Progress week 7

**Current Approach**

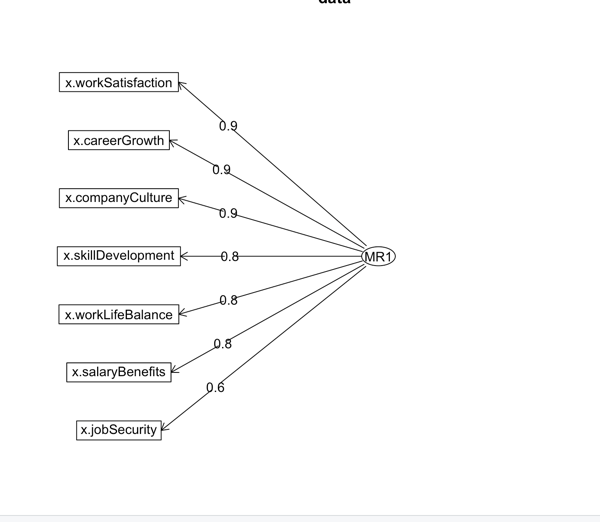
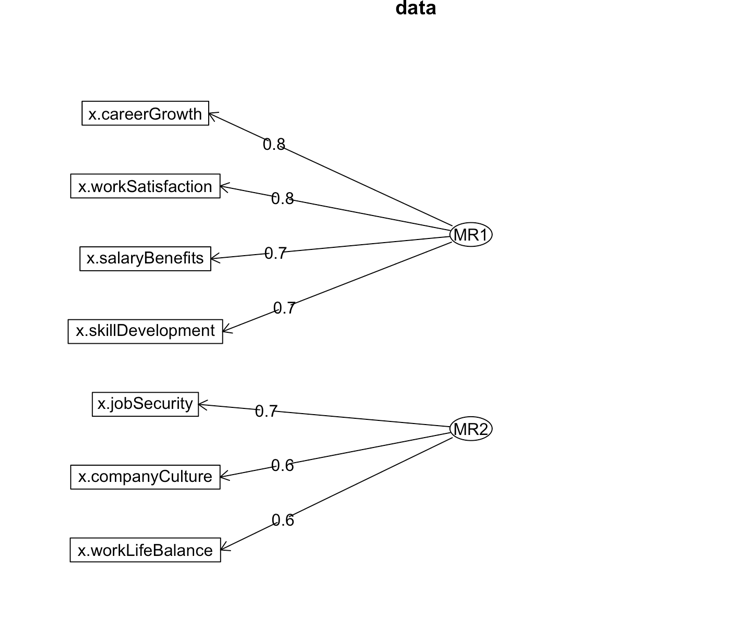
From the above Factor Analysis output we could see that one latent construct (factors) or two latent constructs from the 7 variables are good to have

The basic assumption of factor analysis is that for a collection of observed variables there are a set of underlying variables called factors (smaller than the observed variables), that can explain the interrelationships among those variables.

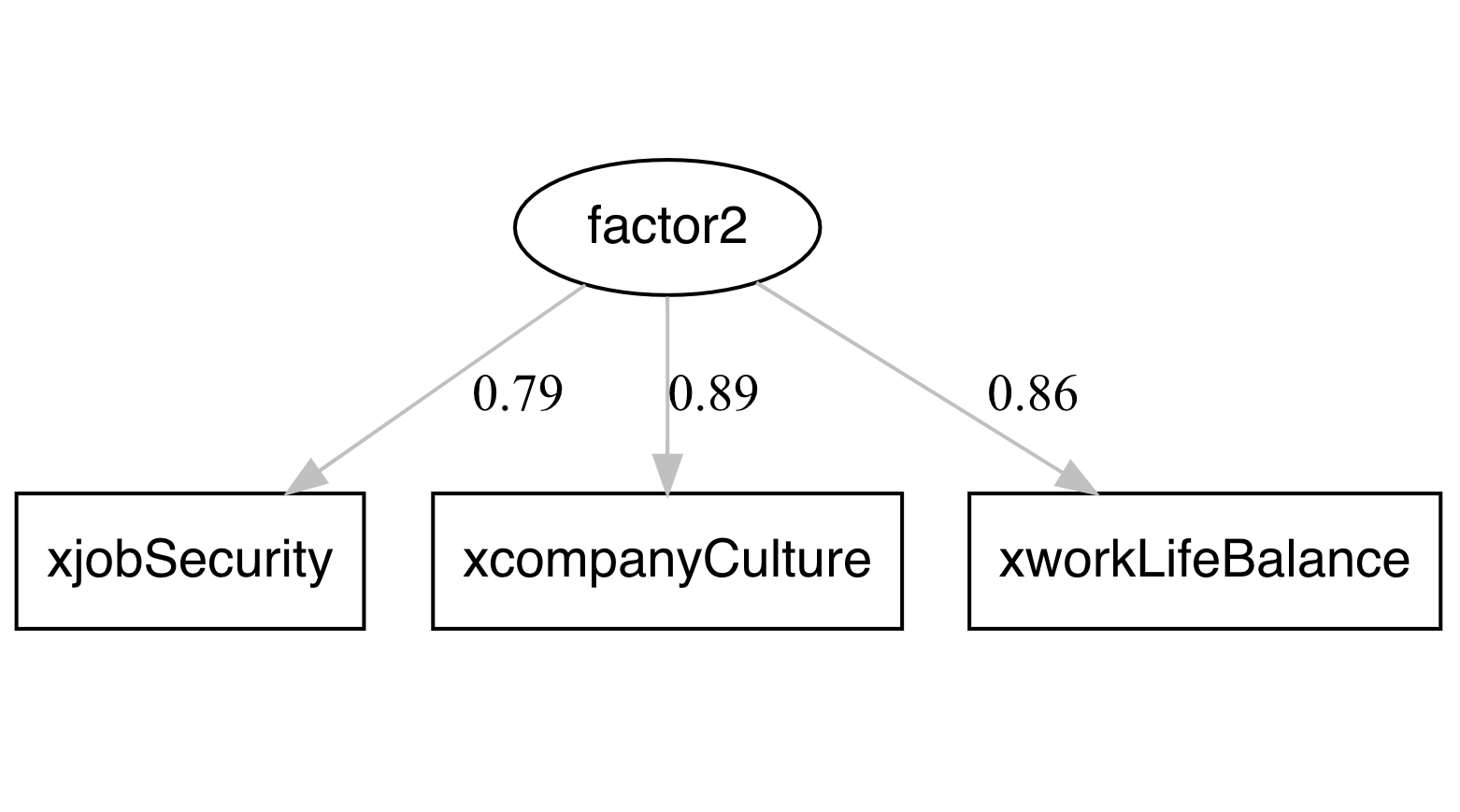
The below two factors are the two axis from the data we have understood from our data

