# Data preprocess

According to the data of Most Recent Cohorts Data that we get from the IPED, we choose about 20 kinds of attributes from the above 200 kinds of attributes. The quality of data affects the quality of the data analysis’s results. In order to help improve the quality of the data and the analysis results, the raw data is preprocessed so as to improve the efficiency and ease of the analysis process. Data preprocessing is one of the most critical steps in our analysis process which deals with the preparation and transformation of the initial dataset. Our data preprocessing methods are divided into following categories:

1, invalid tuples cleaning

2, useless attributes cleaning

3, some sets of mutual substitution attributes cleaning

4, data transformation and data reduction

## Invalid tuples cleaning

Data that extracted from large, real-world data and data warehouses can be lack of attribute values or containing only aggregate data. From our dataset, we find a large number of tuples which contains invalid attributes likes ‘NULL’ or ‘PrivacySuppressed’ more than 60% of the total attributes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | UGDS | PCTFLOAN | NPT44\_PUB | SAT\_AVG | GRAD\_DEBT\_MDN10YR\_SUPP |
| Invalid Ratio(%) | 9.15 | 9.51 | 81.41 | 81.80 | 18.64 |

Incomplete data can occur to a lots of errors. In our algorithm, a tuple is abortion if the item of CURROPER is zero because the CURROPER is a flag for currently operating institution, 0=closed, 1=operating.

After the filtering of the data, we realize the UGDS, HCM and C150\_4\_POOLED\_SUPP/ C200\_L4\_POOLED\_SUPP attributes are vital to the analysis of data. We can’t make a credibly assess to a college or university if the attribute is missing. So we make a risk evaluation system, this system will be related in the Model testing and sensitivity analysis chapter. We can got a dataset which have a stable expression in our next steps after deleting some tuples which have an extremely lower security factor value.

## ‘Useless’ attributes cleaning

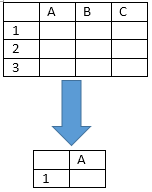
We consider that our data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, we make sure some attributes of data is useless **for us** because some other attributes have a highly performance when we do the research and can replace of them(useless attributes) completely.

In order to simplify our model, we have to throw away these ‘useless’ attributes. For example, gt\_25k\_p6 means share of students earning over $25,000/year (threshold earnings) 6 years after entry, we think this conception is one-sided because of the time interval is short and the standard $25,000/year is non-quantitative comparing to the md\_earn\_wne\_p10 which means median earnings of students working and not enrolled 10 years after entry. So we choose the md\_earn\_wne\_p10 to become a quantization standard among the class of future income attributes.

On the other hand, we must consider the generalization of the data, where raw data are replaced by higher level concepts. For example, categorical attributes, like INSTURL, can be generalized to higher concepts, like city or country. Similarly, the numeric attributes’ values (e.g., age or locale) may be mapper to higher level concepts, like Black, White, Hisp and Asian.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| UGDS\_WHITE | UGDS\_BLACK | | UGDS\_HISP | UGDS\_ASIAN | | UGDS\_AIAN | UGDS\_NHPI |
| UGDS\_2MOR | | **UGDS\_NRA** | | | **UGDS\_UNKN** | | |
| The attributes about distribution of different specials in American universities or colleges | | | | | | | |

This figure describes the data compression from different dimensions



## Some sets of mutual substitution attributes cleaning

As shown in figure, we can conclude that the ACT scores and SAT scores which express the average of the different colleges or universities have a linear relationship. So we can use the ACT\_TOTAL(Average SAT equivalent score of students admitted) or ACT\_TOTAL(Total SAT score of students admitted) to represent the undergraduates’ entrance examination results.

Like this, we can analyse that some other pairs of attributes have such likely features, such as gt\_25k\_p6 and md\_earn\_wne\_p10. We can use this property to make sure the critical factors to our analysis model.

