

Logo

Description automatically generated

Cluster Analysis on LinkedIn Dataset

By

Harshad Subhash - 100345115

Jaspreet Singh - 100343522

Kirandeep Kaur - 100344808

PDD Data Analytics

Langara College

A picture containing drawing

Description automatically generated

Table of Content

|  |  |
| --- | --- |
| 1. Background | 2 |
| 1. Introduction | 3 |
| 1. Descriptive Analysis | 3 |
| 1. Data Reduction | 8 |
| 1. Cluster Analysis | 12 |
| 1. Interpretation | 19 |
| 1. Conclusion | 20 |
| 1. Limitation and Acknowledgement | 20 |

# Background

This study explores a LinkedIn dataset that contains profile information, job history and facial features. Preprocessing the dataset, and understanding the features of this dataset, we used various techniques of dimension reduction to find the key variables, and understand what factors impact the profile of an individual. Cluster analysis is conducted on the reduced dataset that will help us understand certain aspects on how we can improve our engagement on LinkedIn and make our profile stand out for better employment/business opportunities.

LinkedIn is a professional social networking platform that has been around longer than Facebook, Twitter, and Instagram. It is more relevant now than ever, as it has more than 700 million professional profiles, which means unlimited supply of network connections and job opportunities.

Individuals can leverage this platform to build their professional brand online, and think about following questions:

* How do you stand out from others in your industry?
* What makes you marketable?
* Why should someone pay you six figures?

This tool offers a simple and effective way to put you name in the professional map, especially for students completing their education overseas. An individual need to upload a welcoming professional picture, write a professional summary and emphasise on strengths and show case their personality. When hiring manager peruse a profile, they should be able to gain a strong understanding of that individual, and the skills he brings up to the table.

Our main objective is to analyse a LinkedIn profile dataset and segment users based on their profile information. Cluster Analysis is performed on this data set after preprocessing activities and dimension reduction to gain insight on different segments of LinkedIn users.

# Introduction

LinkedIn is the most prominent social networking platform for professionals. It has been a great platform to connect employers with potential employees. It is the steppingstone for every graduate to start a career or an experienced professional to change jobs. It is a great platform to present an individual’s acquired skills, experience, and talent. LinkedIn is a safe place for professionals to connect with like minded people. According to Wikipedia, ” LinkedIn was launched on May 5, 2003, it is mainly used for professional networking, including employers posting [jobs](https://en.wikipedia.org/wiki/Job) and job seekers posting their [CVs](https://en.wikipedia.org/wiki/Curriculum_vitae). As of 2015, most of the company's revenue came from [selling access to information about its members](https://en.wikipedia.org/wiki/Information_broker) to recruiters and sales professionals. Since December 2016, it has been a wholly owned subsidiary of [Microsoft](https://en.wikipedia.org/wiki/Microsoft). [[1]](#footnote-2)As of 2015, LinkedIn had more than 400 million members in over 200 countries and territories. [[2]](#footnote-3)As of May 2020, LinkedIn had 706 million registered members”.

DescriptiveAnalysis

The dataset mainly consists of anonymized data scraped from LinkedIn. The dataset was obtained from Kaggle and is open for public usage. Our dataset has 62709 observations and 52 attributes. The dataset is categorized into 4 aspects: **Profile data, Job Data, Name analysis and Profile Picture analysis.**

Profile data

|  |  |  |
| --- | --- | --- |
| Variable | Class | Description |
| X | Nominal, Discrete, Numerical | Index of the dataset |
| m\_urn | Nominal | Unique reference number of the Individual |
| n\_followers | Numeric, discrete, interval | Number of followers |
| Age | Numeric, discrete, ratio, normally distributed, nominal | 13 is the legal age to work in Australia, all profiles below 13 will not be considered for the analysis.  Range Count  13-22 497  23-32 2372  33-42 4829  43-52 5140  53-62 1919  63-73 450 |

Note:

Age: Most profiles are seen been age group of 30 – 50 years. As shown in the Histogram below.