MR1-MR1 subjectively maps onto reduced personal accomplishment, depersonalization.

MR2-MR2 maps onto emotional exhaustion and also depersonalization

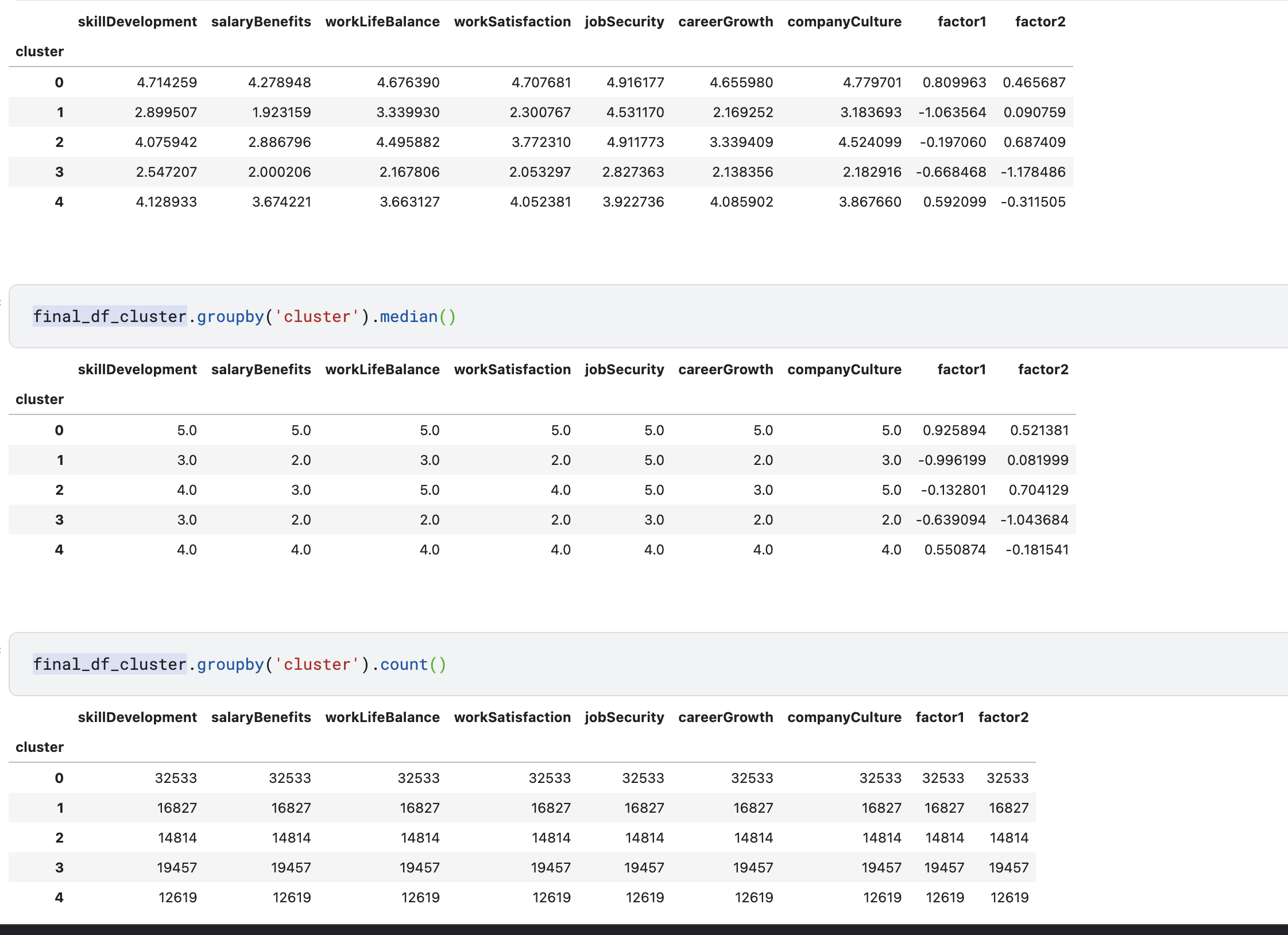


To translate the findings in our dataset to other datasets Confirmatory Factor Analysis can help us

