# KLASTROWANIE LUDZKIEJ AKTYWNOŚCI NA PODSTAWIE SENSORÓW SMARTPHONA

KACPER KUROWSKI BARTOSZ SAWICKI

# OPIS ZBIORU DANYCH

- 7352 obserwacji zebranych od 21 osób
- Każda obserwacja ma przypisaną jedną z 6 aktywności:
  - Leżenie
  - Stanie
  - Siedzenie
  - Chodzenie
  - Chodzenie po schodach w górę
  - Chodzenie po schodach w dół
- Korzystamy z etykiet aktywności przy wyborze podzbioru kolumn, na których będziemy przeprowadzać klastrowanie oraz do weryfikacji wyników klastrowania. W samym klastrowaniu nie korzystamy z labeli.
- 561 kolumn numerycznych z wartościami [-1,1]

# KOLUMNY

```
tBodyAcc-XYZ
tGravityAcc-XYZ
tBodyAccJerk-XYZ
tBodyGyro-XYZ
tBodyGyroJerk-XYZ
tBodyAccMag
tGravityAccMag
tBodyAccJerkMag
tBodyGyroMag
tBodyGyroJerkMag
fBodyAcc-XYZ
fBodyAccJerk-XYZ
fBodyGyro-XYZ
fBodyAccMag
fBodyAccJerkMag
fBodyGyroMag
fBodyGyroJerkMag
```

```
mean(): Mean value
std(): Standard deviation
mad(): Median absolute deviation
max(): Largest value in array
min(): Smallest value in array
sma(): Signal magnitude area
energy(): Energy measure. Sum of the squares divided by the number of values.
iqr(): Interquartile range
entropy(): Signal entropy
arCoeff(): Autorregresion coefficients with Burg order equal to 4
correlation(): correlation coefficient between two signals
maxInds(): index of the frequency component with largest magnitude
meanFreq(): Weighted average of the frequency components to obtain a mean frequency
skewness(): skewness of the frequency domain signal
kurtosis(): kurtosis of the frequency domain signal
```

- Zmienne zaczynające się od t są w domenie czasu, te zaczynające się od f są w domenie częstotliwości (przekształcone Fast Fourier Transform FFT).
- Dane zbierane były w 3 płaszczyznach X, Y, Z

- Powyższe przekształcenia zostały zaaplikowane do danych, aby zagregować 2 sekundowe odcinki.

# PROBLEMY ZE ZBIOREM

- Zduplikowane nazwy kolumn

  Na początku uznaliśmy, że usuniemy zduplikowane kolumny. Później jednak odkryliśmy, że najprawdopodobniej z nazw kolumn usunięte zostały suffixy mówiące o kierunku zmiennej (X, Y, Z). Dodaliśmy suffixy i mogliśmy wczytać wszystkie kolumny do ramki danych.

- Duża liczba kolumn

  Postanowiliśmy wybrać podzbiór kolumn, aby przyspieszyć działanie algorytmów klastrujących. Skorzystaliśmy w tym celu z dywergencji Kullbacka-Leibera (KL), mówiącej o różnicy w rozkładach dwóch cech. Dywergencja KL została wyliczona dla każdej kolumny pomiędzy każdą możliwą parą aktywności. Na tej podstawie mogliśmy wybrać n kolumn najbardziej dzielących dane ze względu na etykiety aktywności. Szczegóły tego procesu są opisane w *feature_selection_by_entropy.ipynb*.

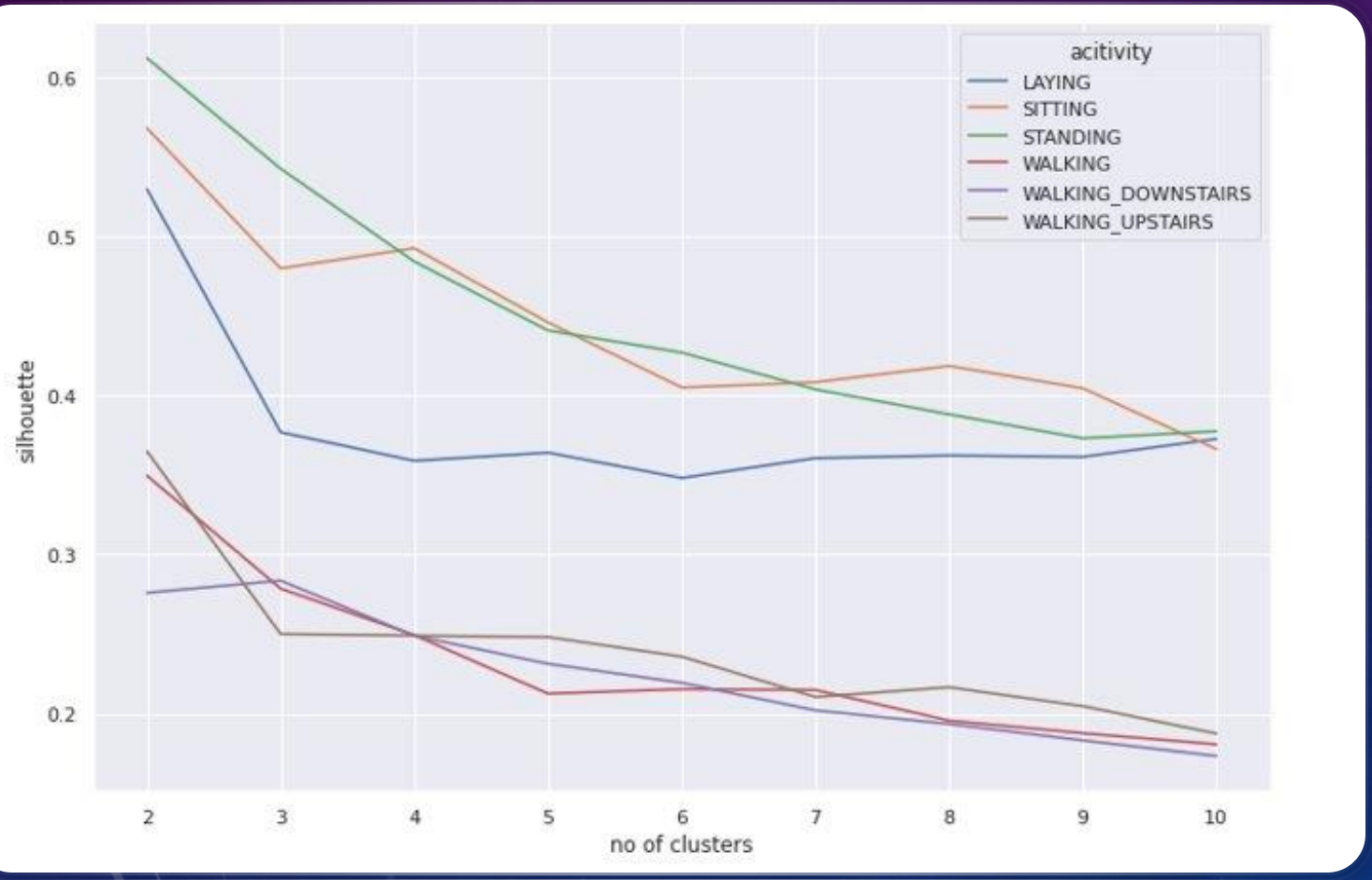
# WNIOSKI Z EDA

- Rozkład cech leżenia często wyróżnia się na tle rozkładów dla innych czynności. Intuicyjnie, jest to najbardziej odróżniająca się czynność.
- Mimo, że wsztstkie kolumny mają pewną fizyczną interpretację, to nabycie intuicji i zrozumienia w przypadku tak licznych cech jest prawie niemożliwe.