Figure1: Age distribution of population

Chart, histogram

Description automatically generated

Job Data

|  |  |  |
| --- | --- | --- |
| Variable | Class | Description |
| avg\_n\_pos\_per\_prev\_tenure | Ratio, Continuous, Numerical | Average number of positions per person when they start using LinkedIn  Range: 1-15 |
| avg\_pos\_len | Numerical, continuous, ratio | Average days worked among all jobs the candidate has done |
| avg\_prev\_tenure\_len | Numerical, continuous, ratio | Average days worked in last 10 years |
| c\_name | Categorical, Nominal | Company name |
| n\_prev\_tenures | Numeric, discrete, right skewed, nominal | Number of jobs worked in the last 10 years |
| tenure\_len | Numeric, discrete, right skewed, nominal | Days spent in current job |
| n\_pos | Numeric, discrete, nominal | Number of positions in the candidate’s career  Range: 1-10 |

Note: WordCloud of company names

Figure2: Frequency distribution of company types

Text

Description automatically generated

Name Analysis

|  |  |  |
| --- | --- | --- |
| Variable | Class | Description |
| Gender | Categorical, nominal | 11,603 Male and 3,565 Female |
| Ethnicity: White, Black, Asian | Numeric, continuous, ratio | 11,439 White, 1225 Black and 2594 Asian candidates |
| Nationality (Region): African, Celtic\_English, East\_Asian, South\_Asian, Hispanic, Nordic, Greek, Jewish, European, Muslim (Middle East) | Numeric, continuous, ratio | Which region do these candidates come from: 120 African, 5760 Celtic\_English, 1549 East\_Asian, 2577 South\_Asian, 144 Hispanic, 331 Nordic, 74 Greek, 51 Jewish, 2433 European, 889 Muslim (Middle East) |

Note:

Figure3: Gender distribution based on race. Figure4: Pie chart distribution of nationality

Chart, bar chart

Description automatically generatedChart, pie chart

Description automatically generated

Profile Picture Analysis

|  |  |  |
| --- | --- | --- |
| Variable | Class | Description |
| Image | Categorical, nominal | The profile picture of the LinkedIn user |
| beauty | Numeric, continuous, interval, normally distributed | Picture quality. Aesthetic attribute of an individual seen in a picture. |
| beauty\_male | Numeric, continuous, interval, normally distributed | Picture quality. Aesthetic attribute of an individual seen in a picture for males |
| beauty\_female | Numeric, continuous, interval, normally distributed | Picture quality. Aesthetic attribute of an individual seen in a picture for females |
| Blur, Blur\_guassian, Blur\_motion | Numeric, continuous, ratio, right skewed | The Blur percentage of an image |
| Glass | Categorical, nominal | Wearing glasses or not. |
| Emotions: Anger, Fear, Happiness, Neutral, Sadness  Surprise (individual column for each variable) | Numeric, continuous, interval | Percentage of which emotion is seen from the picture. |
| Face quality | Numeric, continuous, interval | Measuring the structure of the face using the golden ratio rules and different parameters |
| Head details: Head\_pitch, head\_roll, head\_yaw | Numeric, continuous, normal distribution | 3-D orientation of the head structure of a candidate |
| Mouth details: mouth\_close, mouth\_mask, mouth\_open, mouth\_other | Numeric, continuous, ratio | Percentage of mouth (lips and teeth) seen in the picture |
| Skin details: Skin\_dark, skin\_circle,skin\_stain, skin\_acne, skin\_health, skin\_dark\_circle | Numeric, continuous, ratio | Skin QoL score details derived from image analysis |
| Smile | Numeric, continuous, ratio | Smile score derived from examining posterior and buccal corridor to define smile esthetics. |

Note:

Emotion Detection Face Analysis

Figure5: Face Analysis using an Image Detection tool by Noldus[[3]](#footnote-4)

Graphical user interface, application

Description automatically generated

Blur: Blurriness of the image can be detected using python. OpenCv in python[[4]](#footnote-5) is used to detect blurriness or use blur detection schemes like Wavelet Transform[[5]](#footnote-6).

Smile: Image Analysis done on various muscles of the face region. The smile was split it different frames, in each frame analysed, specific measures on the teeth and soft tissue could be assessed. Gaussian Derivatives[[6]](#footnote-7) are obtained based on pixels of the image.

Beauty: Assessment of Facial Beauty based on non-permanent and acquisition characteristics. From facial aesthetics to make up, the rating is based on how an individual is seen in the image. Golden ratio, symmetry of face and skin health[[7]](#footnote-8) information can be deduced from an image.

Data Cleaning

Total Number of observations before cleaning and preprocessing the data: 62,709

This dataset is large, with 62,709 observations and 52 columns, it is evidence that we have a significant amount of information available in our analysis. In our initial findings, we discovered that Blur column was repeated with 2 different labels and hence we choose to keep only 1 of them. We also noticed the beauty of the candidates is mentioned again in different columns, hence we had to get rid of them, to reduce redundancy. Company name, image and glasses have been excluded from our analysis, as they do provide any input to our analysis. We consider candidates only over age 13 as it is the legal age to work in Australia.

