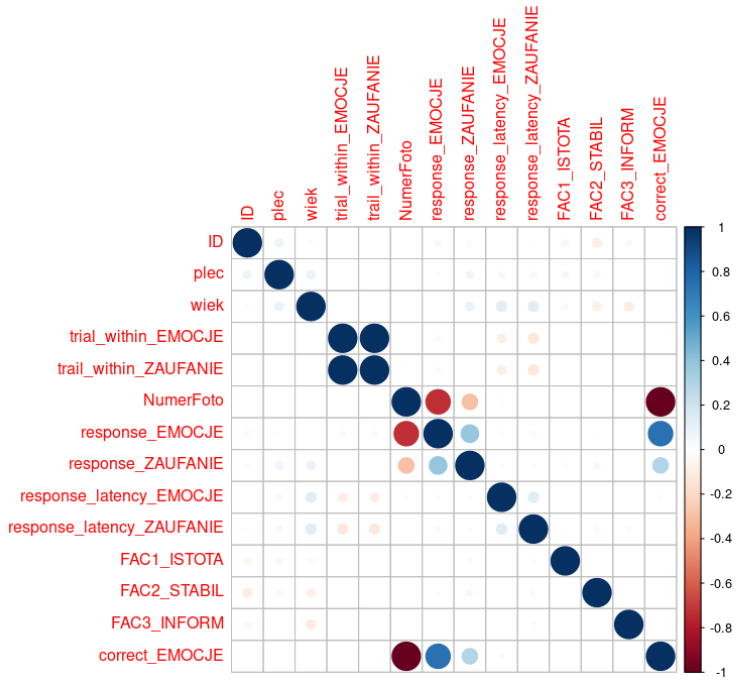
Najbardziej skorelowane są ze sobą zmienne `trial*`. Powodem tego jest to, iż jedna powstaje poprzez dodanie 1 do drugiej. `correct_EMOCJE` jest ujemnie skorelowana z `NumerFoto`, ponieważ numery fotografi określały jaka emocja znajduje się na fotografii. `response_EMOCJE` jest skorelowana z `correct_EMOCJE`. Wygląda na to, że ankietowani dobrze zgadywali. Najciekawsza jest korelacja `response_EMOCJE` i `response_ZAUFANIE`. Oznacza ona, że zaufanie do osoby jest skorelowane z okazywanymi przez nią emocjami.

2.2.3 Rozkład zmiennych numerycznych

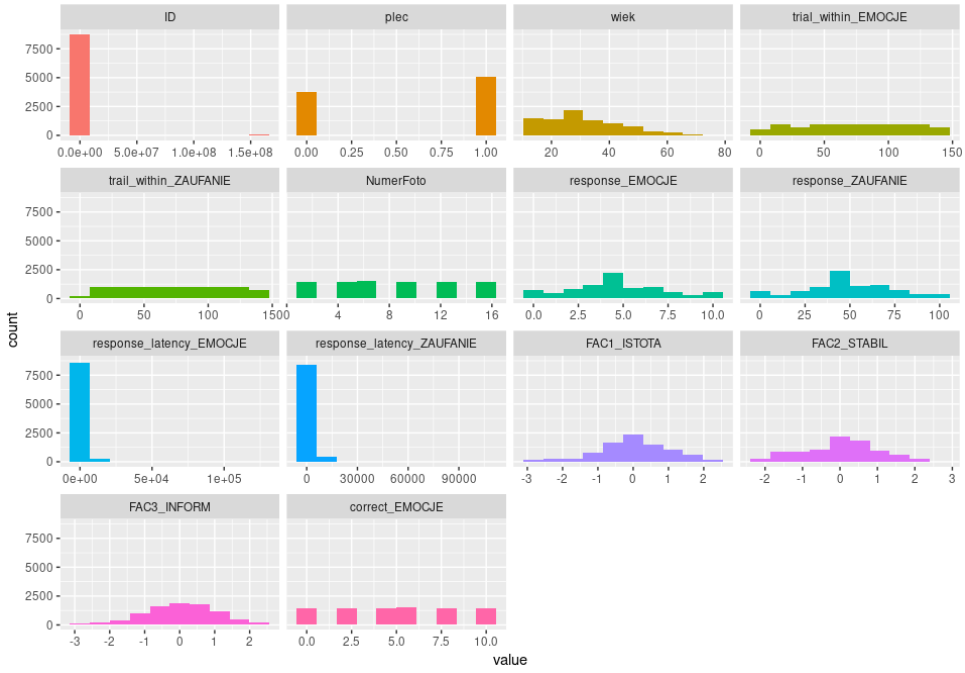
Większość zmiennych jest w miarę równo rozłożona, dzięki czemu unikniem może dużej ilości zakłóceń, które mogłyby się pojawić przy mocno skośnych zmiennych. Jedynymi zmiennymi, które wykazują są wiek oraz zmienne opisujące czas udzielania odpowiedzi. Dzięki zmiennej `wiek`, wiemy, że w badaniu brały udział głównie osoby młode. Zmienne `response*` wskazują, że większość osób odpowiadała szybko, jednakże zdarzyło się kilka dłuższych zastanowień.

