AEDA Report

28 February 2021

Abstract

This exploratory data analysis report was created by the R package AEDA.

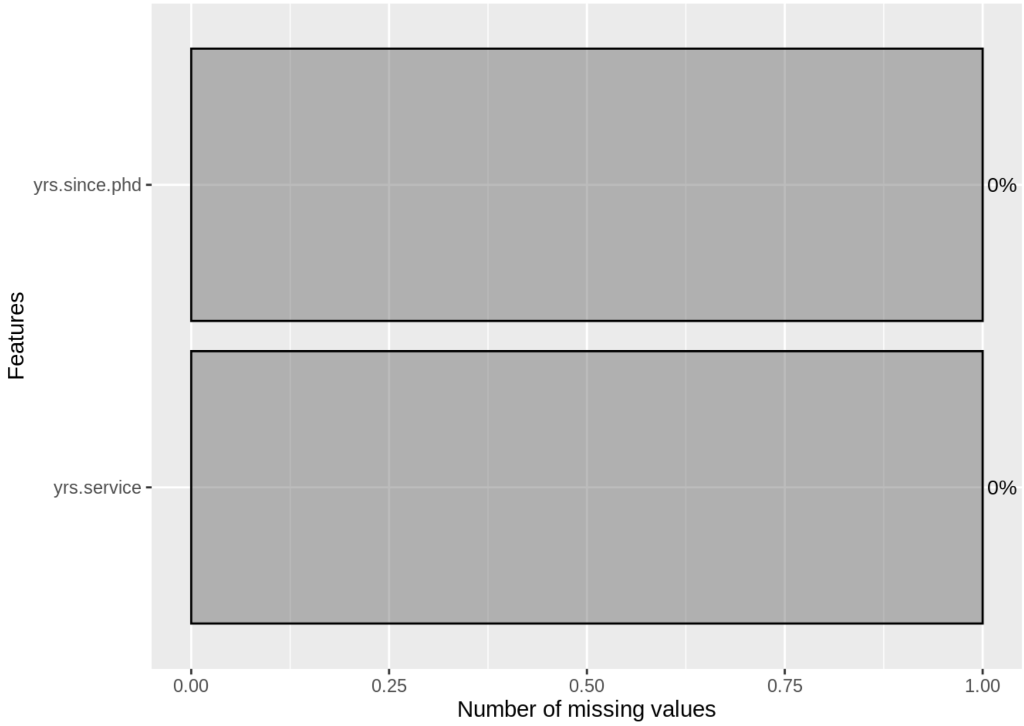
```
## Error : No root directory found in /home/theo/Dropbox/MYR5/R-EDA/regress/salaries or its parent directorie
s. Root criterion: contains a file `DESCRIPTION`
```

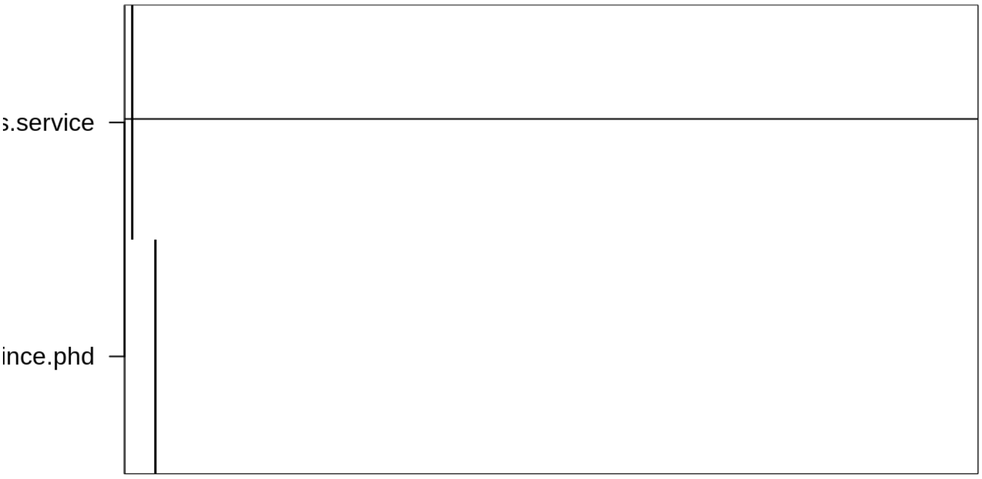
Basic Summary Report

Basic Summary

The dataset data is 1.252810^{4} megabytes in size. In total there are 397 observations, 2 missing values and 6 columns.