Once we have decided the features to consider, we scale the data and use Mahalanobis function to remove outliers from the dataset. At 1% significance level, Chi-square results show that the cut-off score obtained was **84.03713**

Total Number of observations after cleaning and preprocessing the data: 10,610

|  |  |
| --- | --- |
| Actions Performed | Dimension |
| Original data | 62,709 observations and 52 variables |
| Remove duplicate columns | 62,709 observations and 47 variables |
| Remove rows with average position length negative and age less than 13 | 62,623 observation and 47 variables |
| Extract numerical variables | 62,623 observations and 41 variables |
| Groupby data according to URN number as we had multiple observations for same URN number | 15,237 observations and 40 variables |
| Remove outliers with Mahalanobis distance and remove URN number variable as it is not significant for clustering (Alpha =0.01) | 10,610 observations and 40 variables |

Dimension Reduction

The methods of reducing dimensionality of the feature space can be grouped into two categories:

1. Feature Selection Methods
2. Feature Extraction Methods

The feature selection method reduces the dimensionality by selecting a subset of the original feature set, whereas feature extraction methods reduce dimensionality by transforming the dimensional space and producing new values.

The feature selection methods reduce the computational complexity by not computing those features which are not in the reduced feature set, whereas in feature extraction methods all the features are computed before the dimensionality reduction is performed through transformation.[[8]](#footnote-9)

Principal Component Analysis

Our dataset originally had 52 features and more than 60k+ observations, and it was quite difficult to run various clustering methodology both computationally and statistically, so we need to perform dimension and reduction, and principal component analysis is a powerful tool in reducing dimensions in correlated feature sets.

Principal components are nothing but the underlying structure of the data, they are the directions where there is most variance, and directions where the data is most spread out, which means we try to find straight line that best spreads the data out when it is projected along with it, this becomes the first principal component, the straight line that shows the most substantial variation in the dataset, likewise we find second, third and n number of components. The idea is to approximate most of the complexity in the dataset with just few principal components.

We have performed PCA on scaled numeric data and found components that were significant in explaining variance in the original dataset.

Figure5: Plot explaining variance in PCA

Chart, histogram

Description automatically generated

The above Scree plot shows that the first three components explain 25.49% of variance of original data and after 3 components, addition of another component does not enhance the variation explained by a significant value. Hence, first three principal components are important for further analysis of LinkedIn profiles.

For the confirmation of above results, we ran an elbow line plot as well. The Elbow plot shows a large steep/elbow at a component which is considered as the last important principal component.

Figure5: Elbow Line plot showing the proportion of variance

Chart, line chart

Description automatically generated

The above plot shows the elbow at 3rd component and the addition of another component does not significantly increase the proportion of variation explained. Hence, we decide to continue our study with 3 components.

The Significant Components

After we decided to have four principal components, we found out the features that were loaded into them.

PC1 Smile, Neutral emotion, Happiness emotion, Mouth open, Mouth close

PC2 Age, Beauty, Skin health, Skin stain, Tenure length.

PC3 Tenure length, Face quality, Blurriness, Average days/position, Number of positions.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Facial Expressions with mouth gestures

Age and Beauty Factor

Chart, bar chart

Description automatically generated

Experience and Picture Quality

Factor Analysis

Factor analysis consist of procedures for analysing the relations among the set of random features, the main purpose of FA is to account for intercorrelations among n variables, by postulating a set of common factors, considerably fewer in number that the original feature set.8 Factors are the variables which can not be measured directly, hence they are referred to as latent variables.

Factor analysis can also be used for dimension reduction. This technique helps us to find latent variables. The following correlation matrix helps us to estimate the clusters within the dataset. If we would not find any correlated groups in the matrix, we can infer that there are no clusters available in data but here, we can see some correlation visible within the dataset.

Figure9: Correlation plot

Chart, scatter chart

Description automatically generated

We ran parallel analysis on correlated variables of LinkedIn dataset and found latent variables with factor analysis.

Figure10: Parallel Scree Plot representing FA

Chart

Description automatically generated

The above diagram shows that there are 3 latent variables to which we can reduce LinkedIn dataset.

Cluster Analysis

Accessing Clustering Tendency:

Before applying any clustering method to the dataset, it is important to evaluate if the dataset contains meaningful clusters or not. This process is described as accessing the clustering tendency or the feasibility of clustering analysis.

We need to perform this crucial step before applying any type of clustering methodologies because they will always return clusters even if the data does not contain any clusters.

This tendency can be accessed by using[[9]](#footnote-10):

1. Visual Methods
2. Hopkins Test

Visual Method

Figure11: Dissimlarity Matrix



For the visual assessment of the clustering tendency, we shall compute and plot the dissimilarity matrix, red means high similarity and blue means low similarity.

The color level is proportional to the value of dissimilarity between observations, this dissimilarity matrix image confirms there is a cluster structure in the dataset.