# JAK ZWIĘKSZYĆ PRODUKTYWNOŚĆ?

Napisaliśmy skrypt *utils.py,* w którym zawarliśmy fukcje potrzebne do wczytywania danych i selekcji kolumn. Dzięki temu w późniejszych analizach pobranie danych możliwe było w jednej linijce. Może nie jest to innowacyjne odkrycie, ale zapewnia jednakowy format danych przy każdym wczytywaniu. Jest to szczególnie istotne gdy pracuje się w zespole. Szczegół, który warto wykorzystywać w kolejnych projektach.

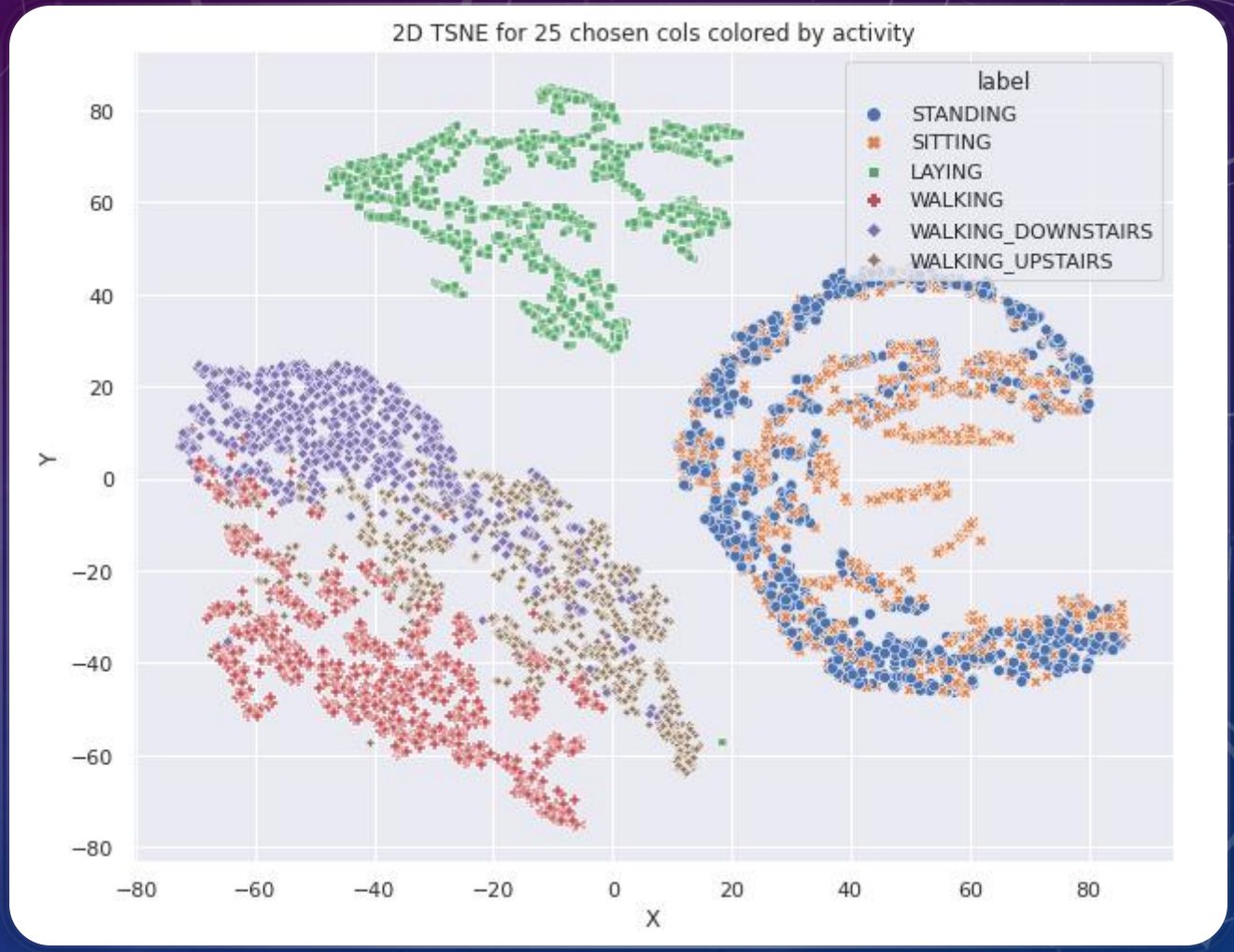
# KLASTROWANIE WEWNĄTRZ AKTYWNOŚCI

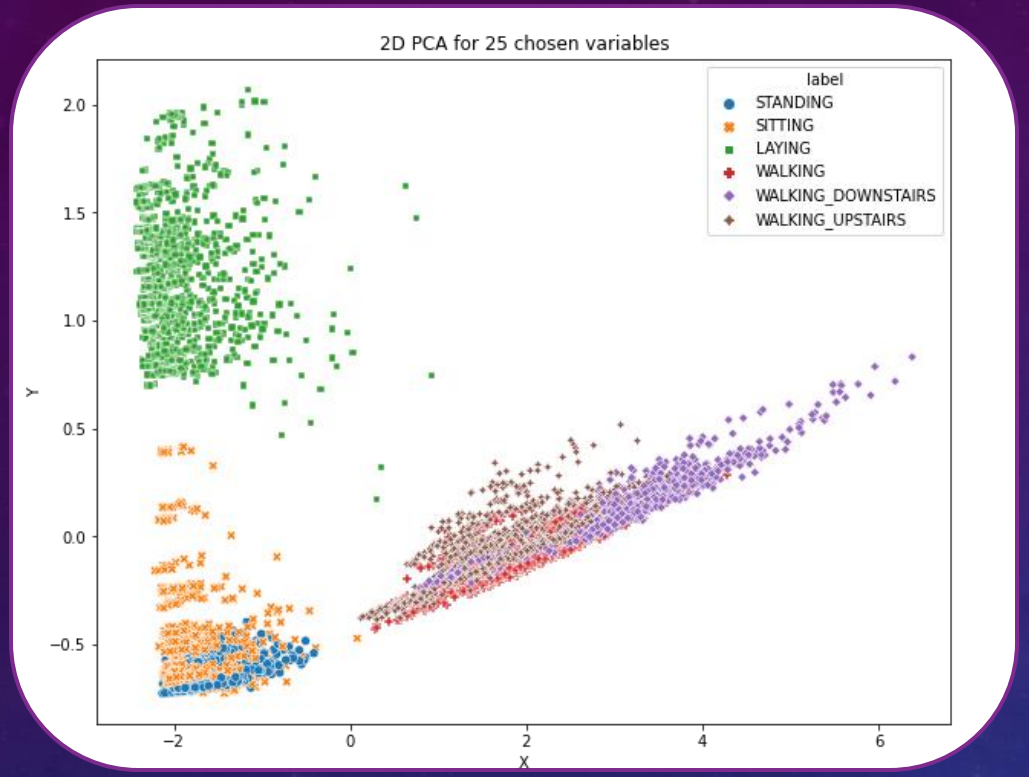


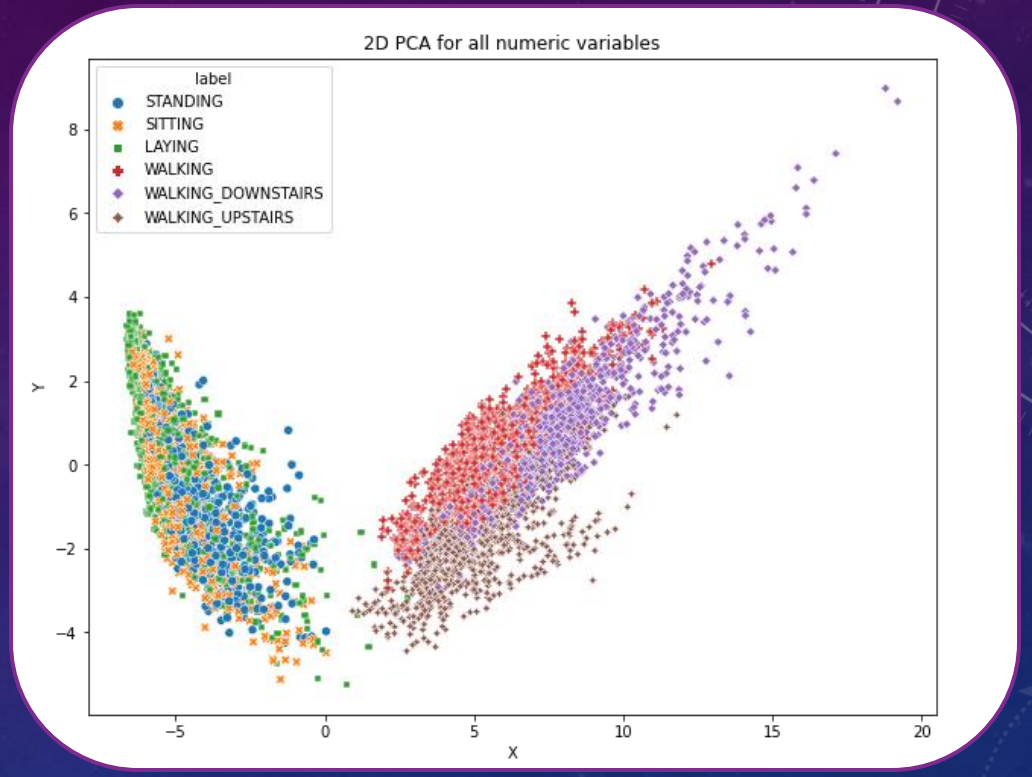

- Niemalże wszystkie czynności mają malejące silheutte wraz ze wzrostem liczby kalstrów.
- Leżenie zdaje się mieć szczególną liczbę 2 klastrów,
- Schodzenie w dół zdaje się móc dzielić na 3 klastry.
- Nieco ciekawie wyglądają 4 klastry dla siedzenia.
- Obecna nieregularnosć sugeruje, że być może czynności dzielą się na fazę wstępną, właściwą i koncową. Spadek w silhuette wówczas tłumaczylibyśmy przez spore okno czasowe uśredniające wyniki.

# GROUND TRUTH

Dwuwymiarowa wizualizacja TSNE. Kolor reprezentuje różne etykiety aktywności. Ogólnie dążymy do tego, aby klastrowanie oddawało podział na czynności.

