Session Clustering with K-Means Algorithm

Elika Araghi (s181740)
Andri Bergsson (s150843)
Emmanouil Chalvatzopoulos (s182324)

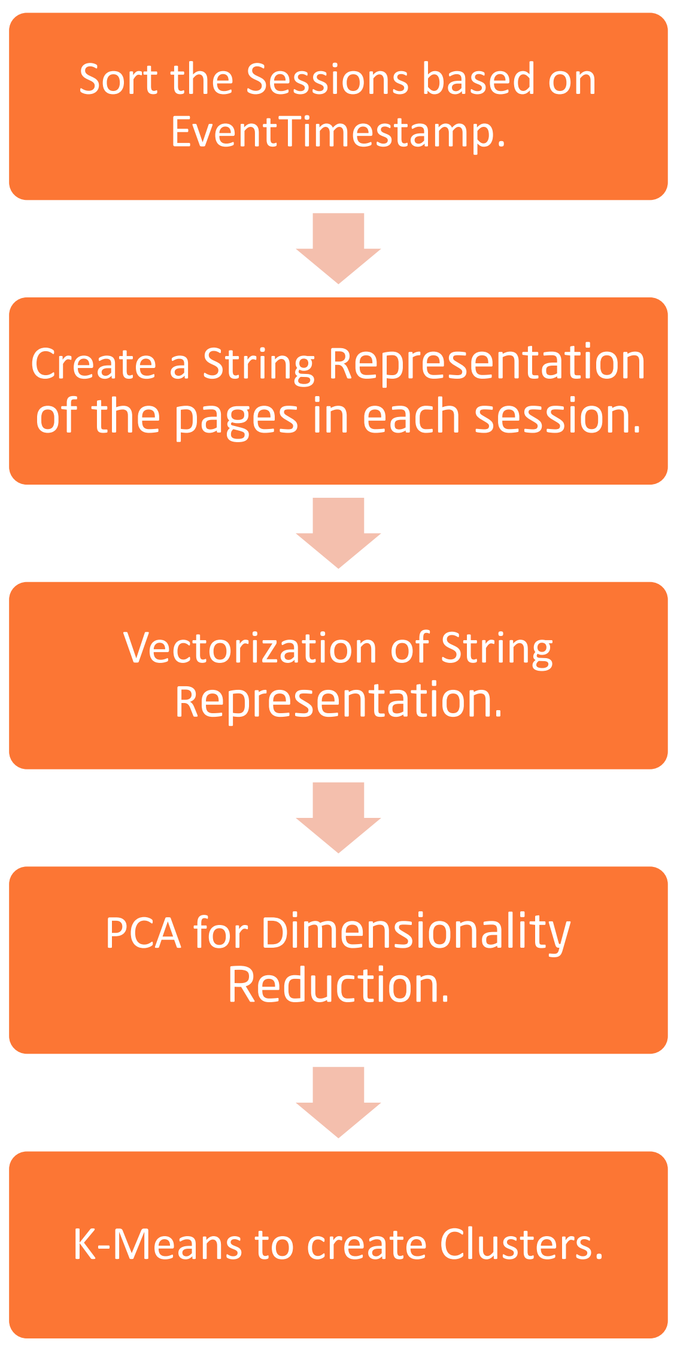
DTU Compute, Technical University of Denmark

Introduction

By applying methods and algorithms learnt in the course “Computational Data Analysis” on a data set of customer web page data provided by Danske Bank, we managed to gain a useful insight of how the customers navigate and interact with Danske Bank’s website. We created clusters to visualize different groups of customers and we provide the methodology and the visualization of our findings.

Methodology

Our main focus was to cluster the customers based on their actions during their visits to the website of Danske Bank. To do that we firstly ordered the sessions by their timestamp, then we created a string representation of the pages that each session contains. Afterwards we vectorized these string representations in order to get a numerical valued vector that represents each session. As a result we ended up with a lot of dimensions as the data set was rather large, so we ran PCA to achieve dimensionality reduction. Finally we used the K-Means algorithm to create clusters based on the preprocessing we did on the data.