Missing Value Summary





Categorical Summary Report

Categorical Summary Results

In the following contingency tables for categorical columns will be displayed:

1-D Contingency Table 1

piece	Freq
AssocProf	64
AsstProf	67
Prof	266

1-D Contingency Table 2

piece	Freq
A	181
B	216

1-D Contingency Table 3

piece	Freq
Female	39
Male	358

2-D Contingency Table 1

	A	B	Sum
AssocProf	26	38	64
AsstProf	24	43	67
Prof	131	135	266

	A	B	Sum
Sum	181	216	397

2-D Contingency Table 2

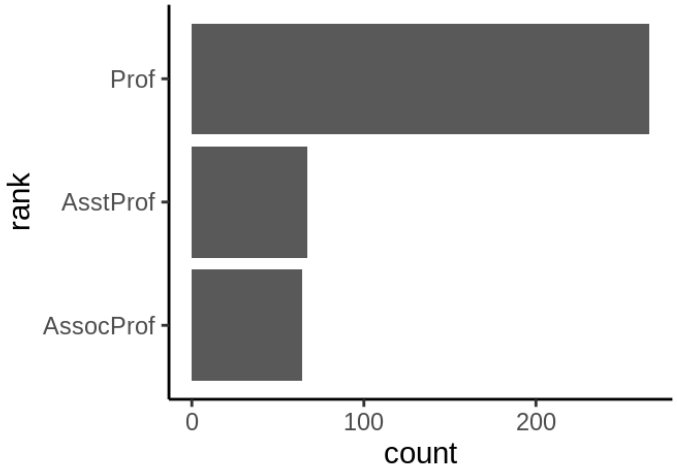
	Female	Male	Sum
AssocProf	10	54	64
AsstProf	11	56	67
Prof	18	248	266
Sum	39	358	397

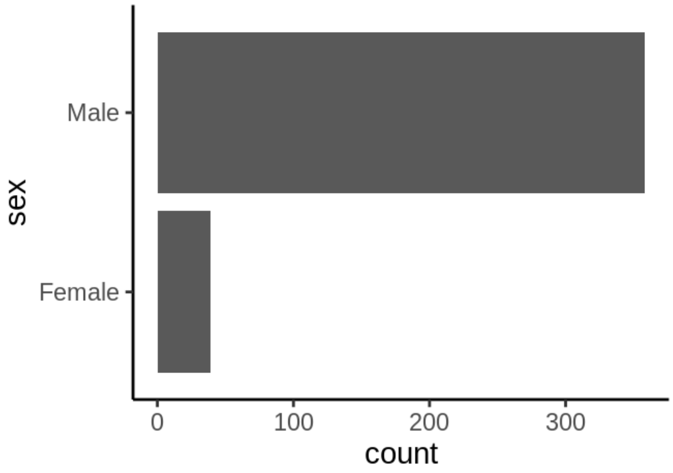
2-D Contingency Table 3

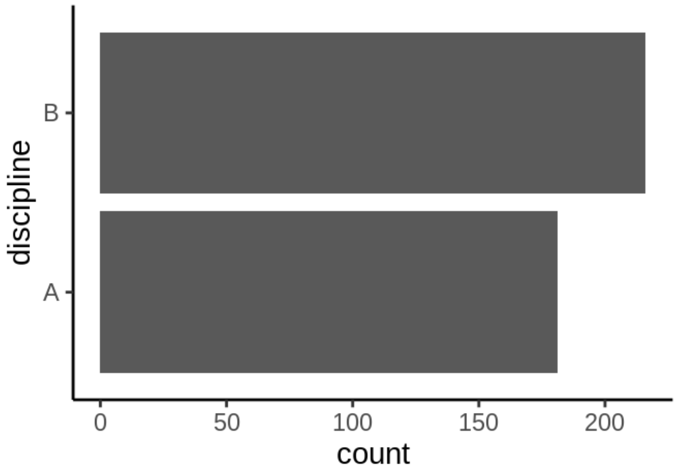
	Female	Male	Sum
A	18	163	181
B	21	195	216
Sum	39	358	397

Categorical Summary Results

In the following bar plots for categorical columns will be displayed:







Numeric Summary Report

Numeric Summary Results

The following data frame shows summary statistics for numeric colums from the dataset:

Numeric Summary

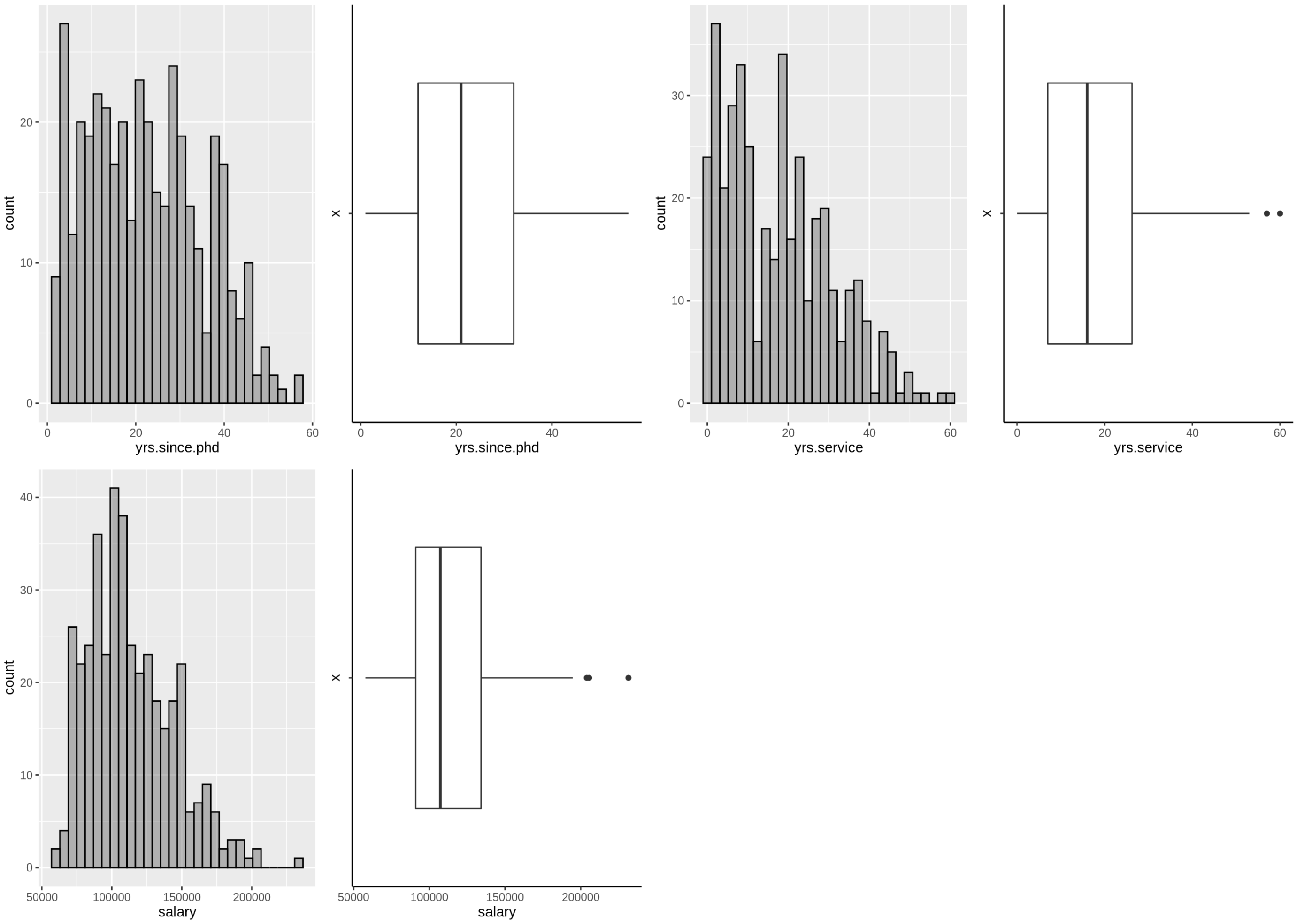
	kurtosis	skewness	mean	sd	min	5%	25%	50%	75%	95%	l.bound	u.bound
yrs.since.phd	2.205	0.301	22.263	12.868	1	4.0	12	21	32.0	44.3	-18.00	62.00
yrs.service	2.693	0.655	17.559	12.994	0	1.0	7	16	26.5	41.6	-22.25	55.75
salary	3.181	0.708	113725.732	30362.426	57800	73289.8	91000	107300	134367.5	169044.5	25948.75	199418.75

General: Following footnotes explain some measure from the table above

Explanation: [a] Kurtosis will be calculated via: $\sum_{ni=1}(x_i - \bar{x})^4/ns^4$; [b] Skewness will be calculated via: $\sum_{ni=1}(x_i - \bar{x})^3/ns^3$; [c] l.bound is defined as: $q_{0.25} - 1.5IQR$; [d] u.bound is defined as: $q_{0.75} + 1.5IQR$; where IQR is defined as $IQR := q_{0.75} - q_{0.25}$

Numeric Summary Plots

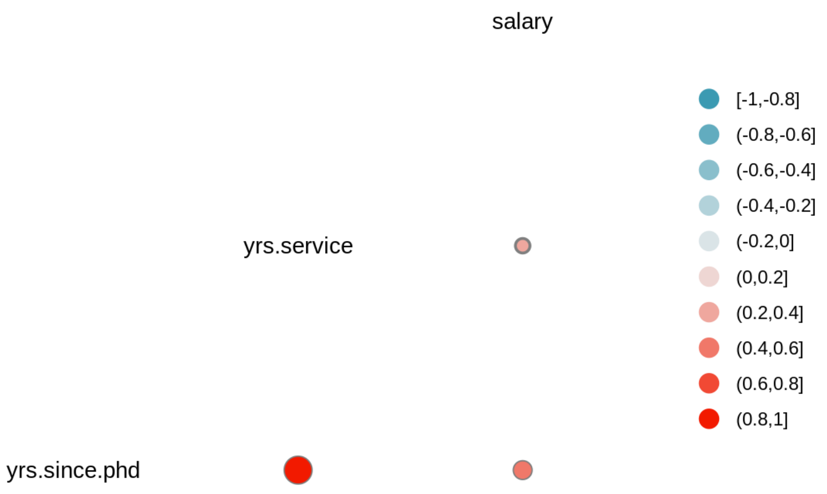
In the following for each numeric column a histogram and box plot will be shown:



Correlation Summary Report

Correlation Plot

This Plot shows the pearson correlation for interval scaled variables. The size and color of the circles indicate the strength of the correlation.



