**PCA/SVD**

For PCA and SVD we chose the same features as for K-Means: «Temperature», «Luminosity», «Radius», «Absolute Magnitude», because they represent different linearly independent characteristics, all of which are important.

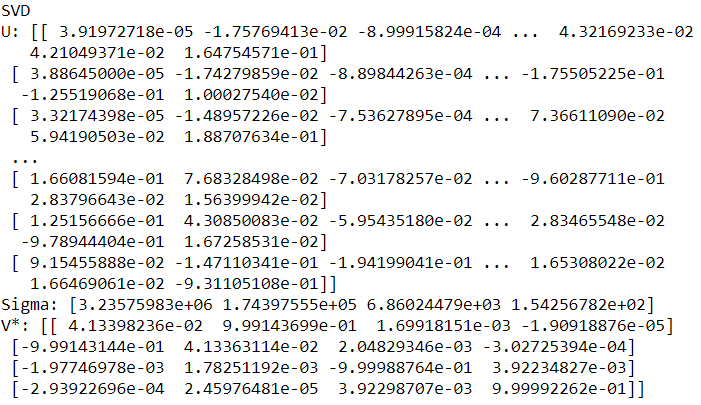
**Standardizing data**

We made three standardization of the data: z-scoring, range and rank. Z-scoring is the same as StandaradScaler from sklearn.preprocessing module. It transforms feature into equivalent with 0 as the mean and 1 as standard deviation. Range standardization is done by substracting the mean (as in z-scoring standardization) and dividing by range (difference between maximum and minimum of the feature). Rank standardization is done by substracting mean and dividing by range. Rank standardization image a feature to a segment [0,1] and it is an equivalent of MinMaxScaler from sklearn.preprocessing module.

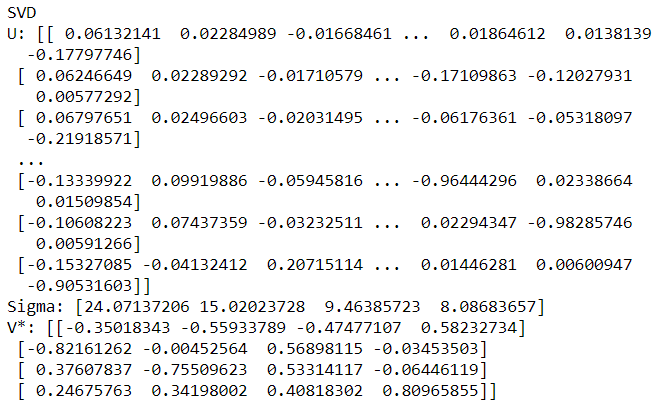
**SVD**

We made SVD for original data and for standardized datasets. We cannot show whole matrices of SVD because their shape is not appropriate for this. But in the following screenshots you can see these matrices partially.

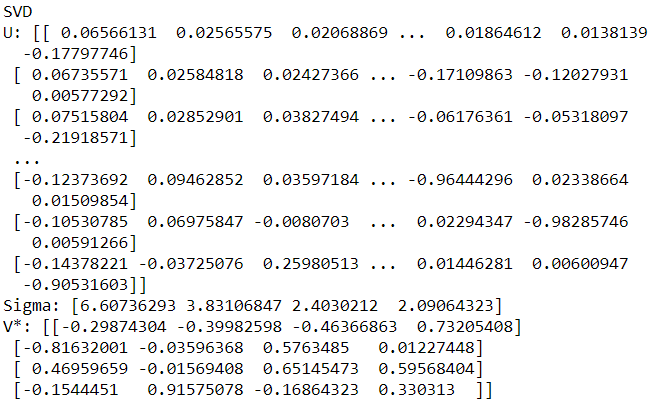
SVD on original data:



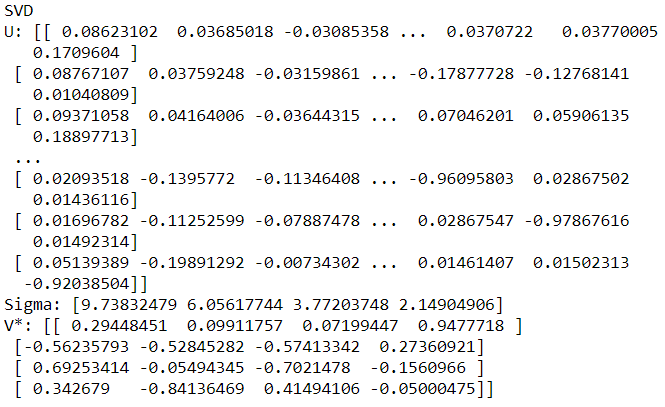
SVD on Z-scored standardized data:



SVD on range standardized data:



SVD on rank standardized data:



The goal of SVD is to find contributions of each component. In the following table you can see natural contribution of each feature in original dataset and in standardized datasets:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Temperature | Luminosity | Radius | Absolute Magnitude |
| Original data | 1.0407e+13 | 3.0415e+10 | 4.7063e+07 | 2.3752e+04 |
| Ranked std data | 94.835 | 36.6773 | 14.2283 | 4.6184 |
| Ranged std data | 43.6573 | 14.677 | 5.7745 | 4.3708 |
| Z-Scored std data | 579.431 | 225.6075 | 89.5646 | 65.3969 |

It shows absolute value of contribution.

In the following table you can see percent contribution of each feature in datasets:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Temperature | Luminosity | Radius | Absolute Magnitude |
| Original data | 99.71 | 0.2897 | 0.000448 | 0.000000227 |
| Ranked std data | 63.07 | 24.39 | 9.46 | 3.072 |
| Ranged std data | 63.75 | 21.43 | 8.43 | 6.38 |
| Z-Scored std data | 60.36 | 23.5 | 9.33 | 6.81 |

The first component («Temperature») contributes the most in each dataset.

In the following table you can see data scatter of each dataset:

|  |  |
| --- | --- |
|  | Data Scatter |
| Original data | 10500603295793.56 |
| Ranked std data | 150.359 |
| Ranged std data | 68.48 |
| Z-Scored std data | 960 |

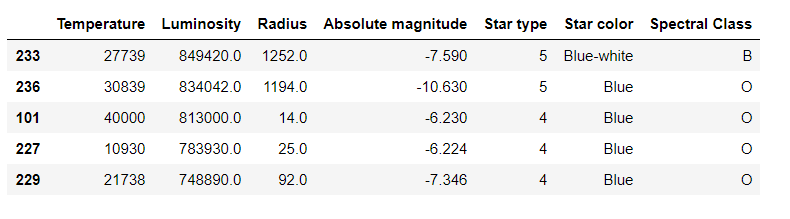
Range standardized data has the least data scatter. Original data has the biggest data scatter because it is not standardized at all, each feature has values of different orders of magnitude.

**Hidden ranking factor**

Hidden factor is a value which characterize an object according to original features. Hidden factor is a value, which is highly correlated with original features.

Hidden ranking factor is rank standardized hidden factor. It has value from 0 to 100.

In the following tables you can see stars with top 5 hidden ranking factors for original data.

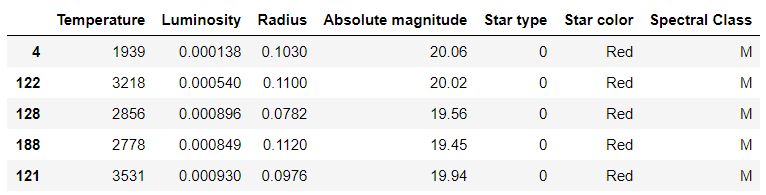


Their hidden ranking factors:

|  |  |
| --- | --- |
| 233 | 1 |
| 236 | 0.982 |
| 101 | 0.958 |
| 227 | 0.922 |
| 229 | 0.882 |

Contribution of this hidden factor is 99.7%

In the following tables you can see stars with top 5 hidden ranking factors for Z-score standardized data.

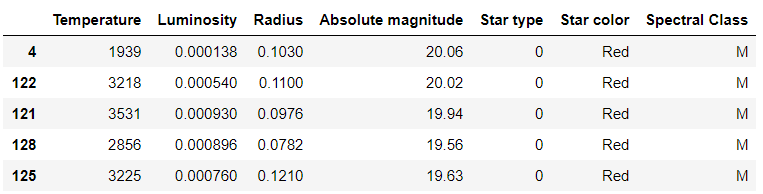


Their hidden ranking factors:

|  |  |
| --- | --- |
| 4 | 1 |
| 122 | 0.992 |
| 128 | 0.9908 |
| 188 | 0.9903 |
| 121 | 0.9902 |

Contribution of this hidden factor is 60.36%

In the following tables you can see stars with top 5 hidden ranking factors for range standardized data.