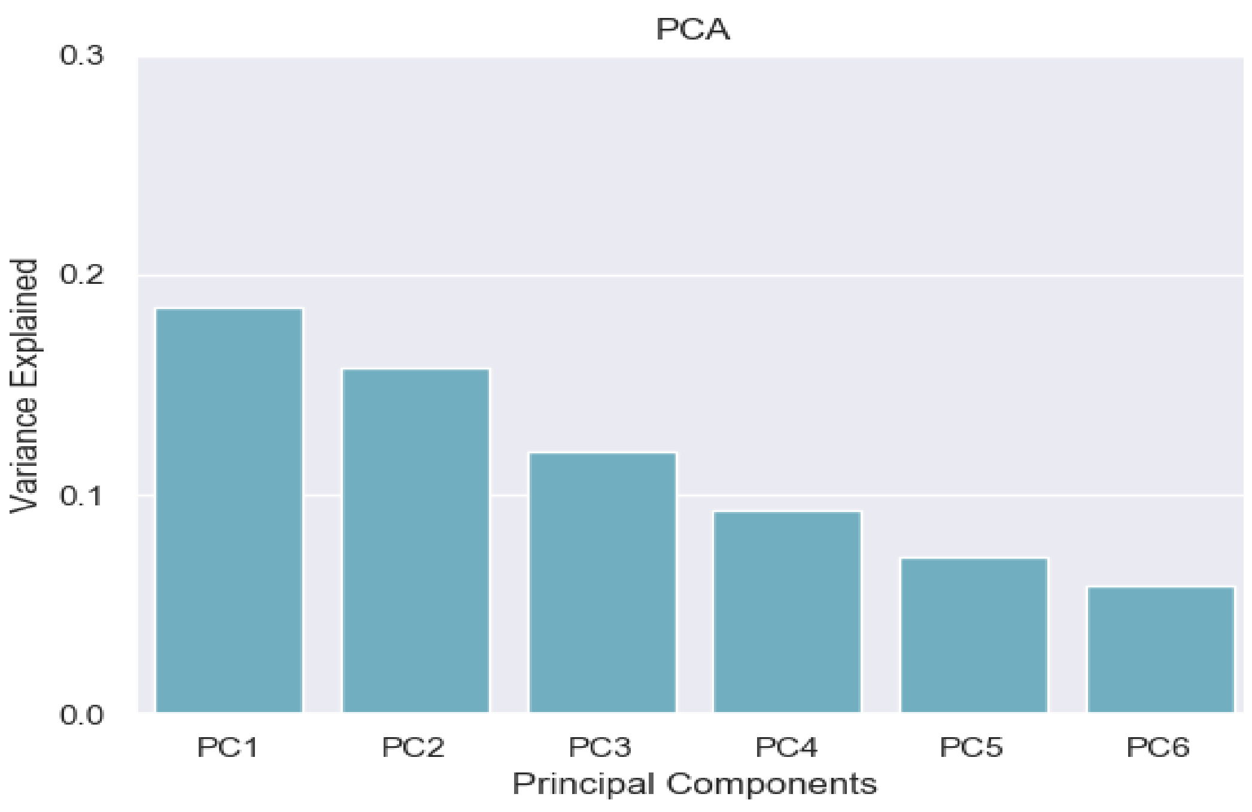

Sort the Sessions based on EventTimestamp.

Create a String Representation of the pages in each session.

Vectorization of String Representation.

