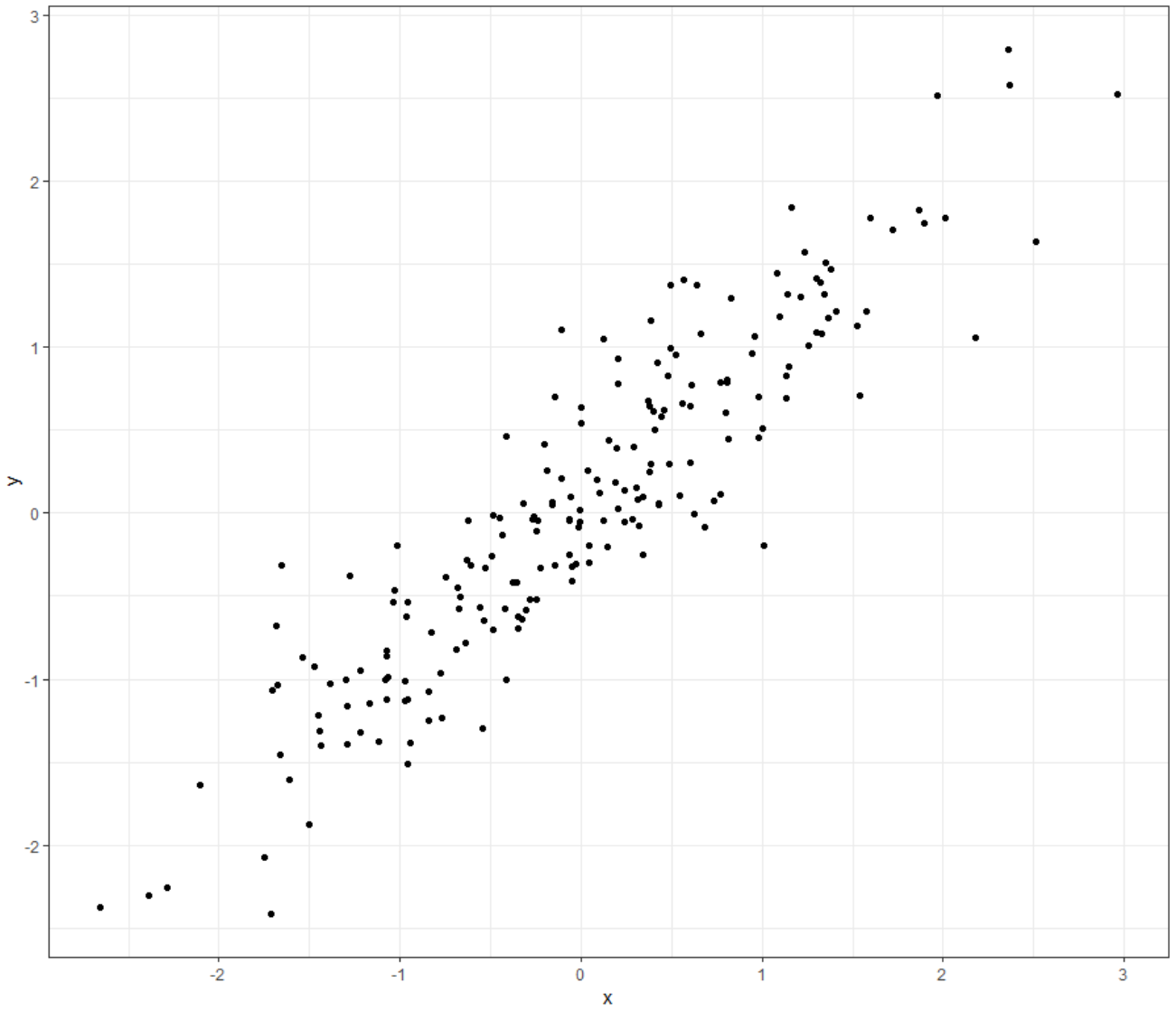
