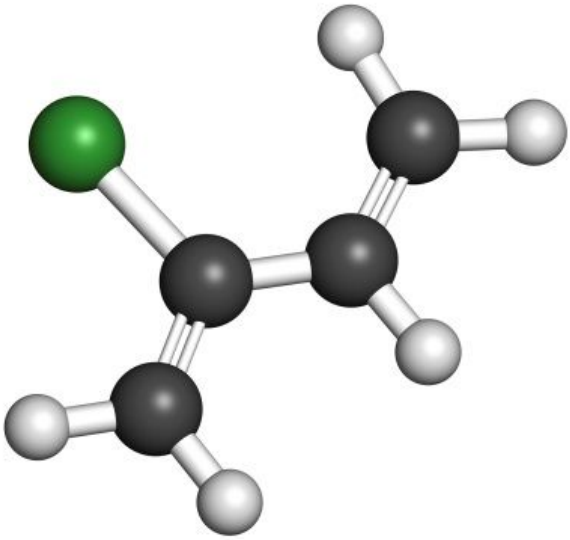
PCA Project Report

Polymer Manufacture



João Viegas

21.05.2018

Student R00157699

Statistical Methods for Big Data

STAT 8007

Higher Diploma in Data Science and Analytics

Table of Contents

Introduction **3**

Methods **4**

Exploratory Data Analysis 4

Principal Component Analysis 4

Results **6**

Part 1 6

Exploratory data analysis 7

Principal Component Analysis (PCA) 11

Analysis of patterns relating principal components and ranking of the polymers 15

Analysis of relation between materials and polymers ranking 16

Part 2 19

Analysis of relation between strength and principal components and polymer materials 19

Analysis of relation between warping resistance and principal components and polymer materials 20

Part 3 23

Analysis of polymers likely performance based on materials composition 24

Strength Analysis 24

Warping resistance analysis 25

Discussion **26**

Introduction

In the process of producing protective coatings, a company is developing a new polymer. This polymer must be of high strength and also resistant to warping. The company is researching these properties, strength and warping resistance, in 60 different polymers, each one with a different combination of materials, labeled as A, B, C, D, E and F. All these polymers were ranked, regarding those properties, and the main purpose of this research is, using Principal Component Analysis (PCA), identify patterns relating the properties, strength and warping resistance, and the materials, in order to define the best possible product.

Methods

This research was conducted using the *R language*, which is considered the *"lingua franca"* of statistics nowadays. It is open source, runs in all major platforms, provides a free development environment and tools and has an extensive support community that creates modules for every possible data science feature that we can think of.

The research is comprised, however, of a set of stages, in which specific techniques and *R* libraries were used.

Exploratory Data Analysis

In the first part of this experiment we've studied the effect of removing observations with missing data, as 15 of the 60 observations had missing data on the material F (Mat_F) column. As alternative approach we've used the *irmi* (Iterative robust model-based imputation) function, from the *VIM* package, to replace the missing values, in that the remaining variables serve as regressors to derive a replacement value, so that we could use all the observations. We then assessed both approaches, A) with incomplete data removed and B), with replacement values provided by the *irmi* function..

To study the correlation between variables we've used the function *pairs*, to cross the scatterplots between each variable. We also used the basic R graphics functions, *hist* and *boxplot*, to verify the assumption of normality of the different variables. We then transformed one variable with the *log* function, Mat_A, which was positively(right) skewed.

We've also tested the dataset approaches for outliers. We've used the *grubbs.test* function, from the *outliers* package, for univariate outliers detection, and the *mahalanobis* function to identify multivariate outliers. We ended up removing a small set of outlier observations.

Principal Component Analysis

In order to perform Principal Component Analysis, we've used the *prcomp* function to calculate the scores, rotation matrix and the eigenvalues. To check the scores for outliers we've plotted a confidence ellipse around the scores plot with the function *dataEllipse*, from the package *car*.

We've also used the function *loadingplot*, from package *pls*, to print the component loadings of each variable, and along with the common plot function, we've used the *biplot* function to overlay the pca scores with the loadings.

In order to select the principal components, we've used the *Kaiser Criterion*, where we tried to select the set of components that could globally reflect more than 95% of the variance in the dataset. We've also inspected the components using the *screeplot*.

To determine the goodness of fit of the selected components we've plotted its residuals in a barplot and thus determining how much the research variables are explained by the selected components model.

In the second part of this experiment, we relate the scores plot to the loadings plot in order to realize the relationship between the loadings of each variable related to each principal component and thus identify a pattern between the rankings and the variables defining every observation.

In the third, and last, part of the experiment we've used the predict function, that along with the PCA model and new data as input, generates scores of the principal components for the new data. With these scores and using the PCA model devised before, we could then infer about the strength and warping resistance properties in this new data.

Results

In this section we describe the results of the experiment with details on the 3 parts involved.

Part 1

In Part 1 we've conducted a parallel exploratory data analysis in order to select the dataset to use. The data comprises 60 observations of different polymers, and its different materials composition, as seen in Table 1.

	Ranking	Mat_A	Mat_B	Mat_C	Mat_D	Mat_E	Mat_F
1	24	558.2576	3.9988	24.9986	15.9495	8.7844	21.6419
2	40	173.1652	4.8525	17.1475	17.7145	4.6662	29.8481
3	46	97.1960	5.7493	16.2507	17.8123	4.4379	32.8661
4	37	160.3977	5.8888	16.1112	16.2386	8.1100	48.9768
5	42	112.6512	6.0177	15.9823	16.9798	6.3804	31.0183
6	51	65.0137	6.3144	15.6856	17.9769	4.0540	35.0962

Table 1: subset of Part 1 dataset

A summary of the data shows the existence of missing values for a variable, Mat_F, as can be seen in Table 2.

Ranking	Mat_A	Mat_B	Mat_C	Mat_D	Mat_E	Mat_F
Min. : 1.00	Min. : 2.225	Min. : 3.999	Min. : 3.018	Min. : 3.001	Min. : 3.999	Min. : 6.041
1st Qu.:15.75	1st Qu.: 13.421	1st Qu.: 8.505	1st Qu.: 7.850	1st Qu.: 6.527	1st Qu.:11.021	1st Qu.:18.873
Median :30.50	Median : 65.266	Median :10.974	Median :11.026	Median :10.857	Median :20.666	Median :28.717
Mean :30.50	Mean :128.821	Mean :11.232	Mean :10.885	Mean :10.847	Mean :20.689	Mean :26.911
3rd Qu.:45.25	3rd Qu.:160.593	3rd Qu.:14.150	3rd Qu.:13.495	3rd Qu.:14.991	3rd Qu.:30.771	3rd Qu.:33.447
Max. :60.00	Max. :627.289	Max. :19.000	Max. :24.999	Max. :18.002	Max. :39.003	Max. :48.977
						NA's :15

Table 2: summary description of Part 1 dataset

In one experiment we've removed the rows with missing data (A), and on the alternative experiment we've applied the *irmi* function in order to replace missing variable values (B). Also in parallel we've conducted a PCA analysis, and then we assessed both in term of goodness of fit.

The outcome of the parallel analysis on the dataset, as described in the R code provided along with this report, was that we've got slightly better outcomes with the alternative experiment (B), hence, we will describe here the analysis that took place with that dataset, with 60 observations, and the missing data replaced by the *irmi* function.

Exploratory data analysis

After applying th irmi function the summary description can be seen in Table 3.

Ranking	Mat_A	Mat_B	Mat_C	Mat_D	Mat_E	Mat_F
Min. : 1.00	Min. : 2.225	Min. : 3.999	Min. : 3.018	Min. : 3.001	Min. : 3.999	Min. :-9.122
1st Qu.:15.75	1st Qu.: 13.421	1st Qu.: 8.505	1st Qu.: 7.850	1st Qu.: 6.527	1st Qu.:11.021	1st Qu.:19.621
Median :30.50	Median : 65.266	Median :10.974	Median :11.026	Median :10.857	Median :20.666	Median :28.772
Mean :30.50	Mean :128.821	Mean :11.232	Mean :10.885	Mean :10.847	Mean :20.689	Mean :26.787
3rd Qu.:45.25	3rd Qu.:160.593	3rd Qu.:14.150	3rd Qu.:13.495	3rd Qu.:14.991	3rd Qu.:30.771	3rd Qu.:33.967
Max. :60.00	Max. :627.289	Max. :19.000	Max. :24.999	Max. :18.002	Max. :39.003	Max. :48.977

Table 3: summary description of Part 1 dataset, with missing data rows replaced by *irmi* function

After this point all the analysis is conducted on the polymers variables data, not on the ranking data, for obvious reasons. The scatterplots across all the variables show that Mat_B is highly correlated with Mat_C and that Mat_D is also highly correlated with Mat_E.