Diagram

Description automatically generated

Clustering using the two factors and segmenting them



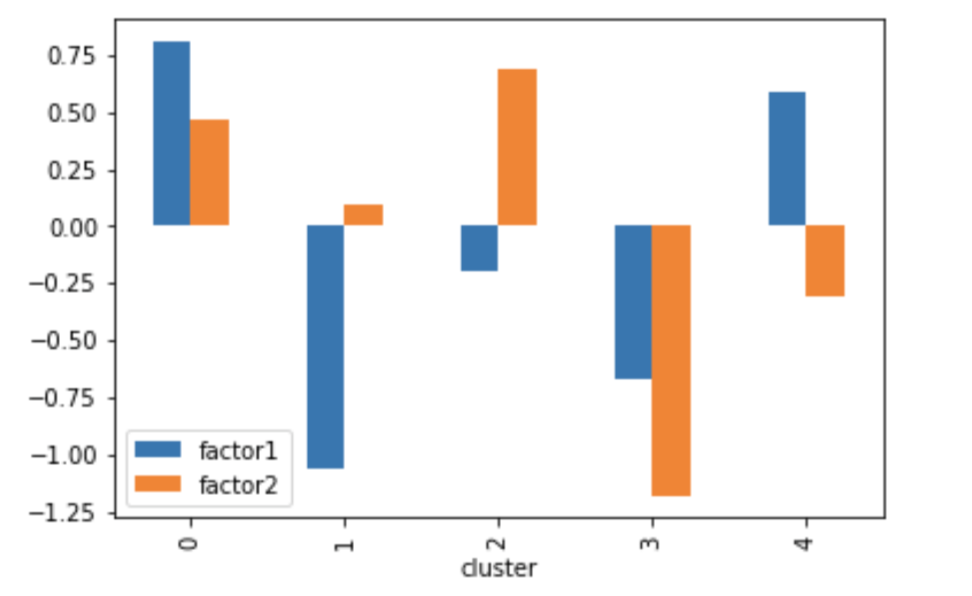
Cluster 0 – 32533 people are very happy with their jobs

Cluster 1 – 16827 people are content with Factor 2 (Job security ,company culture, worklife balance ) and not so content with factor 1 variables, evident from the factor scores

Cluster 2 – 14814 people are happy with Factor2 and content with Factor 1

Cluster 3 – 19457 are not happy with their jobs

Cluster 4 – 12619 are ok with factor 2 and happy with Factor 1 variable components



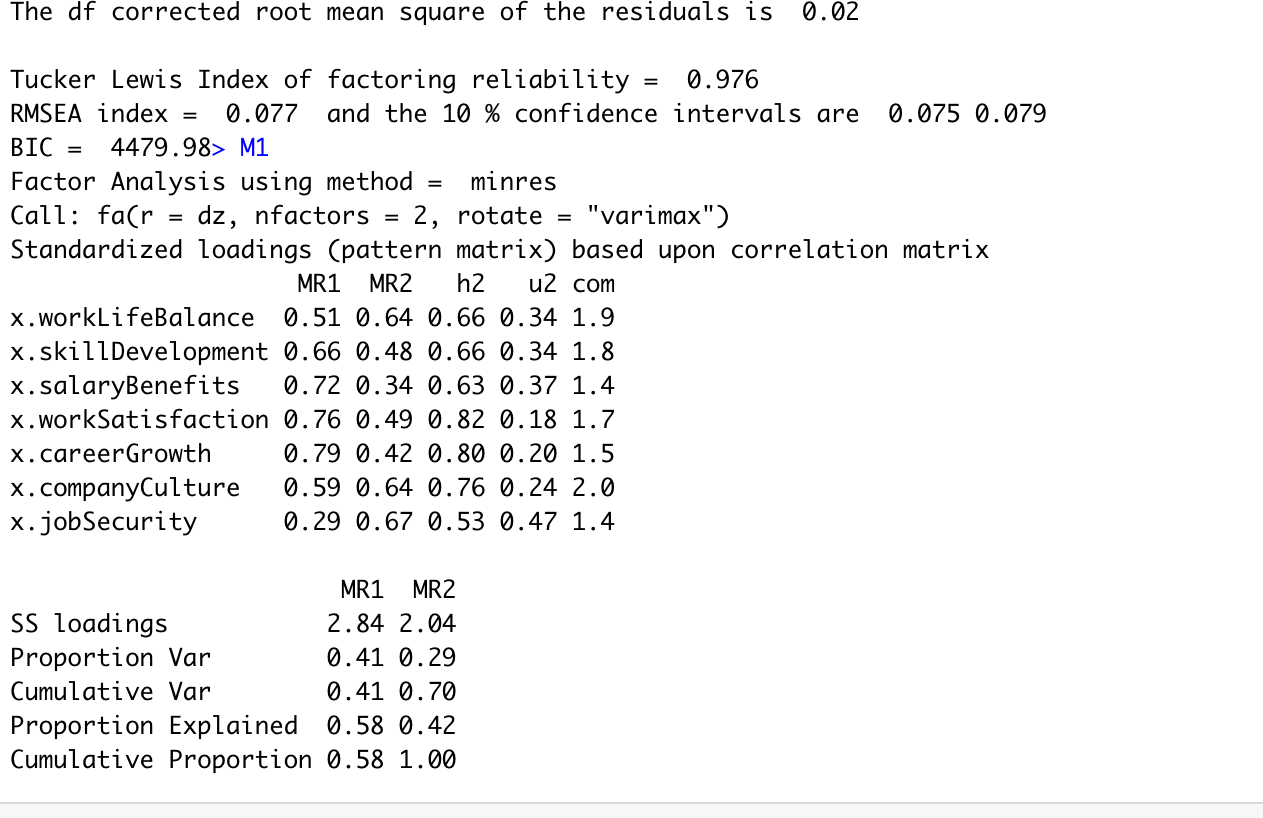
Table

Description automatically generated

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2754990/#:~:text=%22Burnout%22%20is%20a%20syndrome%20of,overextended%20and%20drained%20by%20others.