Cluster Summary Report

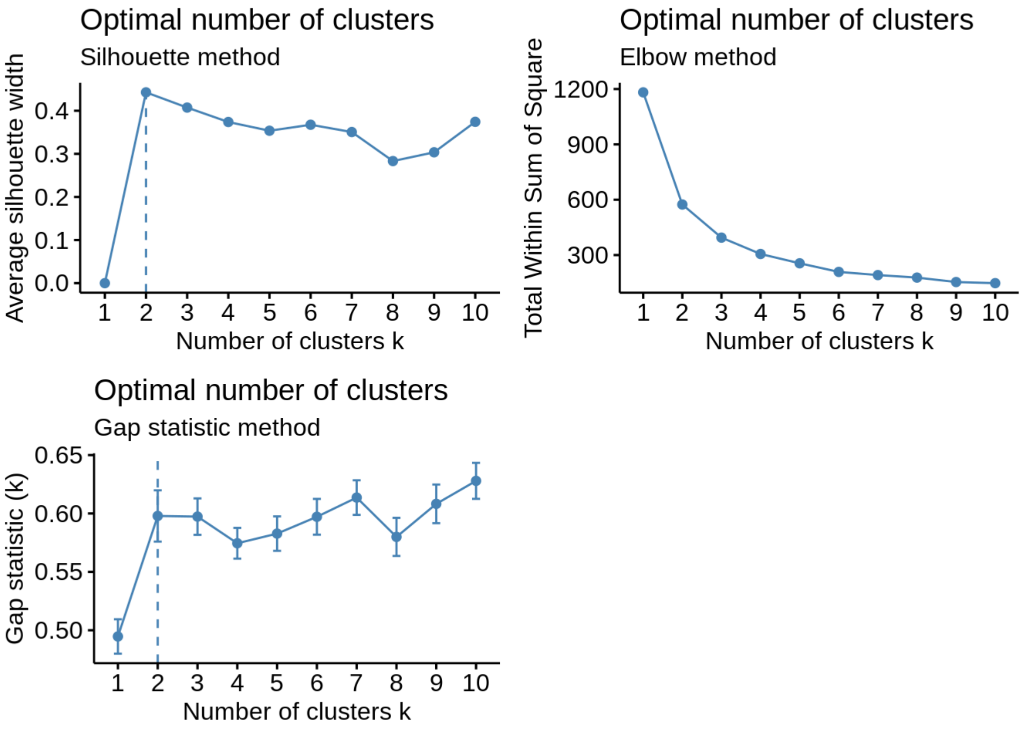
Cluster analysis is a unsupervised learning task, which mainly focuses on grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them.

Overview: All numeric columns

The dataset contains of 3 numeric columns. Since the number of numeric columns is greater than 2, for **vizualization** we compute a principal component analysis and apply the cluster analysis to the respective two principal components:

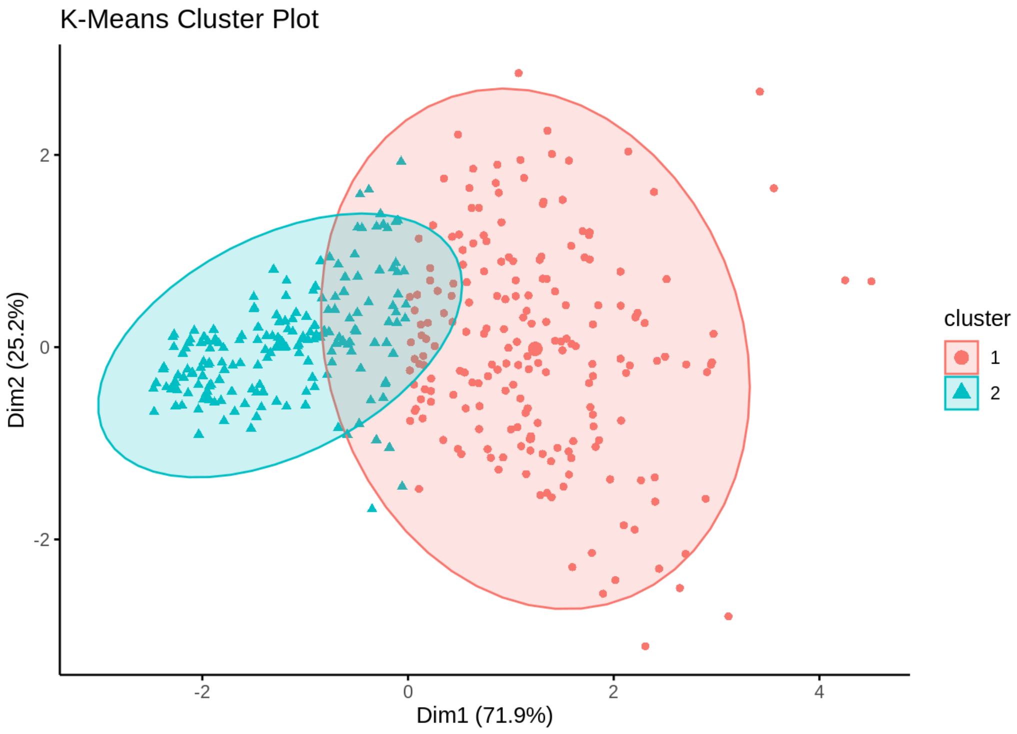
Diagnostics

The following diagnostic plots show how the optimal number of cluster is selected:



Cluster Plot Result

Applying cluster algorithmus we receive following cluster plot:



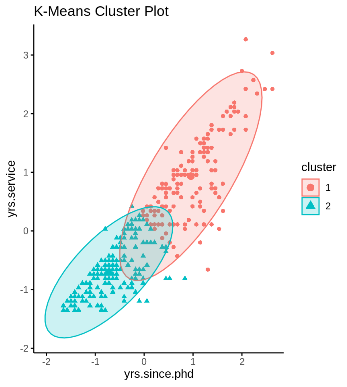
Cluster Result

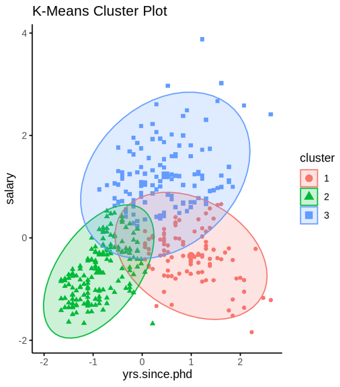
Since Prinicipal components was only for **vizualization** but the clustering algorithm still can handle multidimensional data we receive after transforming the centers from the principal components clusters:

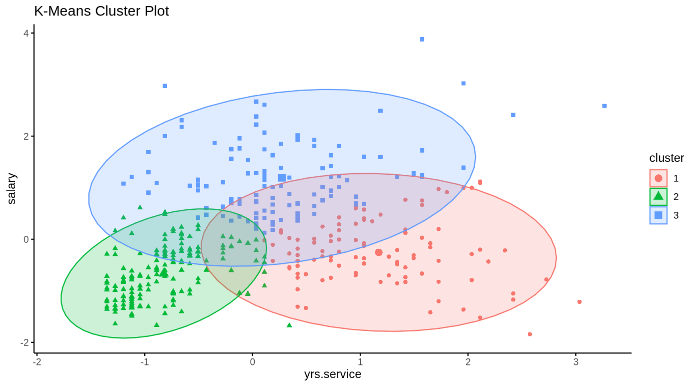
Clustering Centers

yrs.since.phd	yrs.service	salary
0.8139388	0.7848136	0.5034078
-0.8180705	-0.7887974	-0.5059632

Overview: Combinations for numeric columns of dataset







Principal Components Analysis Report for numeric data

Principal Component Analysis (PCA) is a dimensionality reduction method that uses an an orthogonal transformation to reduce a large set of (numeric) variables to a small set of (numeric) variables, called principal components, that still contain most of the information of the large set. Those computed principal components are linearely independent, and hence uncorrelated. The first principal component accounts for as much of the variability/variance in the data as possible, and each succeeding component accounts for as much of the remaining variability/variance as possible.

Prinicipal Component Analysis Summary Results

In the following the principal component rotation (loadings) and the corresponding scree as well as scatterplot will be displayed.

