Name: Thuan Tran
CSS 490
Homework 3

REPORT

NOTE

Double click the winebeer dataset and choose to import the data in the form of numeric array. The first column will be Nan and we will assign the remaining column into another variable. Also, when attaching the picture to the document, the picture might be small. So you can either zoom in the document or take a look at the picture I attached with the document

Mean , Standard Deviation and Scatter Plots

Below is the mean and standard deviation of the winebeer dataset that is in the old space (The original space) and the two variable newStandard and newMean are the new standard deviation and new mean in the transformed space after doing SVD

Discussion about Illustration 1

Looking at the old mean and standard deviation, we can see that there is a huge disperance of how the data is represented. For example, the mean in the first column is only 1.75 but the mean in the last column (last feature) is 153. That is because they are not normalized yet. The same with standard deviation as well.

Also taking a look at the standard deviation of column 2,3 and 5, I expect that there is also some kind of outliners as well since the standard deviation is big for those 3 columns. If there is no outliners, then the standard deviation should be small like column 1 and 4

Taking a look at the covariance matrix in the old space, we can see that all the features are correlated to each other. That means that none of the relationship between 2 features is 0. All of them are either positive correlated or negative correlated.

Discussion about Illustration 2

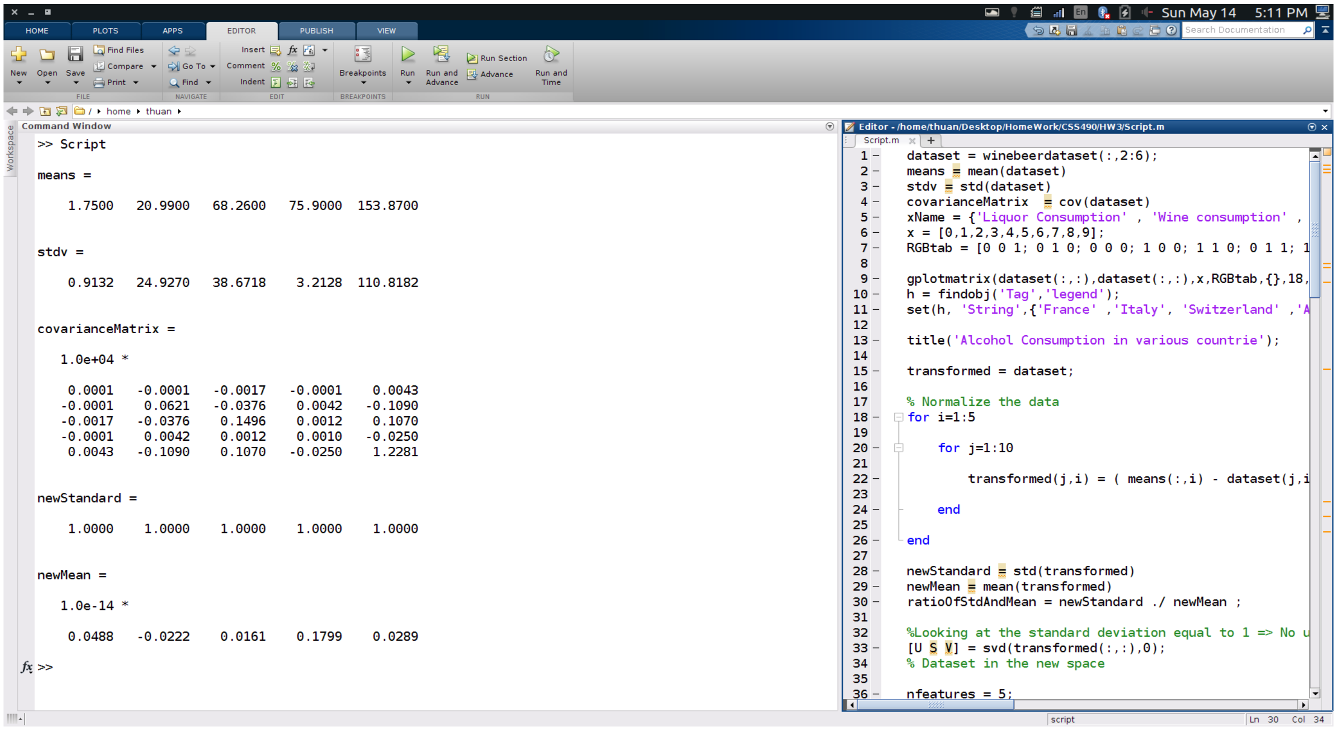
After doing the scatter plot of the old space, I can determine that the features are mixed with each other. There seems to be a slightly correlation between them where if one value of the feature goes up, the other also goes up as well. However, that correlation is not clearly visible in the linear way (maybe in a non-linear way like a quadratic relationship). Or perhaps the features did not follow any pattern at all. Maybe because different countries has different ways to consume alcohol