A picture containing graphical user interface

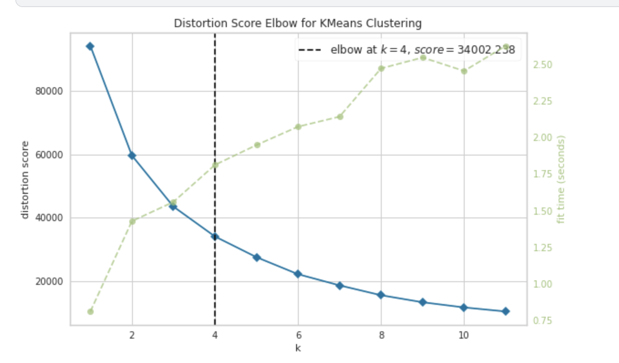
Description automatically generated

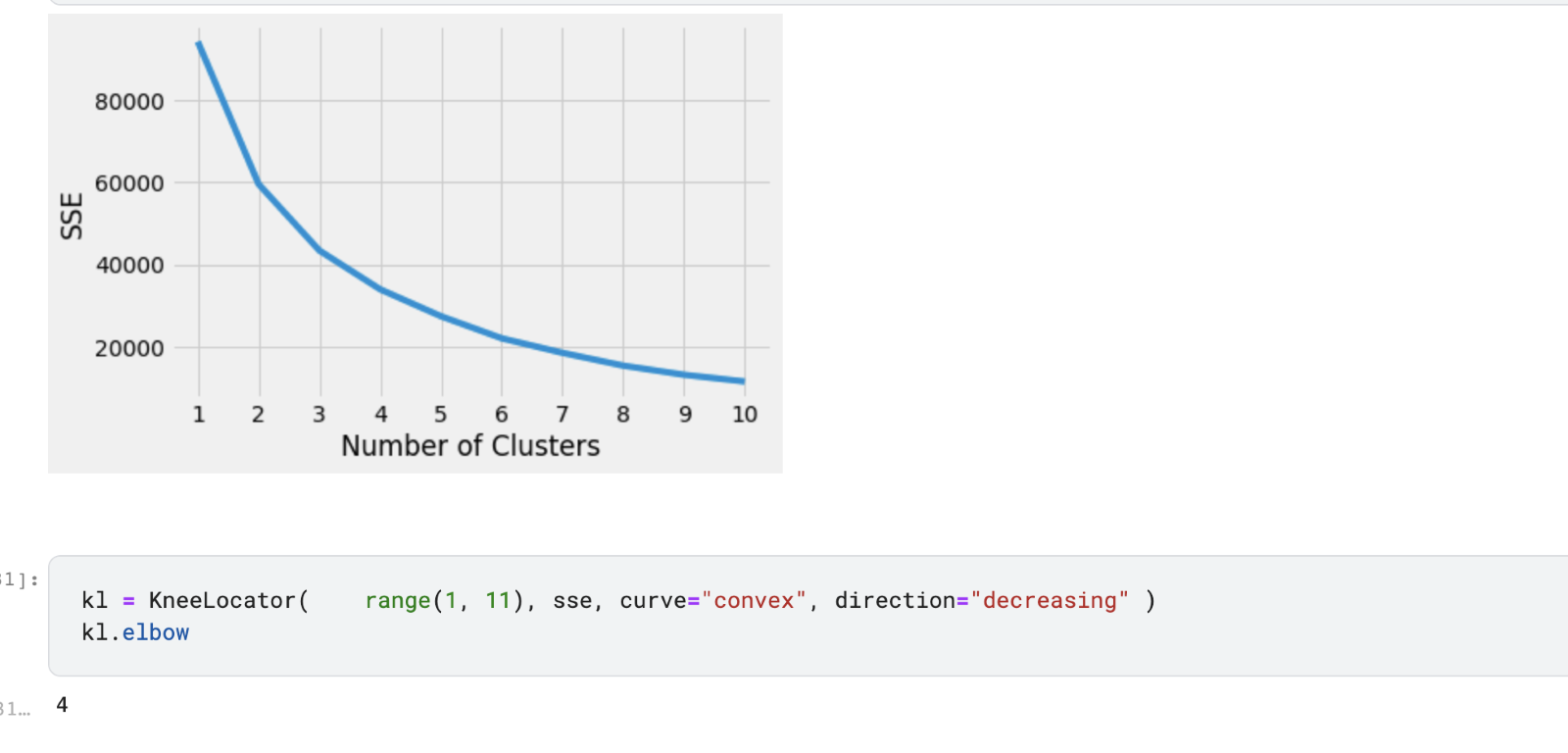


**Cluster analysis post the factor Analysis**

As per our discussions in the previous call we used clustering on the factors of the data

