**K-Means clustering**

For K-Means clustering we chose all 4 quantitative features:

* «Temperature»
* «Luminosity»
* «Radius»
* «Absolute magnitude»

We chose all these features because they represent different characteristics of stars and they are not linearly dependent. All these four features are equally significant for category definition.

We standardized all columns and then applied K-Means algorithm to split the data into 4 clusters and into 7 clusters. K-Means was applied 10 times with following parameters: maximum number of iterations 500, random initial cluster centers, relative tolerance 0.0001, EM-style algorithm and different random state, which equals the number of iteration. From 10 results of clustering we chose the best according to the parameter «inertia»: the less inertia, the better clustering.

In the following table you can see results of K-Means initializations, where «i» is a number of initialization and a value of random state.

*Table 1.*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | i=0 | i=1 | i=2 | i=3 | i=4 | i=5 | i=6 | i=7 | i=8 | i=9 |
| 4 clusters | 281.5 | 257.21 | 283.93 | 277.05 | 264.86 | 282.05 | 281.51 | 281.51 | 257.21 | 256.9 |
| 7 clusters | 140.63 | 147.51 | 136.51 | 137.01 | 137.01 | 137.01 | 198.68 | 137.06 | 192.15 | 137.01 |

The best result for 4 clusters was received in the initialization with i=9. The best result for 7 clusters was received in the initialization with i=2.

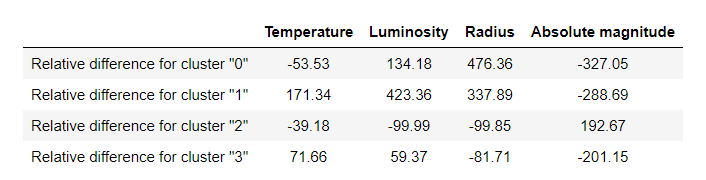
Now let us do a more thorough interpretation of all four clusters.

**Relative difference**

Now we are going to count relative difference for all clusters and all quantitative features. Relative difference shows how much cluster mean of a feature differs from grand mean of the same feature.

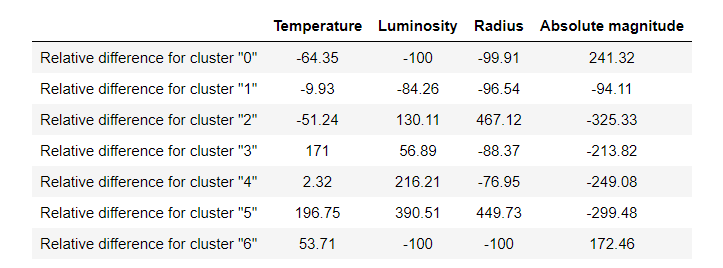
In the following table you can see relative differences for every feature and for every cluster of 4-cluster K-Means.

*Table 2.*



In the following table you can see relative differences for every feature and for every cluster of 7-cluster K-Means.

*Table 3.*



The absolute value of relative difference in these tables shows the percent of deviation from the grand mean of the feature. The more value the more deviation. The sign shows the direction of deviation. Minus means that the mean of the cluster is less than grand mean and plus means that the mean of the cluster is greater than grand mean.

**Quetelet indeces**

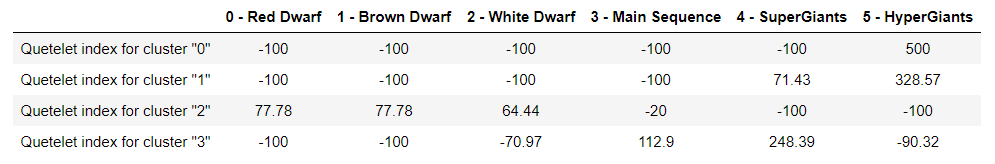
Quetelet index shows what part of category belongs to a cluster. It doesn’t show an exact part of category but show it comparably: the less absolute value the bigger part of cluster belongs to a category.