Hopkins Test

The Hopkins statistic is used to assess the clustering tendency of a data set by measuring the probability that a given data set is generated by a uniform data distribution. In other words, it tests the spatial randomness of the data.

The Null and alternative hypothesis is defined as follows:

**Ho -** the dataset is uniformly distributed (no meaningful clusters)

**Ha -** the dataset is not uniformly distributed (contains meaningful clusters)

We conducted the Hopkins test iteratively, using 0.5 as the threshold to reject alternative hypotheses, i.e if H < 0.5, then it is unlikely that D has significant clusters, we got the H value of 0.91 which highlighted the presence of clusters in the data.

Partition Clustering

Partition clustering classifies observations based on similarities and requires the analyst to specify the number of clusters to be generated.

K- Means Clustering

It is the unsupervised machine learning algorithm for partitioning a given data set into a set of k clusters, which are pre specified by the analyst. Objects in the same clusters have high intra-class similarity and objects in different clusters have low inter-class similarity.

Since this method requires the optimal number of clusters in advance so we must determine the value of k prior to executing clustering. There are different methodologies to find out the optimal number of k, we have generated the elbow plot, silhouette width plot and the gap statistic to select the optimal number of clusters and both elbow and silhouette width plot suggest the three is the optimal number of clusters.

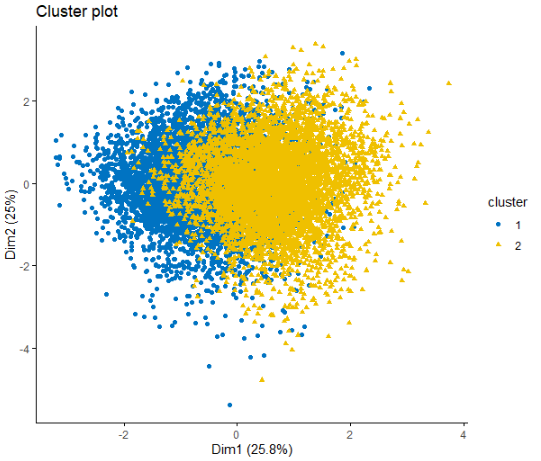
Figure12: Methods for selecting Optimal k

Chart

Description automatically generated

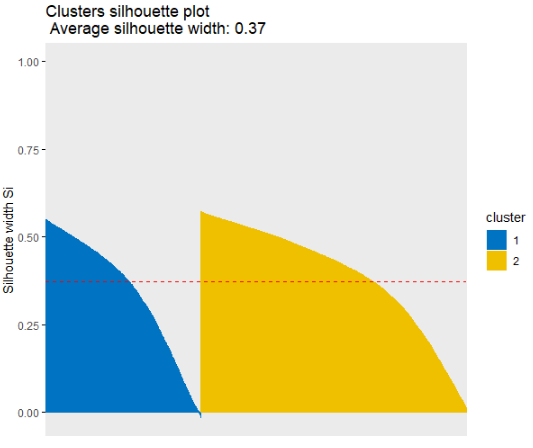
After determining the optimal number of clusters, we performed k-means and plotted two clusters. It can be observed that clusters are not completely distinct and there exist a certain level of overlapping in the clusters.

Figure13: Cluster plot of obtained clusters



Then we plotted the silhouette plot, and the average width came out to be 0.37 for k means clustering with three clusters.

Figure14: Silhouette plot of clusters obtained



Hierarchical Clustering

Hierarchical clustering is an alternative approach to partitioning clustering for grouping objects based on similarity, and it does not require the number of clusters in advance.

Agglomerative Clustering

It starts by treating each object as a single-ton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram.

This algorithm requires linkage criterion that determines the distance between sets of observations as a function of pairwise distances between observations.[[10]](#footnote-11)