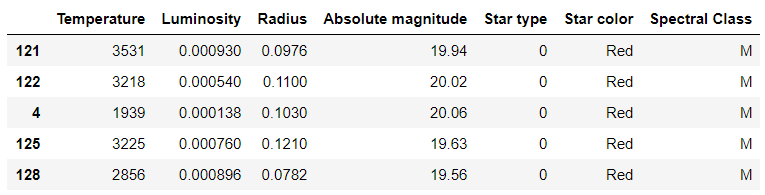


Their hidden ranking factors:

|  |  |
| --- | --- |
| 4 | 1 |
| 122 | 0.994 |
| 121 | 0.991 |
| 128 | 0.989 |
| 125 | 0.988 |

Contribution of this hidden factor is 63.75%

In the following tables you can see stars with top 5 hidden ranking factors for rank standardized data.



Their hidden ranking factors:

|  |  |
| --- | --- |
| 121 | 1 |
| 122 | 0.9999 |
| 4 | 0.9898 |
| 125 | 0.9866 |
| 128 | 0.9809 |

Contribution of this hidden factor is 63.07%

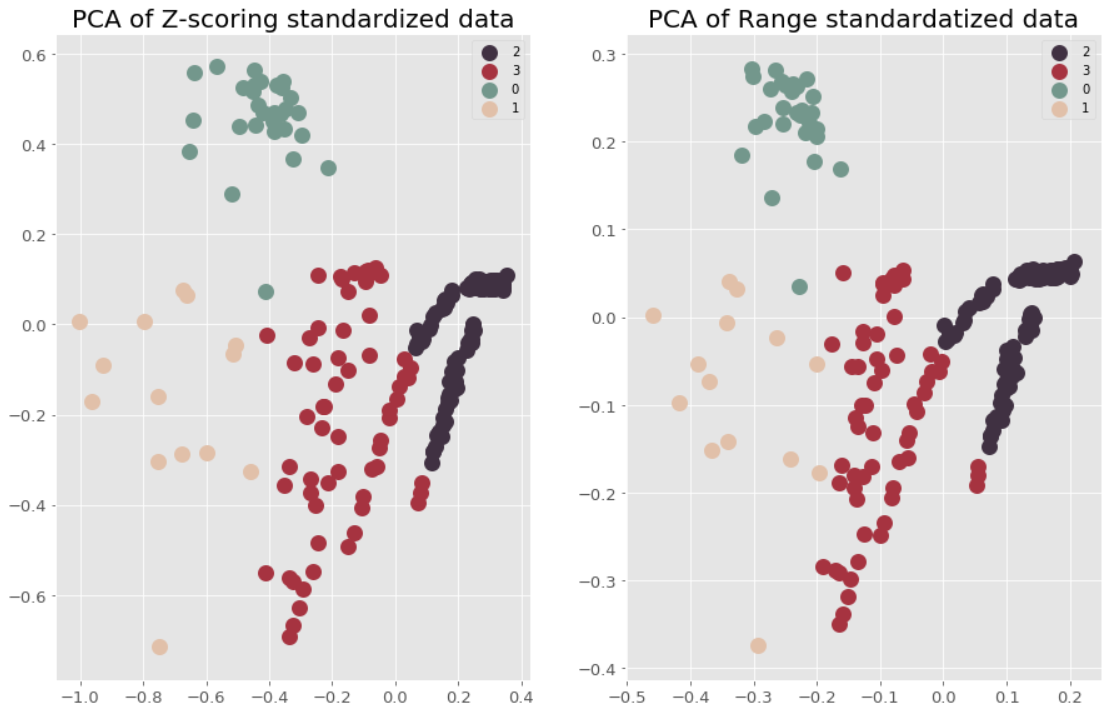
Ranks of the same object from different dataset are not equal. Hidden factors of standardized datasets are quite similar. Objects of M class, red color, and 0 star type has the highest hidden factors. In the original dataset stars of other types have the highest hidden factors.

**Data visualization**

Using two first components (principal components) we made data visualization in coordinates of these two components.

Objects of different colors belong to different clusters, labels of clusters are mentioned in graph legend.