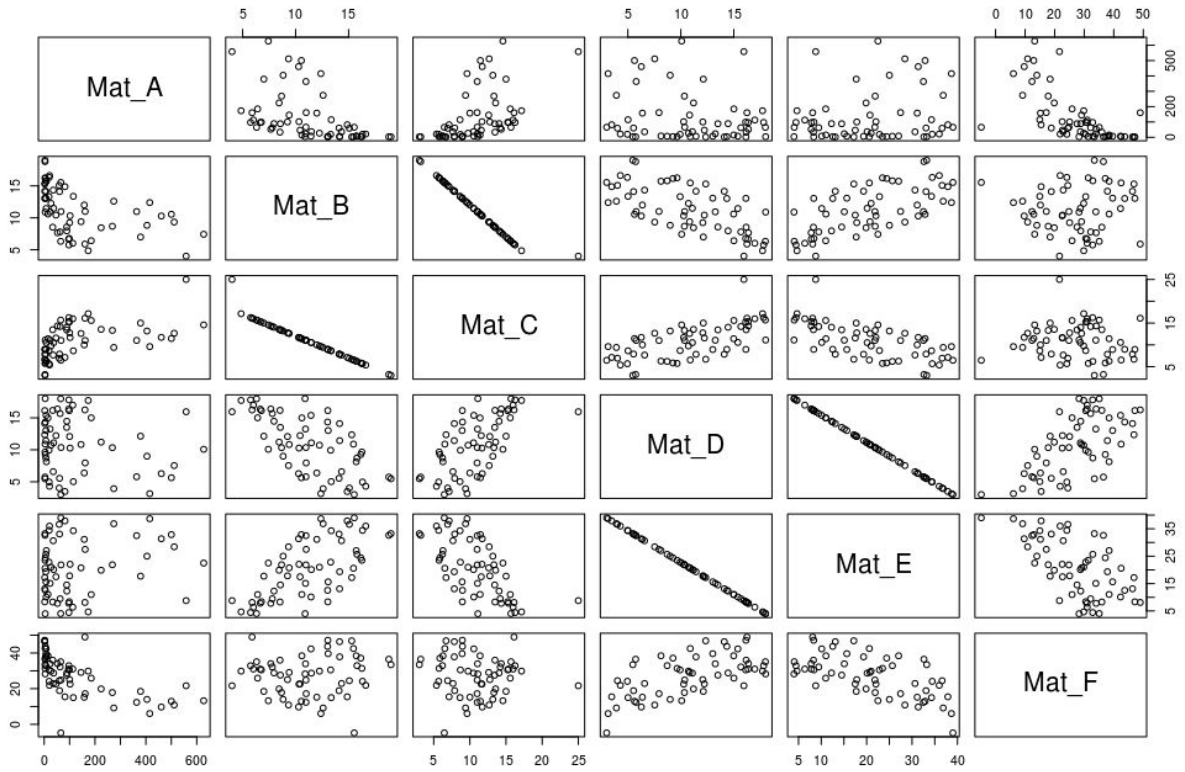


Figure 1: dataset variables scatterplots

We then assessed the normality assumption of the data, as required for PCA analysis.

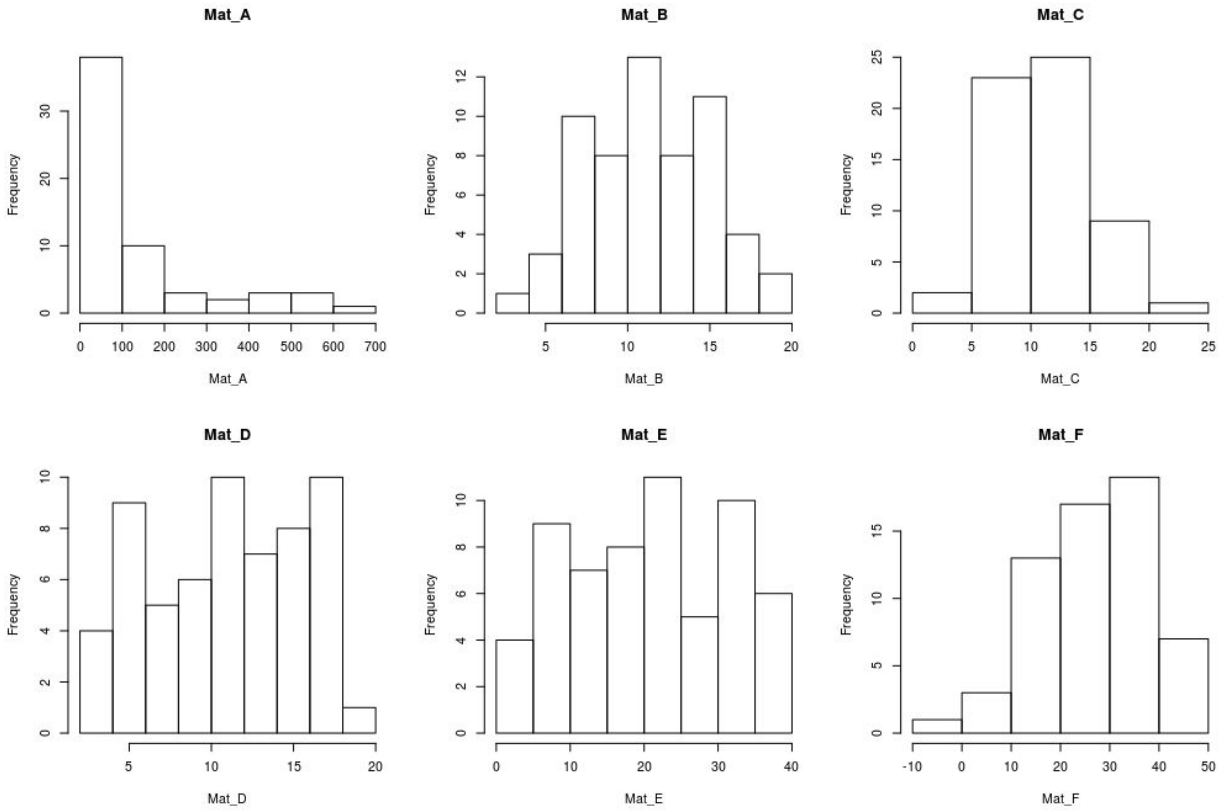


Figure 2: histograms for the dataset variables

Analysing the histograms, Mat_A revealed to be positively skewed, so we decided to transform it with the log function.