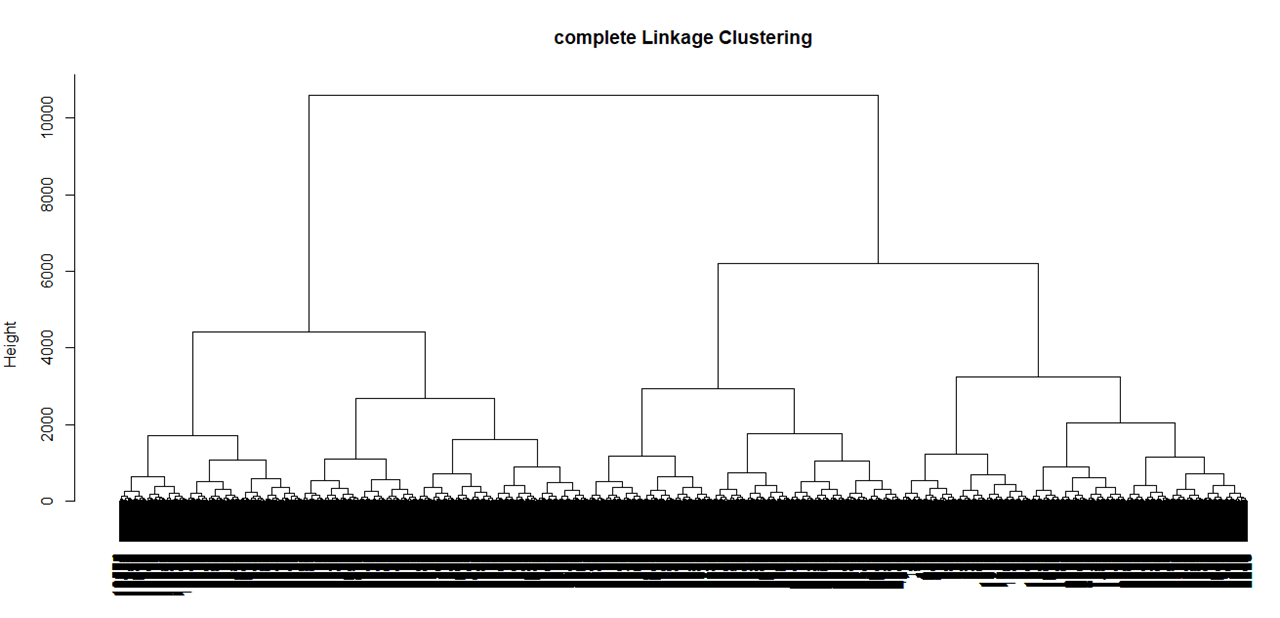
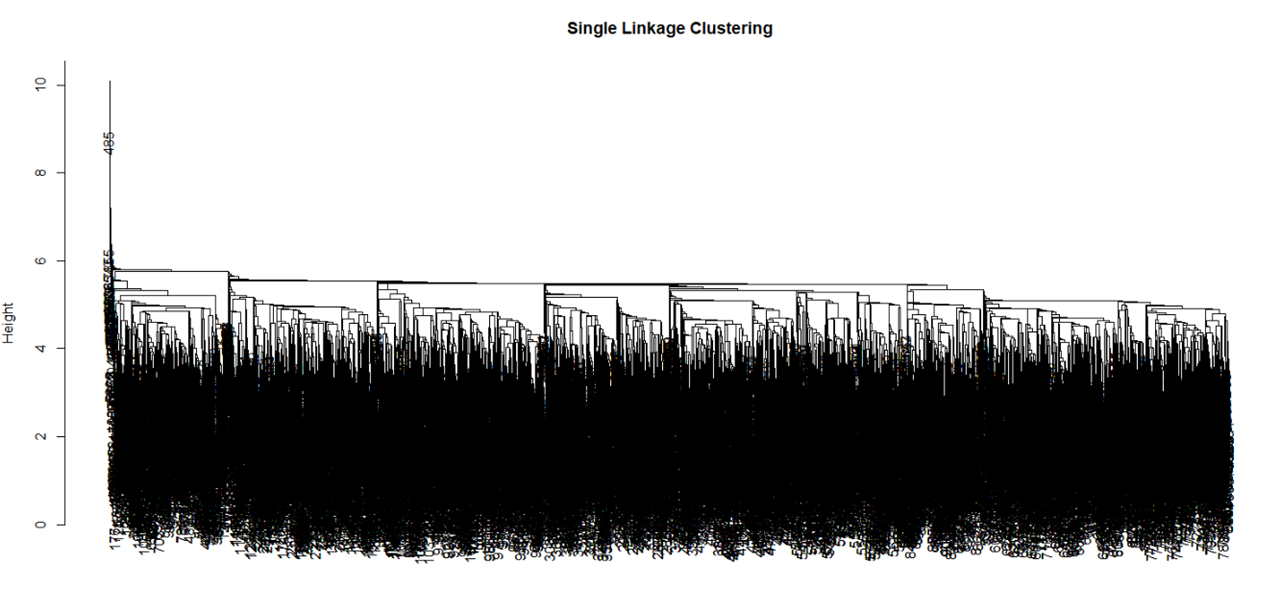
There are different types of linkage functions which can be used in this methodology, namely single, complete, average and ward.

|  |  |
| --- | --- |
| Linkage methods | Coefficients |
| Single | 0.8952 |
| Complete | 0.9734 |
| Average | 0.9600 |
| Ward | 0.9973 |

For this dataset single linkage generates the least clustering coefficient, it means a single linkage method is not appropriate for hierarchical clustering for our dataset. Whereas, ward linkage criterion generates the maximum clustering coefficient, hence we will be using that in our analysis.