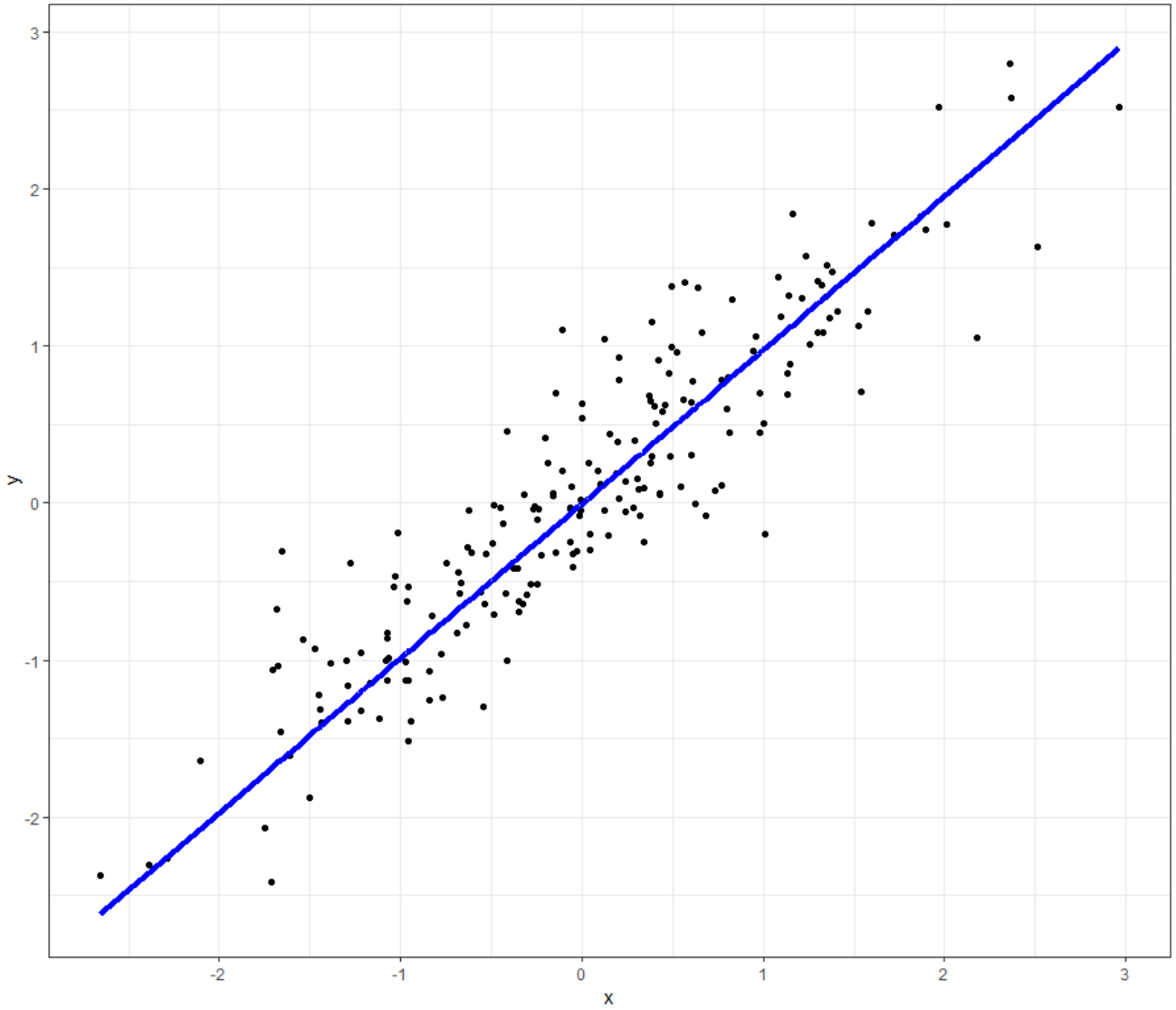
PCA