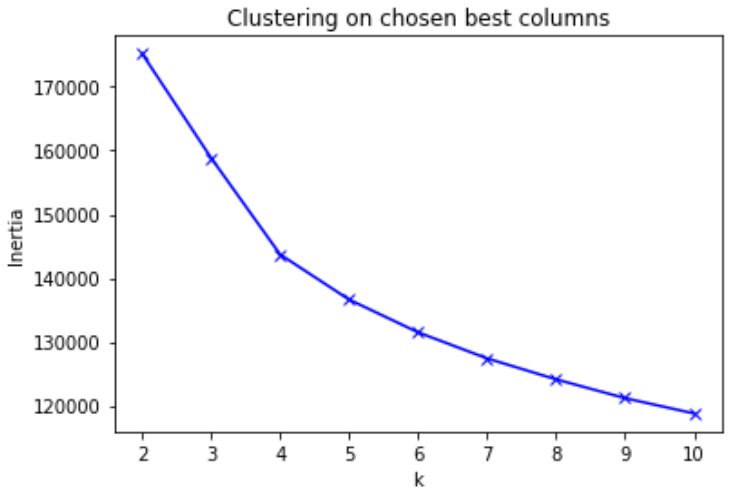


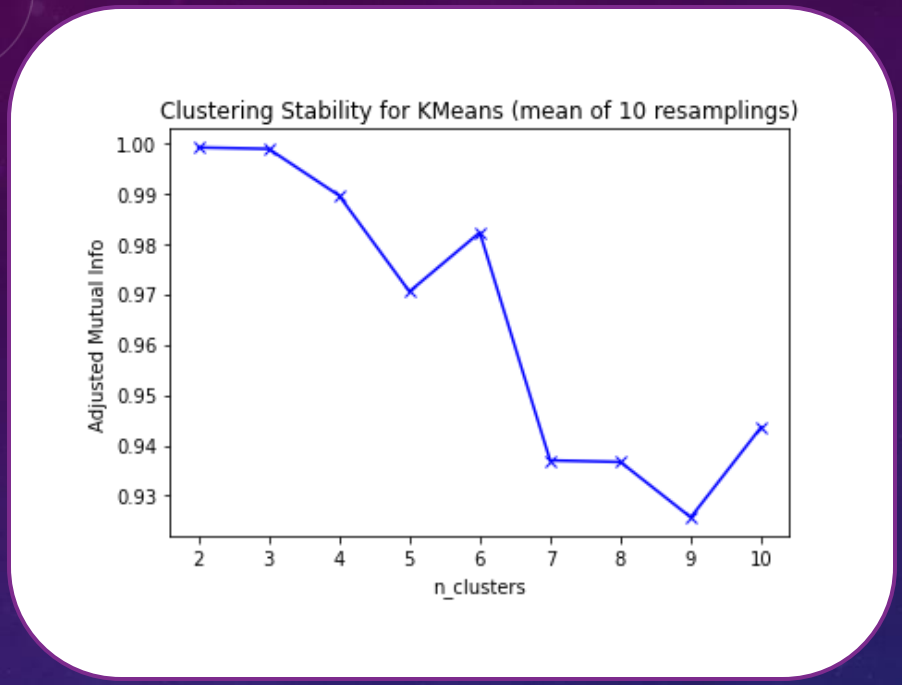


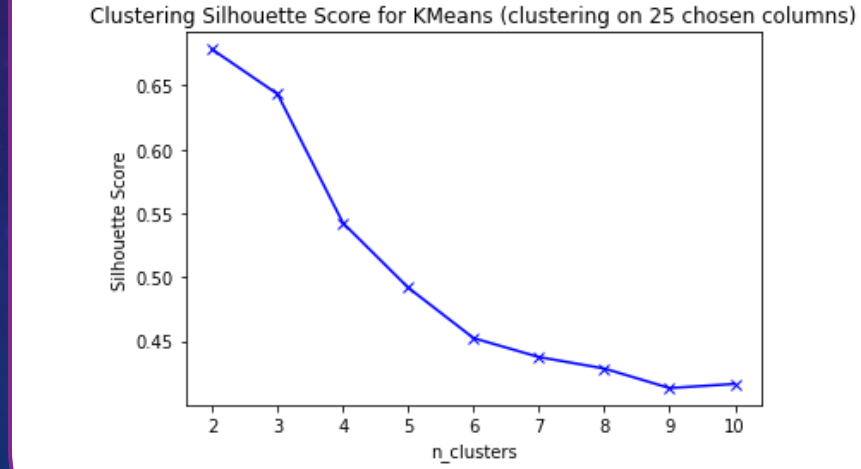




PCA dla 25 najbardziej rozdzielających kolumn bardziej oddziela czynności o różnych labelach. Leżenie jest zauważalnie inną grupą.

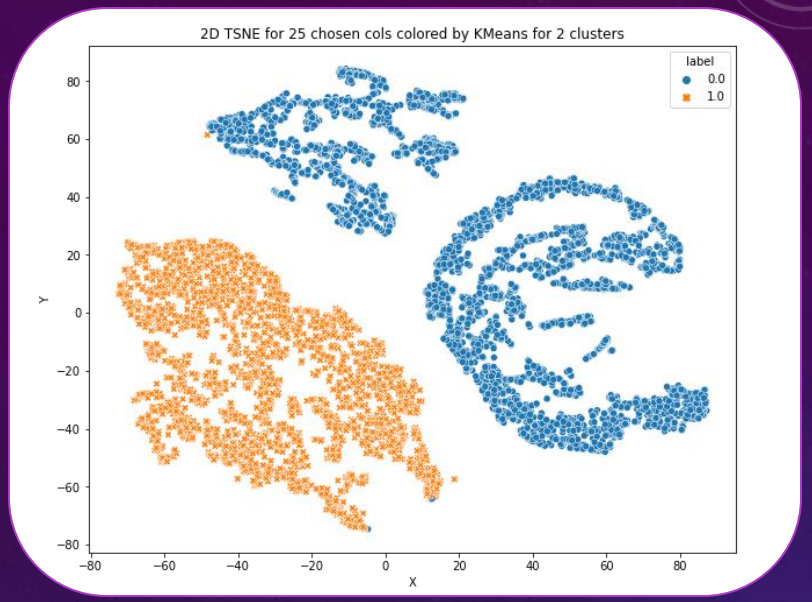
# KMEANS

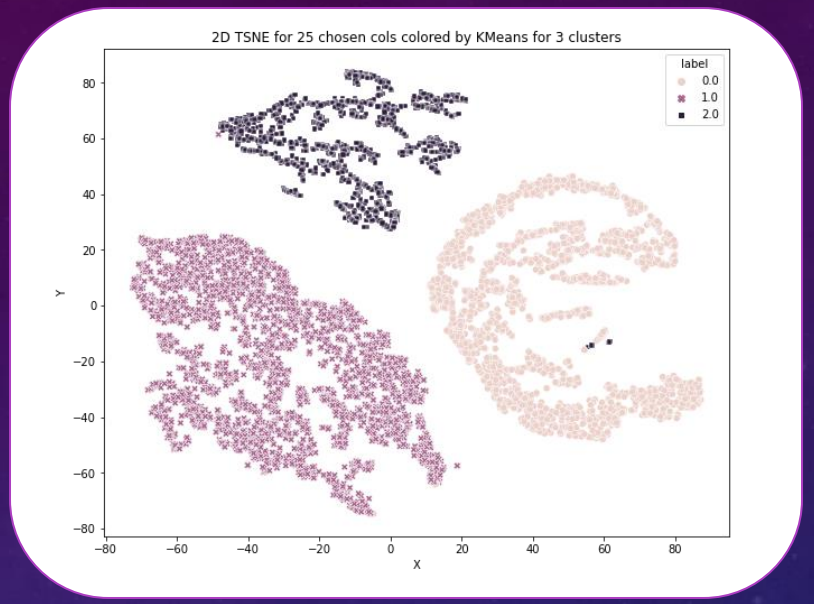


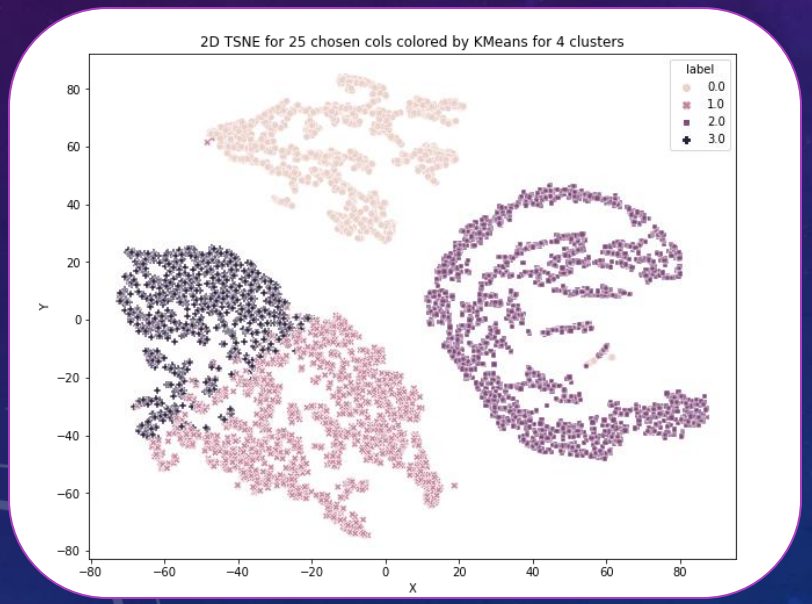




- Wysoka stabilność klastrowania dla 2, 3 i 6 klastrów
- Silhouette najwyższe dla 2 lub 3 klastrów