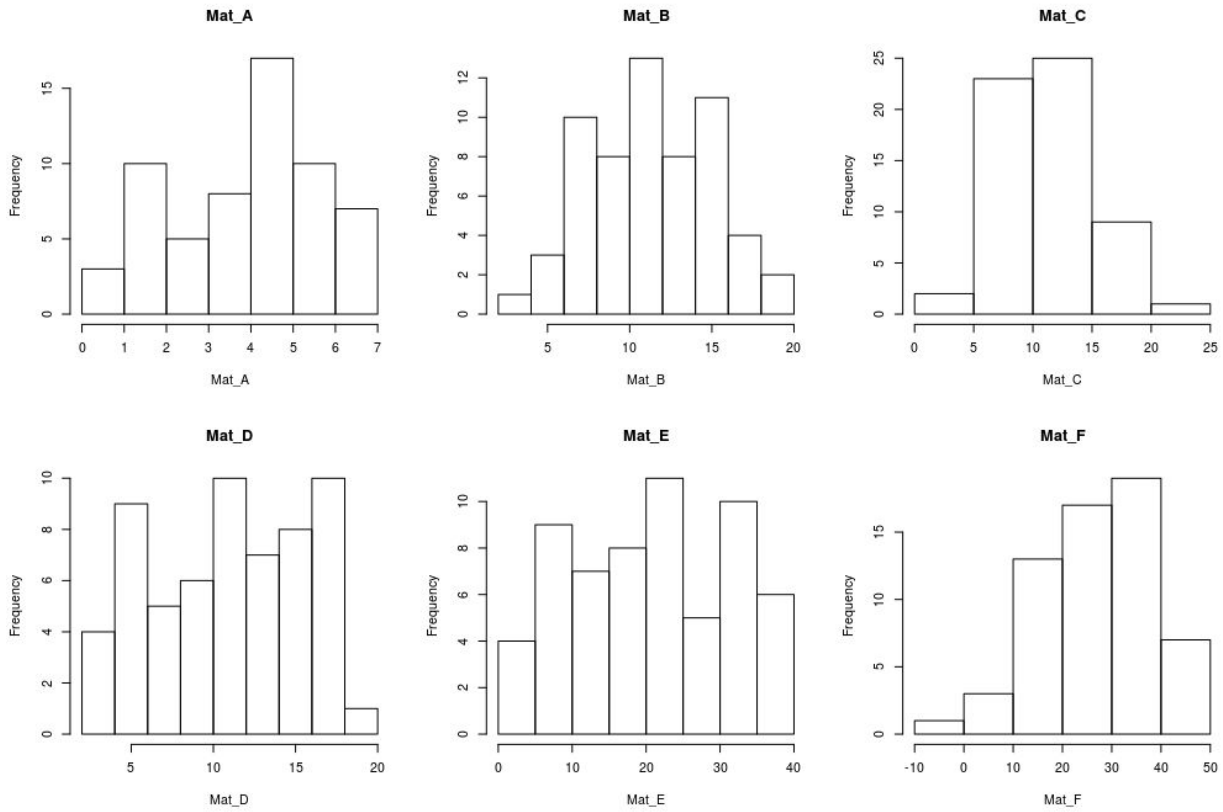


Figure 3: histograms of the dataset variables after transforming variable Mat_A

We then checked for outliers in the dataset.

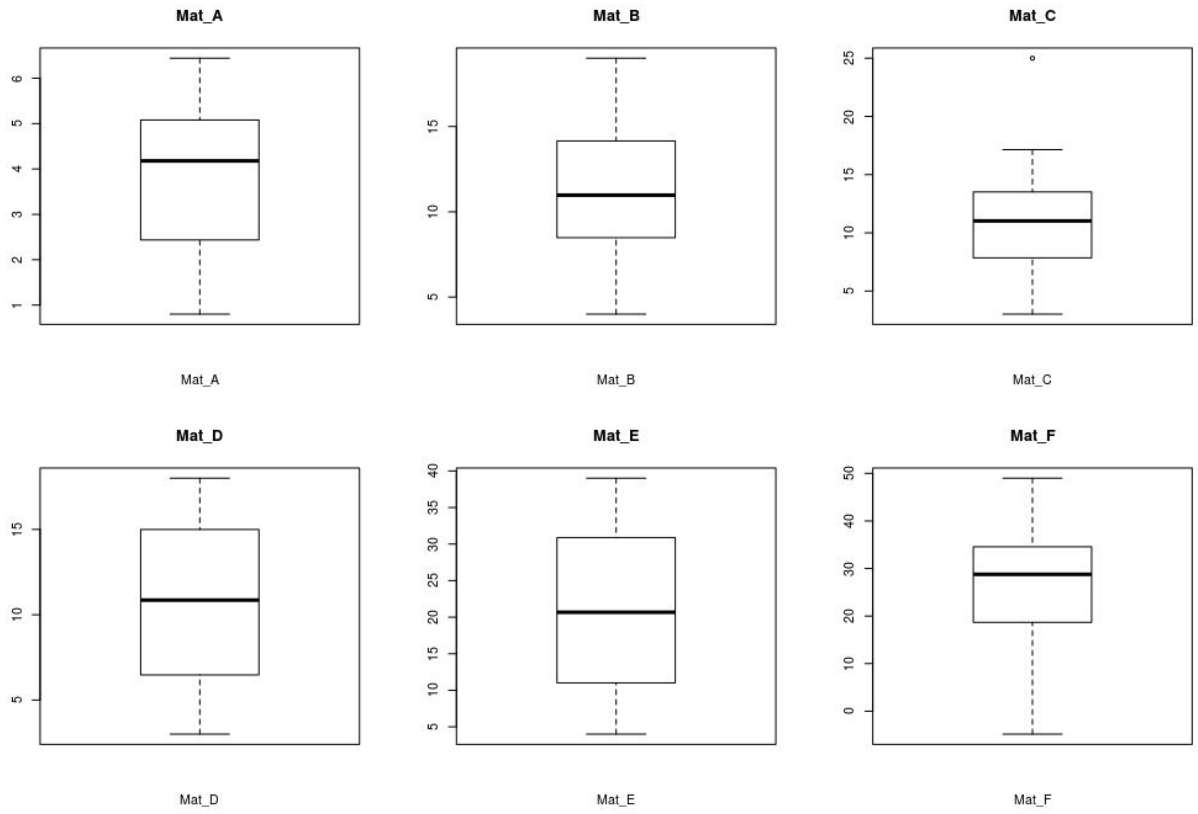


Figure 4: boxplots for the dataset variables

Accordingly to the boxplots in Figure 4, we decided to investigate the Mat_C outlier, in line 1, using the Grubbs test:

Grubbs test for one outlier

data: d$Mat_C
G = 3.54550, U = 0.78333, p-value = 0.005356
alternative hypothesis: highest value 24.9986 is an outlier

As the p-value is smaller than 0.01, we decided to confirm this with the multivariate outliers analysis with the *mahalanobis* function.

The *mahalanobis* distance for line 1, where the Mat_C outlier was located, was 58.0163932, and compared to the test statistic χ^2 with 95% confidence level and 6 degrees of freedom, 12.59159, we can then assume this is an outlier and remove this observation from the dataset. After this we assume there are no more outliers in the dataset.

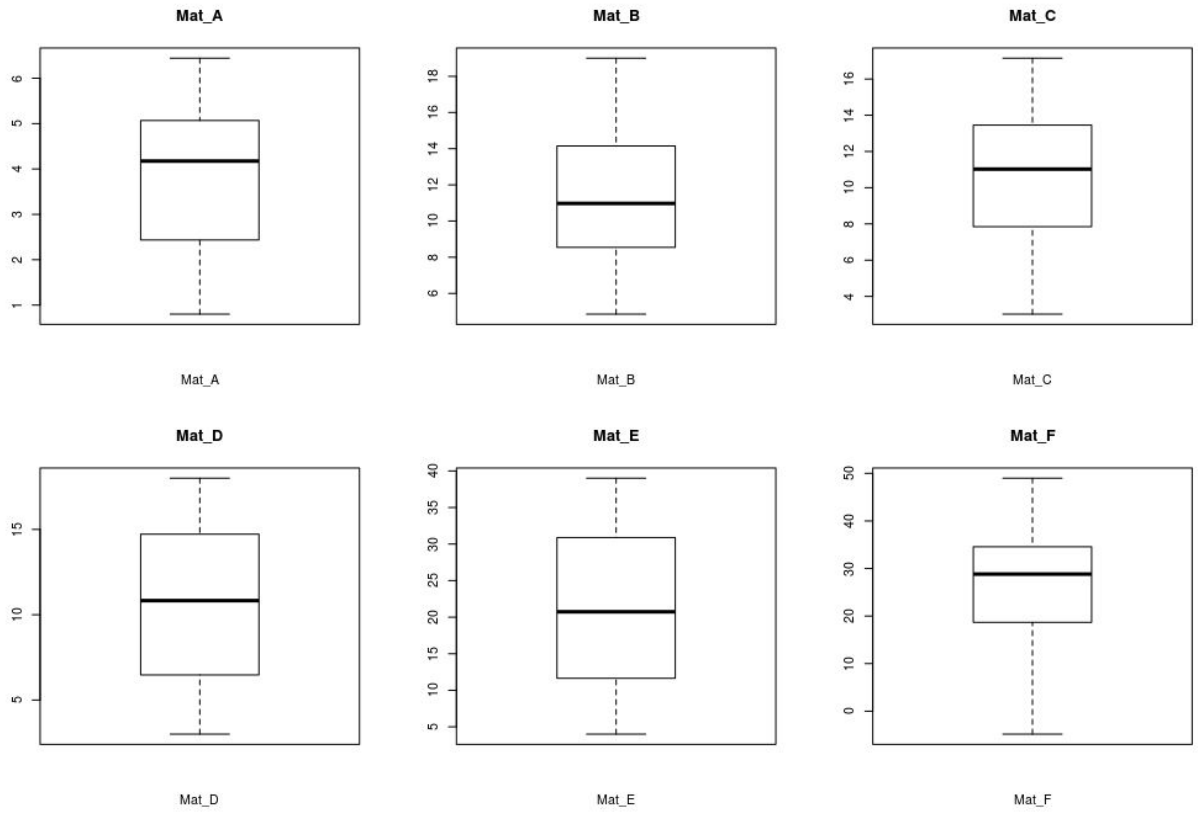


Figure 5: boxplots for the dataset variables, after removing of the outlier observation

At this stage we are ready to perform the PCA analysis. We store the rankings information in the row names and remove the variable from the dataset.