Problem:
- w wysokowymiarowej przestrzeni dane mogą wyglądać podobnie jak w tym niskowymiarowym przykładzie:

część wymiarów może mieć duży rozrzut (oś x), a część mały (y)



Rozwiązanie:

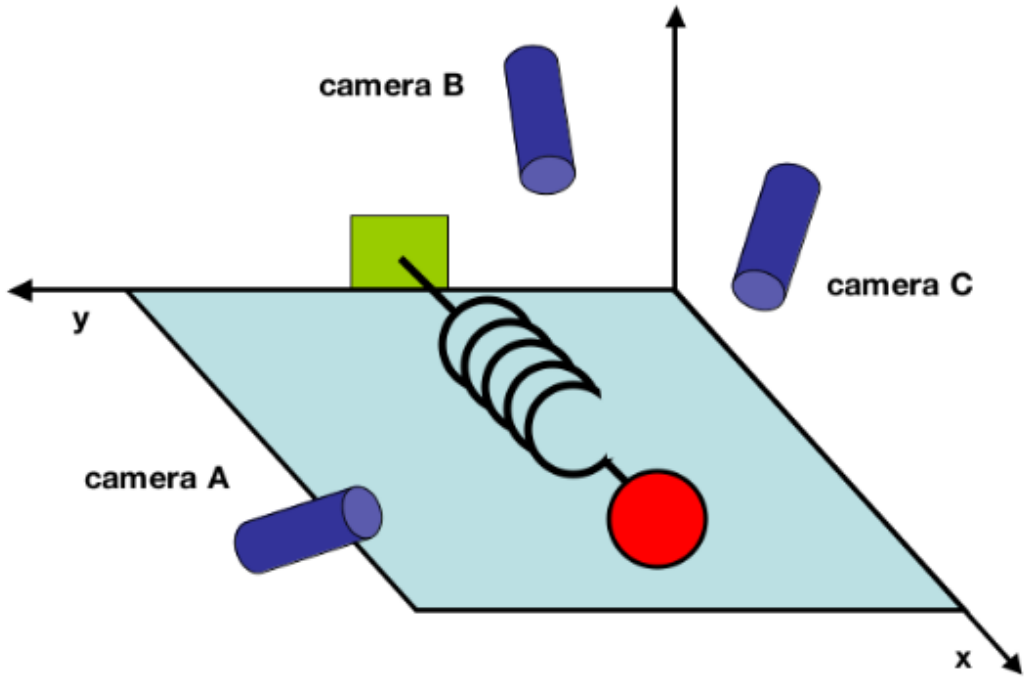
-możemy zredukować opis danych do mniejszej liczby cech.

-> Jak znaleźć nowe cechy?

-> Przekształcenie liniowe: rzut

Jak znaleźć dobrą przestrzeń do rzutu?

- W przykładzie: widać, że na osi y dane mają mały rozrzut, więc intuicyjnie wiemy, że w ten kierunek nie jest ważny.
- Czyli warto szukać kierunków, w których dane mają największy rozrzut.
- Analogia z cieniem



https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf

Rozwiązanie

With this assumption *PCA* is now limited to re-expressing the data as a *linear combination* of its basis vectors. Let $\mathbf{X}$ and $\mathbf{Y}$ be $m \times n$ matrices related by a linear transformation $\mathbf{P}$. $\mathbf{X}$ is the original recorded data set and $\mathbf{Y}$ is a re-representation of that data set.

$$\mathbf{PX} = \mathbf{Y} \tag{1}$$

Also let us define the following quantities[2].

- $\mathbf{p_i}$ are the *rows* of $\mathbf{P}$
- $\mathbf{x_i}$ are the *columns* of $\mathbf{X}$
- $\mathbf{y_i}$ are the *columns* of $\mathbf{Y}$.