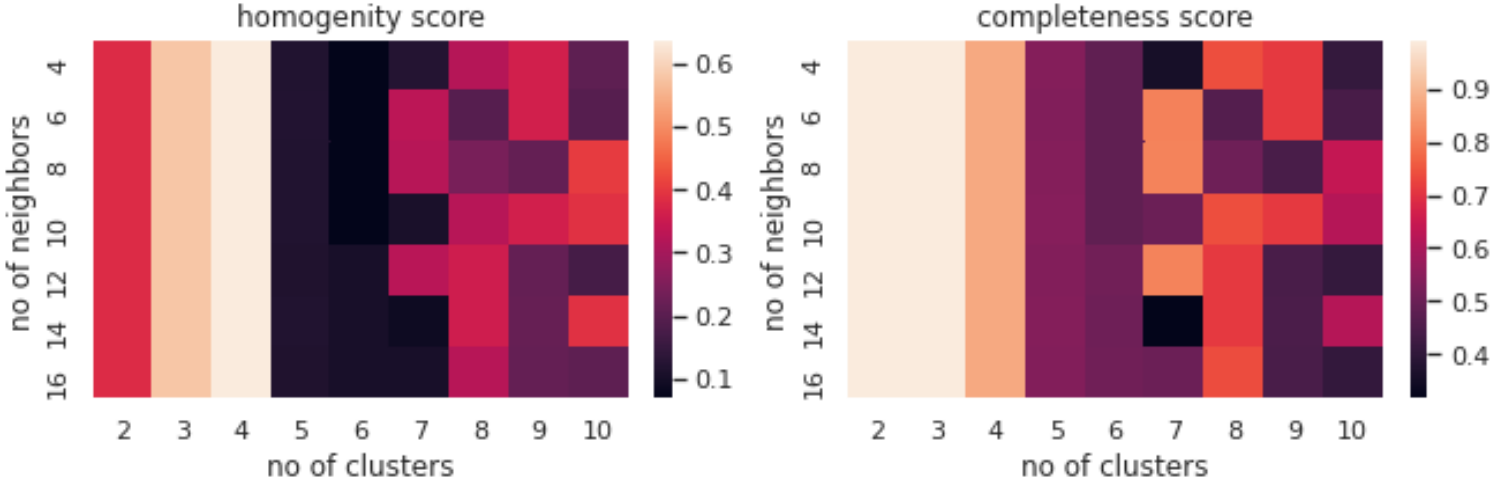


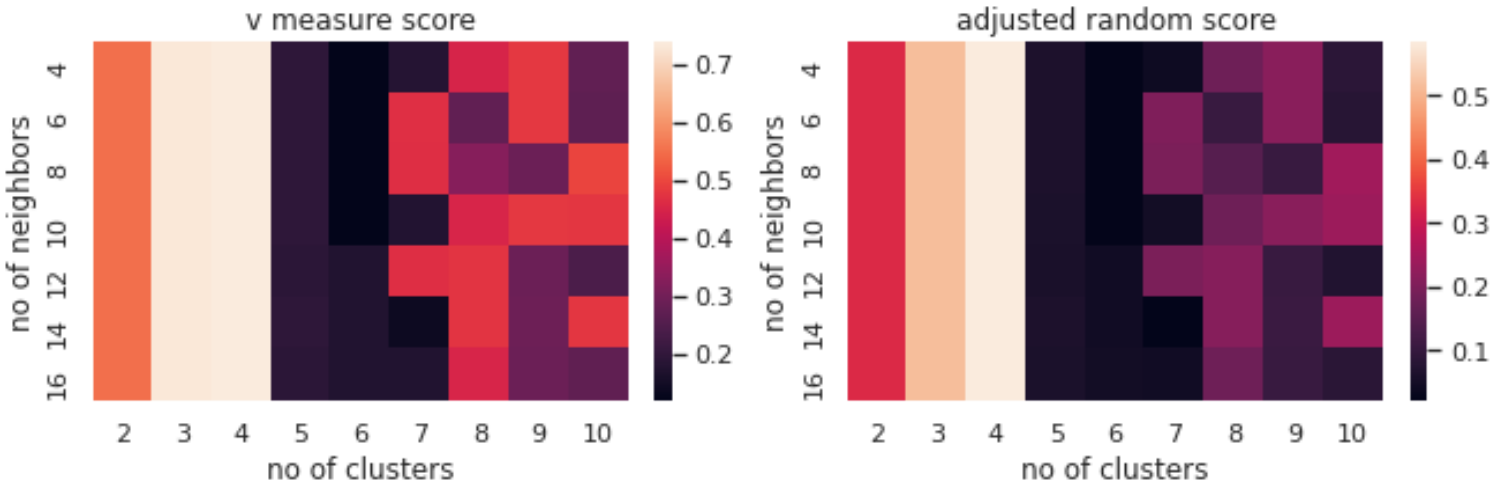


- Dla 2 i 3 klastrów etykiety aktywności pokrywają się z klastrami.
- Dla 4 i 6 klastrów uzyskujemy podział nie do końca zgodny z etykietami.
- Leżenie jest łatwo odróżnialne od innych czynności.
- Siedzenie i stanie tworzą jedną grupę, którą trudno rozdzielić
- Podobnie, grupa chodzenia jest trudna do podziału.