Discussion about Illustration 3

In addition, after I transformed the old space into the new space, I can see that all of means are now normalized. In addition, the standard deviation of all features are now 1. Meaning that there is a variance in each features. Also that the new data in the new space is close to each other instead of having a big standard deviation that suggest outliners

In addition, the ratio of the Standard Deviation and Mean also suggest that the features are helpful as well. We can see that the ration of the standard deviation and mean is more than 0. Meaning that the standard deviation is large and the mean is small, suggesting variance in the data

In conclusion, none of the features are useless. They all contribute to the new space because of their ratio



```
>> Script

means =

    1.7500   20.9900   68.2600   75.9000  153.8700


stdv =

    0.9132   24.9270   38.6718    3.2128  110.8182


covarianceMatrix =

   1.0e+04 *

    0.0001   -0.0001   -0.0017   -0.0001    0.0043
   -0.0001    0.0621   -0.0376    0.0042   -0.1090
   -0.0017   -0.0376    0.1496    0.0012    0.1070
   -0.0001    0.0042    0.0012    0.0010   -0.0250
    0.0043   -0.1090    0.1070   -0.0250    1.2281


newStandard =

    1.0000    1.0000    1.0000    1.0000    1.0000


newMean =

   1.0e-14 *

    0.0488   -0.0222    0.0161    0.1799    0.0289

fx >>
```

```
dataset = winebeerdataset(:,2:6);
means = mean(dataset)
stdv = std(dataset)
covarianceMatrix  = cov(dataset)
xName = {'Liquor Consumption' , 'Wine consumption' ,
x = [0,1,2,3,4,5,6,7,8,9];
RGBtab = [0 0 1; 0 1 0; 0 0 0; 1 0 0; 1 1 0; 0 1 1; 1

gplotmatrix(dataset(:,:),dataset(:,:),x,RGBtab,{},18,
h = findobj('Tag','legend');
set(h, 'String',{'France' ,'Italy', 'Switzerland' ,'A

title('Alcohol Consumption in various countrie');

transformed = dataset;

% Normalize the data
for i=1:5

    for j=1:10

        transformed(j,i) = ( means(:,i) - dataset(j,i

    end

end

newStandard = std(transformed)
newMean = mean(transformed)
ratioOfStdAndMean = newStandard ./ newMean ;

%Looking at the standard deviation equal to 1 => No u
[U S V] = svd(transformed(:,:),0);
% Dataset in the new space

nfeatures = 5;
```