2.2.4 Badanie za pomocą t-SNE

Po użyciu `t-SNE` nie otrzymaliśmy wyodrębnionych konkretnych klastrów, ale zarysowują nam się pewne małe podgrupki. Postawnowiliśmy pokolorować te punkty w zależności od wieku ankietowanego. Widać pewną zależność tych wzorów od wieku. Nasuwają nam się dwie prawdopodobne teorie. Jedna jest taka, że osoby w podobnym wieku podobnie reagują i odbierają emocje. Druga natomiast zakłada, że mamy

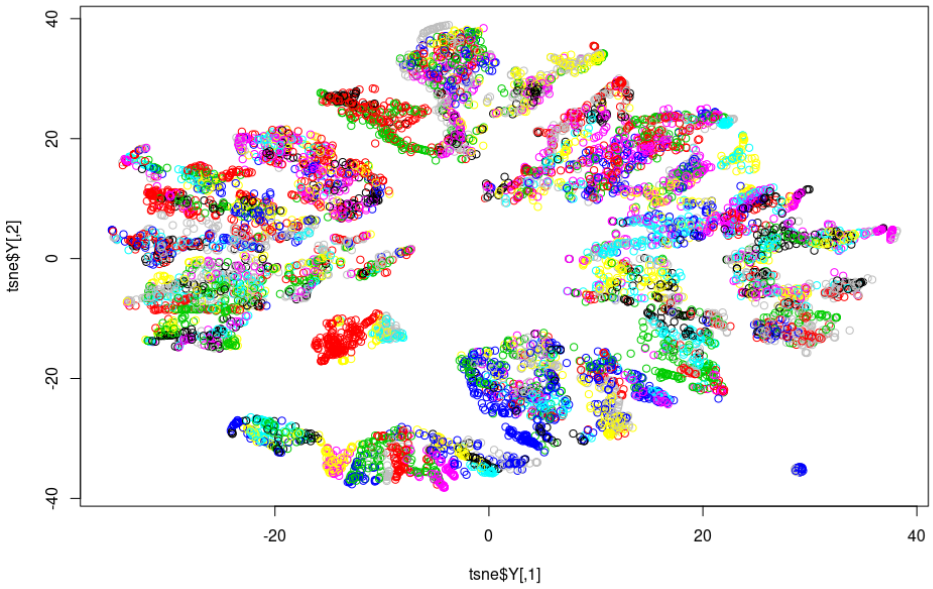


Rysunek 7: Wykres korelacji zmiennych w zbiorze danych z eksperymentu



Rysunek 8: Wykres rozkładów zmiennych numerycznych w zbiorze danych z eksperymentu

na tyle mało danych, że te miniklasterki, które nam się zarysowują to po prostu odpowiedzi jednej osoby.



Rysunek 9: Wykres rozkładów zmiennych numerycznych w zbiorze danych z eksperymentu

2.3 Połączone zbiory i anomalie

Ponieważ chcieliśmy dokonywać klasteryzacji po osobach, konieczne dla nas było połączenie obu ramek w jedną, zawierającą wszystkie informacje na temat poszczególnych osób. Szczegóły procesu przedstawimy w kolejnej sekcji. Tutaj wspomnimy tylko, że w wyniku tych operacji wykryliśmy kilka anomalii w zbiorze - jeden z wierszy zawierał brakujące wartości, natomiast jedna osoba na wszystkie pytania odpowiadała tak samo. Te dwie osoby wyrzuciliśmy z rozważań, ponieważ mogłyby zaburzać wyniki klasteryzacji.

3 Inżynieria cech

3.1 Łaczenie wierszy

Zanim połączyliśmy ramki, musieliśmy w jakiś sposób zagregować dane na temat jednej osoby – w ramce ‘photo‘ jednej osobie odpowiadało kilkanaście wierszy oznaczające reakcje na poszczególne zdjęcia.

W szczególności interesowały nas kolumny:

1. **response_EMOCJE** – odpowiedź w skali od 0 do 10 na pytanie o emocję wyrażoną na zdjęciu,
2. **NumerFoto** – numer prezentowanego zdjęcia (czyli prawdziwa emocja wyrażona w skali od 15 do 1; skala była odwrotna niż skala odpowiedzi),
3. **response_latency_EMOCJE** – czas odpowiedzi na pytanie o emocję wyrażoną na danym zdjęciu,
4. **response_ZAUFANIE** – odpowiedź w skali od 0 do 100, jak bardzo ankietowany byłby skłonny zaufać osobie z danym wyrazem twarzy,
5. **response_latency_ZAUFANIE** – czas odpowiedzi na pytanie o zaufanie.

Odpowiedź na każde zdjęcie była przechowywana w osobnym wierszu. Każda z osób oceniała cały zestaw zdjęć, po kilka na poszczególne poziomy emocji (tzn. każda osoba widziała kilka zdjęć o danej wartości NumerFoto, oczywiście, nie znając jej).

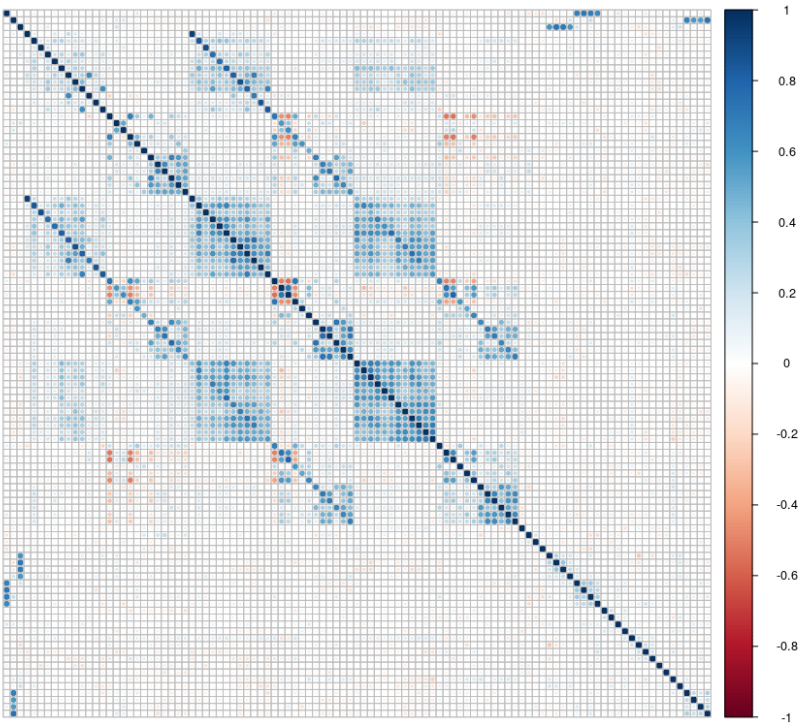
Policzyliśmy dla każdej osoby dla każdej wartości NumerFoto następujące wartości:

- `mean_response_EMOCJE_k` – uśredniona wartość odpowiedzi na pytanie o emocje dla zdjęć o NumerFoto równym k,
- `max_response_EMOCJE_k` – maksymalna wartość odpowiedzi na pytanie o emocje dla zdjęć o NumerFoto równym k,
- `min_response_EMOCJE_k` – minimalna wartość odpowiedzi na pytanie o emocje dla zdjęć o NumerFoto równym k.