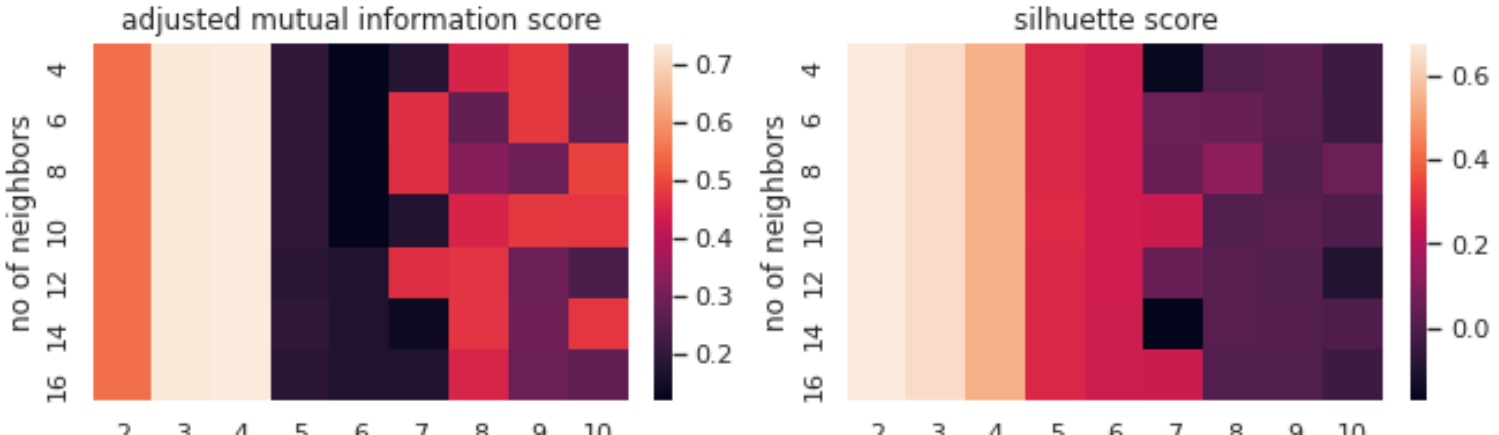
















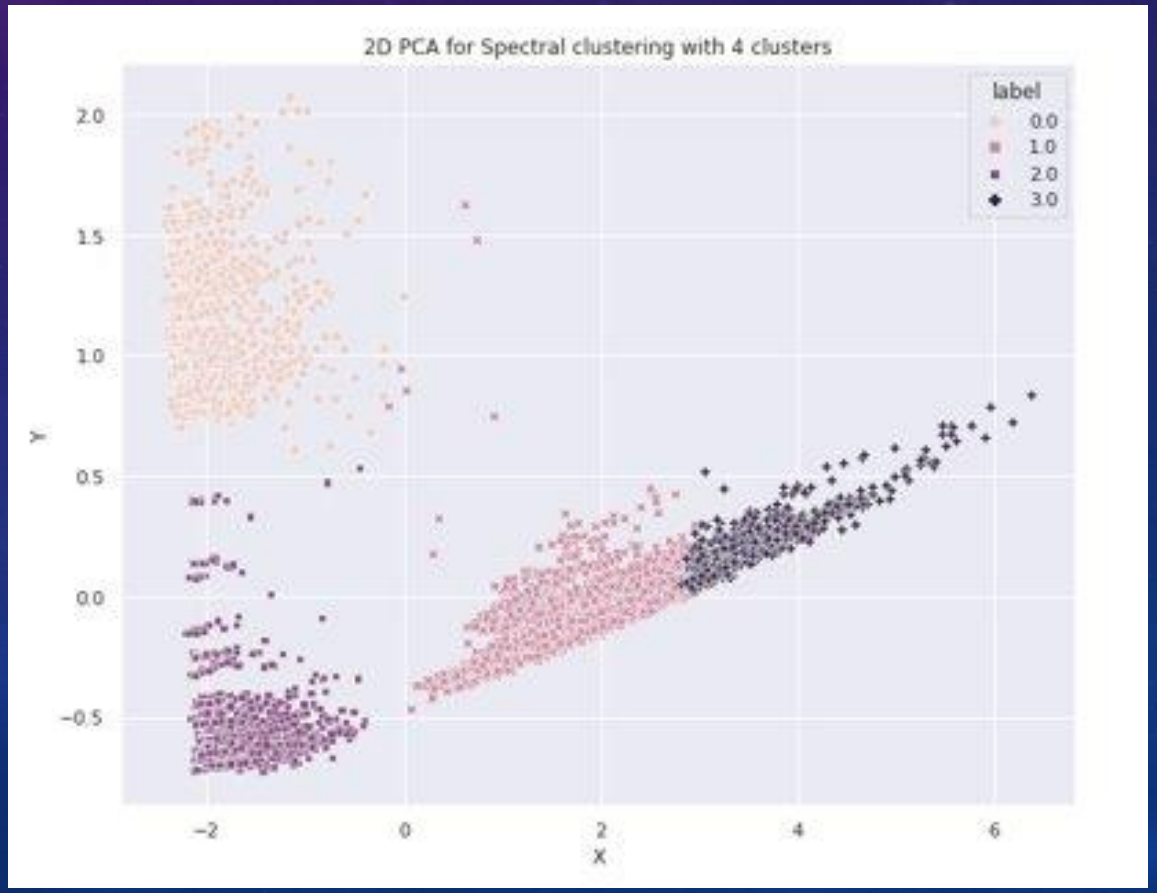
# SPECTRAL CLUSTERING

- Score'y oprócz silhuette porównują klastrowanie z labelami od aktywności
- 3, 4 klastry zdają się być preferowane,
- Silhuette początkowo maleje wraz ze zwiększaniem klastrów,
- Wartość n_neighbors zdaje się nie wpływać na jakość "dobrych" klastrowań.
- Największe "podobieństwo" do czynności jest dla klastrowania na 3, 4.