Illustration 1: Mean and Standard Deviation

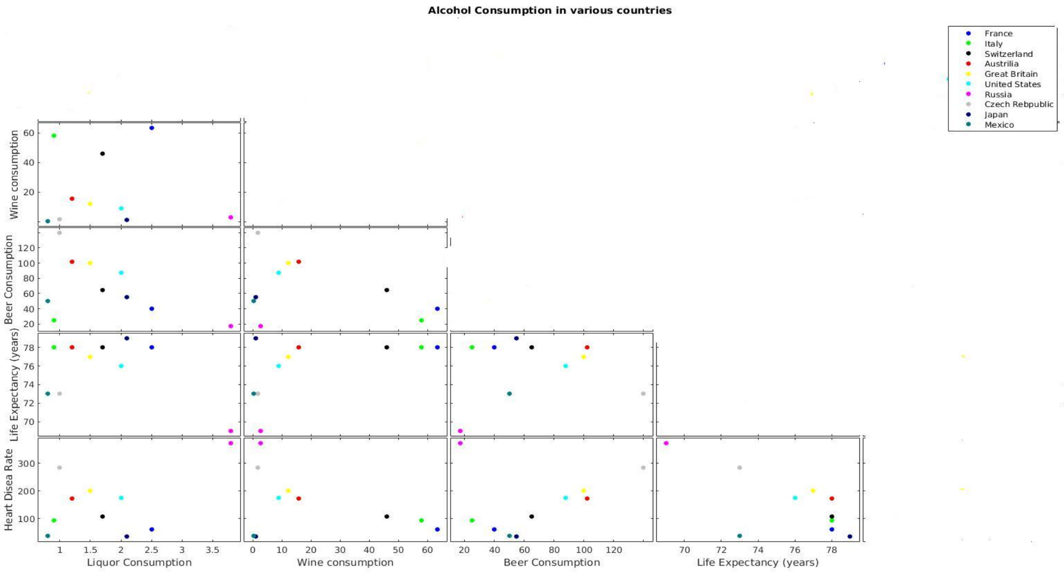
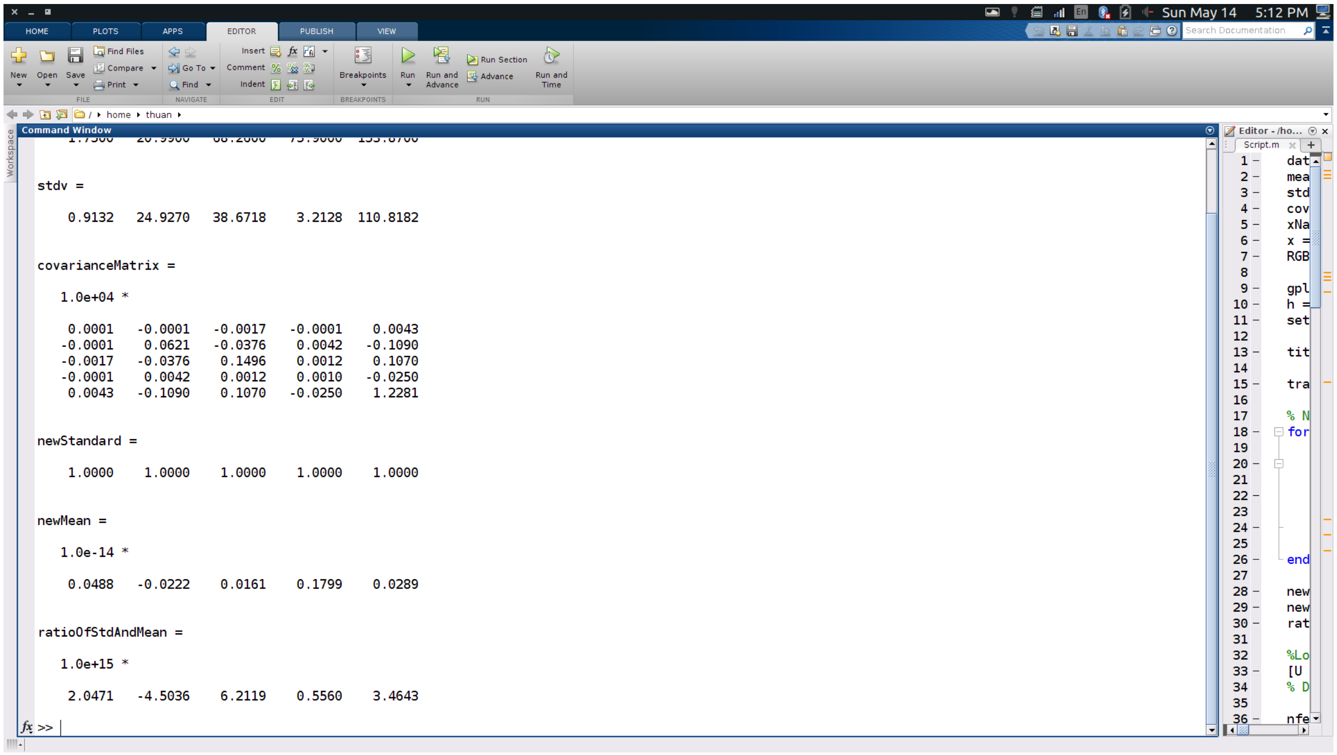