Quetelet index is calculated by the formula:

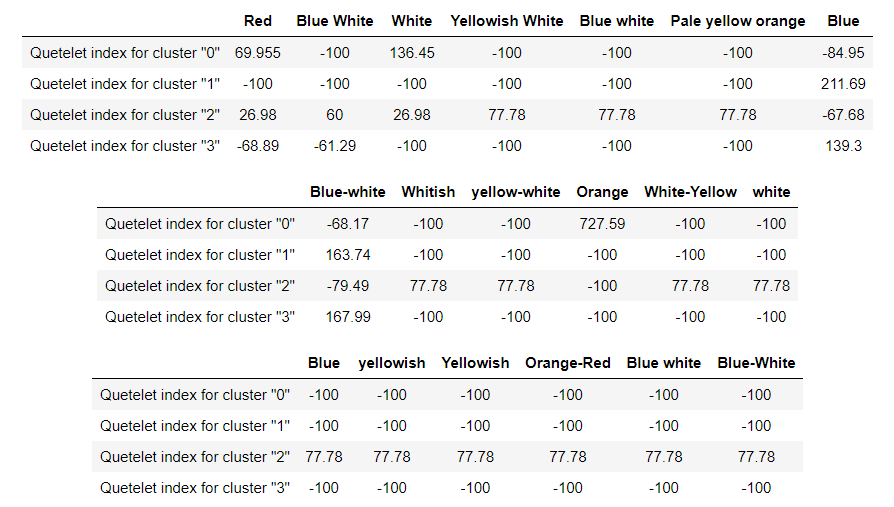
where – the proportion of objects belonging to a cluster and a category simultaneously, – the proportion of objects belonging to a cluster, – the proportion of objects belonging to a category.

In the following tables you can see quetelet indeces for each category of features «Star type», «Star color» and «Star Class» for 4-cluster splitting.

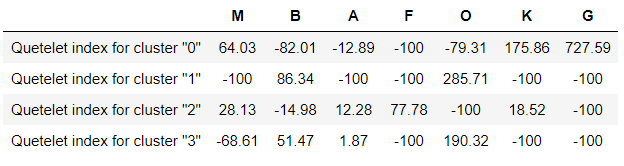
*Table 4. Quetelet indeces for star types*



*Table 5. Quetelet indeces for star colors*

**

*Table 6. Quetelet indeces for spectral classes*



-100 means that no object belongs to a cluster and a category simultaneously.

The more number (the less absolute number) the more objects from a cluster belong to a category simultaneously and the more objects from a category belong to a cluster.

0 is possible only when all objects belong to one category and to one cluster.

These tables show the differences in clusters.

**Making sets and .**

is a set, which is formed by features and categories, values of which is greater than 35%.

is a set, which is formed by features and categories, values of which is less than -35%.

**Sets for 4-cluster split.**

**Cluster «0»**

for cluster 0 consist of following numeric features and categories: «Luminosity», «Radius», «HyperGiants», «Red», «White», «Orange, «M», «K», «G».

for cluster 0 consist of following numeric features and categories: «Temperature», «Absolute magnitude», *«Red Dwarf», «Brown Dwarf», «White Dwarf», «SuperGiants»*, *«Blue White», «Yellowish White», «Blue white», «Pale yellow orange»*, «Blue», «Blue-white», *«Whitish», «yellow-white»*, «Orange», *«White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White»*, «B», «F», «O».

Cluster «0» is less than average in these features. But we want to highlight that most color categories are equal -100 (features in italics), which means that there is no objects of such cluster and this is true for every cluster.

In general cluster «0» is very much greater in sizes and has red, orange and blue colors.

**Cluster «1»**

for cluster 1 consist of following numeric features and categories: «Temperature», «Luminosity», «Radius», «SuperGiants», «HyperGiants», «Blue», «Blue-white», «B», «O».

for cluster 1 consist of following numeric features and categories: «Absolute magnitude», *«Red Dwarf», «Brown Dwarf», «White Dwarf», «Main Sequence»,* *«Red», «Blue White», «White», «Yellowish White», «Blue white», «Pale yellow orange», «Whitish», «yellow-white», «Orange», «White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White», «M», «A», «F», «K», «G».*

In general cluster «1» is greater in sizes and has more blue and blue-white colors.