# SPECTRAL PCA

## 3 klastry


## 4 klastry



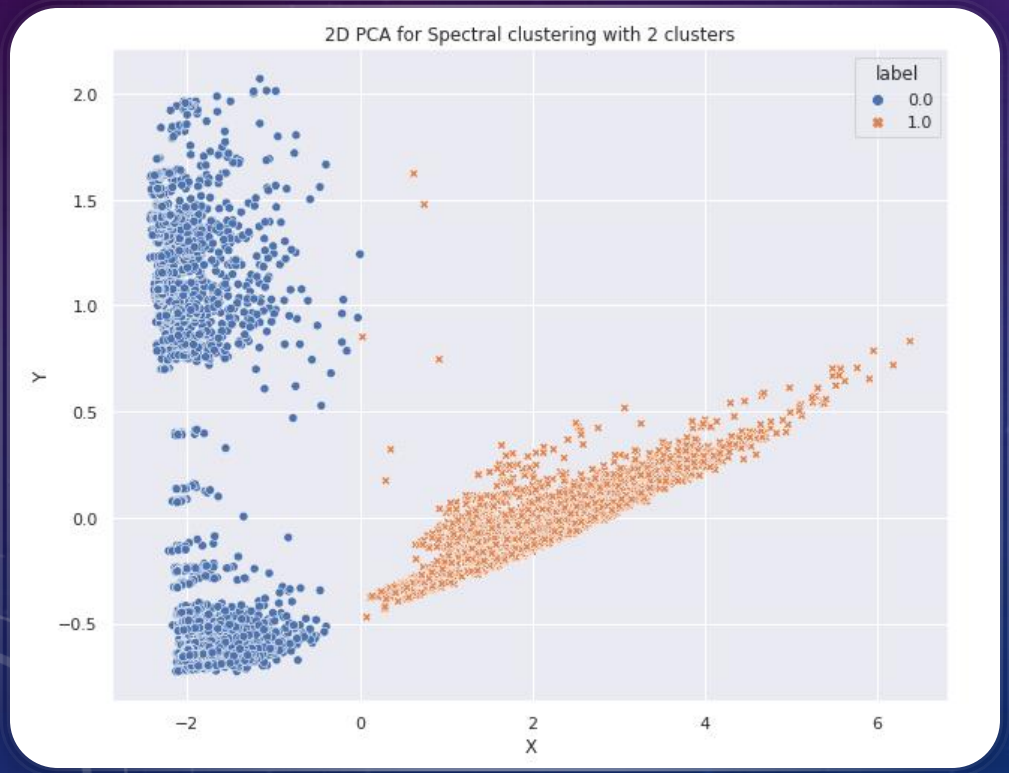
# SPECTRAL TSNE

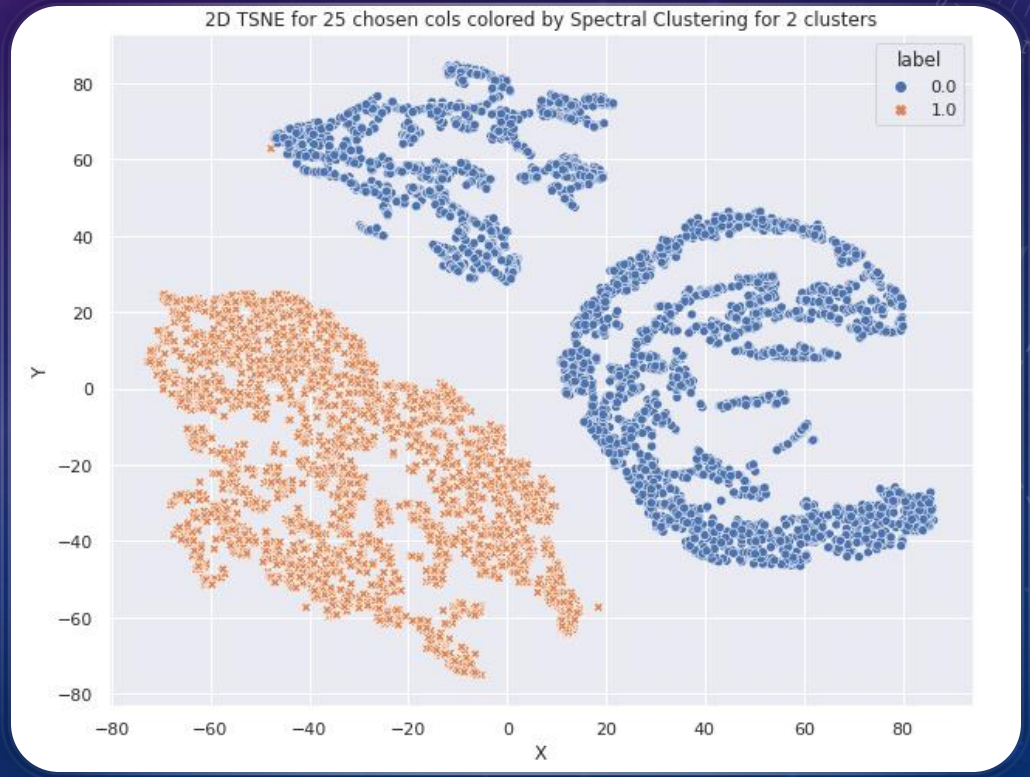
3 klastry

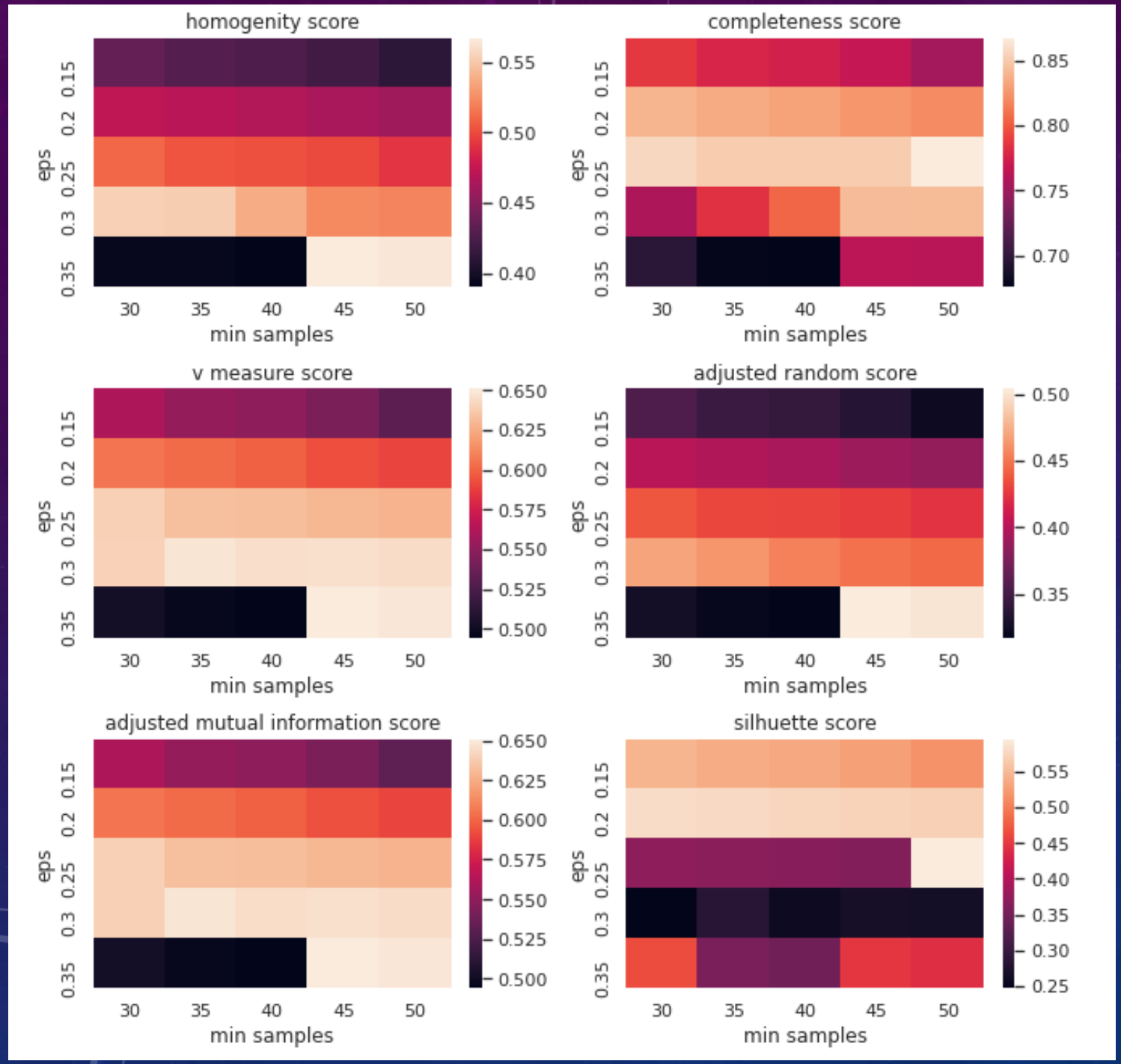

4 klastry

# WNIOSKI:

- Klastrowania na 3, 4 mają dobre PCA I tSNE.
- Klastrowanie na 4 klastry zdaje się widzieć część obserwacji "schodzenia w dół".
- Przypadek na 2 klastry nie daje ładnego PCA, czy tSNE.