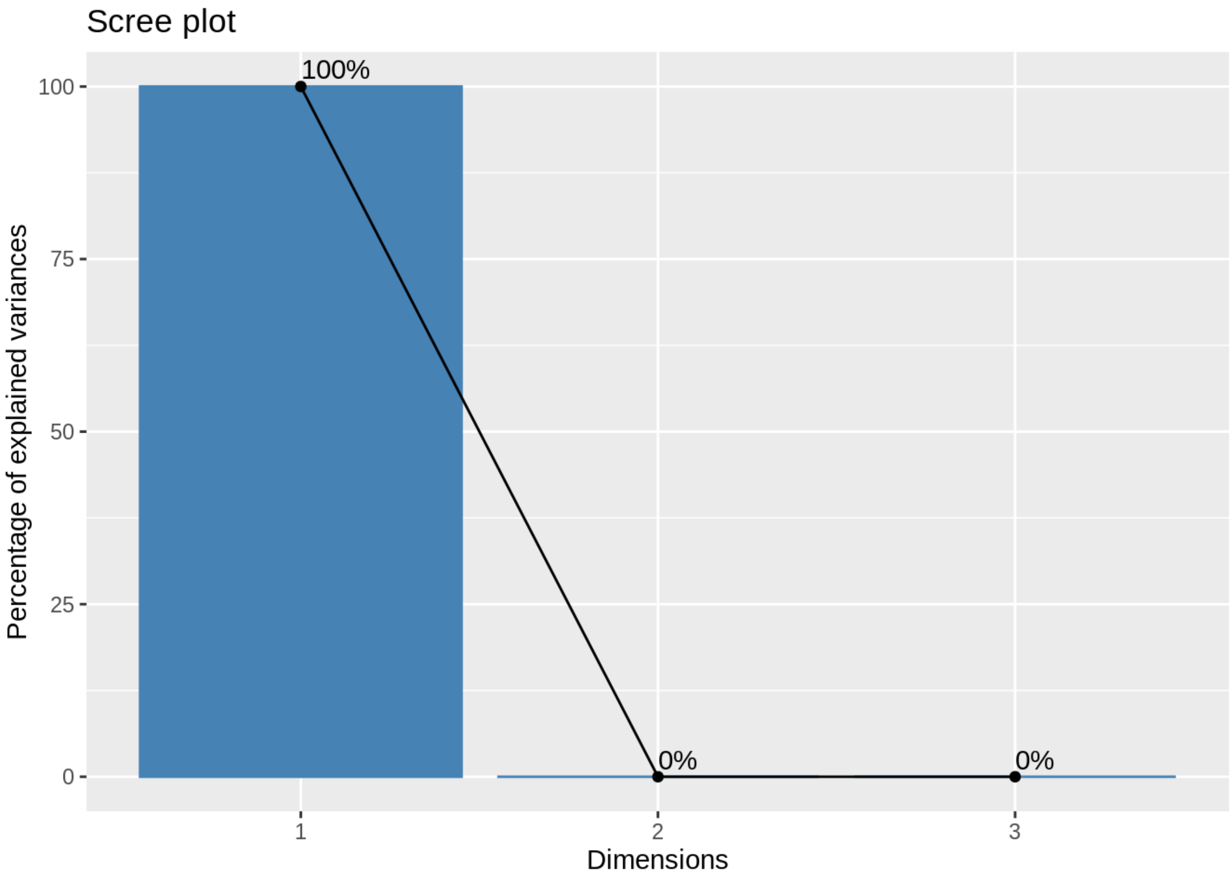
Rotation Loadings and Standard Deviation

The following datatable shows the matrix of variable loadings:

The variances of the 1. to 3. principial component are: 3.036242710^{4}, 16.480986, 3.806264 .

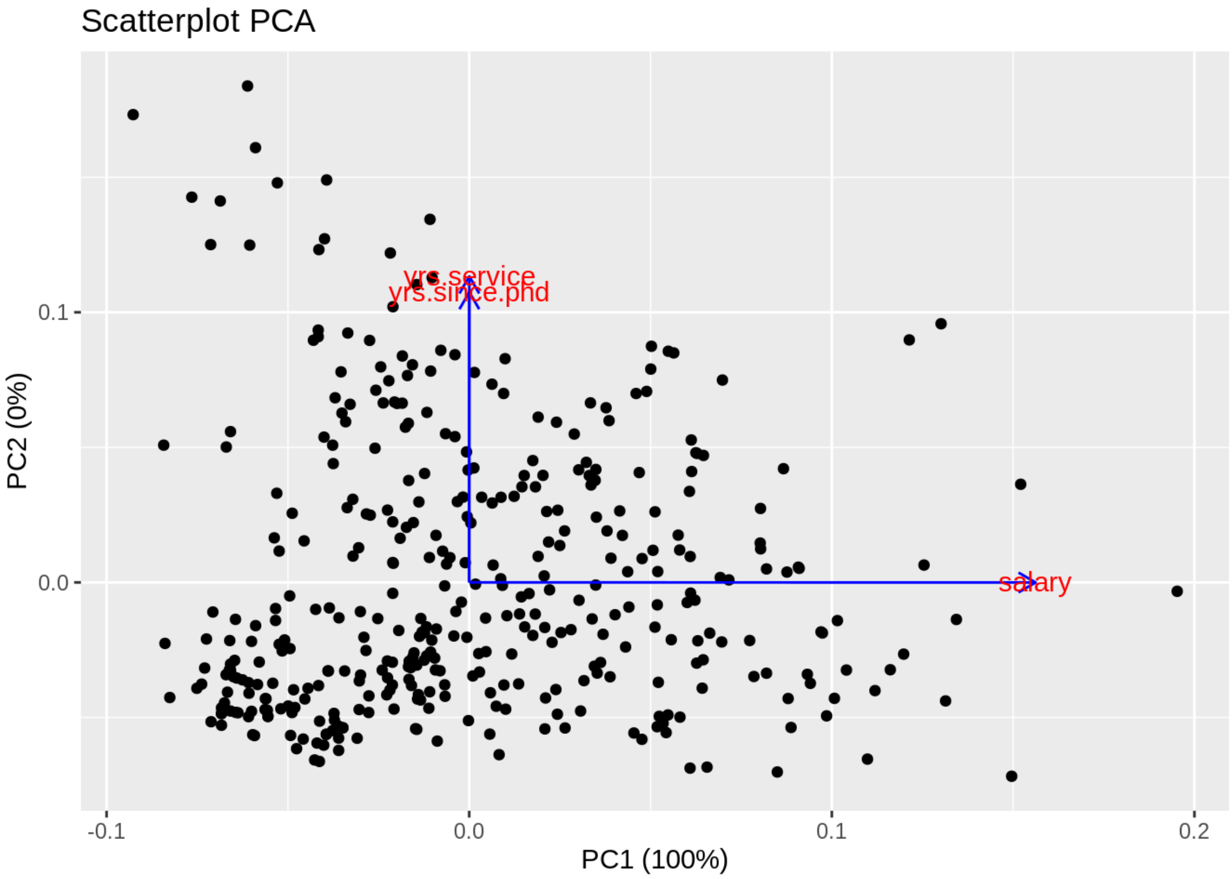
Screeplot for principal component analysis