## Data transformation reduction

In data transformation and reduction, the data are transformed into appropriate forms for the cluster analysis, information mining or attribute correlation. In our related works, we mainly do three ways of transformation:

1, Transform the attribute’s value to an normalization index

We utilize a simple but efficient function to make the attribute’s value clarity. The attribute data are scaled so as to fall within a small specified range 0 to 1(float number).

Obviously, the maximum one’s index is 1, and the minimum one’s index is 0. We realize that this way can give a clearly description for our programs.

2, Aggregation the data

We use the summary or aggregation operation are applied to the data. For a instance, the items such as ACTENMID(Midpoint of the ACT English score), ACTMTMID(Midpoint of the ACT math score) and ACTWRMID(Midpoint of the ACT writing score) can summarize a newly ACT\_TOTAL to weigh the concerns of the ACT scores.

Like this, we can reduce the data into a generality index to measure the different influence factor and decrease the amount of the attributes.

# Cluster analysis

Cluster analysis contains a broad suite of designed to find groups of similar attributes within a dataset. Partitioning methods divide the data set into a number of groups predesigned by the user. In our model, we must do a coarse-grained classification to simplify our system. After the process of cluster analysis, we will do qualitative analysis to our clusters and research the distance between the various clusters.

We choose the machine learning algorithm to complete our research because machine learning have three advantages: effectiveness, online learning ability and efficiency.

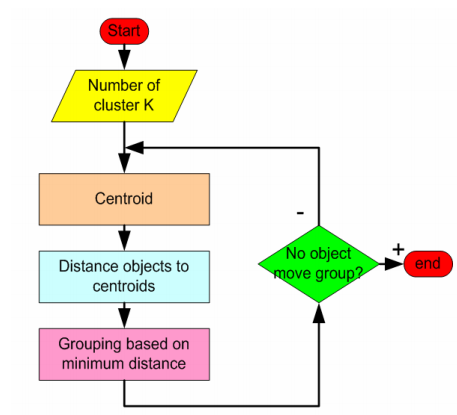
Between the all kinds of outstanding algorithms in machine learning, we find an almost excellent algorithm that which is called k-means clustering algorithm. k-means clustering that is a method of [vector quantization](https://en.wikipedia.org/wiki/Vector_quantization), aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) n observations into k clusters in which each observation belongs to the cluster with the nearest [mean](https://en.wikipedia.org/wiki/Mean), serving as a [prototype](https://en.wikipedia.org/wiki/Prototype) of the cluster.

We can describe our algorithm in a concise way containing three steps:

1, Give an initial set of k means , the algorithm proceeds to next two steps.

2, Assign each observation to cluster whose mean yields the least within cluster sum of squares (WCSS). This is the nearest mean after calculating the squared Euclidean distance (Assign different weight to different dimensions)

3, Calculate the new means to be the centroids of the observations in the new clusters. Then do the second step until the clusters tend to be stable.



In order to take an optimal K-value, we use an improvement algorithm of k-means algorithm. This kind of improvement algorithm called k-means++ chooses initial centers in a way that gives a provable upper bound on the cluster number. After the running of our program, we find an upper bound between 4 and 6, so we assign an integer 4 to K.

We have ran the program we have designed above, and gained two useful datasets.

1, the attribute values of centroid of clusters.

2, the K sets of different university or colleges.

We running this program several times because we randomly initialize the clusters’ centroid when the first step of our program. And then, we received a set of average value after handling the several similar data. We make sure the different automatic initialization can gain an identical result. From the figure, we can infer the number of universities or colleges in clusters have a similar quantity.

|  |  |  |  |
| --- | --- | --- | --- |
| Time  cluster | 1 | 2 | 3 |
| 1 | 297 | 292 | 293 |
| 2 | 352 | 352 | 352 |
| 3 | 145 | 145 | 145 |
| 4 | 604 | 609 | 608 |