- Zakładamy, że P jest ortonormalna
- P jest rzutem

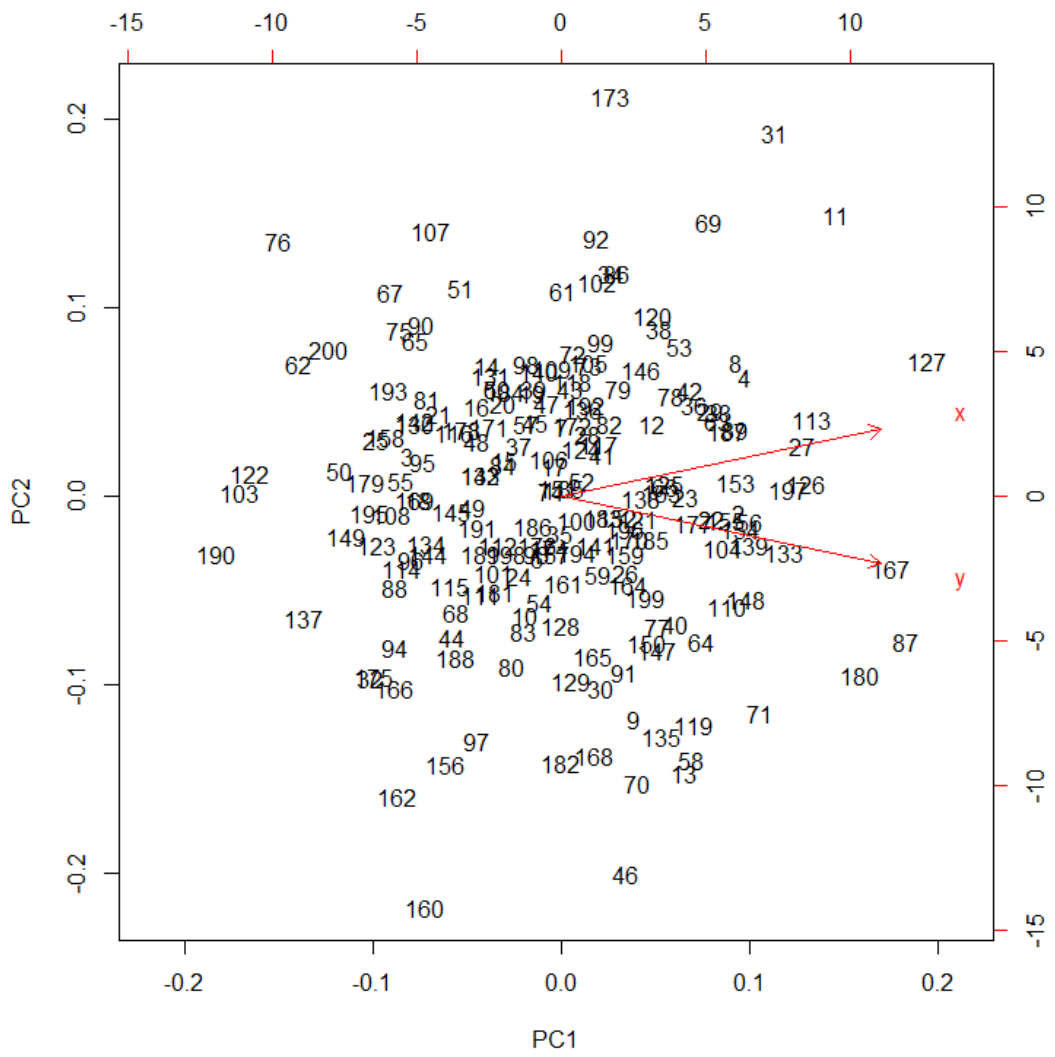
$$\begin{aligned}
\sigma^2_{\vec{w}} &= \frac{1}{n}\sum (\vec{x_i} \cdot \vec{w})^2 \\
&= \frac{1}{n}(\mathbf{Xw})^T(\mathbf{Xw}) \qquad \frac{\partial u}{\partial w} = 0 = \frac{\partial f}{\partial w} - \lambda\frac{\partial g}{\partial w} \\
&= \frac{1}{n}\mathbf{w}^T\mathbf{X}^T\mathbf{Xw} \qquad \frac{\partial u}{\partial \lambda} = 0 = -(g(w) - c) \\
&= \mathbf{w}^T\frac{\mathbf{X}^T\mathbf{X}}{n}\mathbf{w} \qquad u = \mathbf{w}^T\mathbf{Vw} - \lambda(\mathbf{w}^T\mathbf{w} - 1) \\
&= \mathbf{w}^T\mathbf{Vw} \qquad \frac{\partial u}{\partial \mathbf{w}} = 2\mathbf{Vw} - 2\lambda\mathbf{w} = 0 \\
&\qquad\qquad \mathbf{Vw} = \lambda\mathbf{w}
\end{aligned}$$

Co dalej?

- Wniosek: najlepsza baza dla danych w tym problemie jest rozpięta przez wektory własne macierzy kowariancji.
- Ile wektorów należy wziąć?

$$R^2 \equiv \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

- Procent objaśnionej wariancji:
  (Pytanie pomocnicze: jaka jest wariancja nowych zmiennych?)
- Tradycyjne kryterium: wybór tylu składowych głównych, żeby błąd przybliżenia był mniejszy niż zadana wielkość = wyjaśniona wariancja większa niż ustalony odsetek.



Biplot

Zadania

- Przeprowadzić PCA z pomocą funkcji eigen
- Porównać wyniki z wynikami działania funkcji prcomp i princomp.
- Sprawdzić znaczenie elementów obiektów zwracanych przez te funkcje.
- Bonus: jaka jest korelacja między składową główną a oryginalną zmienną?