Here are the dendrograms for the four different linkages and it can be observed visually that ward criterion is the best among them.

Figure15: Dendrograms representing 4 different linkages



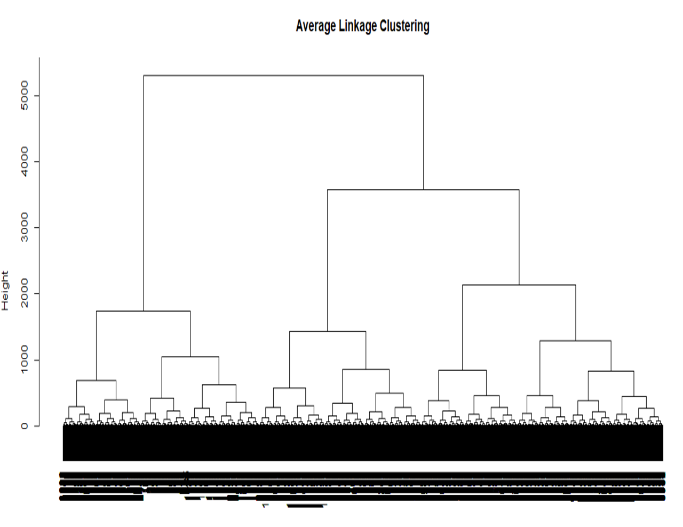
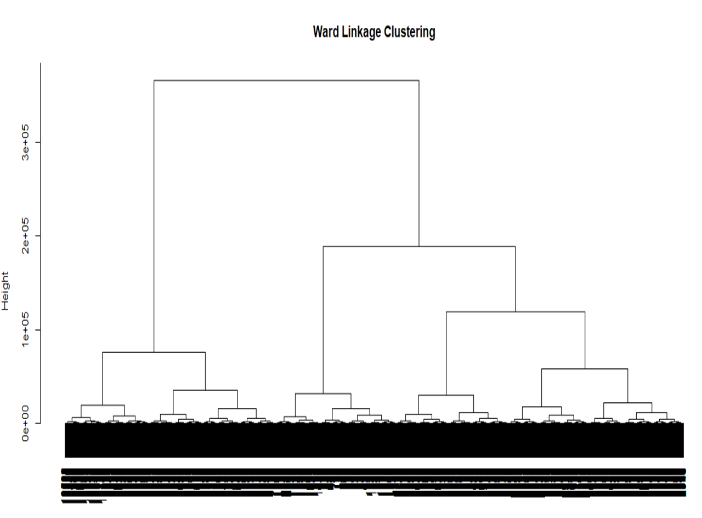
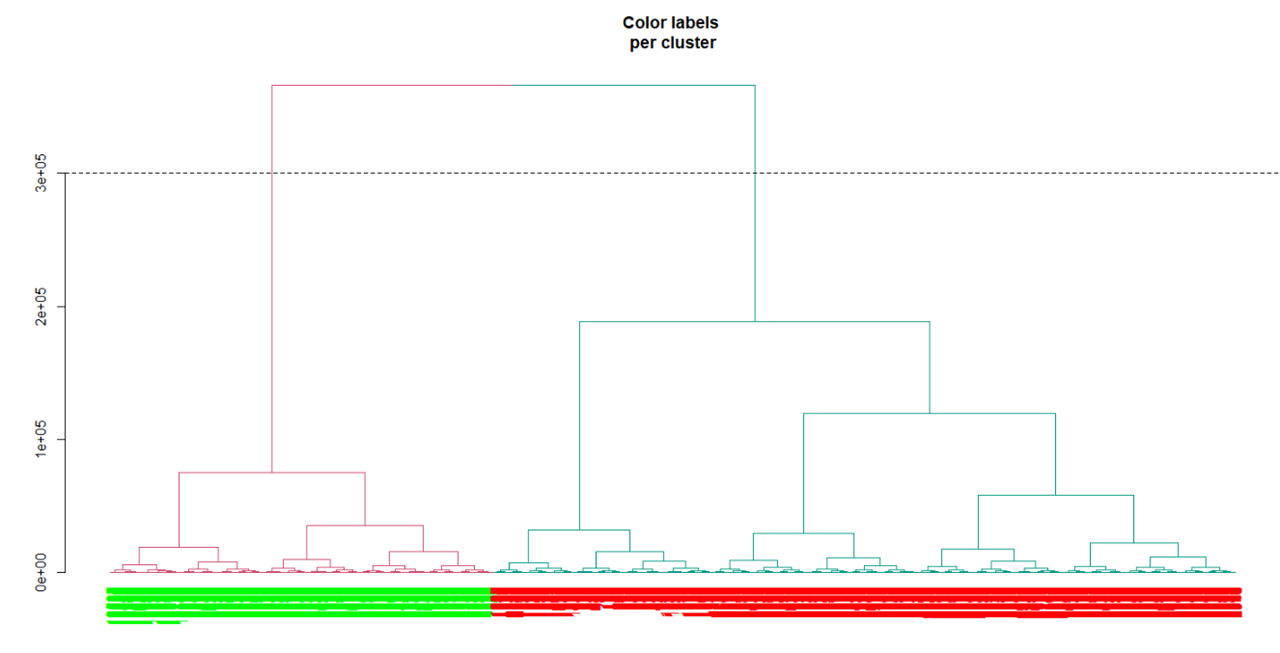
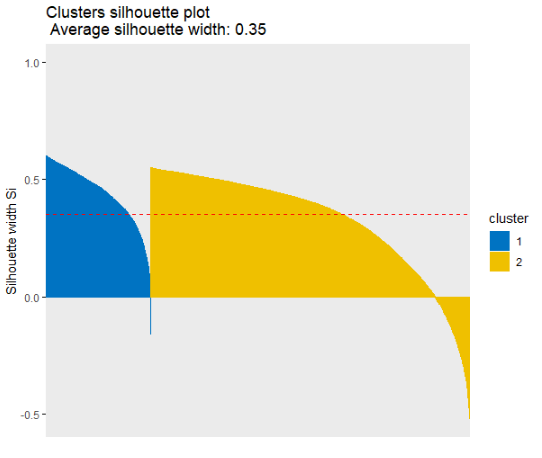