**Cluster «2»**

for cluster 2 consist of following numeric features and categories: «Absolute magnitude», «Red Dwarf», «Brown Dwarf», «White Dwarf», «Blue White», «Yellowish White», «Blue white», «Pale yellow orange», «Whitish», «yellow-white», «White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White», «F».

for cluster 2 consist of following numeric features and categories: «Luminosity», «Radius», *«SuperGiants», «HyperGiants»,* «Blue», «Blue-white», «Orange», *«O», «G».*

Cluster «2» is very much less in sizes and has more white, yellow and blue colors.

**Cluster «3»**

for cluster 3 consist of following numeric features and categories: «Temperature», «Luminosity», «Main Sequence», «SuperGiants», «Blue», «Blue-white», «B», «O».

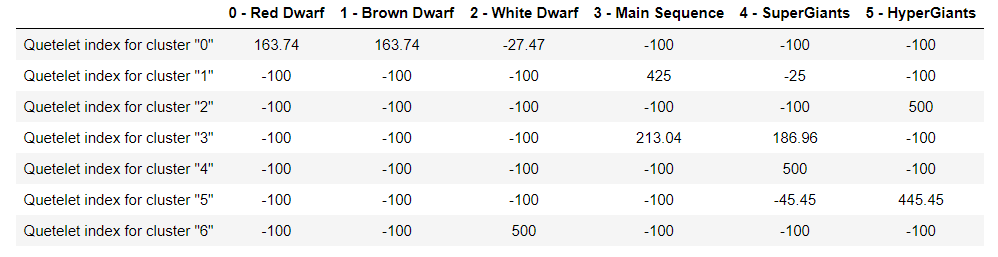
for cluster 3 consist of following numeric features and categories: «Radius», «Absolute magnitude», *«Red Dwarf», «Brown Dwarf»,* «White Dwarf», «HyperGiants», «Red», «Blue White», *«White», «Yellowish White», «Blue white», «Pale yellow orange», «Whitish», «yellow-white», «Orange», «White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White»,* «M», *«F», «K», «G».*

Cluster «3» is greater in temperature and luminosity, has more blue colors, less in size and magnitude.

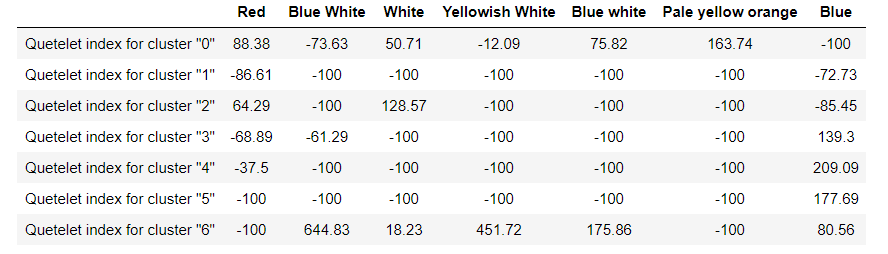
**Sets for 7-cluster split.**

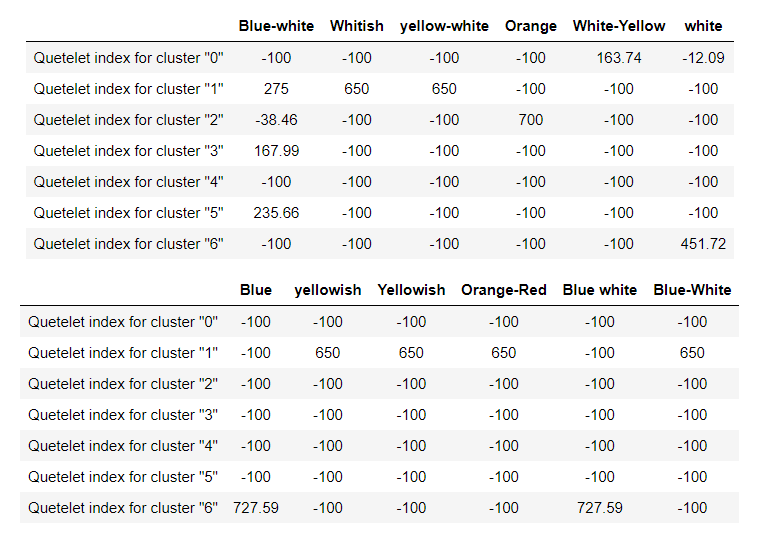
In the following tables you can see quetelet indeces for each category of features «Star type», «Star color» and «Star Class» for 7-cluster splitting.