Principal Component Analysis (PCA)

The PCA analysis gave us the following information regarding the principal components found:

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.864	1.5483	0.35864	0.0005154	4.458e-05	1.501e-05
Proportion of Variance	0.579	0.3995	0.02144	0.0000000	0.000e+00	0.000e+00
Cumulative Proportion	0.579	0.9786	1.00000	1.0000000	1.000e+00	1.000e+00

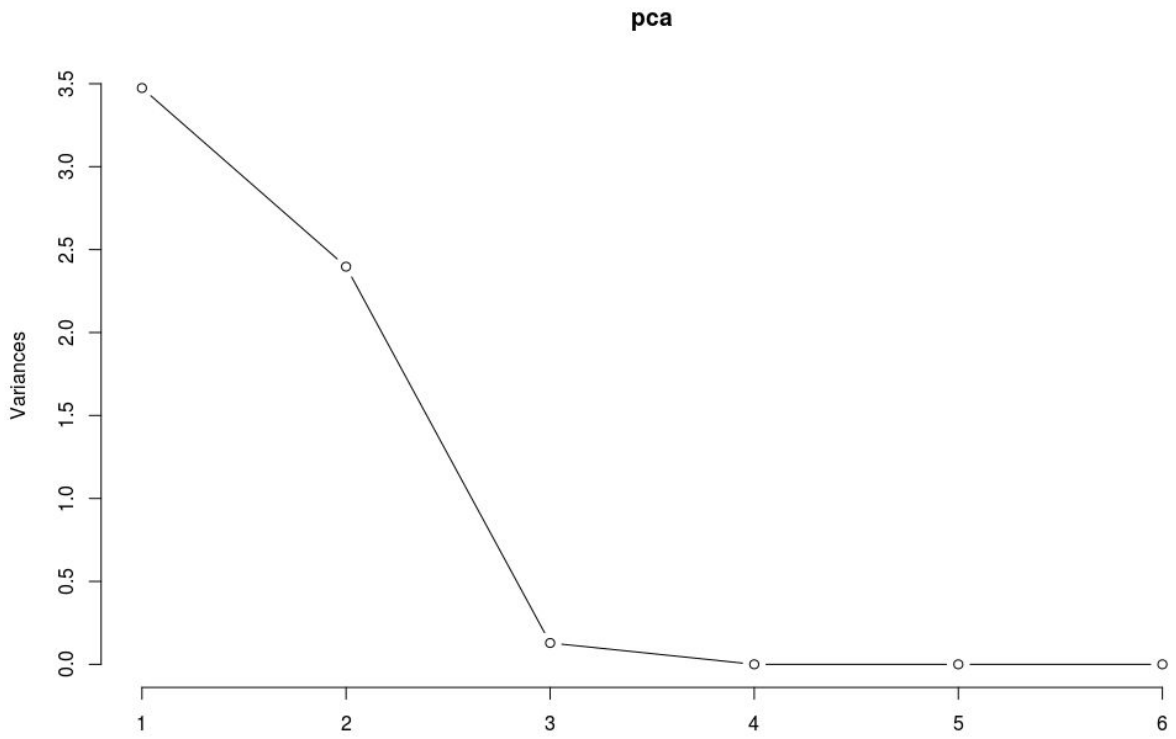


Figure 6: PCA analysis screeplot

According to the screeplot and using the Kaiser Criterion to select the number of principal components with eigenvalues above 1, where the sd is the square root of the eigenvalue associated with the component, we would choose PC1 and PC2, also because they are responsible respectively for 57.9% and 39.95% of the variance and for the aggregate value of 97.86% of variance os the model.

We can also study the goodness of fit of the model, with one or two components, on how it can explain the variance of the polymer variables.

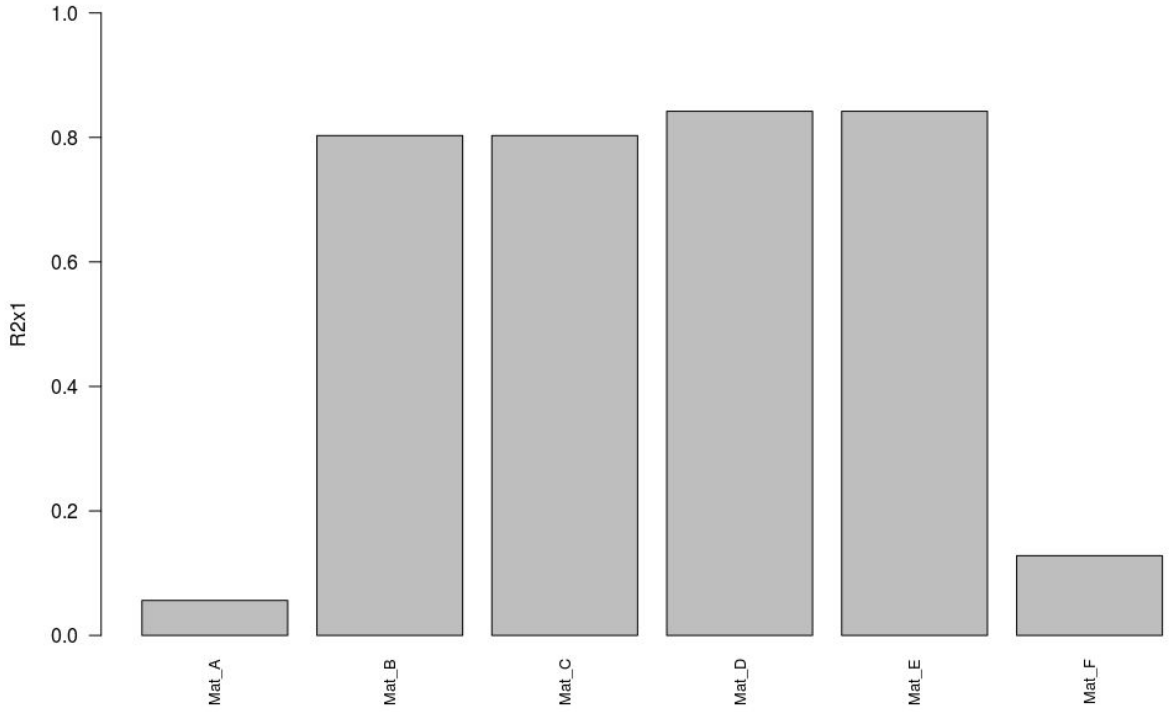


Figure 7: goodness of fit analysis for PCA model with one component, PC1

Regarding the model with one component, PC1, it can explain most of the variance of the variables Mat_B,Mat_C,Mat_D and Mat_E, but can only explain a small part of Mat_F and even smaller part of Mat_A.

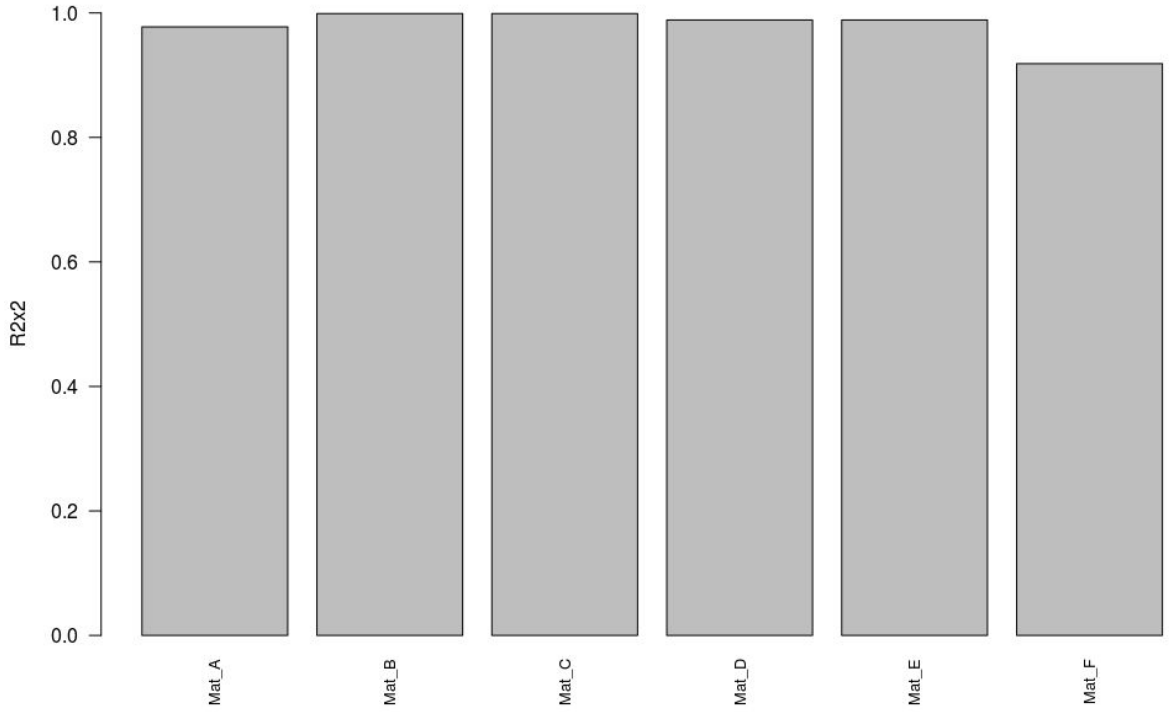


Figure 8: goodness of fit analysis for PCA model with two components, PC1 and PC2

The 2 components model, PC1 and PC2, does explain almost entirely the variance of all the variables.