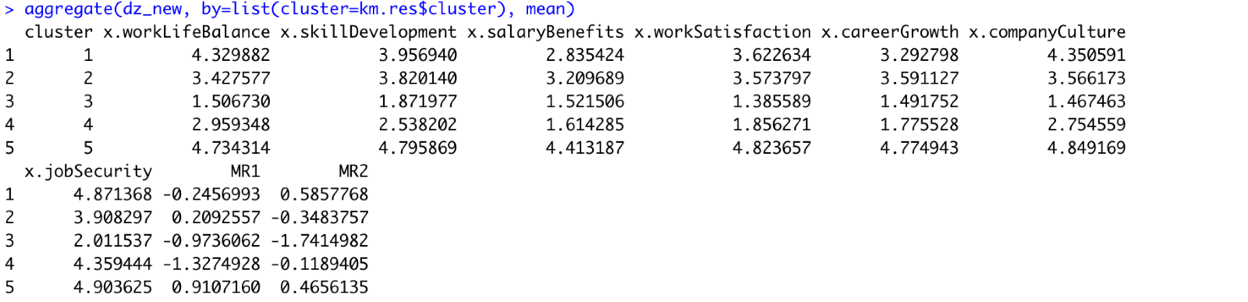
Finding out the best number of k for clustering from elbow plot and silhouette plot

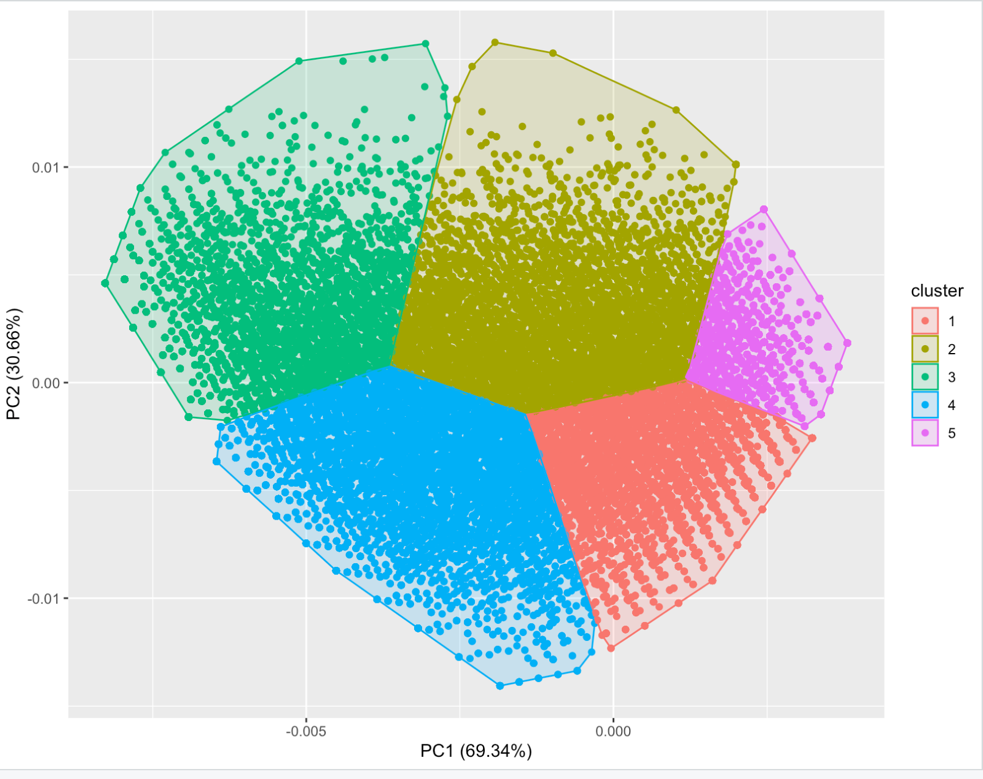




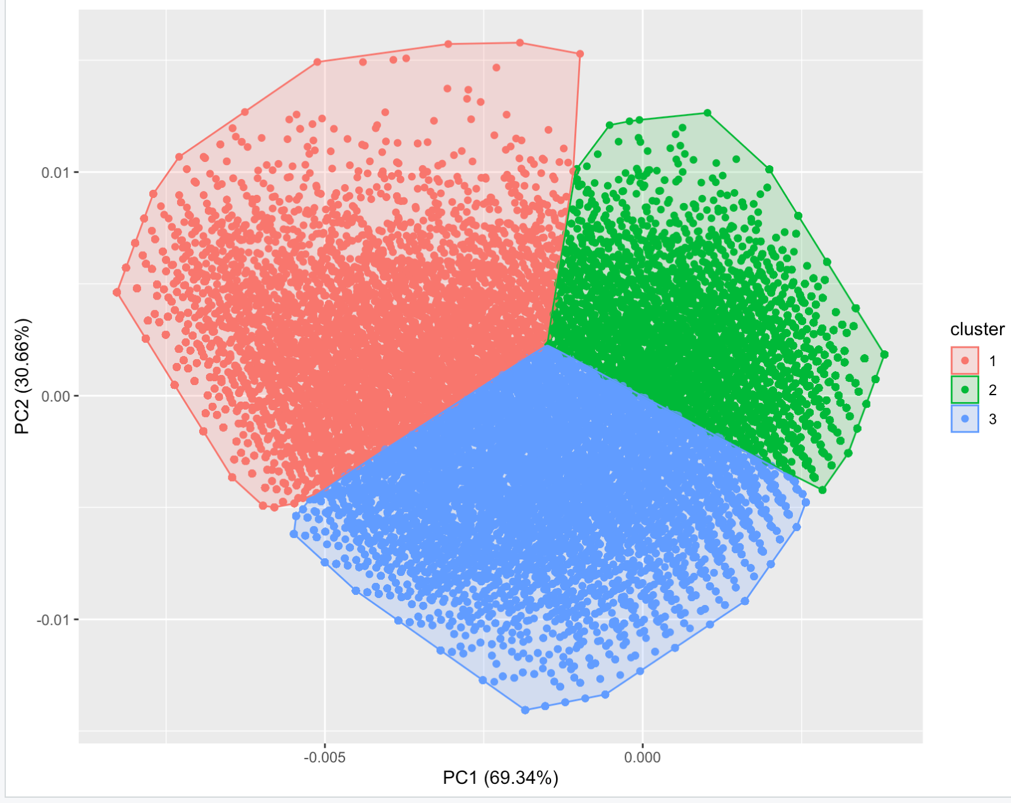
**2 factors and 5 clusters** /**3 clusters**