Illustration 2: Scatter Plots of the old space



```
stdv =

    0.9132   24.9270   38.6718    3.2128  110.8182


covarianceMatrix =

   1.0e+04 *

    0.0001   -0.0001   -0.0017   -0.0001    0.0043
   -0.0001    0.0621   -0.0376    0.0042   -0.1090
   -0.0017   -0.0376    0.1496    0.0012    0.1070
   -0.0001    0.0042    0.0012    0.0010   -0.0250
    0.0043   -0.1090    0.1070   -0.0250    1.2281


newStandard =

    1.0000    1.0000    1.0000    1.0000    1.0000


newMean =

   1.0e-14 *

    0.0488   -0.0222    0.0161    0.1799    0.0289


ratioOfStdAndMean =

   1.0e+15 *

    2.0471   -4.5036    6.2119    0.5560    3.4643

fx >>
```

```
dat
mea
std
cov
xNa
x =
RGB

gpl
h =
set

tit

tra

% N
for


end

new
new
rat

%Lo
[U
% D

nfe
```

Illustration 3: New Mean and Standard Deviation

After doing SVD

Scree Plots

After doing the SVD, I obtained the following Scree Plots

Discussion about Illustration 4 : Normal Scree Plot

Looking at the normal scree plot, we can see that we only need the first 4 features in the new transformed space to capture the information. Since at feature 5, it only captures less than 0.05 about the information. The most important is the first new feature where it captures about 0.45 information of the new space and the second where it captures 0.32 of the new space.

Discussion about Illustration 5: Cumulative Scree Plot

Just like the normal scree plot, we can immediately see that we do not need the 5th feature , at the 4th feature , we already capture 0.98 information. So adding the 5th feature is not really necessary