# Explanation and analysis of the clusters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | UGDS | SAT\_AVG\_ALL | md\_earn\_wne\_p10 | GRAD\_DEBT\_MD  N\_SUPP | NPT4 | NUM |
| 1 | 4524.39 | 1200.92 | 58147.35 | 23348.05 | 27181.92 | 297 |
| 2 | 4453.09 | 1006.38 | 37117.85 | 21110.74 | 12165.72 | 352 |
| 3 | 22374.66 | 1116.39 | 45974.83 | 20767.00 | 14348.20 | 145 |
| 4 | 2114.93 | 1035.04 | 40271.24 | 26136.16 | 20713.19 | 604 |
| AVE | 8366.77 | 1089.68 | 45377.82 | 22840.49 | 18602.26 | 1398 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| clusters | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 |
| 1 | B | A+ | A+ | A | A+ |
| 2 | B | B | C | B | C |
| 3 | A+ | A | A | B | B |
| 4 | C | B | B | A+ | A |
| A+ means highest, A means higher, B means middle, C means lower | | | | | |

Factor1 (UGDS): the student scale of the school. TO DO

Factor2 (SAT\_AVG\_ALL): the average SAT (Scholastic Assessment Test) equivalent score admitted. TO DO

Factor3 (md\_earn\_wne\_p10): Median earnings of students working and not enrolled 10 years after entry. TO DO

Factor4 (GRAD\_DEBT\_MDN\_SUPP): Median debt of completers, suppressed for n=30. TO DO

Factor5 (NPT4): Average net price for Title IV institutions (public institutions, private for-profit and nonprofit institutions) TO DO

After the analysis of clusters’ different features, we research that No.3 cluster is the best investment objection for the Goodgrant Foundation.

# Risk and security analysis

Large and complex investment projects have always needed a substantial management structure to insure that our different parts in an origanized fashion to achieve the tasks at hand. We think that different university or colleges like different stocks. And anyone of the ‘stocks’ exists systematic risks and unsystematic risks. So we design a model to calculate the risk-factor index to evaluate the risks in a comprehensive method.

We can consider lacking of vital data is a kind of systematic risk, so we evaluate various attributions to conclude an experimental formula and assign weight to the factors. Then we analysis two sides of risk and give an appropriate value 0.4 for coefficient of d.

#####################################################################################