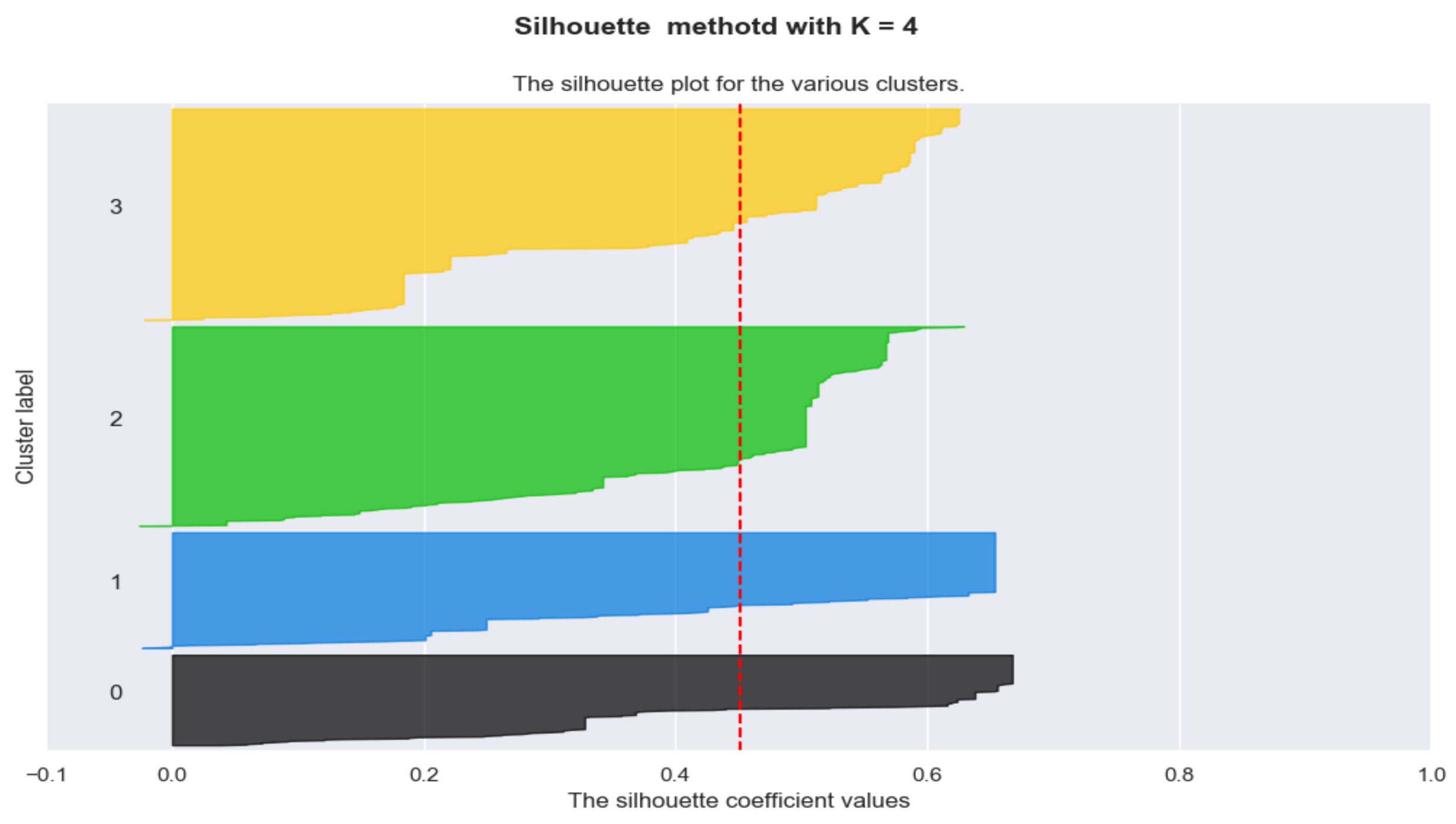
PCA for Dimensionality Reduction.