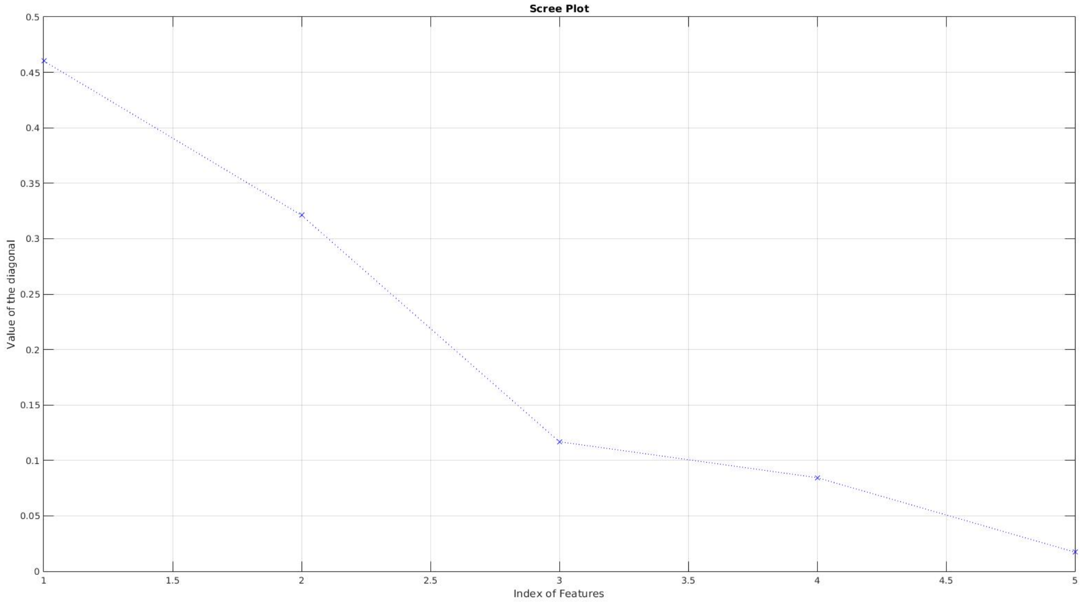


Illustration 4: Normal Scree Plot

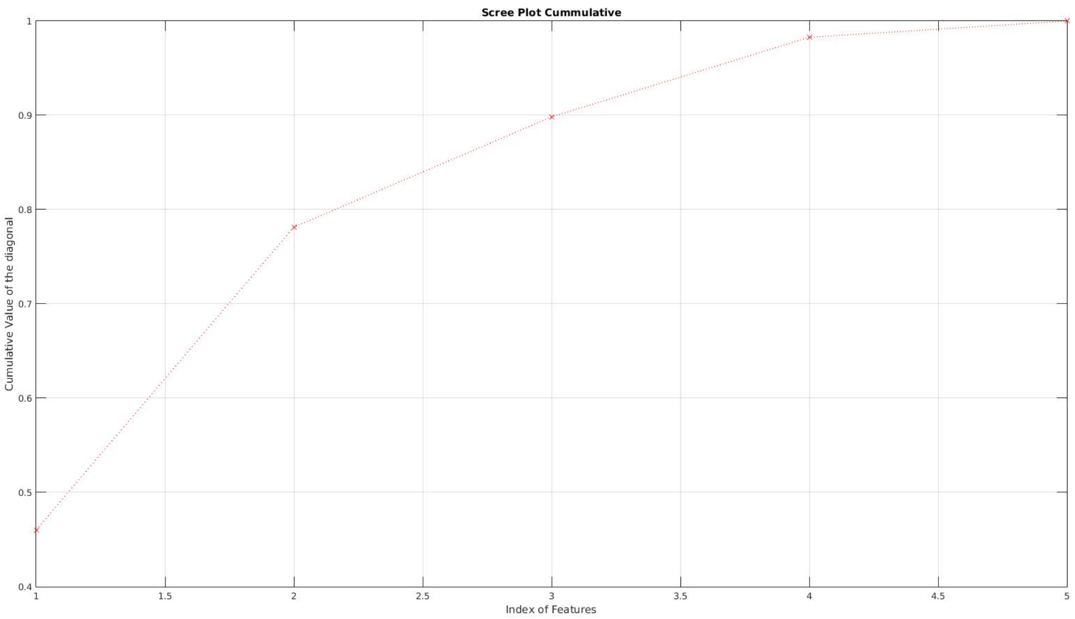


Illustration 5: Cumulative Scree Plot

Loading Vectors

Below are the illustration of the loading vector of the new space. In my opinion, only loading vectors 1 to 4 are important because of their value in the scree plot

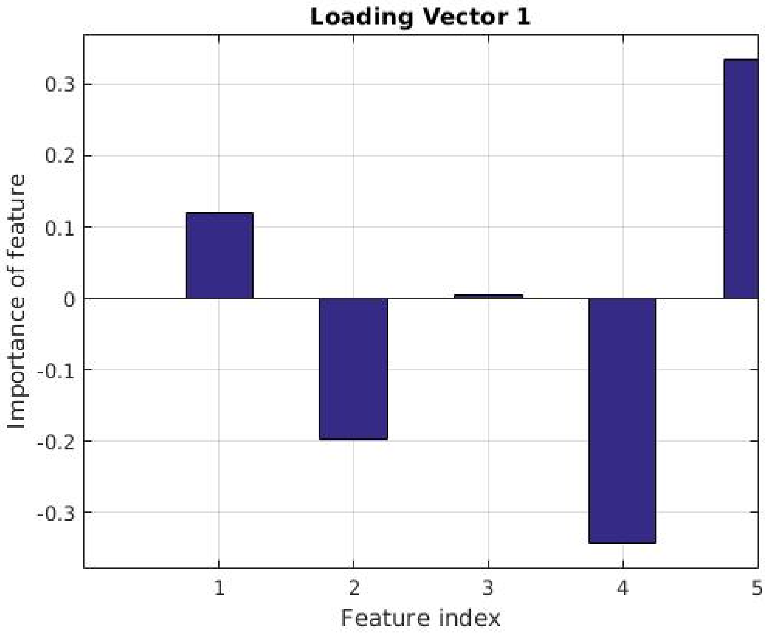


Illustration 6: Loading Vector 1

Loading Vector 1 suggests that the feature at index 3 in the old space do not contribute much this Loading Vector. The vectors that contribute much are vector at index 4 and 5 (negatively and positively). While the features at index 1 and 2 have a minor contribution