*Table 7. Quetelet indeces for star types*

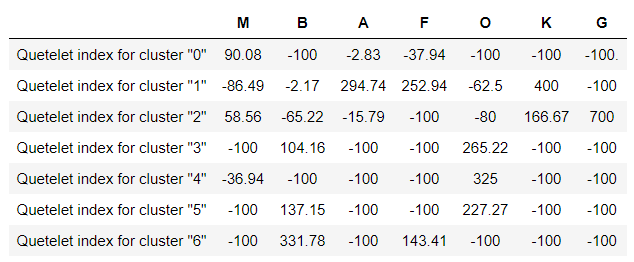
**

*Table 8. Quetelet indeces for star colors*





*Table 9. Quetelet indeces for spectral classes*

**

**Cluster «0»**

for cluster 0 consist of following numeric features and categories: «Absolute magnitude», «Red Dwarf», «Brown Dwarf», «Red», «White», «Blue white», «Pale yellow orange», «White-Yellow», «M»,

for cluster 0 consist of following numeric features and categories: «Temperature», «Luminosity», «Radius», *«Main Sequence», «SuperGiants», «HyperGiants»*, «Blue White», *«Blue», «Blue-white», «Whitish», «yellow-white», «Orange», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White», «B», «O», «K», «G».*

Cluster «0» is less in sizes, temperature and luminosity, has more red, orange and white colors.

**Cluster «1»**

for cluster 1 consist of following numeric features and categories: «Main Sequence», «Blue-white», «Whitish», «yellow-white», «yellowish», «Yellowish», «Orange-Red», «Blue-White», «A», «F», «K».

for cluster 1 consist of following numeric features and categories: «Luminosity», «Radius», «Absolute magnitude», *«Red Dwarf», «Brown Dwarf», «White Dwarf», «HyperGiants»,* «Red», *«Blue White», «White», «Yellowish White», «Blue white», «Pale yellow orange»*, «Blue», *«Orange», «White-Yellow», «white», «Blue», «Blue white»,* «M», «O», *«G».*

Cluster 1 has more size and more yellow colors.

**Cluster «2»**

for cluster 2 consist of following numeric features and categories: «Luminosity», «Radius», «HyperGiants», «Red», «White», «Orange», «M», «K», «G».

for cluster 2 consist of following numeric features and categories: «Temperature», «Absolute magnitude», *«Red Dwarf», «Brown Dwarf», «White Dwarf», «Main Sequence», «SuperGiants»,* *«Blue White», «Yellowish White», «Blue white», «Pale yellow orange», «Whitish», «yellow-white», «White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White»,* «B», *«F»*.

Cluster 2 has much greater luminosity and size and more red and orange.

**Cluster «3»**

for cluster 3 consist of following numeric features and categories: «Temperature», «Luminosity», «Main Sequence», «SuperGiants», «Blue», «Blue-white», «B», «F».

for cluster 3 consist of following numeric features and categories: «Radius», «Absolute magnitude», *«Red Dwarf», «Brown Dwarf», «White Dwarf», «HyperGiants»,* «Red», «Blue White», *«White», «Yellowish White», «Blue white», «Pale yellow orange», «Whitish», «yellow-white», «Orange», «White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White», «M», «A», «O», «K», «G».*

Cluster 3 has much higher temperature and luminosity and more blue colors.

**Cluster «4»**

for cluster 4 consist of following numeric features and categories: «Luminosity», «SuperGiants», «Blue», «O».

for cluster 4 consist of following numeric features and categories: «Radius», «Absolute magnitude», *«Red Dwarf», «Brown Dwarf», «White Dwarf», «Main Sequence», «HyperGiants»,* «Red», *«Blue White», «White», «Yellowish White», «Blue white», «Pale yellow orange»,* *«Blue-white», «Whitish», «yellow-white», «Orange», «White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White»,* «M», *«B», «A», «F», «K», «G».*

Cluster 4 has a very much greater size and luminosity and more blue colors.