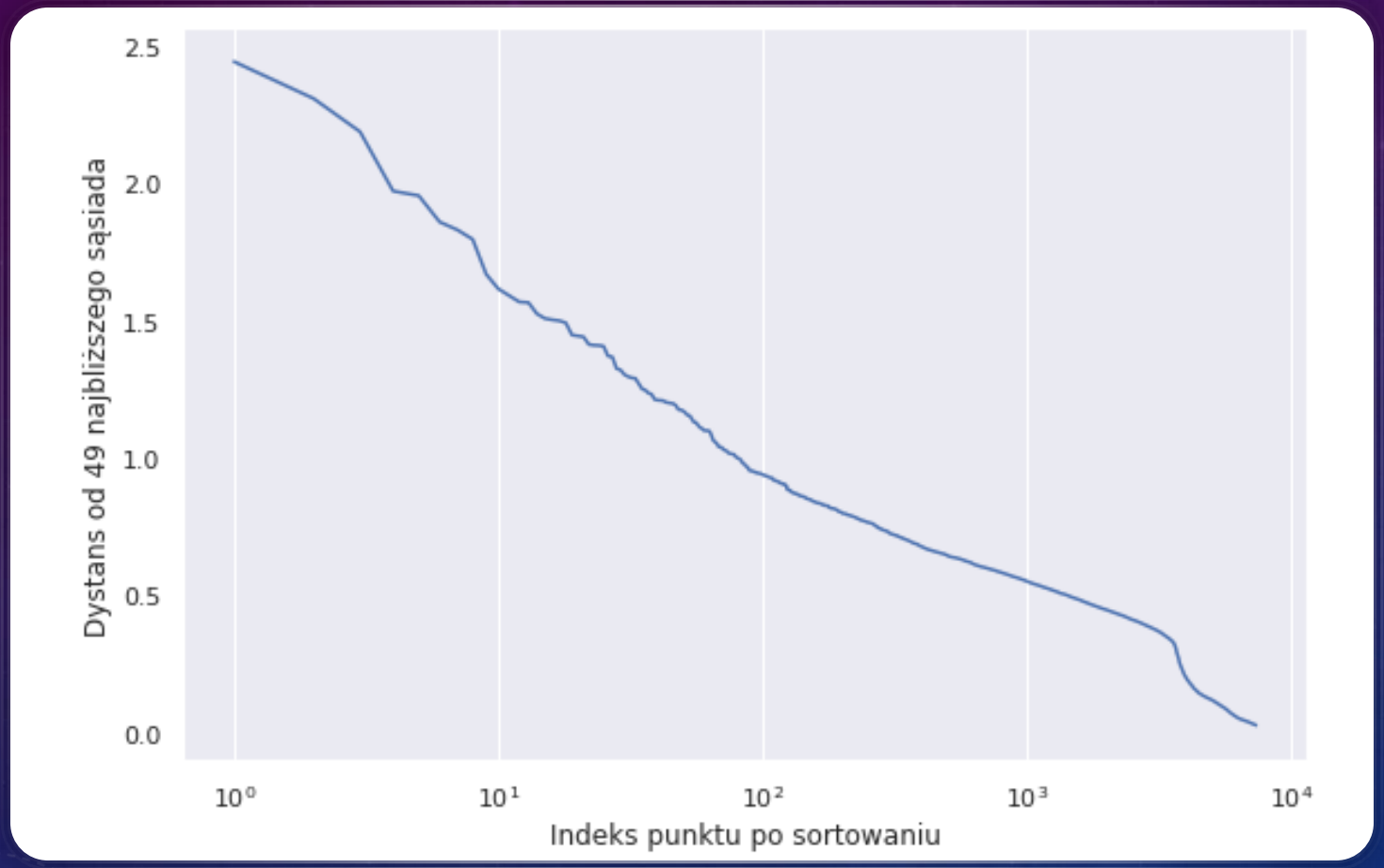






# DBSCAN

- Score'y oprócz silhuette porównują klastrowanie z labelami od aktywności
- Eps = 0.20, 0.25 (dla 50 min_samples) zdają sie być dobre
- Najlepsze Silhuette I podobieństwo do czynności mamy dla eps = 0.25 oraz min_samples = 50.

# TEORIA NIE ZAWODZI

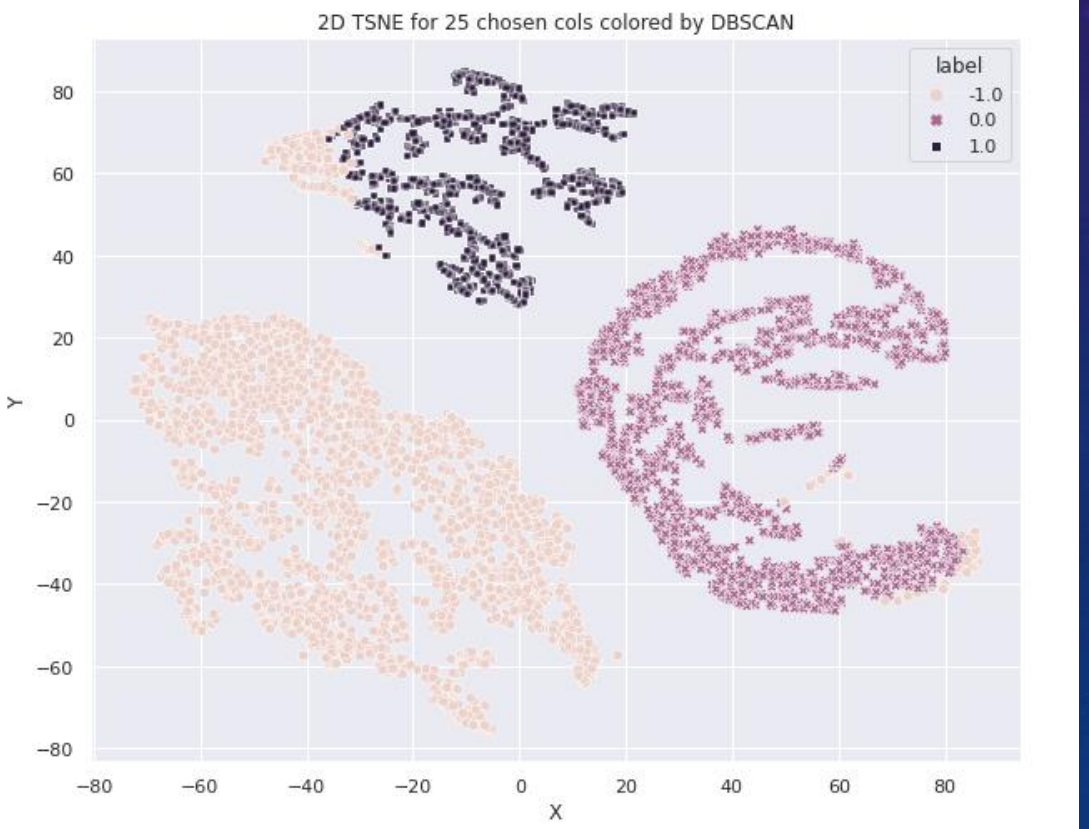



- Badamy dla 25 kolumn, wiec min_samples = 2(25) = 50
- Dystans do najbliższego sąsiada (eps) zdaje sie mieć ostatni łokieć mniej wiecej dla eps = 0.25

# PCA

# tSNE


2D PCA for 25 chosen variables for DBSCAN
label
-1.0
0.0
1.0
2.0
1.5
1.0
0.5
0.0
−0.5
Y
−2
0
2
4
6
X


2D TSNE for 25 chosen cols colored by DBSCAN
label
-1.0
0.0
1.0
80
60
40
20
0
−20
−40
−60
−80
Y
−80
−60
−40
−20
0
20
40
60
80
X

# WNIOSKI

- DBSCAN gorzej sobie radzi niż kMEANS, czy Spectral Clustering. Wytłumacznie tworzonych klastrów jest też znacznie trudniejsze.

# WYNIKI

- KMeans i Spectral Clustering okazały się porównywalnie dobre. DBSCAN wypadł gorzej.
- W uzyskanych klastrowaniach poszczególne klastry odpowiadają grupom czynności. Najbardziej wyróżnia się leżenie. Stanie i siedzenie często są w jednym klastrze. Chodzenie, chodzenie po schodach w górę i w dół przy n_clusters = 3 są w jednym klastrze. Możemy również wyróżnić Spectral Clustering na 4 klastry, w którym znacząca większość obserwacji w nowym, czwartym klastrze pochodzi z czynności schodzenia w dół.
- Dalsze zwiększenie liczby klastrów powoduje podzielenie grupy chodzenia, ale w tak powstałych klastrach nie mamy pełnej spójności etykiet. Nie znaleźliśmy algorytmu, który dzieliłby grupę chodzenia ze względu na etykiety. Stąd wynika, że pewne momenty chodzenia po schodach mogą być bardziej podobne do chodzenia po poziomej powierzchni, niż do innych momentów chodzenia po schodach.