Figure16: Dendrogram with 2 clusters (Ward Linkage)



Also, the silhouette plot indicates the average width with hierarchical clustering is 0.35 with two clusters.

Figure17: Silhouette plot of clusters



Cluster Validation

Now that we have determined that the optimal number of clusters is three, we were interested in finding that the clusters identified by partitioning and hierarchical methods are the same.

So, we created a confusion matrix between the identified clusters, and it explained 89.63% of accuracy, which means 90% of the data points were identified in the same clusters by both algorithms.

|  |  |  |
| --- | --- | --- |
| Clusters | 1st | 2nd |
| 1st | 668 | 0 |
| 2nd | 110 | 283 |

|  |  |
| --- | --- |
| Accuracy | 0.8963 |
| Sensitivity | 0.8506 |
| Specificity | 1.0000 |

This further supported our approach that we followed for cluster analysis and optimal number of k.

Figure17: Metrics Comparison

A picture containing chart

Description automatically generated

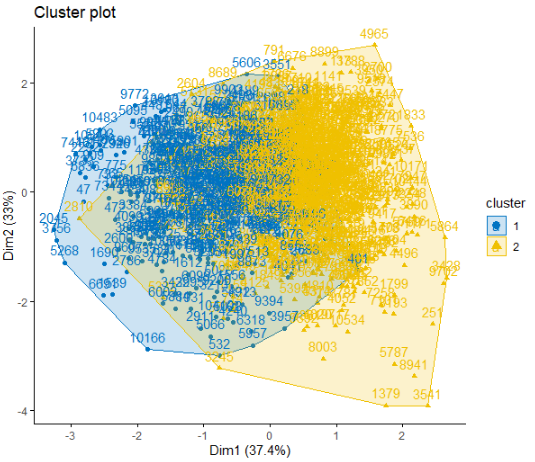
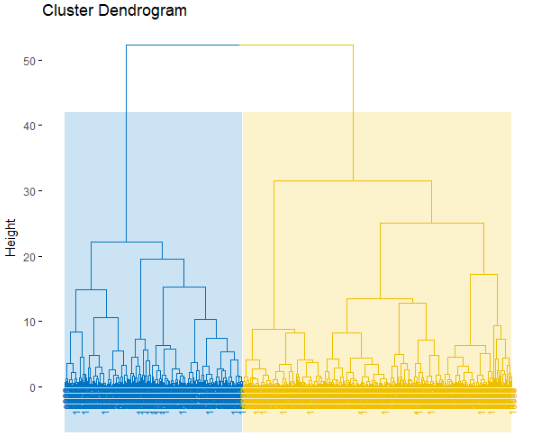
Advanced Clustering

Hierarchical K-Means Clustering

K-means has some limitations: it requires the user to specify the number of clusters in advance and selects initial centroids randomly. The final k-means clustering solution is very sensitive to this initial random selection of cluster centers. The result might be different each time you compute k-means.

This approach computes hierarchical clustering first and cuts it into k clusters, then computes the center of each cluster and utilizes them as initial cluster center for the k means.