Sensowne materiały

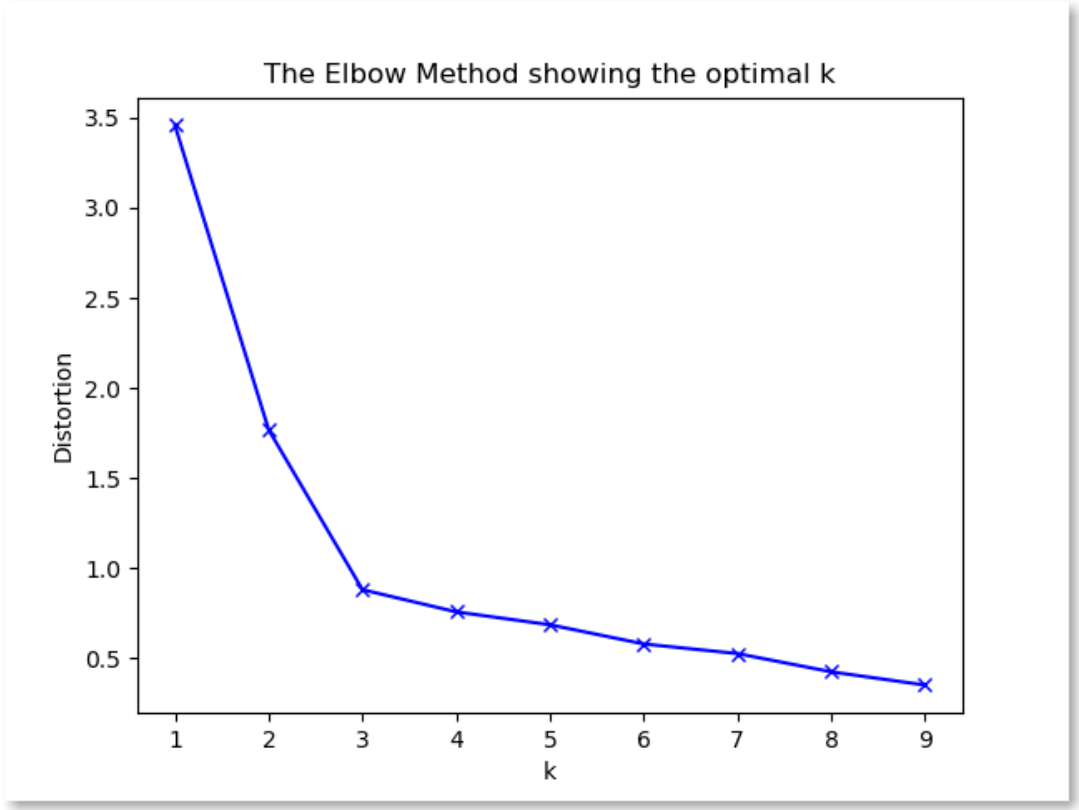
- http://stat.cmu.edu/%7Ecshalizi/350/lectures/10/lecture-10.pdf
- https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf
- http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