Analogiczne wartości policzyliśmy dla odpowiedzi na pytanie o zaufaniem, dla czasu odpowiedzi na jedno oraz na drugie. W efekcie uzyskaliśmy $72 = 3 * 4 * 6$ nowe kolumny (ponieważ było sześć możliwych wartości NumerFoto).

3.2 Redukcja korelacji

Sprawdziliśmy następnie, czy zmienne te są skorelowane:

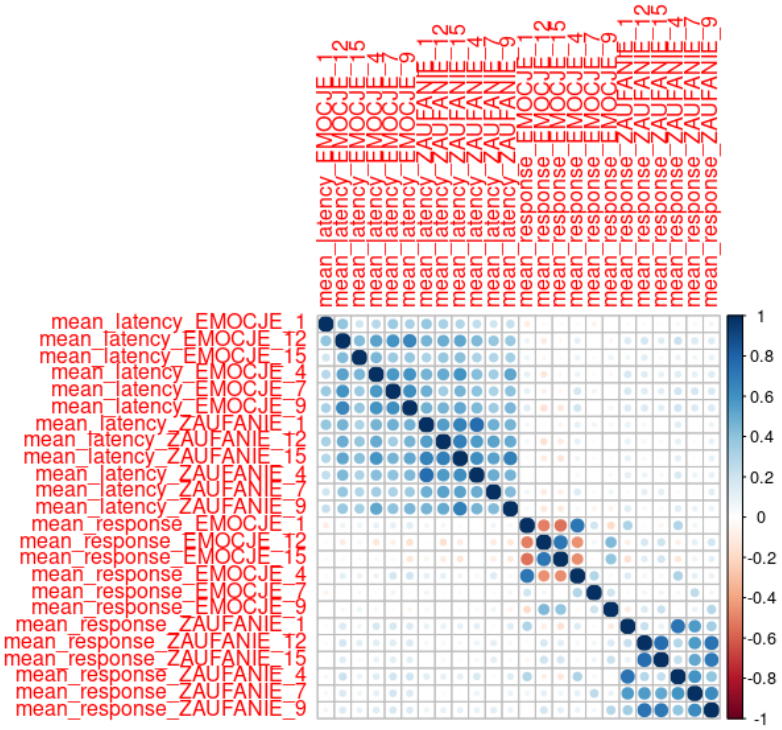


Rysunek 10: Korelacje w pierwszej połączonej ramce

Zobaczywszy silnie skorelowane całe grupy zmiennych, sprawdziliśmy, że były to kolumny zawierające czasy odpowiedzi na poszczególne zdjęcia. Tak silne korelacje pozwalają wnioskować, że nie był to wcale istotny czynnik, więc nie musimy ich w ten sposób agregować. Zamiast tego postanowiliśmy więc zachować tylko dwie kolumny – `mean_response_latency_EMOCJE` oraz `mean_response_latency_ZAUFANIE`.

Więcej uwagi poświęciliśmy odpowiedziom na pytania, między którymi również występowały korelacje. Występowały one zarówno pomiędzy kolumnami `mean*`, `min*`, `max*` jak i pomiędzy poszczególnymi

numerami zdjęć. Dla samych kolumn `mean*` występowało tych korelacji najwięcej, co można dojrzeć na wykresie **??**.



Rysunek 11: Korelacje w pierwszej połączonej ramce między średnimi

Z wykresu możemy wywnioskować, że silnie skorelowane są odpowiedzi na dwa najniższe numery zdjęć, podobnie jak dwa najwyższe. W związku z tym postanowiliśmy je zgrupować – teraz mamy już tylko cztery możliwe numery zdjeć. Jako że jednak nie byliśmy pewni, czy grupowanie nie sprawi, że utracimy za dużo informacji, postanowiliśmy zachować zbiory zarówno w takiej jak i takiej wersji.

Poza tym po połączeniu występowały bardzo silne korelacje pomiędzy kolumnami **FAC1_ISTOTA**, **FAC2_STABIL**, **FAC3_INFORM** a częścią kolumn z ramki 'esence'. Wynika to z faktu, że kolumny te powstały jako uśrednienie drugich.

Ponieważ jednak nie mogliśmy się zdecydować, zachowanie których zmiennych – czy oryginalnych, czy przetworzonych (FAC) – będzie lepsze, utworzyliśmy dwa zbiory z dwoma rozwiązaniami – ostatecznie otrzymaliśmy cztery zbiory, których krótką charakterystykę prezentuje tabela **??**.

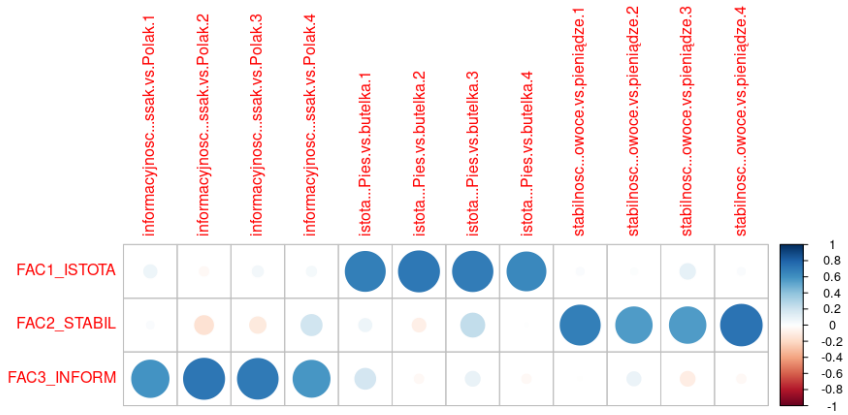
Tablica 1: Tabela opisująca zbiory użyte do klasteryzacji

	zachowane wszystkie numery zdjęć	zgrupowane numery zdjęć
zachowane przetworzone FAC	*wo_latencies_w_FAC*	*pgrouped_w_FAC*
zachowane nieprzetworzone zmienne	*wo_latencies_wo_FAC*	*pgrouped_wo_FAC*

4 Trenowanie modeli

4.1 Metodyka ewaluacji

Z inżynierii cech dostaliśmy 4 zbiory danych o różnej charakterystyce, ale niewiadomej skuteczności. Jak wybrać najlepszy?