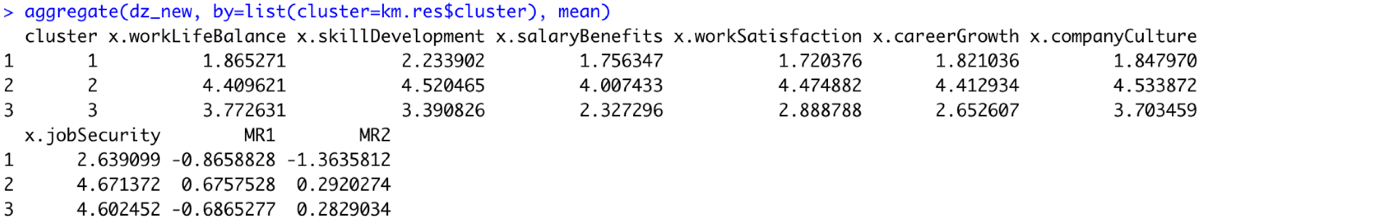
**(The better clustering out of all the experiments done)**





2 factors and 3 clusters





Having factors more than 2 resulted in poor clustering of the data

Table

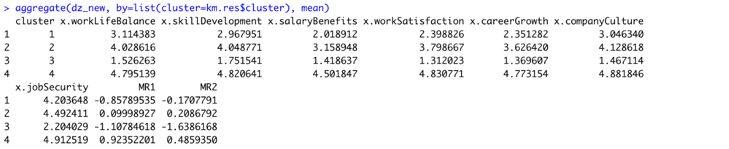
Description automatically generated with medium confidence

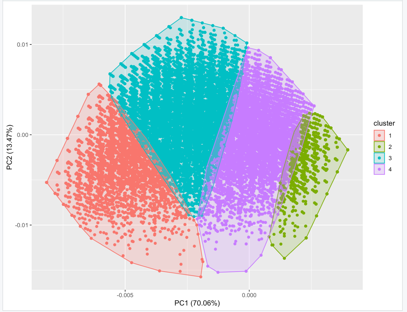
Chart, surface chart

Description automatically generated

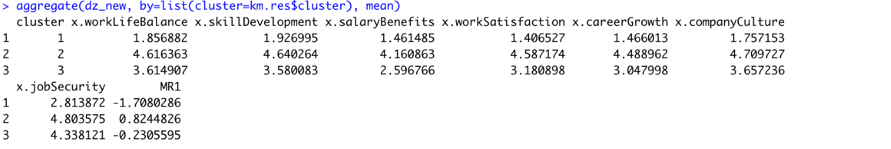
Screenshots of the experiments we rejected out due to overlapping of the segments

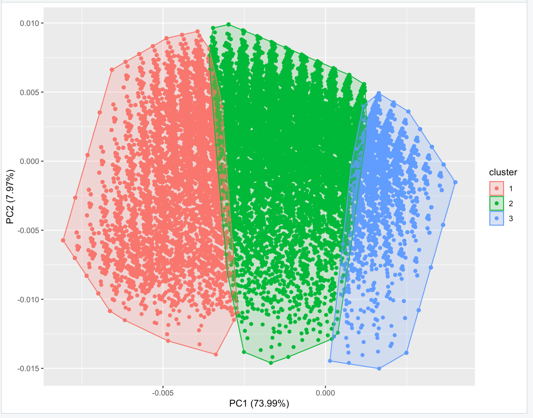
2 factors and 4 clusters (overlapping is present)



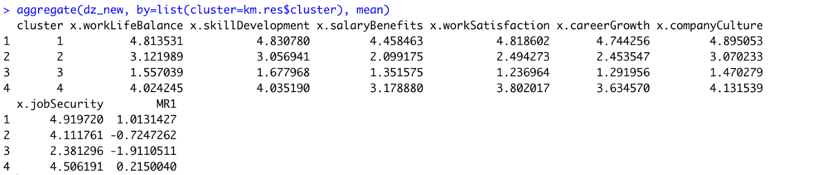


One factor and 3 clusters )overlapping is present



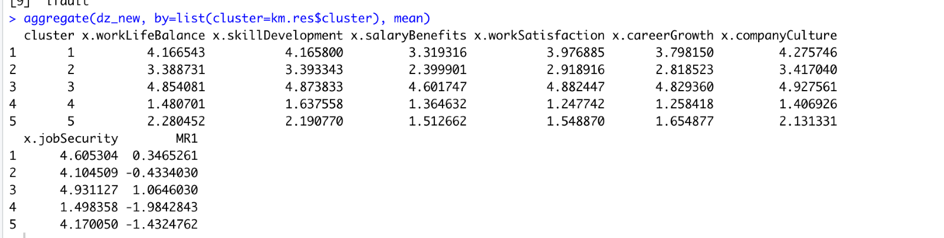


One factor 4 clusters (overlapping is present)





One factor 5 clusters (overlapping is present)





FUTURE STEPS

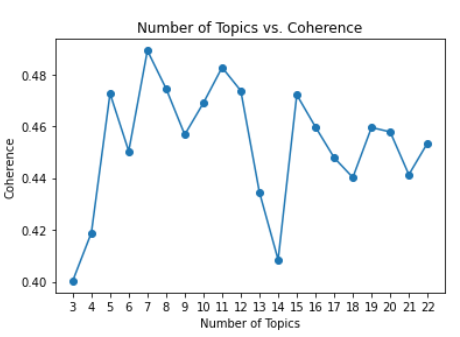
* Given our approach of clustering into 3 segments and the factor analysis .we would like to proceed using the clustered category text for further analysis.