Analiza skupień

Liczba klastrów + ocena konkretnego pogrupowania

Elbow plot

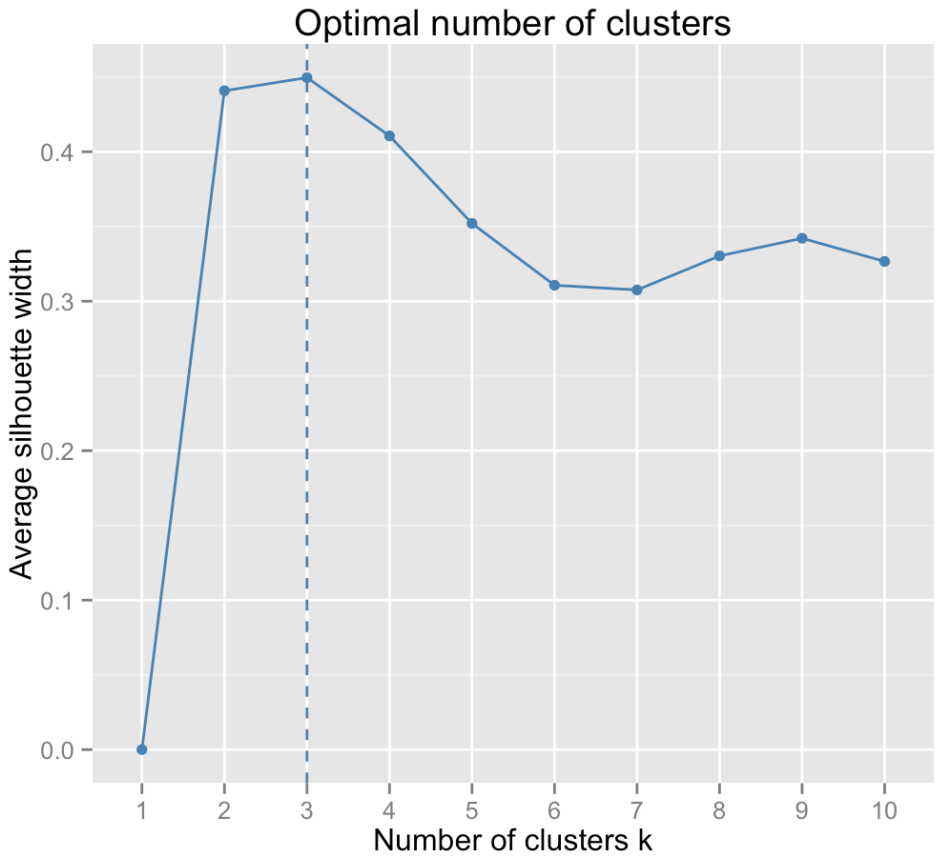
- Obliczyć klastry dla wielu wartości k
- Dla każdego k obliczyć łączną sumę odchyleń wewnątrz klastrów SSE
- Narysować wykres zależności SSE od k.
- Miejsce, od którego zmiany w SSE są malutkie uznajemy za dobrą liczbę klastrów



https://pythonprogramminglanguage.com/kmeans-elbow-method/

Average silhouette method

- Obliczyć podział na k klastrów dla wielu k.
- Obliczyć średnią wartość stat. silhouette dla każdego k (sil).
- Narysować wykres zależności sil od k.
- Najlepsza liczba klastrów jest położona w maksimum wykresu.



https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

Gap statistic

1. Cluster the observed data, varying the number of clusters from k = 1, ..., k_{max}, and compute the corresponding total within intra-cluster variation W_k.
2. Generate B reference data sets with a random uniform distribution. Cluster each of these reference data sets with varying number of clusters k = 1, ..., k_{max}, and compute the corresponding total within intra-cluster variation W_{kb}.
3. Compute the estimated gap statistic as the deviation of the observed W_k value from its expected value W_{kb} under the null hypothesis: $Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}^*) - log(W_k)$. Compute also the standard deviation of the statistics.
4. Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: $Gap(k) \geq Gap(k+1) - s_{k+1}$.

https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#silhouette-coefficient

Silhouette statistic

For each observation i, the silhouette width s_i is calculated as follows:

1. For each observation i, calculate the average dissimilarity a_i between i and all other points of the cluster to which i belongs.

2. For all other clusters C, to which i does not belong, calculate the average dissimilarity $d(i, C)$ of i to all observations of C. The smallest of these $d(i, C)$ is defined as $b_i = \min_C d(i, C)$. The value of b_i can be seen as the dissimilarity between i and its "neighbor" cluster, i.e., the nearest one to which it does not belong.

3. Finally the silhouette width of the observation i is defined by the formula: $S_i = (b_i - a_i)/max(a_i, b_i)$.

https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#silhouette-coefficient

Dunn index

1. For each cluster, compute the distance between each of the objects in the cluster and the objects in the other clusters
2. Use the minimum of this pairwise distance as the inter-cluster separation (*min.separation*)
3. For each cluster, compute the distance between the objects in the same cluster.
4. Use the maximal intra-cluster distance (i.e maximum diameter) as the intra-cluster compactness
5. Calculate the *Dunn index* (D) as follow:

$$D = \frac{min.\,separation}{max.\,diameter}$$

https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#silhouette-coefficient

Davies-Bouldin index

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \text{ where}$$

$$R_i = \max_{j=1\ldots nc, i\neq j}\left(R_{ij}\right), \; i = 1\ldots n_c$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$d_{ij} = d\left(v_i, v_j\right), \; s_i = \frac{1}{\|c_i\|}\sum_{x\in c_i} d\left(x, v_i\right)$$

Where,

- d(x,y) is the Euclidean distance between x and y.
- c_i is the cluster i.
- v_i is the centroid of cluster c_i
- $\|c_i\|$ refers to the norm of c_i

https://www.hackerearth.com/problem/approximate/davies-bouldin-index/