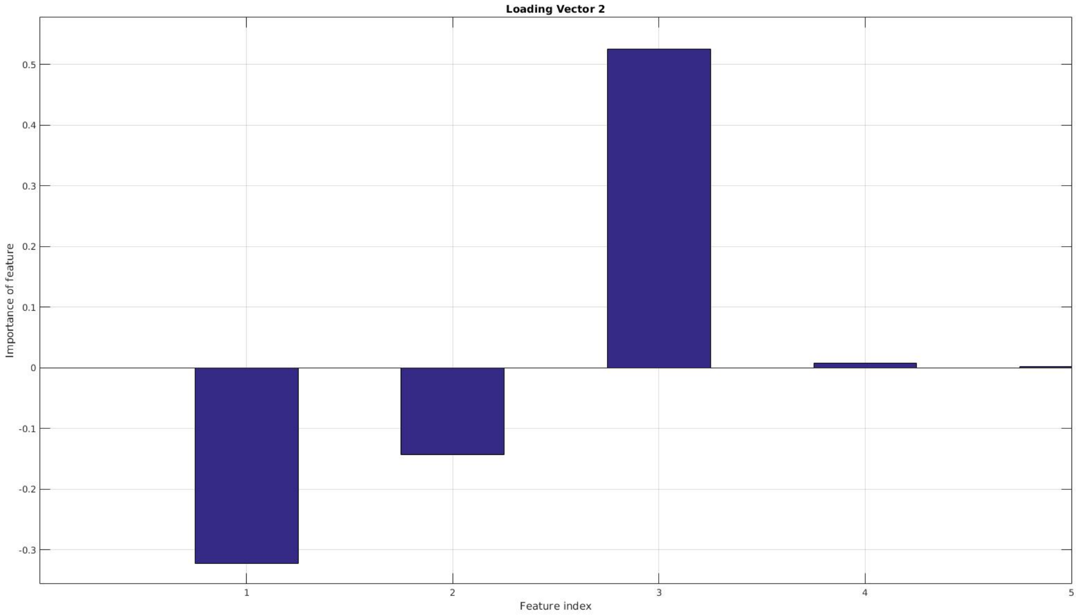


Illustration 7: Loading Vector 2

Loading vector 2 suggests that the features at index 4 and 5 do not contribute to the loading vector 2. And the one that contribute the most is the feature at index 3 (positively). The features at index 1 and 2 also contribute but now so much like the one at 3

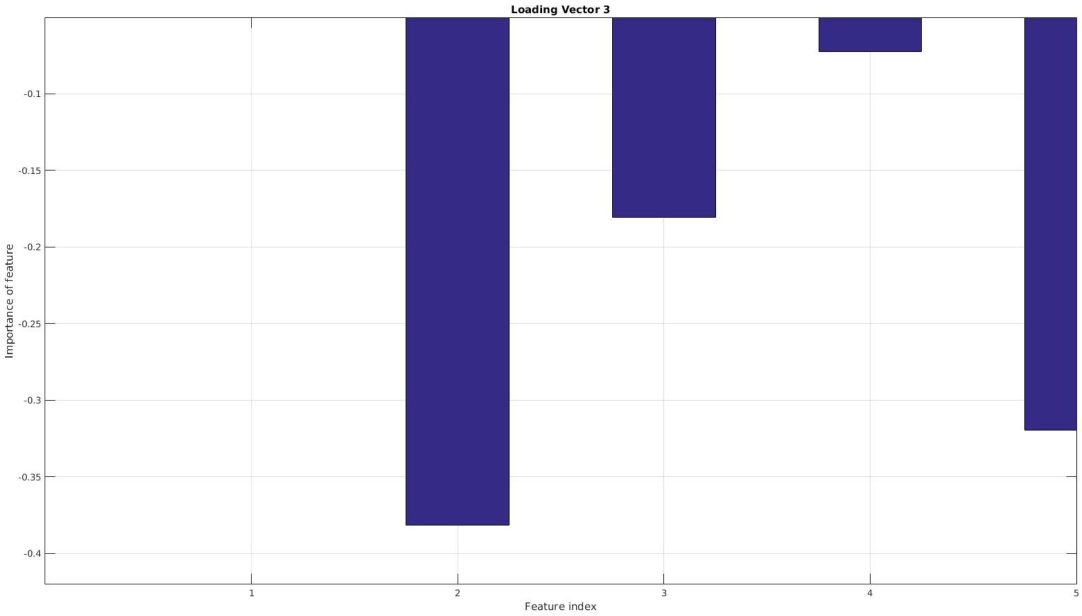


Illustration 8: Loading Vector 3

For the loading vector 3, the feature at index 1 is useless since it did not contribute anything at all. All the remaining features contribute to the loading vector 3 (negatively) with the smallest is feature 4 and the largest is feature 2