(Ex.Analysing cluster 1 text and topic modelling of cluster one helps us in understanding the sentiment of the people who have given very less ratings to the company )

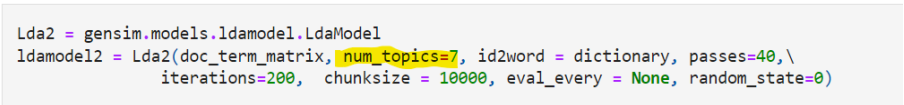
* We would like to translate this to a recommendation system or a Multiclass classification approach (As we have data that is labelled into clusters now)
* Converting the text to a sentiment score and validating our approach with text and without text scores

Fine Tuning LDA:

LDA requires that we specify the number of topics that exists in a corpus of text. There are several common measures that can be optimized, such as predictive likelihood, perplexity, and coherence. Much literature has indicated that maximizing coherence, particularly a measure named Coherence Value  (https://svn.aksw.org/papers/2015/WSDM\_Topic\_Evaluation/public.pdf), leads to better human interpretability. This measure assesses the interpretability of topics given the set of words in generated topics. Therefore, we will optimize this measure. Result of Coherence analysis on text corpus



LDA Model Selected basis the ‘CONS’ column:



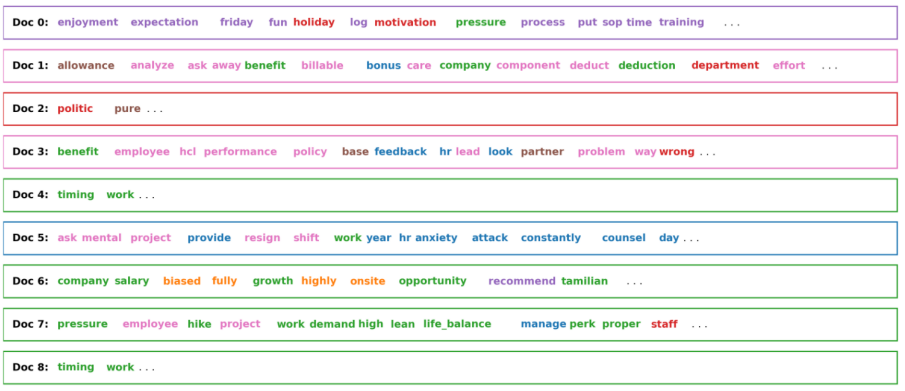
LDA Output basis the ‘CONS’ column:



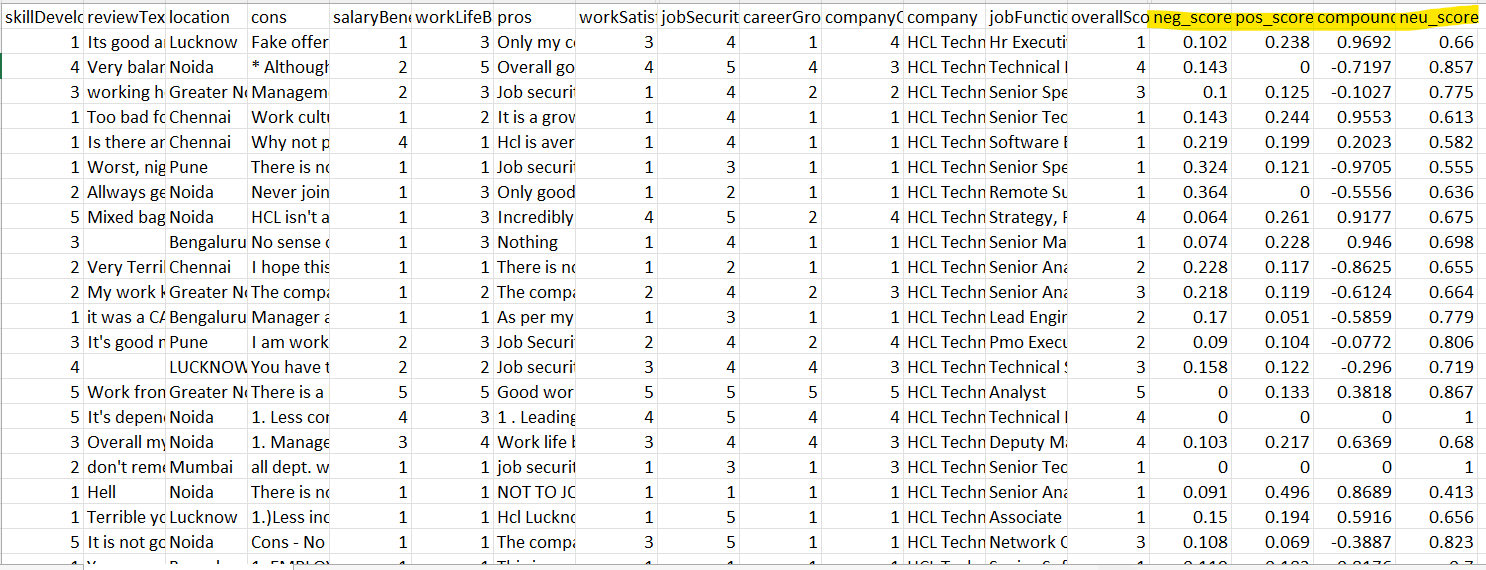
* Need to name these topics on basis of subjective and domain expertise