Rysunek 12: Korelacje w pierwszej połączonej ramce między średnimi

Postanowiliśmy sprawdzić jak różne modele klasteryzacyjne zachowują się na tych zbiorach danych i wybrać najlepszy zbiór w zależności od otrzymanych wyników. Dla każdego zbioru zbudowaliśmy model klasteryzacyjny z różną liczbą klastrów: od 2 do 11. Aby porównać wyniki wykorzystaliśmy 4 różne indeksy mierzące jakoś klasteryzacji: 2 z rodziny indeksów Dunna, znormalizowany indeks Γ oraz indeks sepracji klastrów. Indeksy Dunna wybraliśmy ze względu na ich jakość detekcji optymalnej liczyby klastrów. Statystykę Γ wybraliśmy, ponieważ był on polecany przez różne źródła internetowe. Natomiast indeks separacji wybraliśmy ze względu na jego intuicyjność.

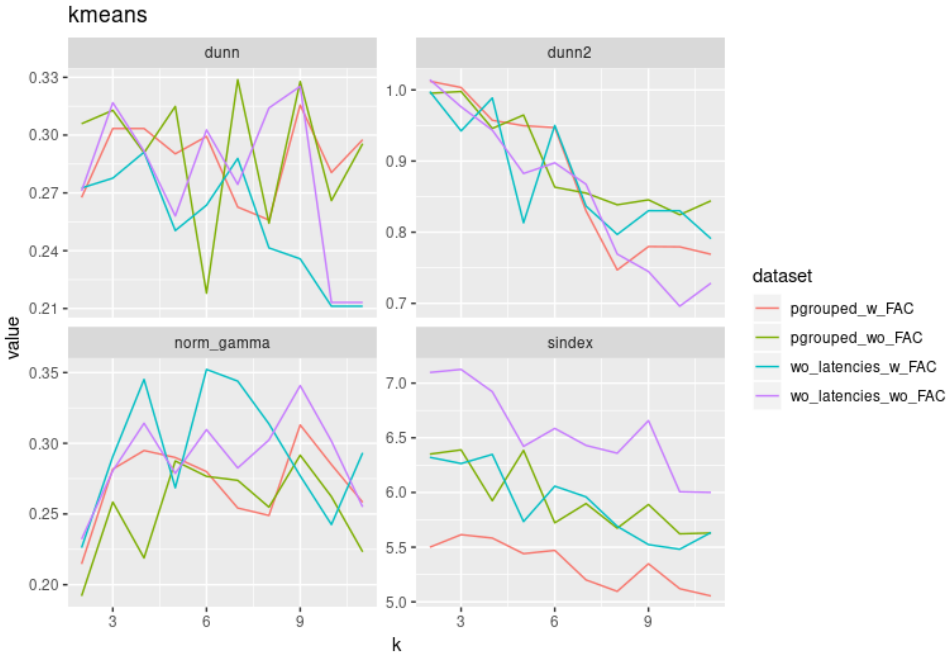
Po przeanalizowaniu wykresów zdecydowaliśmy się dalej pracować na zbiorze `wo_latencies_wo_FAC`. Był to najbardziej ogarniczony ze wszystkich naszych zbiorów względem zbioru wejściowego. W następnej fazie porównywaliśmy różne modele na wybranym zbiorze używając wcześniej stosowanych indeksów. Każdy z modeli trenowaliśmy na $k = 2, 3, ..., 11$ liczbie klastrów i wyciągaliśmy z nich statystyki porównawcze. Jako uzupełnienie analizy jakości klastrowania przez wybrane modele stworzyliśmy wykresy wizualizujące rozkład naszych klastrów po użyciu narzędzia `tSNE`, aby wzrokowo zbadać jak algorytmy rozłożyły klastry. Do tej analizy wykorzystaliśmy tylko te liczby klastrów, które uznaliśmy za bliskie optymalnej.

4.2 Użyte modele

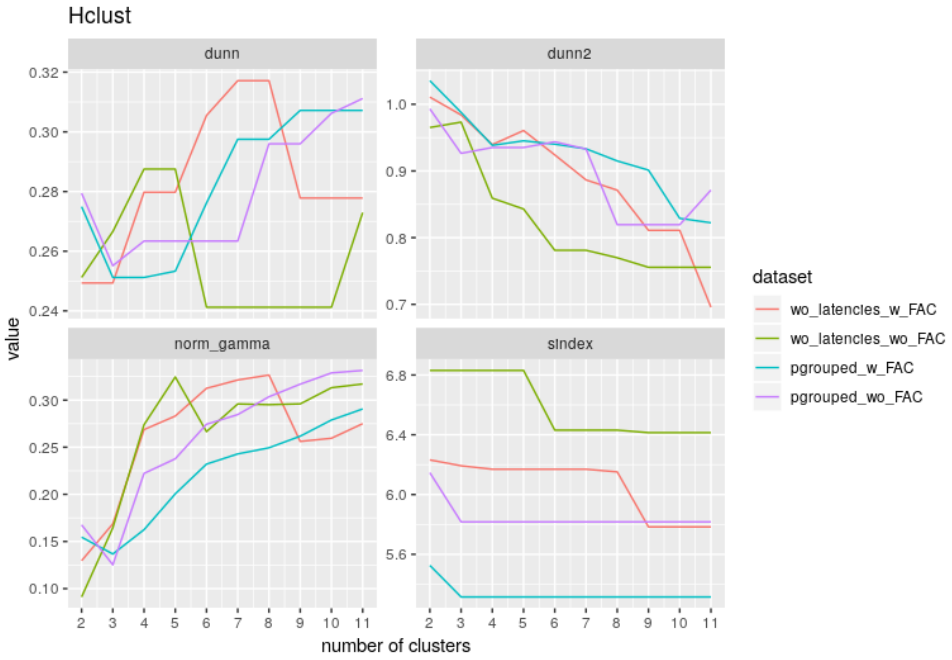
Do testowania zdecydowaliśmy się wybrać algorytmy:

1. k means (`kmeans`)
2. hierarchical clustering z kryterium Warda (`hclust`)
3. partitioning around medioids (`pam`)
4. gaussian mixture modelling (`mclust`)

Zaczeliśmy od wyliczenia wybranych wcześniej statystyk dla różnych zbiorów, liczb klastrów i modeli. Wyniki przedstawiają wykresy 13, 14, 15, 16.



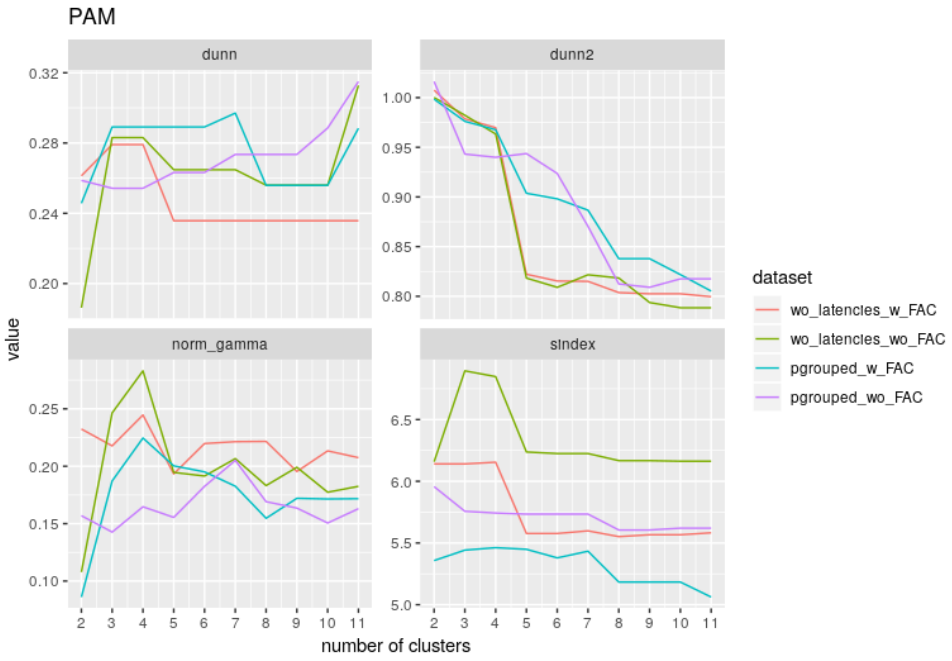
Rysunek 13: Wykres rozkładów statystyk dla algorytmu `kmeans`



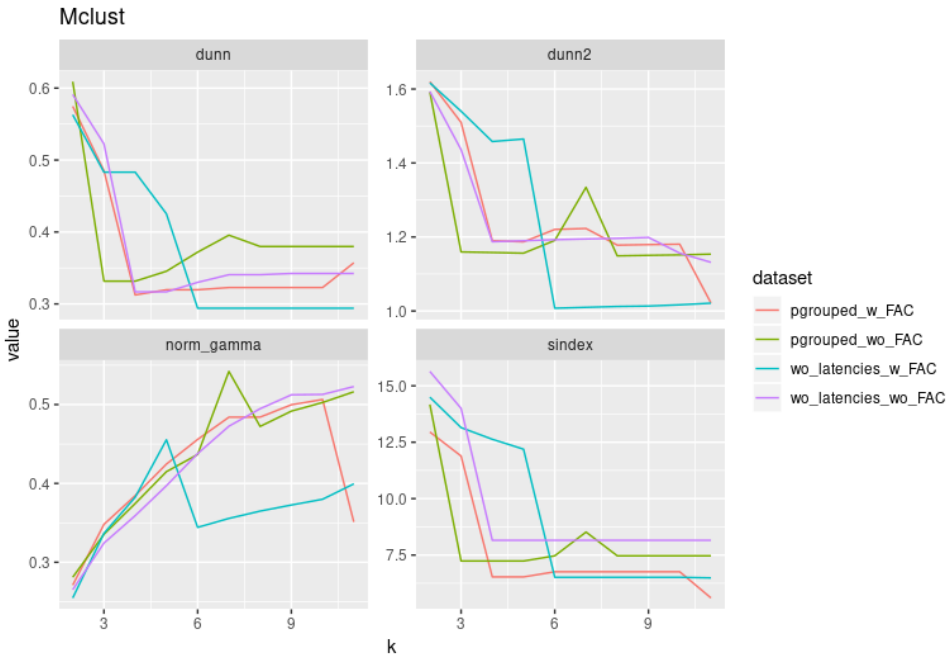
Rysunek 14: Wykres rozkładów statystyk dla algorytmu `hclust`

Ponieważ najlepsze wyniki otrzymywaliśmy na zbiorze `wo_latencies_wo_FAC`, postanowiliśmy porównać między sobą wszystkie algorytmy na tym zbiorze. Przedstawia to rysunek 17.