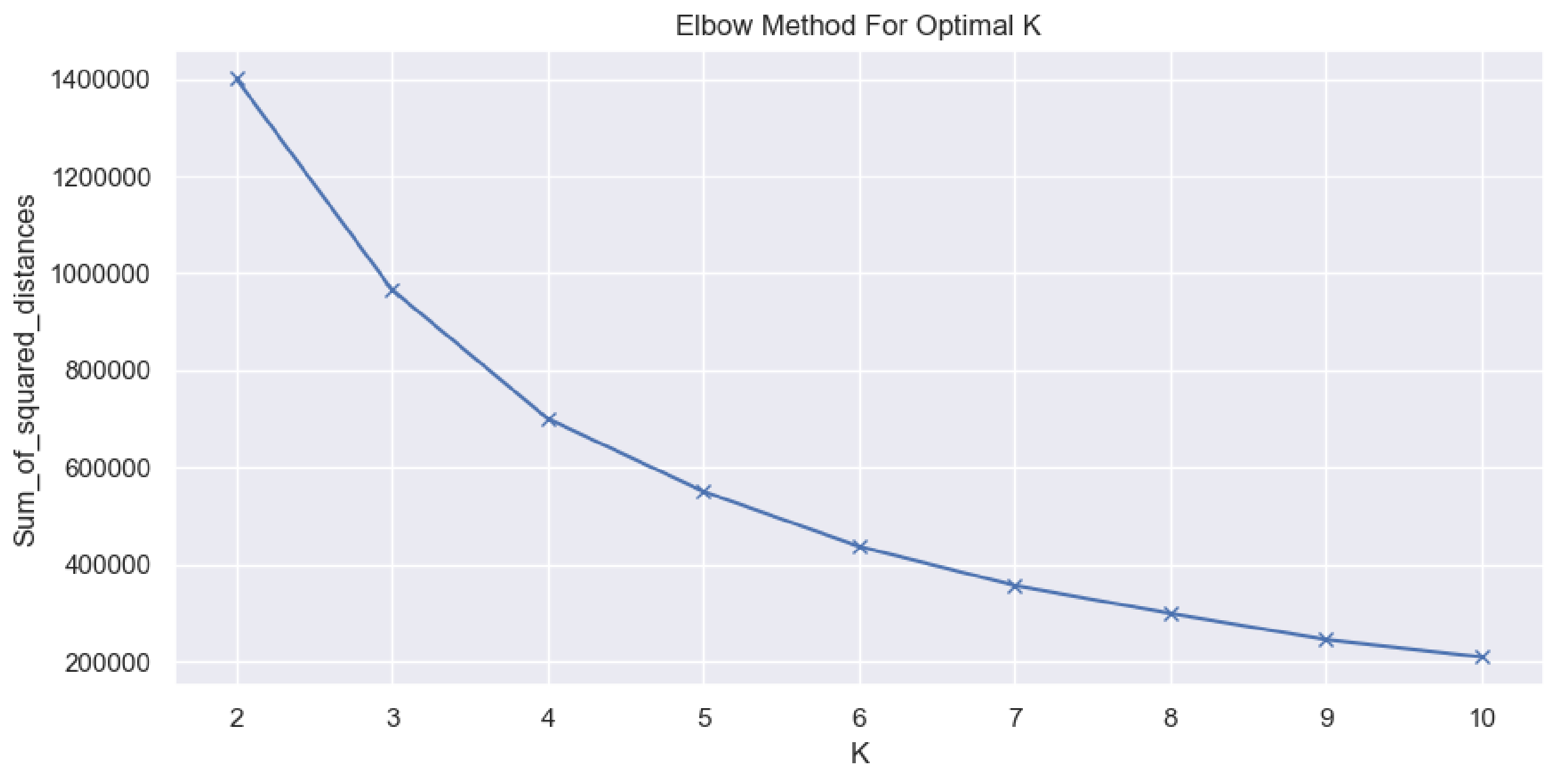
K-Means to create Clusters.