We could then check for score outliers, plotting a confidence ellipse around the model scores.

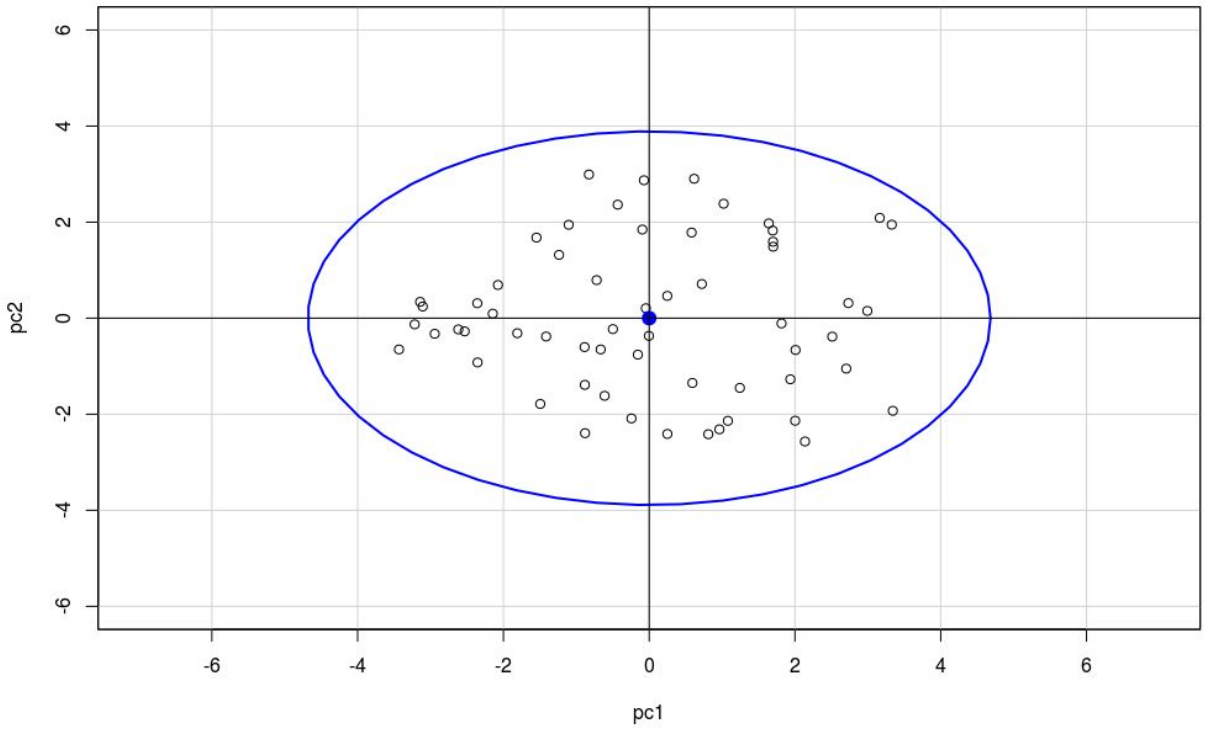


Figure 9: confidence ellipse on the scores plot for components PC1 and PC2

At this point, we define our model comprising both principal components PC1 and PC2.

Analysis of patterns relating principal components and ranking of the polymers

We can now analyse the rankings and its correlation with the principal components.

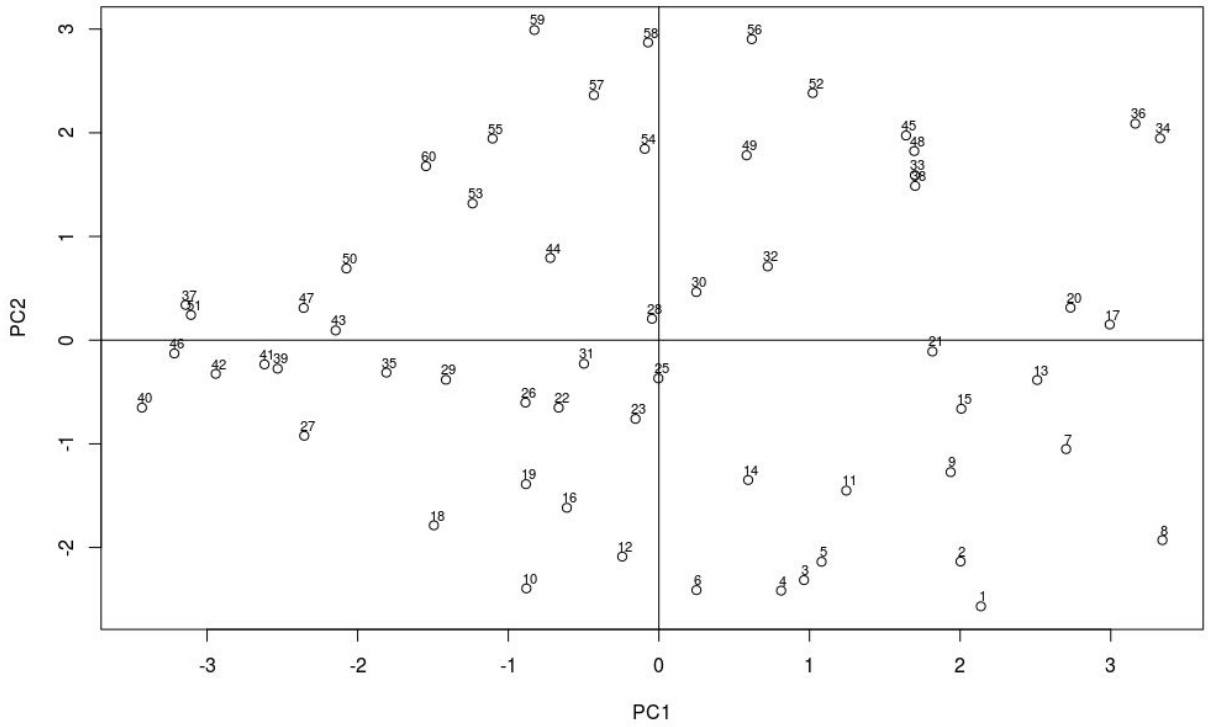


Figure 10: PCA score plot with rankings labelled

Clearly there is a strong positive correlation between PC2 and the ranking. It is specially clear the fact that the higher rankings are located on the higher values of PC2, higher than 1, and around the value 0 of PC1, but mostly in the negative region of PC1, or maybe we can just say that most of the higher ranking observations are on the upper left quadrant of the plot.

We can also say that the lower rankings happen to be located in the right lower quadrant, although it seems there is again a slightly stronger correlation with PC2 than with PC1.

Analysis of relation between materials and polymers ranking

Here we can inspect the loadings plot and identify any relationship between material variables and polymers ranking.