**Cluster «5»**

for cluster 5 consist of following numeric features and categories: «Temperature», «Luminosity», «Radius», «HyperGiants», «Blue», «Blue-white», «B», «O».

for cluster 5 consist of following numeric features and categories: «Absolute magnitude», *«Red Dwarf», «Brown Dwarf», «White Dwarf», «Main Sequence», «Red», «Blue White», «White», «Yellowish White», «Blue white», «Pale yellow orange», «Whitish», «yellow-white», «Orange», «White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White», «M», «A», «F», «K», «G».*

Cluster 4 has a very much greater size, luminosity and temperature and more blue colors.

«Temperature», «Luminosity», «Radius», «Absolute magnitude», «Red Dwarf», «Brown Dwarf», «White Dwarf», «Main Sequence», «SuperGiants», «HyperGiants», «Red», «Blue White», «White», «Yellowish White», «Blue white», «Pale yellow orange», «Blue», «Blue-white», «Whitish», «yellow-white», «Orange», «White-Yellow», «white», «Blue», «yellowish», «Yellowish», «Orange-Red», «Blue white», «Blue-White», «M», «B», «A», «F», «O», «K», «G».

**Cluster «6»**

for cluster 6 consist of following numeric features and categories: «Temperature», «Absolute magnitude», «White Dwarf», «Blue White», «Yellowish White», «Blue white», «Blue», «white», «Blue», «Blue white»,

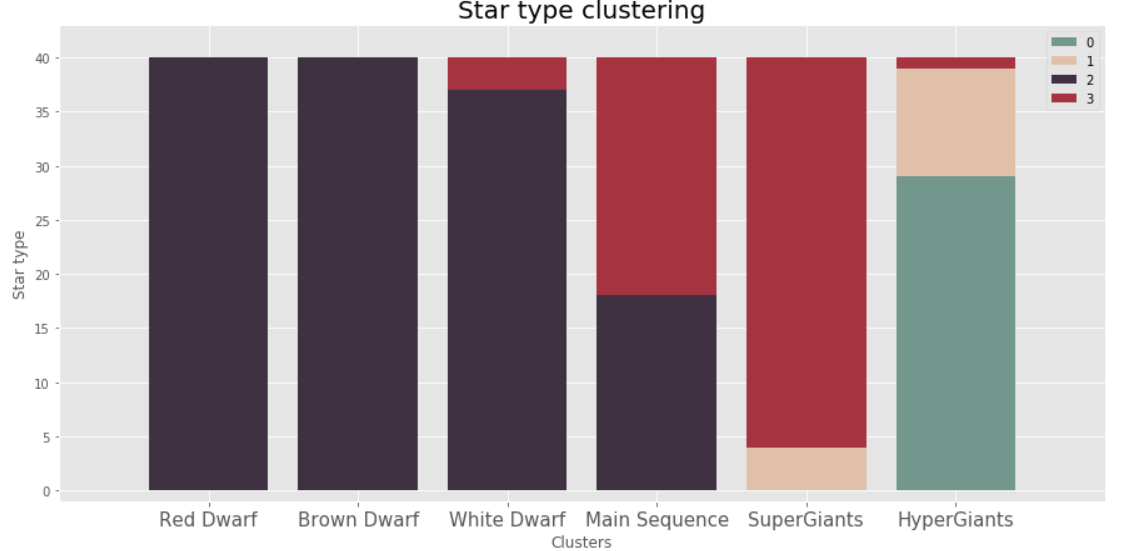
for cluster 6 consist of following numeric features and categories: «Luminosity», «Radius», *«Red Dwarf», «Brown Dwarf», «Main Sequence», «SuperGiants», «HyperGiants», «Red», «Pale yellow orange», «Blue-white», «Whitish», «yellow-white», «Orange», «White-Yellow», «yellowish», «Yellowish», «Orange-Red», «Blue-White».*

Cluster 6 has less size, higher temperature and more blue and white colors.