The following plot shows the percentage amount of variance each principal component explaines in descending order:



Scatterplot for individual observations

The following plot shows how the observations are transformed onto a 2-dimensional space using the rotation/loadings for the first two principal components applied to each origin numeric variable:



Multidimensional Scaling Summary Report

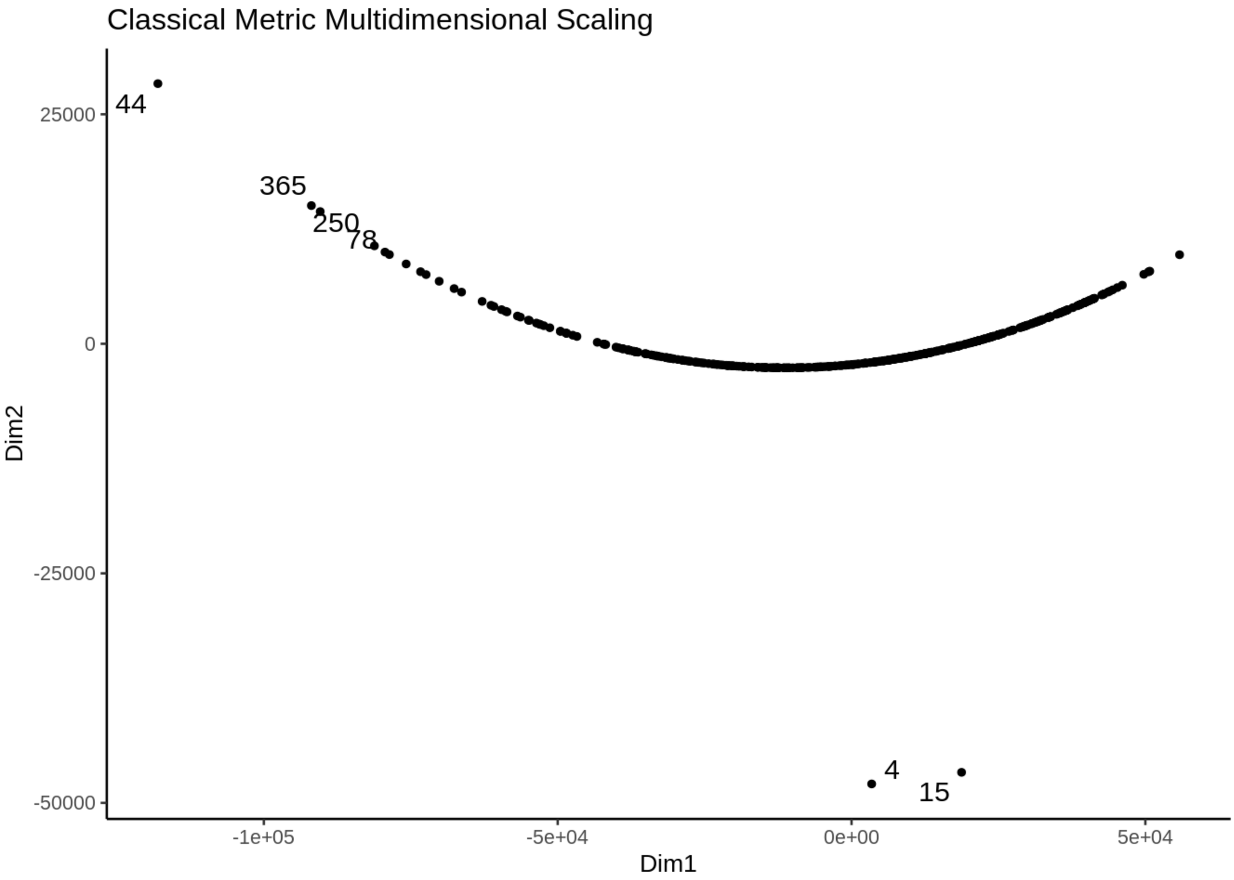
Multidimensional scaling (MDS) is a **visual representation of distances or dissimilarities between sets of objects**. Objects can be colors, faces, map coordinates, political persuasion, or any kind of real conceptual stimuli (Kruskal and Wish, 1978). Objects that are more similar (or have shorter distances) are closer together on the graph than objects that are less similar (or have linger distances). As well as interpreting dissilarities as distances on a graph, MDS can also serve as a dimension reduction technique for high-dimensional data. An MDS algorithm aims to place each object in $K-$dimensional space such that the between-object distances are preserved as well as possible. The approach of AEDA is to set $K = 2$ in order to create a two-dimensional scatterplot to represent the objects.