Statystyki wskazywały, że nalepiej radził sobie algorytm `mclust`, najgorzej `pam`, a `kmeans` i `hclust` otrzymywały podobne wyniki. Najlepsze wyniki otrzymywaliśmy dla małej liczby podziałów, więc zdecy-

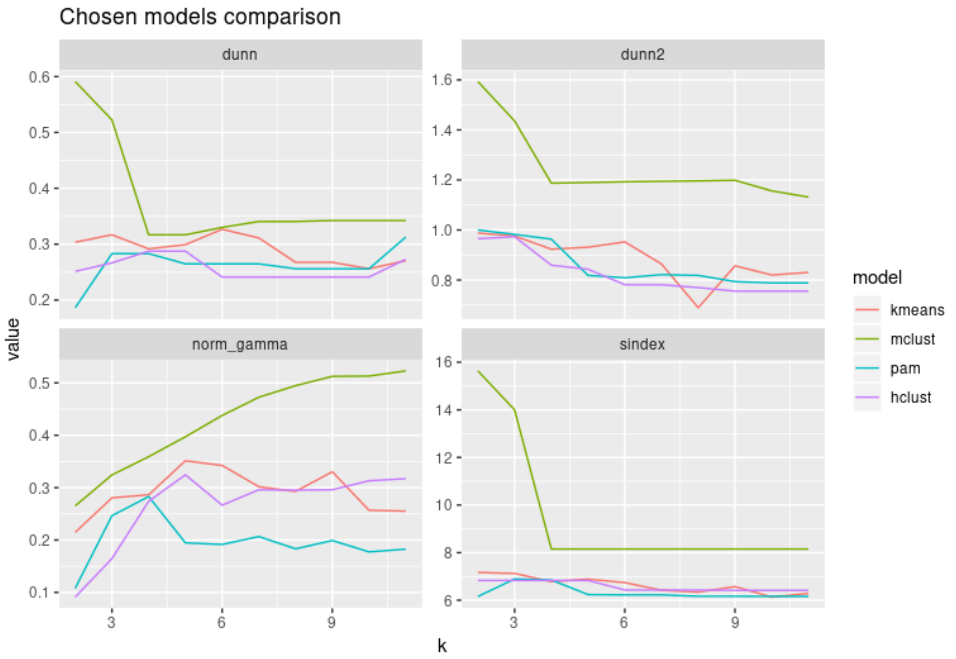


Rysunek 15: Wykres rozkładów statystyk dla algorytmu `pam`



Rysunek 16: Wykres rozkładów statystyk dla algorytmu `mclust`

dowaliśmy się dokonać porównania dla dwóch i trzech klastrów. Postanowiliśmy zwizualizować utworzone klastry za pomocą algorytmu t-SNE 18. Okazało się, że `mclust` umieścił prawie wszystkie obserwacje w jednym klastrze. Z kolei klastry stworzone przez `kmeans` i `hclust` są bardzo podobne. Nieco zbliżone, choć bardziej nieregularne podziały dał algorytm `pam`.



Rysunek 17: Wykres rozkładów statystyk dla wszystkich algorytmów na wybranym zbiorze

Rysunek 18: Wizualizacja utworzonych klastrów

5 Wybrany model

Po przeanalizowaniu wyników zaczęliśmy od odrzucenia modelu mclust, jako że utworzył skrajnie niezbalansowany podział wyróżniając kilka obserwacji w osobnych klastrach. Patrząc na wyznaczone statystyki zdecydowaliśmy się wskazać trzy, jako optymalną liczbę klastrów dla naszego zbioru. Przy tym wyborze kierowaliśmy się głównie statystykami `dunn2` i `sindex`, które powyżej tej wartości wykazywały spadek oraz rozbieżności między algorytmami. Następnie porównaliśmy ze sobą podziały na trzy klastry stworzone przez modele. Ostatecznie wybraliśmy algorytm k-średnich. Otrzymał on najwyższe wartości statystyk dla trzech klastrów i stworzył najbardziej regularny podział zbioru.

5.1 Wybrane cechy

Wybrany model był wytrenowany na zbiorze `wo_latencies_wo_FAC` kolumnach:

- dotyczących danych osobowych (dwie kolumny: płeć i wiek),
- dotyczących opóźnienia w odpowiedzi (dwie kolumny: opóźnienie przy pytaniu o emocje i o zaufanie),
- dotyczących minimalnej, średniej, maksymalnej odpowiedzi na pytanie o emocje i o zaufanie przy sześciu różnych numerach zdjęć ($3 \cdot 2 \cdot 6 = 72$ kolumny),
- dotyczących odpowiedzi na siedem grup po cztery pytania w ankiecie ($7 \cdot 4 = 28$ kolumn).

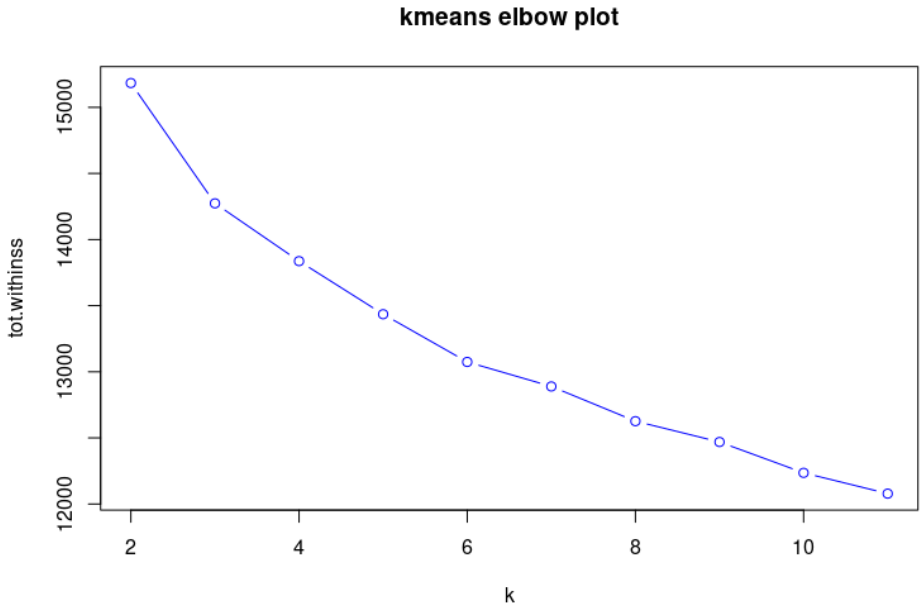
5.2 Skuteczność modelu

Przyjrzeliśmy się dokładniej jak `kmeans` zachowuje się dla kolejnych liczb klastrów. Chcieliśmy wybrać optymalną liczbę, klastrów dla naszego zbioru, więc postanowiliśmy jeszcze raz spojrzeć na używane przez nas wcześniej indeksy **??**. Średnio rzecz biorąc kmeans osiągał według nas najlepsze wyniki dla 3 klastrów. Widać to głównie na indeksie `dunn2` oraz `sindex`. `dunn` także osiąga tam lokalne maksimum.