**Conceptualization**

Now let us have a look on the results of clustering according to the types of stars. We will assume that a type of star belongs to a cluster, which contains most of the objects of this type.

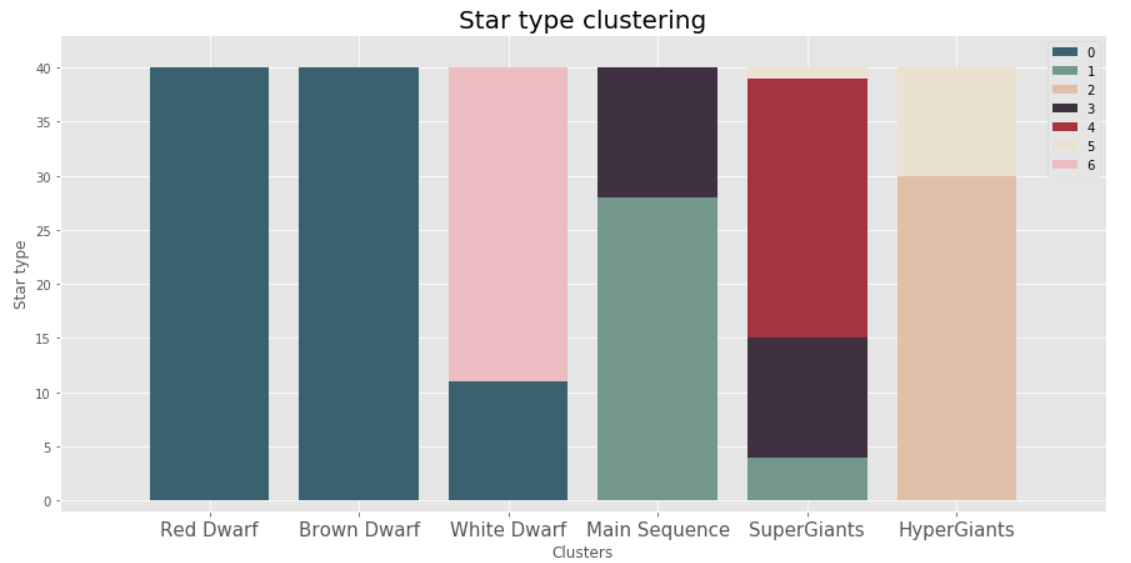
In the following graph we can see distribution of different types of stars for 4 clusters.



According to our assumption, we can say that «Red Dwarf», «Brown Dwarf» and «White Dwarf» are in the cluster «2», «SuperGiants» and «HyperGiants» are in the cluster «3». «HyperGiants» are in the cluster «0». The cluster «1» contains some stars of «SuperGiants» type and some stars of «HyperGiants».

Certainly, objects are not splitted perfect. There are few objects of «WhiteDwarf» and «HyperGiants» belong to the cluster «3» and almost a half of objects of «Main Sequence» belong to the cluster «2».

In the following graph we can see distribution of different types of stars for 4 clusters.



According to our assumption, we can say that «Red Dwarf», «Brown Dwarf» are in the cluster «0», «White Dwarf» is in cluster «6», «Main Sequence» is in cluster «1» «SuperGiants» are in the cluster «4». «HyperGiants» are in the cluster «2». The cluster «5» contains some stars of «HyperGiants» and «SuperGiants» and cluster «3» contains some stars of «Main Sequence» and «SuperGiants».

K-Means in both cases (4 clusters and 7 clusters) split stars according to their type. We think that 4-cluster splitting is better. Because it is more interpretable with given features in dataset for people without astronomic education. It splits dwarf stars, stars with average / a little bit bigger size and giant stars, while 7-cluster splitting select more subtypes of stars which are not selected in usual classifications.

Also 4-cluster splitting cope with big number of color categories better. There are a lot of the same colors which are written differently (for example: «Blue White», «Blue-white», «Blue white», «Blue-White») but in fact they are the same, little number of clusters let to ignore this thing.