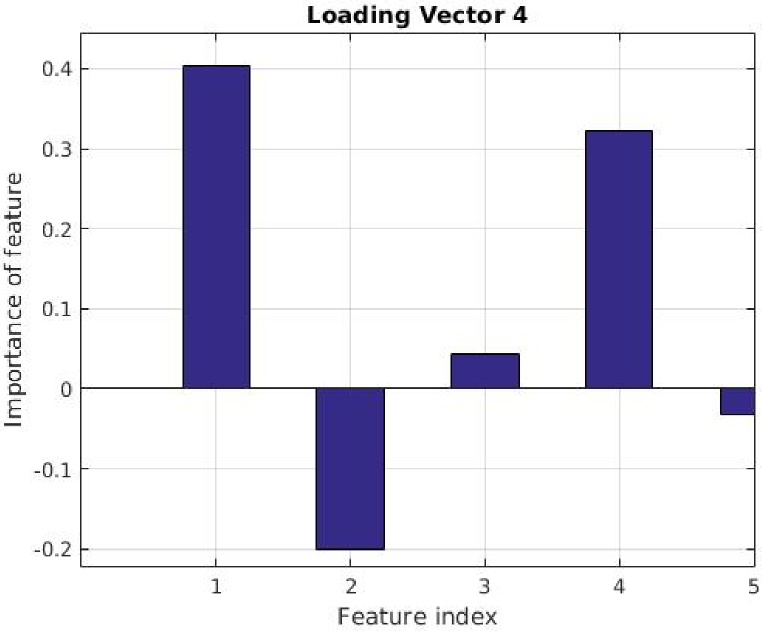


Illustration 9: Loading Vector 4

The loading vector 4 have all of the features in the old space contribute to it. The largest contributing factor is feature at index 1 and the smallest is feature at index 5

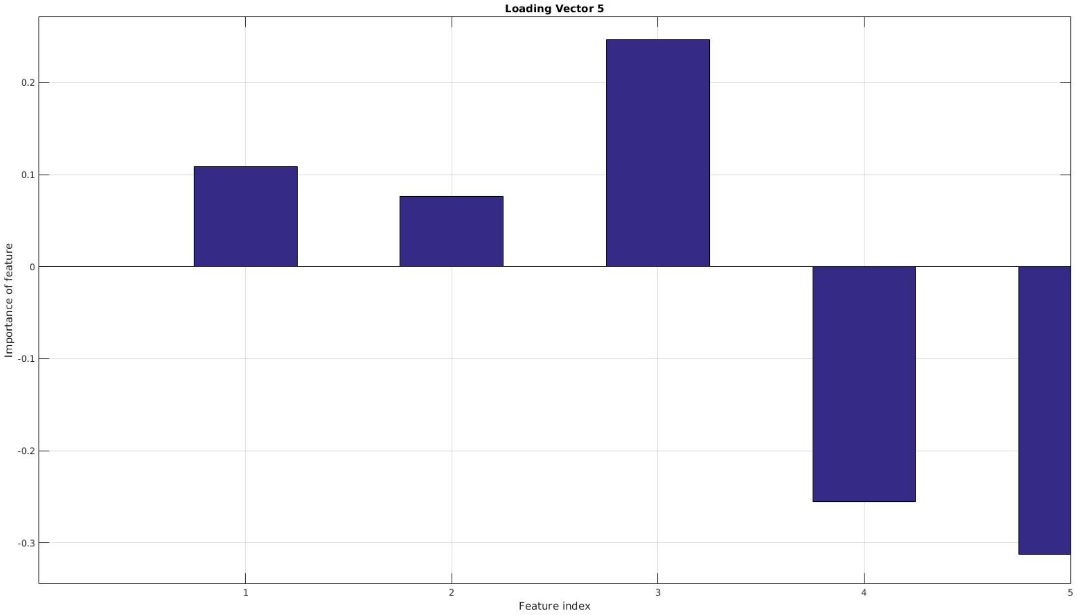


Illustration 2: Loading Vector 5

The loading vector 5 also have all of the features in the old space contribute to each. The biggest contribution is at index 5 and the smallest is at index 2

Looking at the heights of each index in each loading vectors, I can see that they all contribute differently to the new axis of each loading vectors. Depending on the loading vectors, some of the indexes contribute more than the other.