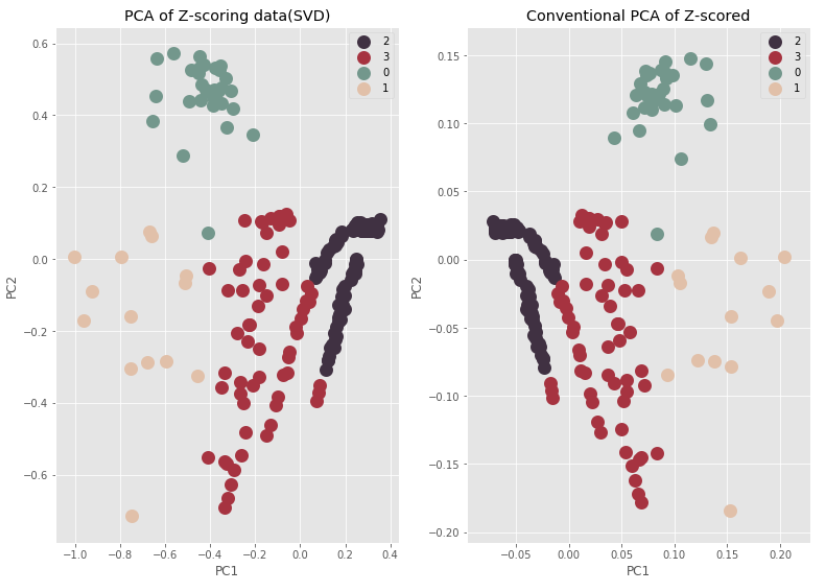
In the following graphs you can see visualizations of PCAs of Z-scored standardized data and range standardized data:



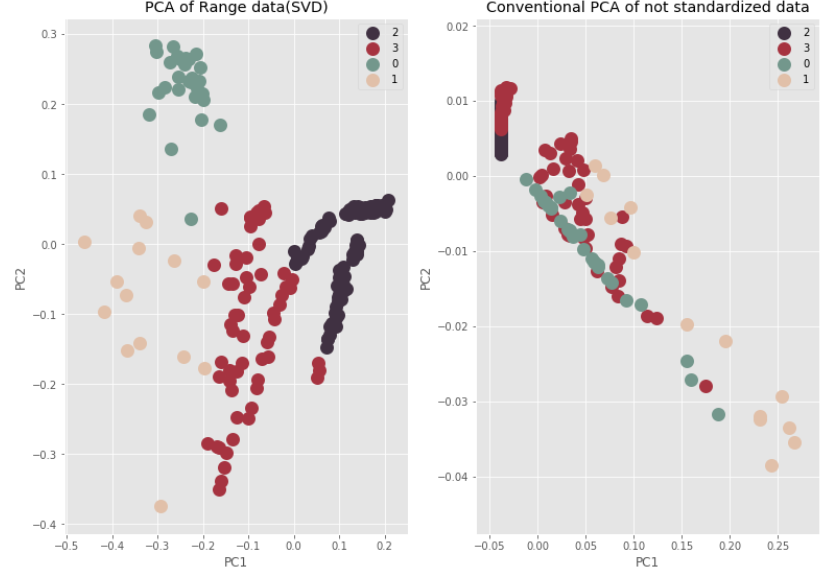
Visualizations are quite similar, they just differ a bit in scale. Clusters, which were obtained by K-Means are well separated in both graphs. Also PCAs have different axes scales.

After that we applied conventional PCA for original data and for Z-score standardized data. Also we applied PCA from sklearn.decomposition module.

Now lets compare the results.



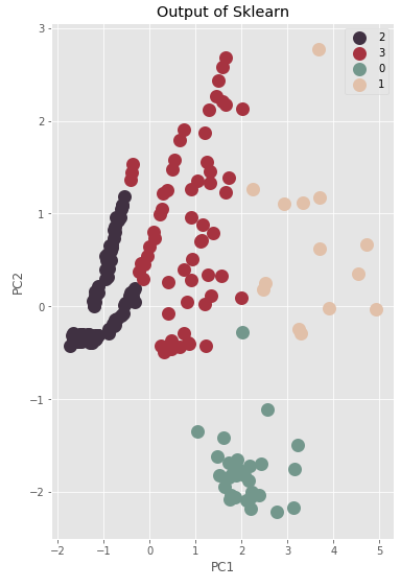
Graphs are mirrored and have different axes scales. Clusters are well separated in both graphs.

In the following graphs you can see visualizations of PCAs of range standardized data and original dataset.

Visualization of PCA of range standardized data is quite similar with previous visualizations. Clusters are well separated.

Visualization of PCA of original data differs. Clusters are not separated, this visualization is not interpretable. This fact means that standardization is obligatory for clustering and other methods of data analysis.

In the following graph you can see sklearn PCA visualization:



The graph is mirrored visualization of PCA of range standardized data.

It is hard to say exactly which method of standardization is better, because they have very similar results. Z-score standardization can be easily interpreted. If the value equals 0, it means that the value is the mean. If the value is less or bigger than 0, it means that value differs from the mean, and absolute value shows how much and in what side it differs from the mean. In our point of view Z-scoring is better in the context of interpreting values in a dataset. But in fact it doesn’t influence on results greatly.