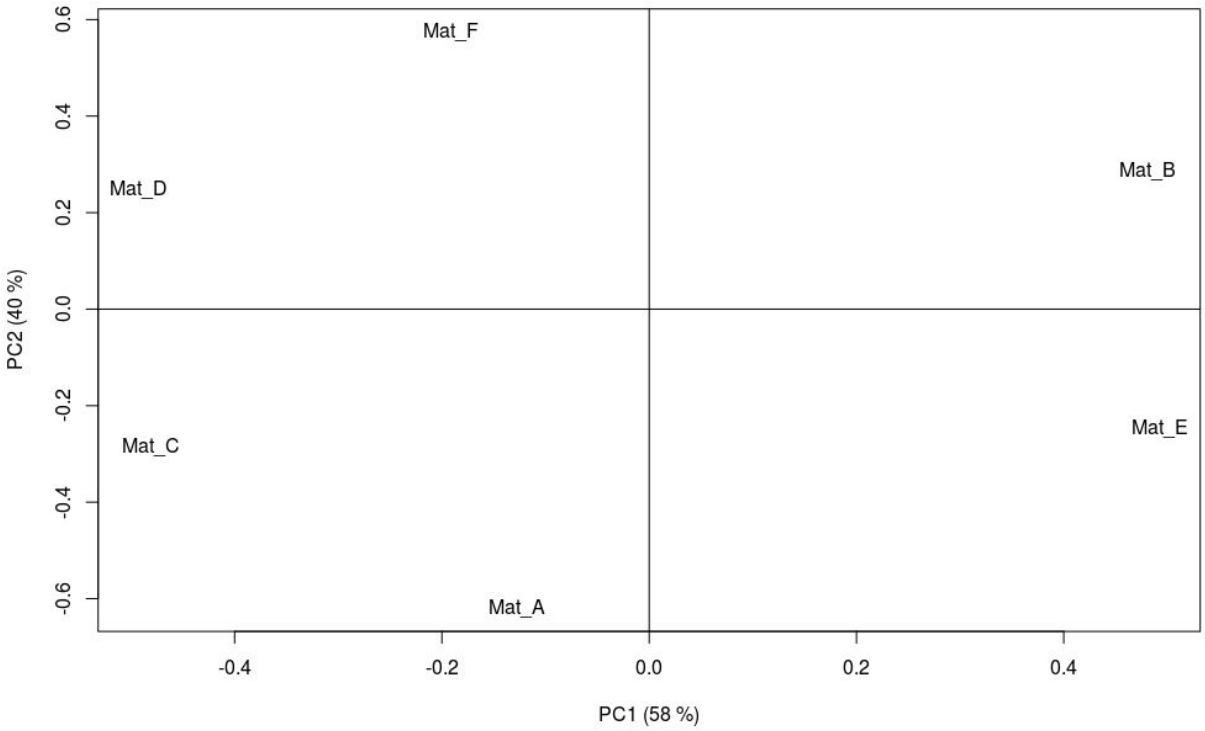


Figure 11: PCA loadings plot relating principal components and dataset variables

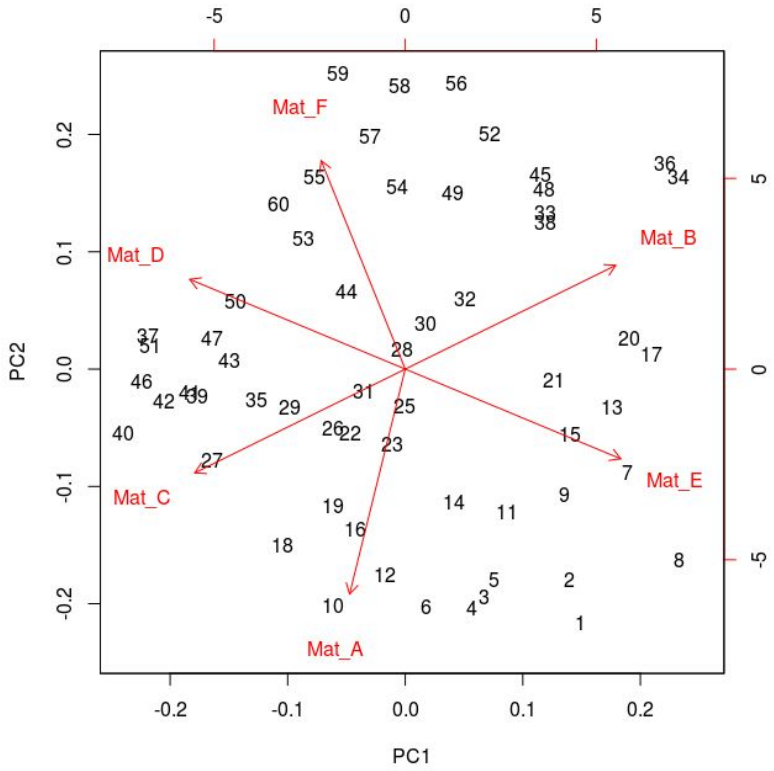


Figure 12: PCA biplot, overlaying scores plot and loadings plot

From the loadings plot i Figure 11, but specially from the biplot in Figure 12, we can see that Mat_E and Mat_A are strongly related with the higher ranked polymers whereas Mat_F, and less significantly Mat_D, are the variables/materials related with the lower ranked polymers.

Part 2

In this Part 2, we are using a dataset with the same variables data, but instead of a general ranking information, we have now two additional columns with rankings related with Strength and Warping resistance, and we want to establish relationships between our previous PCA analysis, and the new rankings information.

Analysis of relation between strength and principal components and polymer materials

We now inspect the scores plot labelled with the strength rankings and the biplot.

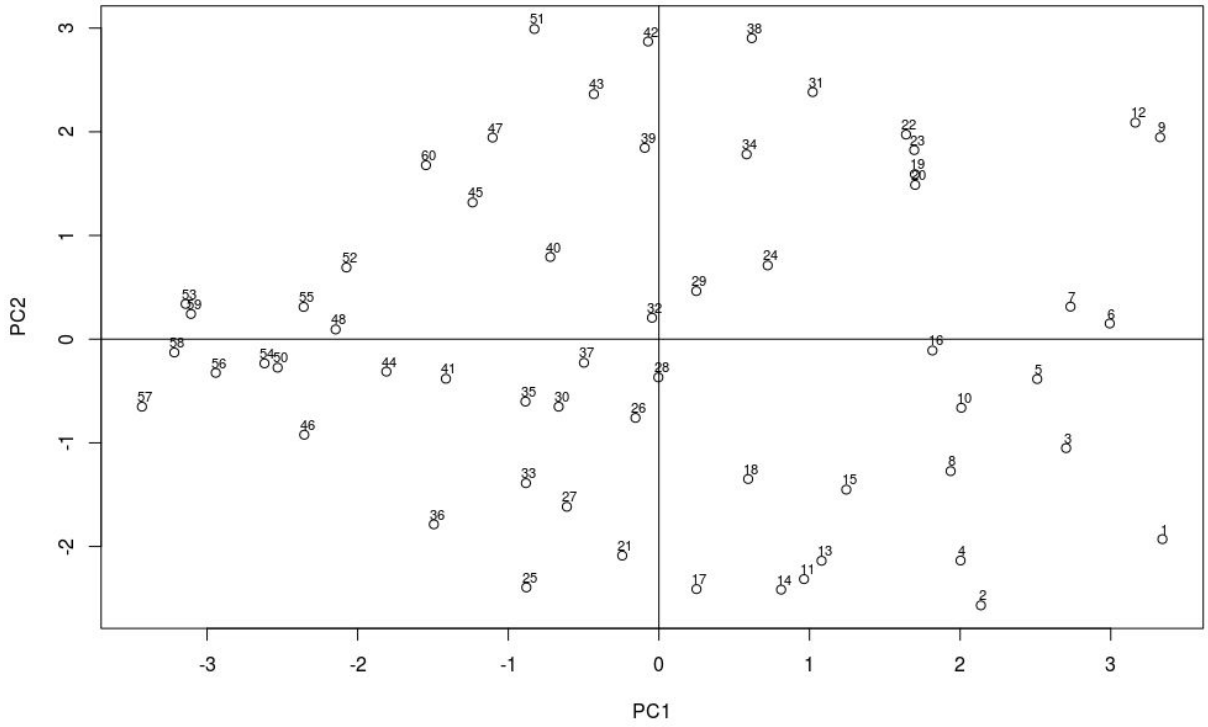


Figure 13: PCA score plot with strength rankings labelling

From the scores plot we can see the that the observations with higher strength ranking are positively correlated with PC1, and in a slight degree, negatively correlated with pc2. The higher ranked observations are thus distributed in the right quadrants with the extreme values in lower right quadrant. Consequently, the lower ranked observations are almost in its entirety negatively correlated with PC1, with the correlation with PC2

not being that much clear in this case, so most of the lower ranked observations are placed on the most negative values of PC1 and in between values [-1:1] of PC2.