I also did not see any pattern of the height of the indexes in these loading vectors. This indicate that there is no correlation between any indexes so that we can eliminate one and only use the other. Thus, all the indexes are necessary

2D Scatter Plots

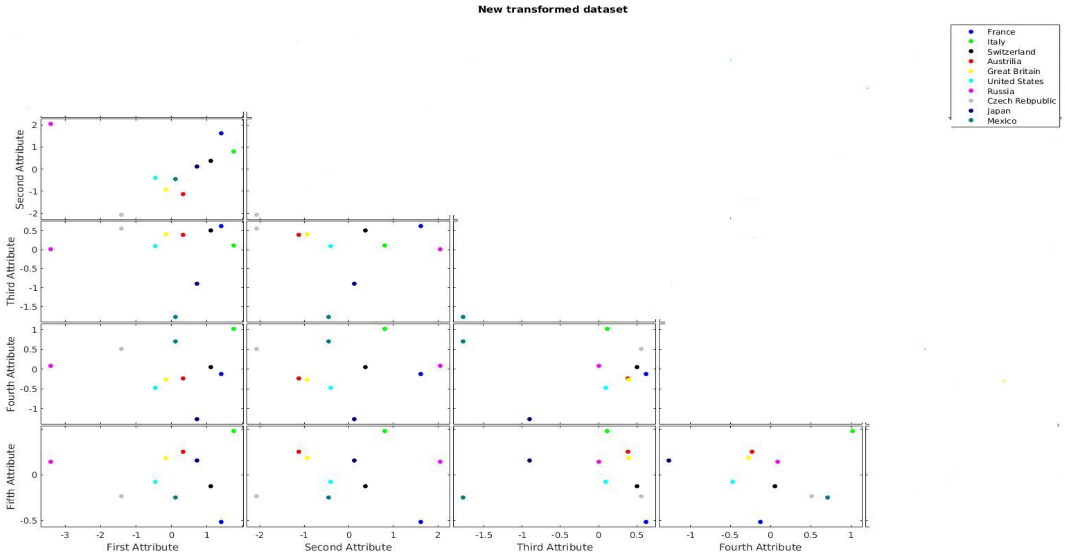


Illustration 3: 2D Scatter Plot of the transformed

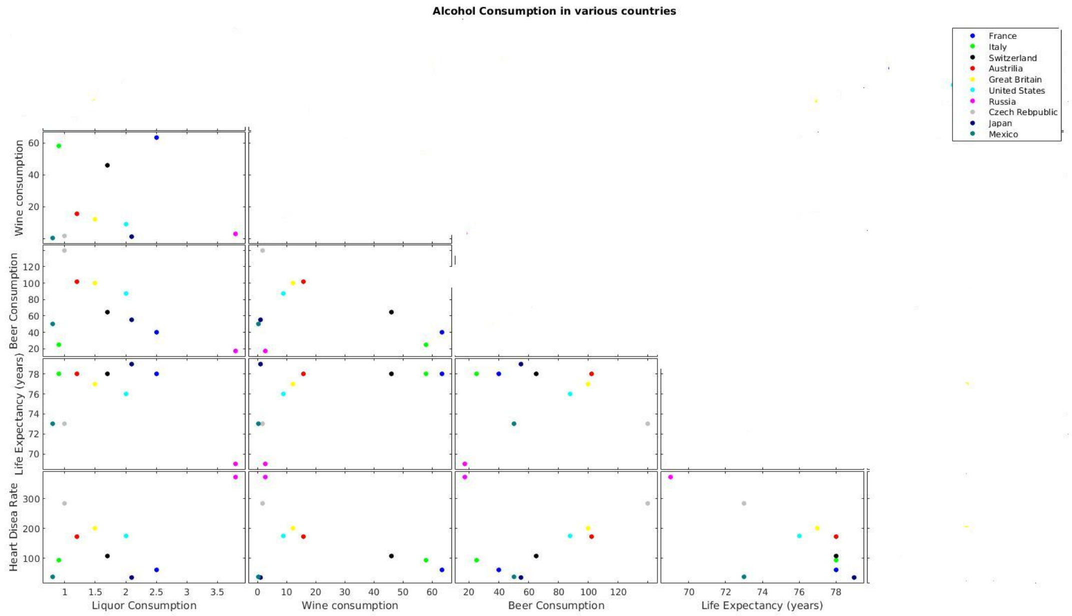


Illustration 4: 2D Scatter Plot of the original dataset

Looking at the new 2D Scatter Plot of the new space, I did not see much difference from the old one. There is not distinction that is able to identify which country belong to which region. However, I noticed that the new scatter plot somehow group the countries more in the middle of the graph than the old scatter plot.

In conclusion, looking at 2D Scatter Plot only is not enough to determine if the new transformed data set was able to separate the data. We might need to switch to 3D plots and rotate the plots to see if they are separated.