Tablica 2: Tabela przedstawiająca wyniki różnych miar dla kolejnych liczb klastrów kmeans

k	dunn	dunn2	norm_gamma	sindex
2	0.27	1.01	0.23	7.10
3	0.32	0.98	0.28	7.13
4	0.29	0.94	0.31	6.92
5	0.26	0.88	0.28	6.42
6	0.30	0.90	0.31	6.59
7	0.27	0.87	0.28	6.43
8	0.31	0.77	0.30	6.36
9	0.33	0.74	0.34	6.66
10	0.21	0.70	0.30	6.01
11	0.21	0.73	0.25	6.00

Spojrzeliśmy także uważnie na wykres łokciowy 19. Wykres jest dość liniowy, nie widać wyraźnych punktów mogących wskazać na optymalną liczbę klastrów. Jednakże, gdybyśmy mieli wybierać jakiś punkt, w którym mogłaby być wartość optymalna to byłoby to właśnie 3.



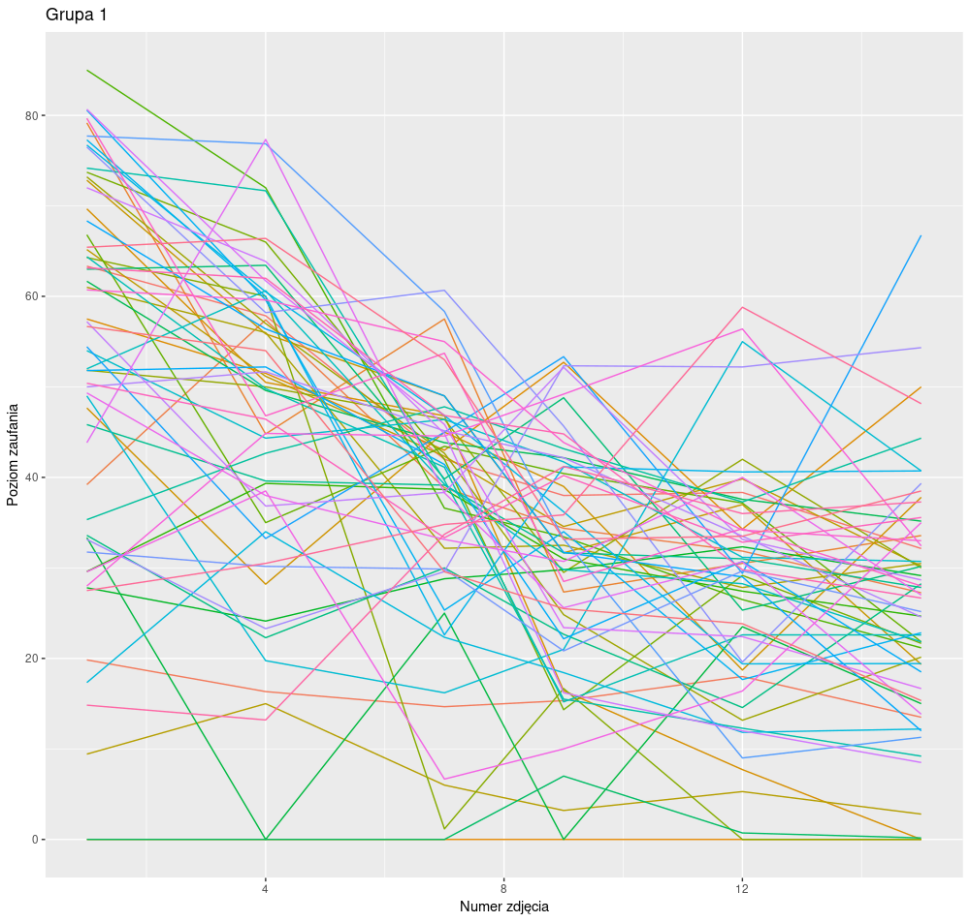
Rysunek 19: Wykres łokciowy kmeans

5.3 Utworzone klastry

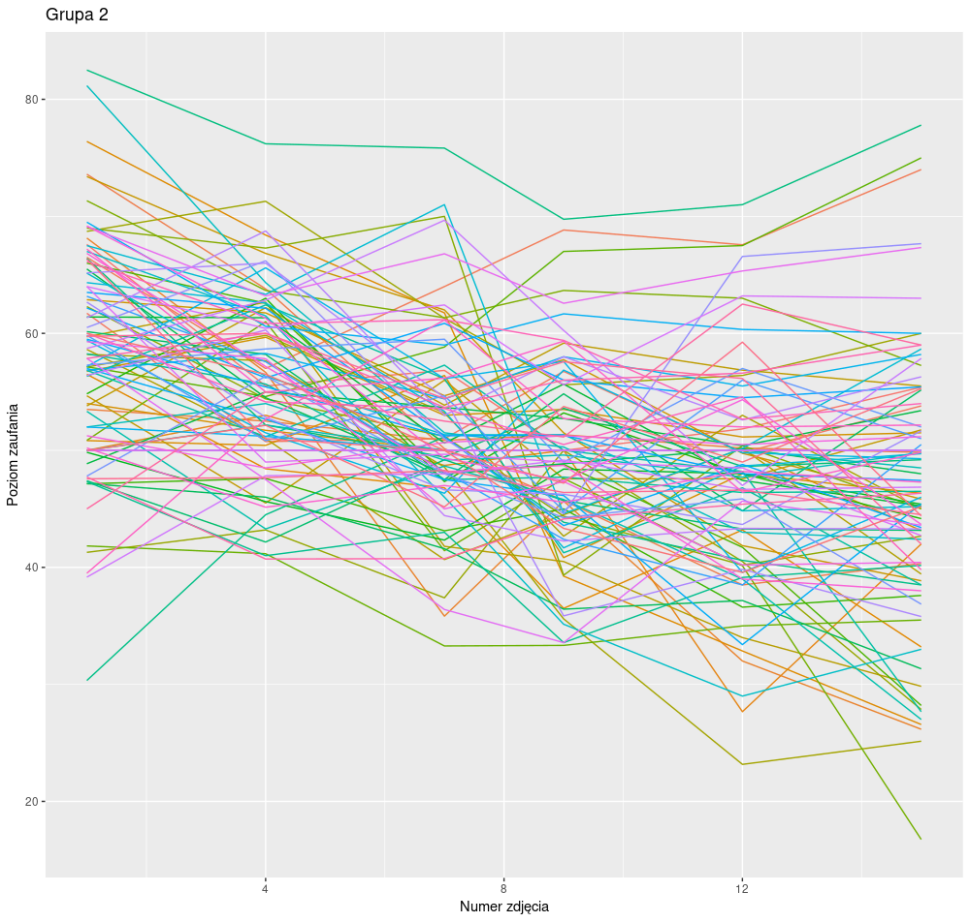
Na wykresach 20, 21, 22 prezentujemy wykresy poziomu zaufania od numeru zdjęcia dla poszczególnych osób w poszczególnych klastrach. Widzimy, że rozkłady te znacząco się od siebie różnią i częściowo są zgodne z tezą badania - ludzie różnie oceniają swe zaufanie w zależności od mimiki drugiej osoby.