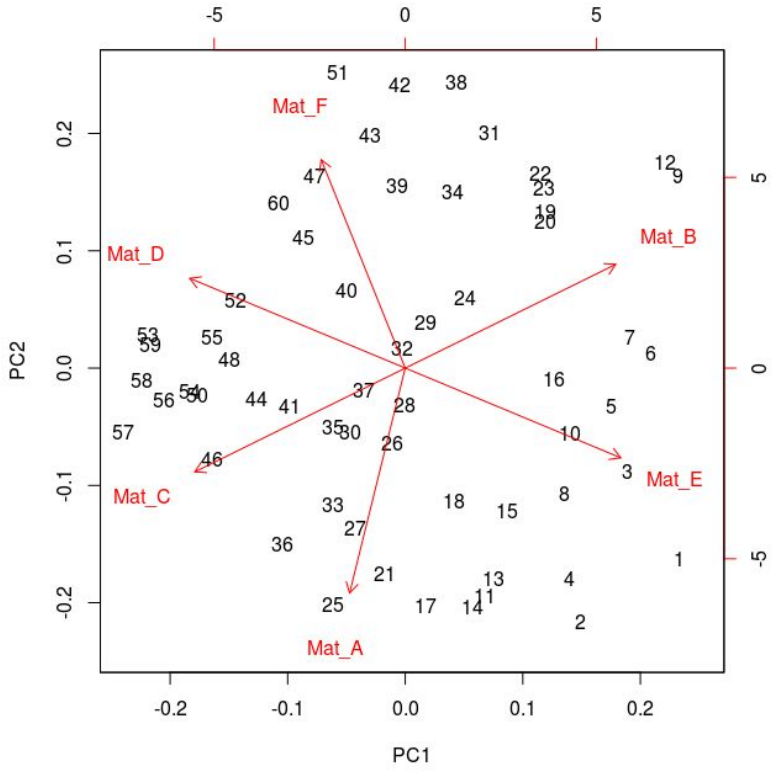


Figure 14: PCA biplot, overlaying scores plot and loadings plot

The biplot, reflecting the loadings and the scores, shows that Mat_E and in a lesser extent, Mat_B, are the most important materials when it comes to strength performance as are the most correlated with higher ranked observations. In the same way we can say that Mat_C and Mat_D are also related to lower Strength outcomes.

Analysis of relation between warping resistance and principal components and polymer materials

We now inspect the scores plot labelled with the warping resistance rankings and the biplot.

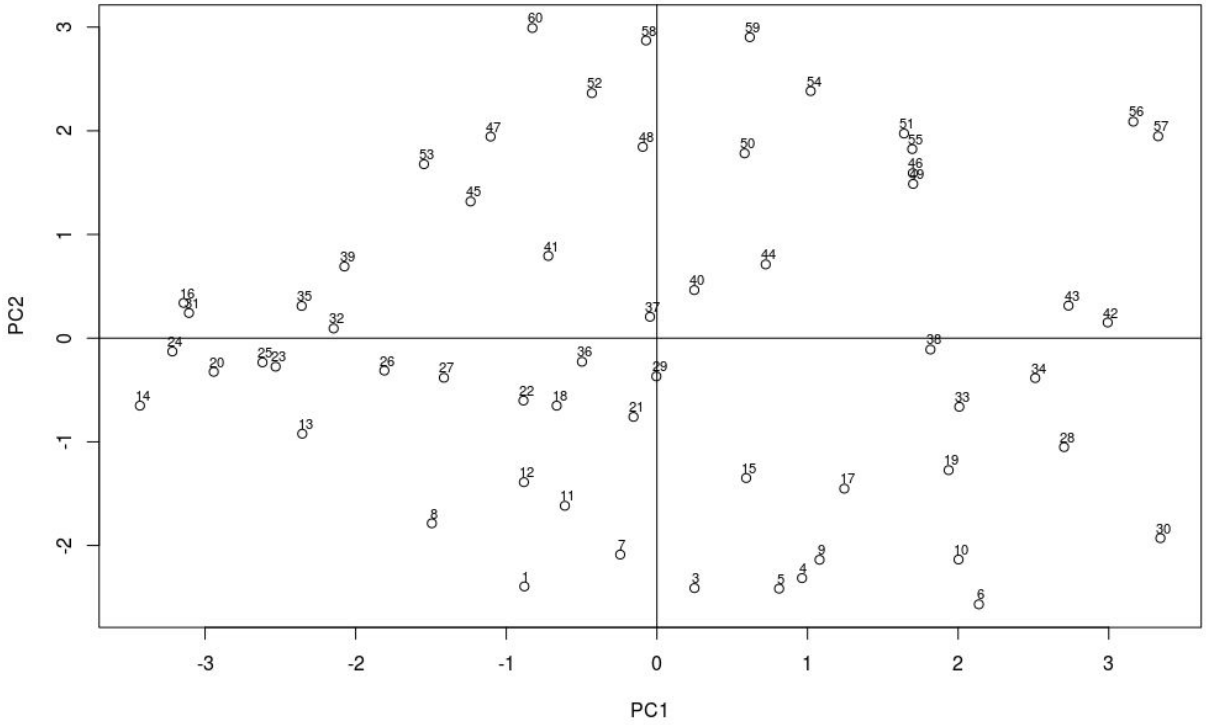


Figure 15: PCA score plot with warp resistance rankings labelling

From this scores plot we can see that the observations with higher warp resistance ranking are negatively correlated with pc2, and we can't clearly realize the relationship between pc1 and warp resistance. Hence, the higher ranked observations are located in the lower values of pc2, and the lower rankings in the higher values of pc2, both around the area where PC1 has zero value.

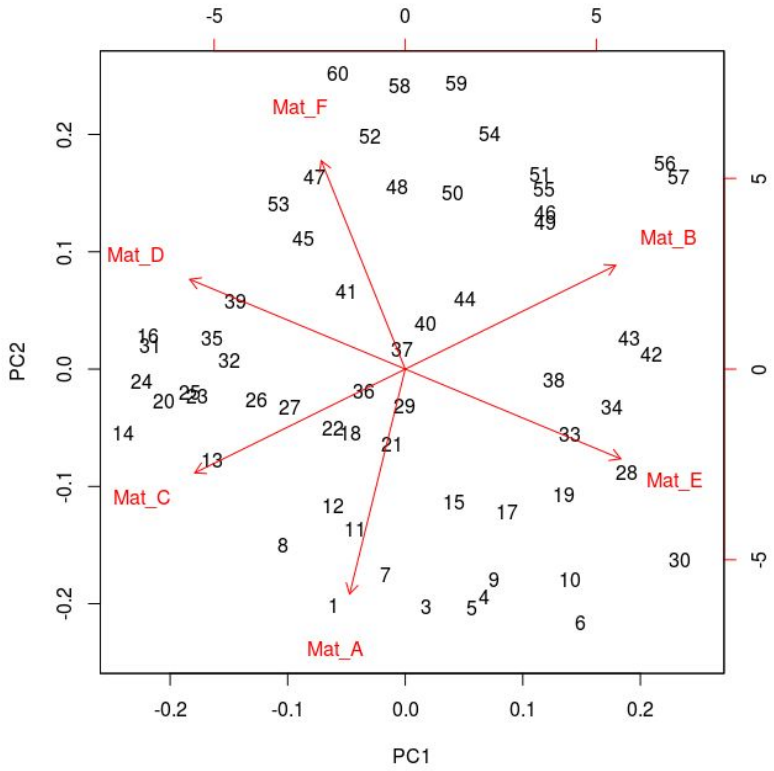


Figure 16: PCA biplot, overlaying scores plot and loadings plot

This biplot in Figure 16, reflecting the loadings and the scores, shows that higher ranked polymers in warping resistance are strongly related with Mat_A and to a lesser extent to Mat_C and Mat_E. Mat_F and Mat_B seem to be related to the lower ranked observations.

Part 3

In this Part 3, we handle a new dataset, with 10 new observations comprising different composition of materials, the same 6 we've been analysing before, and we want to assess is likely performance. So in this case, instead of rankings we have a label, identifying each observation.