Multidimensional Scaling Summary Results

The following data frame shows the created dimensions for numeric columns from the dataset:

Multidimensional Scaling Plot

In the following the result of MDS will be shown:



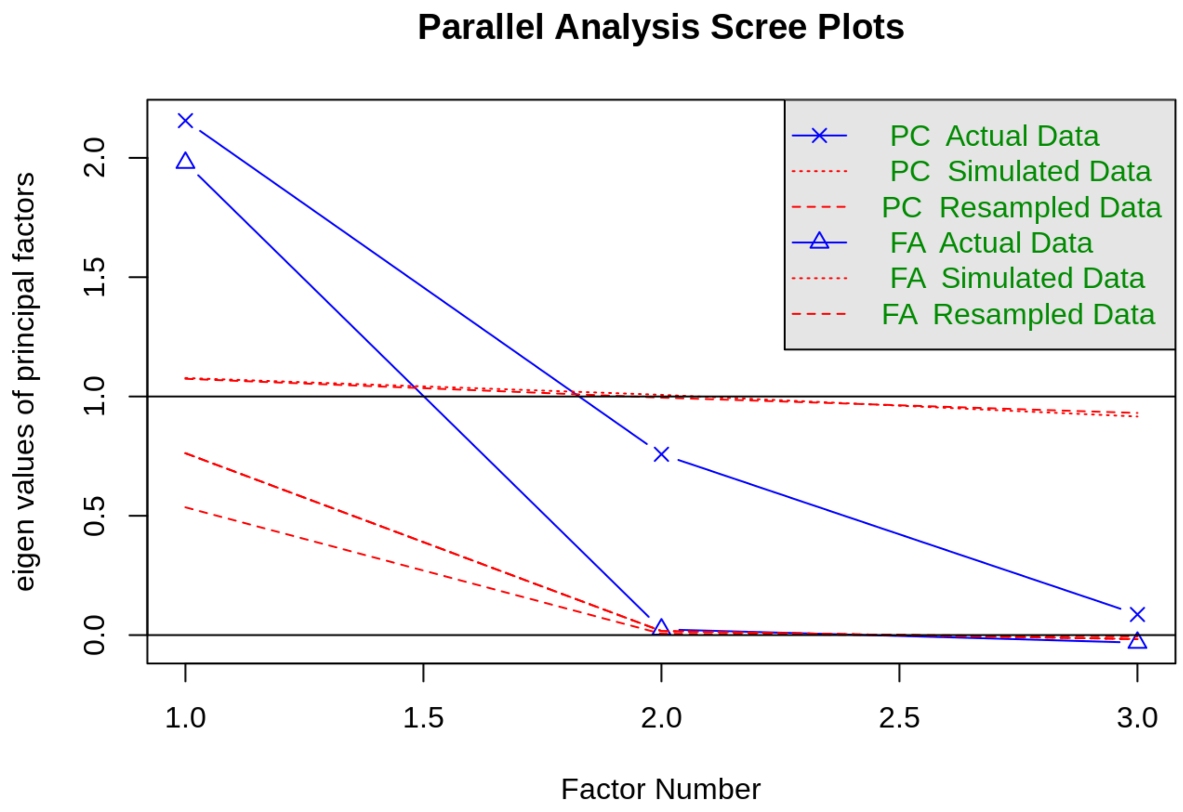
Factor Analysis Summary Report

Factor analysis is a way to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables, called **factors**.
Like principal component as well as multidimensioanl scaling its purpose is dimension reduction. It is a way to find hidden patterns, show how those patterns overlap and show what characteristics are seen in multiple patterns. These **factors** each embody a set of **observed variables** that have similar response patterns.

Factor Analysis Summary Results

Pre-Analysis for optimal number of factors

In the following a pre analysis for the optimal number of factors will be displayed:



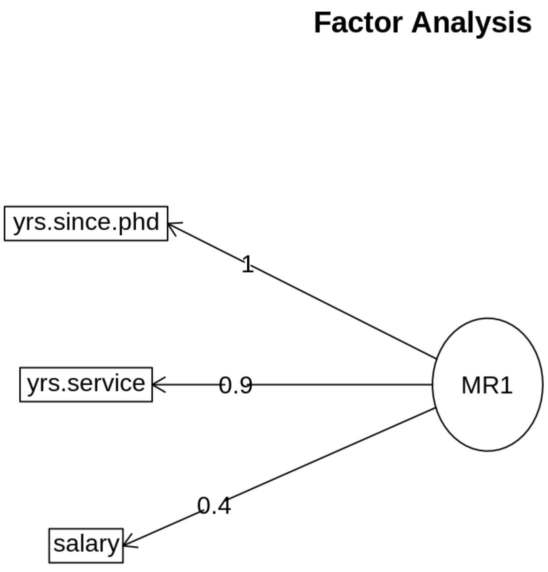
According to the parallel analysis 1 factors should be taken.

Factor Loadings

The following data frame shows the factor loadings for each numeric column. Note that the rows are the numeric columns.

Factor Analysis Graph

The following graphs visualizes the factor loadings. A cutoff of 0.3 will be applied:



Processing math: 100%