5.4 Podsumowanie

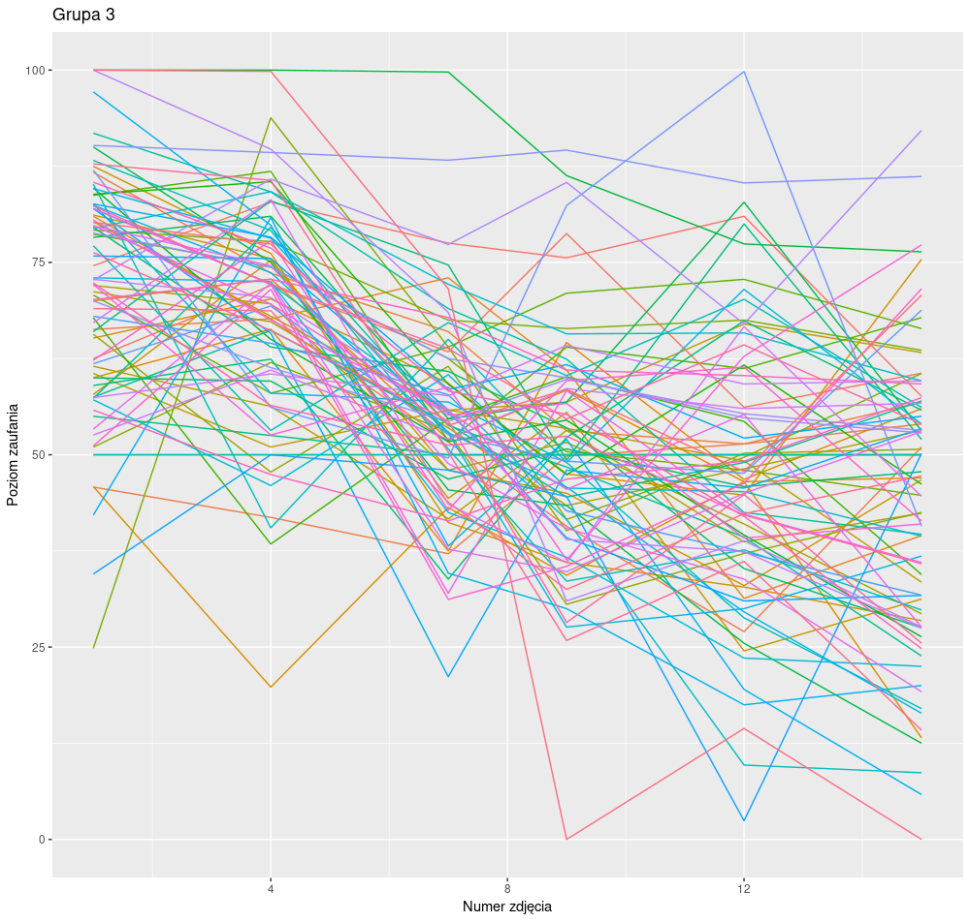
Celem projektu było stworzenie modelu dokonującego klasteryzacji osób biorących udział w eksperymencie na temat postrzegania emocji przeprowadzonym w Centrum Nauki Kopernik. W fazie eksploracyjnej analizy danych poznaliśmy zestaw danych, jakim dysponowaliśmy, następnie przekształciliśmy je odpowiednio w fazie inżynierii cech, aby uzyskać z nich jak najwięcej informacji i na ich podstawie zbudować modele, a następnie wybrać ten, który najlepiej się sprawdzi. Przetestowaliśmy kilka modeli na kilku różnych zestawach kolumn, badając różne możliwe liczby klastrów. Ostatecznie wybraliśmy algorytm k-średnich z trzema klastrami, który dawał zadowalające wyniki w poszczególnych miarach skuteczności. Klastry utworzyły grupy, które prezentują podział częściowo zgodny z tezą badania.



Rysunek 20: Wizualizacja klastra 1



Rysunek 21: Wizualizacja klastra 2



Rysunek 22: Wizualizacja klastra 3