We performed our analysis with this technique to confirm if the claim of two optimal clusters is correct or not and here are the results of the analysis.



And it generated two optimal clusters, which supported our analysis so far.

Interpretation

We find two different clusters from hierarchical as well as partitioning clustering. We segmented all the LinkedIn profiles into two clusters according to different characteristics with similar features within the clusters.

Cluster 1:

Asian people form majority of this cluster. They uploaded good quality pictures on LinkedIn and have good number of followers. Image analysis showed high score for beauty in their profile picture and seemed professional. The length of their previous positions is large which indicates loyalty; and it also means they were satisfied from their previous jobs.

Cluster 2:

This cluster is made up of individuals with a greater amount of experience. They are senior professionals with longer careers.  One thing that surprised us is that they are not very concerned about their profile picture quality on LinkedIn as we found that their pictures are either blurred or we can notice deteriorating skin health from our analysis. This cluster mostly comprises of Celtic English.

Conclusion

We successfully identified the underlining clusters in the dataset generated by web scraping 15k LinkedIn profiles using techniques such as Partitioning, Hierarchical and Advanced Clustering. We were able to interpret a pattern of professional images in a cluster with young people who are mostly immigrant to Australian subcontinent. In other cluster, we could see more mature adults who were not concerned about the picture quality and are mostly native to Australia.

Limitations

* This study is restricted to Australian profiles of LinkedIn users.
* Categorical columns were not considered because of limitations of algorithms
* The gender distribution was unbalanced in our sample data.
* Density based clustering could not be applied because of low computational power

Acknowledgment

We would like to record the appreciation of Andrew Truman of Sydney, New South Wales, Australia for providing the web scraped profiles and its information open to public access on Kaggle database.[[11]](#footnote-12)

1. “LinkedIn”, Wikipedia, 27 November 2020, “https://en.wikipedia.org/wiki/LinkedIn” [↑](#footnote-ref-2)
2. “How LinkedIn’s 400 Million Members are Helping Build the Economic Graph”, Aatif Awan, LinkedIn Official Blog, 29 October, 2015,” https://blog.linkedin.com/2015/10/29/400-million-members” [↑](#footnote-ref-3)
3. “Emotion Analysis FaceReader”, Noldus Information Technology, Beijing China,” https://www.noldus.com/facereader” [↑](#footnote-ref-4)
4. “Blur detection with OpenCV”, Adrian Rosebrock, PyImageSearch , 7 September 2015,” https://www.pyimagesearch.com/2015/09/07/blur-detection-with-opencv/” [↑](#footnote-ref-5)
5. “Blur Processing Using Double Discrete Wavelet Transform”, Yi Zhang, Keigo Hirakawa, Department of Electrical & Computer Engineering, University of Dayton, 2013,” https://openaccess.thecvf.com/content\_cvpr\_2013/papers/Zhang\_Blur\_Processing\_Using\_2013\_CVPR\_paper.pdf” [↑](#footnote-ref-6)
6. “Smile Detection Using Multi-scale Gaussian Derivatives”, Varun Jain, James L. Crowley12th WSEAS International Conference on Signal Processing, Robotics and Automation, Feb 2013, Cambridge, United Kingdom, hal-00807362f, ” https://hal.inria.fr/hal-00807362/document” [↑](#footnote-ref-7)
7. # “Assessment of female facial beauty based on anthropometric, non-permanent and acquisition characteristics”, Antitza Dantcheva, INRIA, Jean-Luc Dugelay, EURECOM , September 2014,” https://www.researchgate.net/publication/271659255\_Assessment\_of\_female\_facial\_beauty\_based\_on\_anthropometric\_non-permanent\_and\_acquisition\_characteristics/”

   [↑](#footnote-ref-8)
8. # ”Dimensionality Reduction using Factor Analysis”, Nitin Khosla, School of Microelectronics Engineeing Griffith University, Australia December, 2004,” https://research-repository.griffith.edu.au/bitstream/handle/10072/366058/02Main.pdf?sequence=1”

   [↑](#footnote-ref-9)
9. “Methods for assessing clustering tendency”, Alboukadel Kassambara, STDHA, 2017,” Practical Guide to Cluster Analysis in R” [↑](#footnote-ref-10)
10. # ” Hierarchical Clustering”, Wikipedia, 7 October 2020,” https://en.wikipedia.org/wiki/Hierarchical\_clustering#Linkage\_criteria”

    [↑](#footnote-ref-11)
11. “LinkedIn Profile Data”, Andrew Truman, Sydney, New South Wales, Australia, 2017, “https://www.kaggle.com/killbot/datasets” [↑](#footnote-ref-12)