LDA Current State and Next Steps

• We were able to split each review by topic



* Extract the most dominant topic in each review text and perform sentiment analysis to get a final sentiment score



* Fine-tuning the sentiment analysis results
* Tried TextBlob and VADER to arrive at the sentiment score on ‘REVIEW TEXT’



* Sentiment score can be used as additional feature. It will also lay foundation to Suggestion Box, Water Cooler etc. – Next Steps

Articles referred to estimate the factor scores

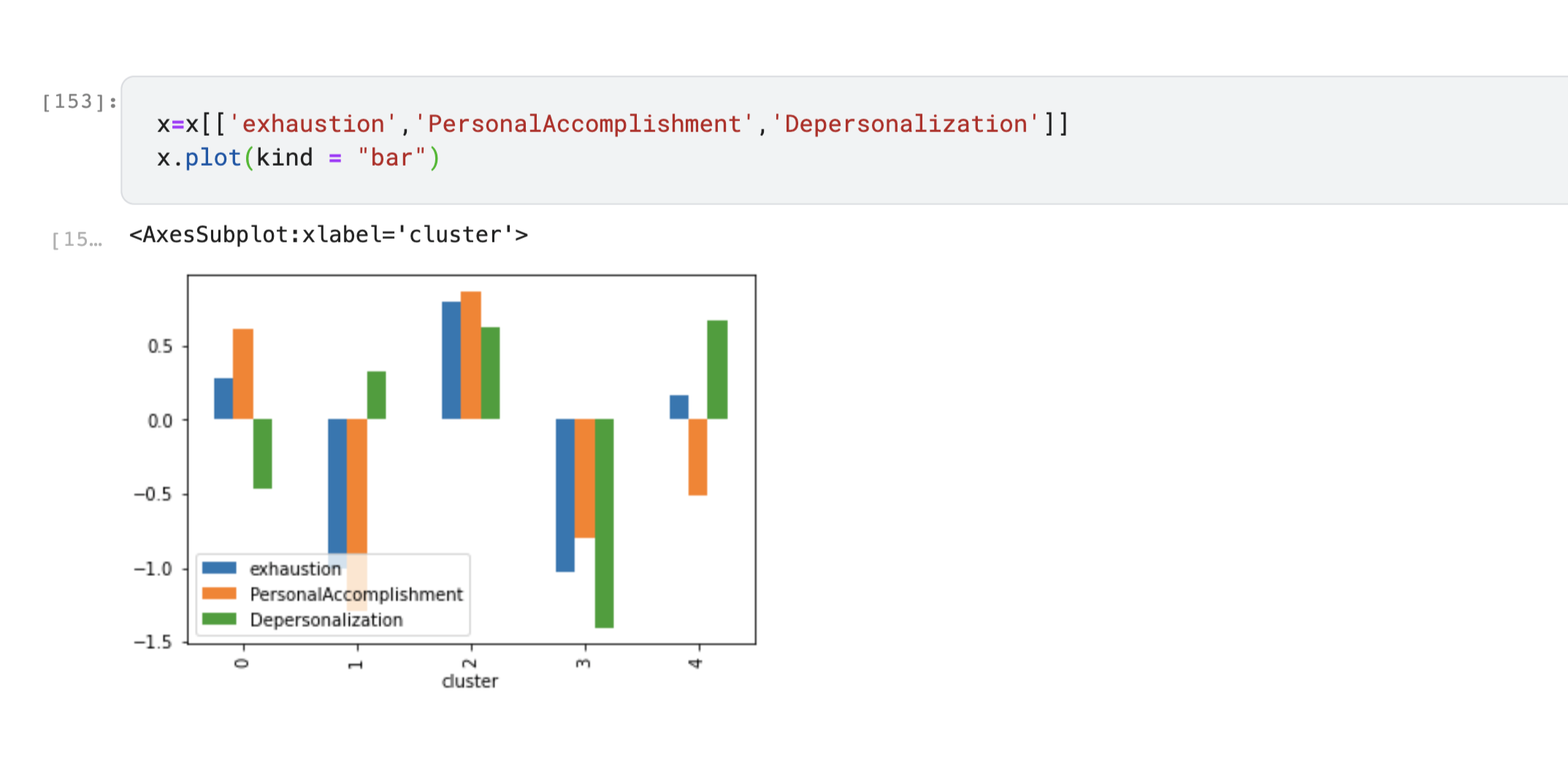
<https://www.researchgate.net/publication/255643537_Understanding_and_Using_Factor_Scores_Considerations_for_the_Applied_Researcher>

<https://medium.com/@Gabrielefrattini/clustering-market-segments-with-k-means-and-factor-analysis-9b3876581b16>

<https://rpubs.com/pjmurphy/758265>

<https://github.com/Dhamodaran-Babu/Machine-Learning-Exercises/tree/master/12.Factor%20Analysis>

<https://www.tqmp.org/RegularArticles/vol09-2/p079/p079.pdf>



Table

Description automatically generated

Table

Description automatically generated

Diagram

Description automatically generated