	Label	Mat_A	Mat_B	Mat_C	Mat_D	Mat_E	Mat_F
1	A	25.6911	4.0011	17.9997	16.0686	8.5067	NA
2	B	5.9963	18.9799	3.0003	3.3188	38.2561	NA
3	C	20.7652	13.1259	8.8741	3.6690	37.4391	12.453
4	D	12.2748	16.7757	5.2243	3.0110	39.0009	NA
5	E	17.2606	11.4210	10.5790	8.8969	25.2405	NA
6	F	26.1649	5.4029	16.5971	13.3524	14.8443	NA

Table 4: subset of Part 3 dataset

	Label	Mat_A	Mat_B	Mat_C	Mat_D	Mat_E	Mat_F
A	:1	Min. : 4.001	Min. : 4.001	Min. : 3.000	Min. : 3.011	Min. : 4.006	Min. :12.45
B	:1	1st Qu.: 7.566	1st Qu.: 6.389	1st Qu.: 7.699	1st Qu.: 4.976	1st Qu.: 9.717	1st Qu.:13.82
C	:1	Median :16.090	Median :11.151	Median :10.849	Median :12.684	Median :16.405	Median :15.19
D	:1	Mean :15.705	Mean :10.877	Mean :11.121	Mean :10.850	Mean :20.688	Mean :15.19
E	:1	3rd Qu.:24.051	3rd Qu.:14.301	3rd Qu.:15.611	3rd Qu.:15.550	3rd Qu.:34.389	3rd Qu.:16.57
F	:1	Max. :26.165	Max. :18.980	Max. :18.000	Max. :18.000	Max. :39.001	Max. :17.94
(Other)	:4						NA's :8

Table 5: summary description of Part 3 dataset

The dataset also has missing data in the variable Mat_F, and like previously, we've replaced the missing data using the *irmi* function, as most of the data in that variable is missing and because we need that variable to match the PCA analysis we've just performed.

Analysis of polymers likely performance based on materials composition

Strength Analysis

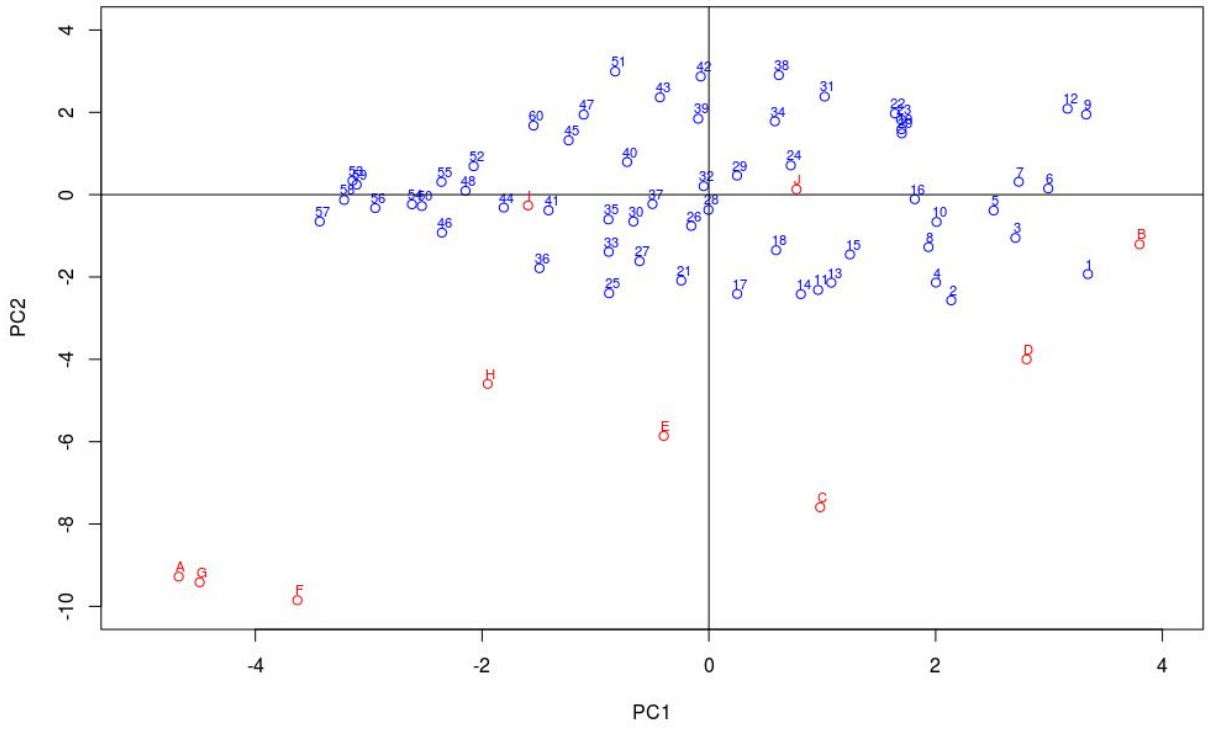


Figure 17: PCA score plot with strength rankings and new data

Based on the new data provided and the PCA model derived previously we can see that the polymer observations labelled as D and B will have good performance regarding strength of the polymer, as they are placed in the lower right quadrant of the plot with strong positive correlation with PC1 and strong negative correlation with PC2.

Warping resistance analysis

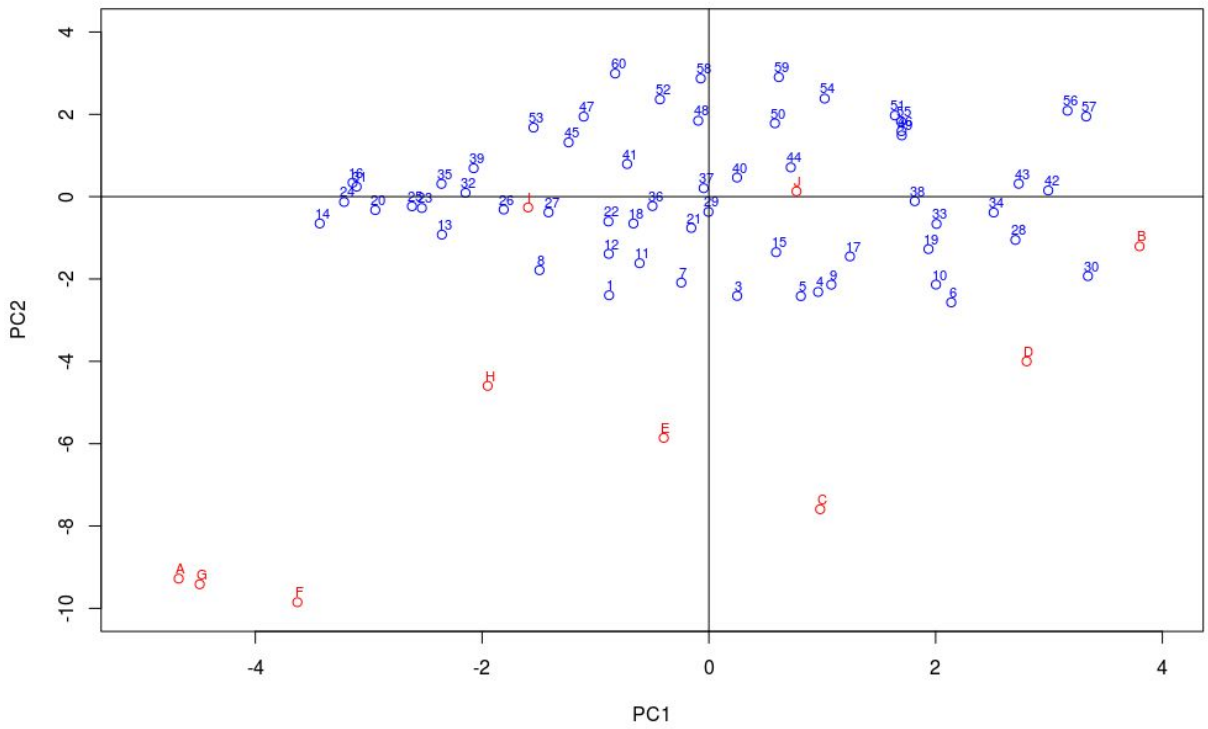


Figure 18: PCA score plot with warping resistance rankings and new data

Based on the new data and the PCA model, we can say that every new polymer observation that has negative values of PC2 will probably be highly resistant to warping, as the correlation with PC1 is not clear, although the higher ranked values don't take extreme values of PC1, so in this case we could probably say that H, E, C and D will be polymers with high warping resistance.

Discussion

In this analysis, we devised a PCA model with just two components, that is able to explain 97.86% of the variance of the dataset variables, e.g., polymer materials.

Based on the principal component scores and loadings, we concluded about the relevant variables/material once trying to understand polymer performance regarding strength, Mat_E and Mat_B, and warping resistance, Mat_A and to a lesser extent Mat_C and Mat_E.

We then used the PCA model to try and classify new polymer materials combinations in terms of these properties.

This data experiment would clearly benefit of a more comprehensive dataset, as we've used a 60 observations long dataset, that was further reduced by one observations to 59.

Although the PCA plots wouldn't clearly give us a neat separation between the observations regarding principal components and loadings, so that it would ease the selection of right variables to explain the strength and warping resistance properties, the biplots were somewhat telling in providing reasonable insight into possible combinations to accomplish that goal.