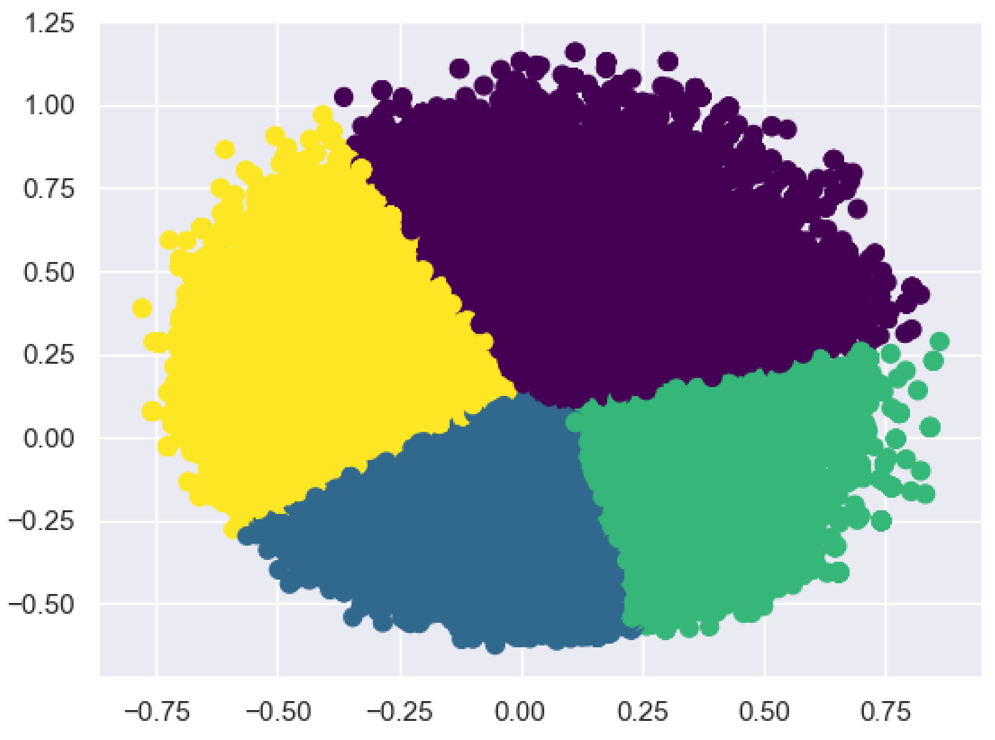
Number of Clusters

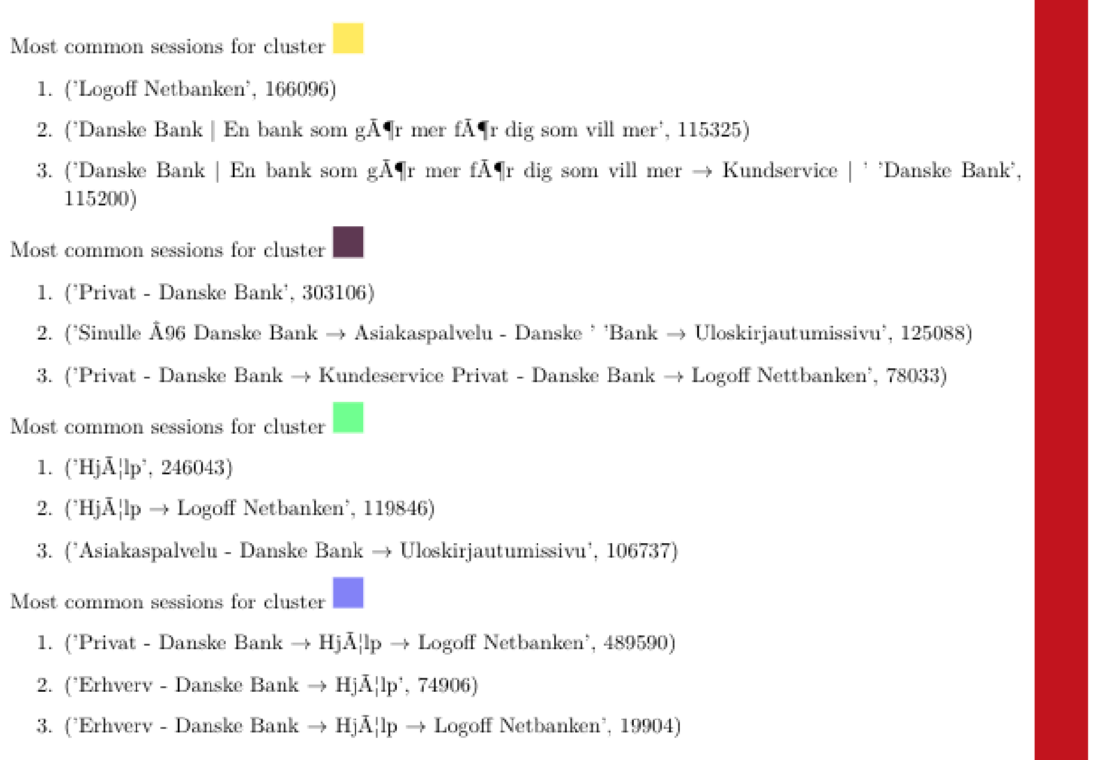
To determine the optimal number of clusters, the Elbow Method and the Silhouette Method were used. We concluded the optimal number of clusters to be K=4, as can be seen by the plots below. The main advantages using K-means are it’s efficiency in large datasets and it’s easy to implement while some major disadvantages are that it’s computationally expensive and it lacks consistency.





Visualization




Most common sessions for cluster 🟨

1. ('Logoff Netbanken', 166096)
2. ('Danske Bank | En bank som gÃ¶r mer fÃ¶r dig som vill mer', 115325)
3. ('Danske Bank | En bank som gÃ¶r mer fÃ¶r dig som vill mer → Kundservice | ' 'Danske Bank', 115200)

Most common sessions for cluster 🟪

1. ('Privat - Danske Bank', 303106)
2. ('Sinulle Â96 Danske Bank → Asiakaspalvelu - Danske ' 'Bank → Uloskirjautumissivu', 125088)
3. ('Privat - Danske Bank → Kundeservice Privat - Danske Bank → Logoff Nettbanken', 78033)

Most common sessions for cluster 🟩

1. ('HjÃ¦lp', 246043)
2. ('HjÃ¦lp → Logoff Netbanken', 119846)
3. ('Asiakaspalvelu - Danske Bank → Uloskirjautumissivu', 106737)

Most common sessions for cluster 🟦

1. ('Privat - Danske Bank → HjÃ¦lp → Logoff Netbanken', 489590)
2. ('Erhverv - Danske Bank → HjÃ¦lp', 74906)
3. ('Erhverv - Danske Bank → HjÃ¦lp → Logoff Netbanken', 19904)