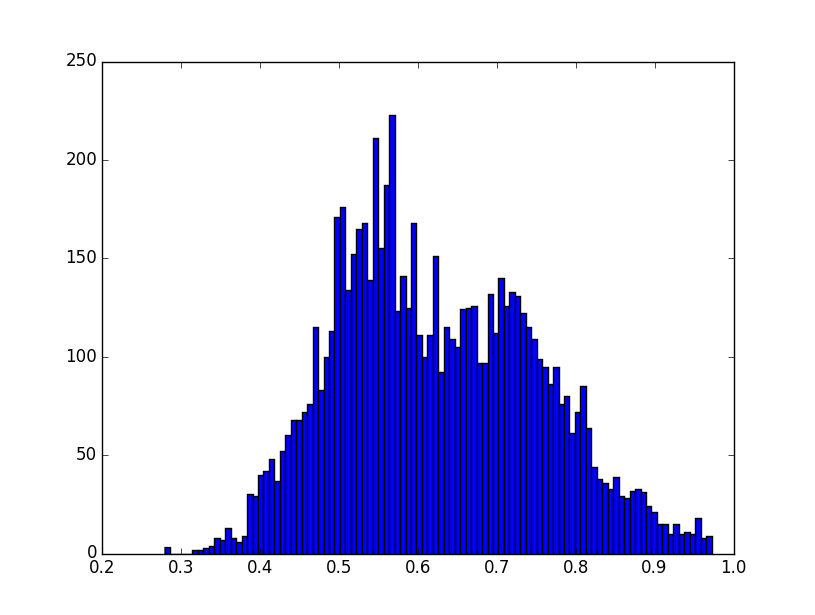
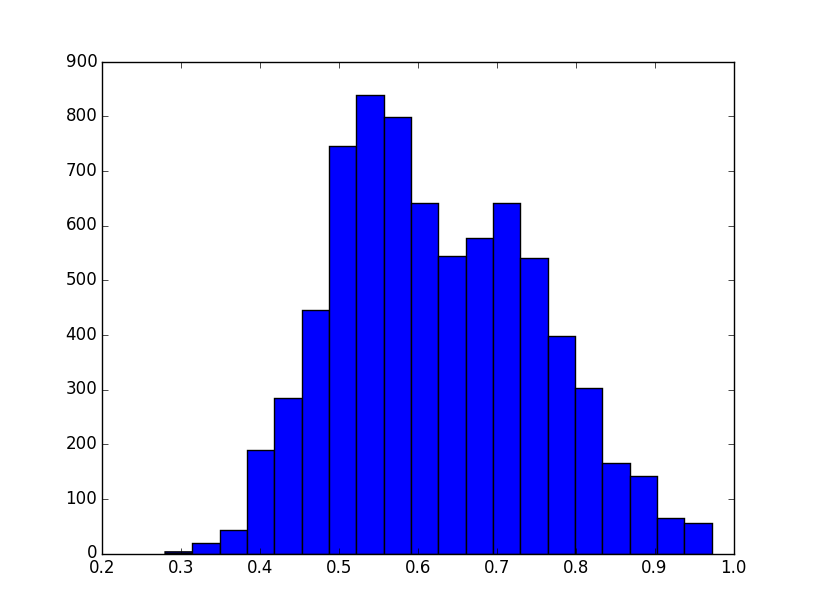
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | HCM2 | UGDS | SAT\_AVG\_ALL | md\_earn\_wne\_p10 | GRAD\_DEBT\_MDN\_SUPP |
| Weight | 7 | 3 | 3 | 2 | 2 |
| Attribute | NPT4 | PCT | UG25abv | POOLED\_SUPP | RPY\_3YR\_RT\_SUPP |
| Weight | 1 | 1 | 2 | 4 | 4 |

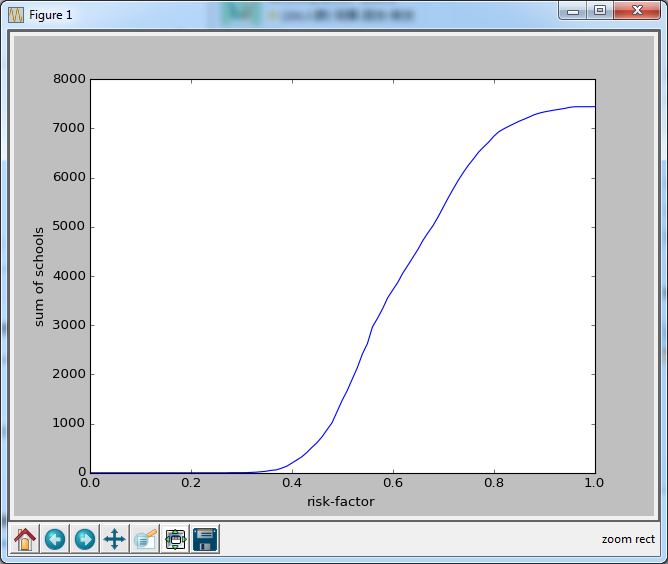
With the systematic risks, we give a set of weights to the major ten attributes which play a vital role in our model. So we can estimate an appropriate value of the systematic risks.

With the unsystematic risks, we choose RPY\_3YR\_RT\_SUPP (3-year repayment rate, suppressed for n=30) and POOLED\_SUPP (a new attribute measures the graduation rate by merging C150\_4\_POOLED\_SUPP and C200\_L4\_POOLED\_SUPP In the previous chapter) to become our measure standard, and we adopt an equal weight method to estimate the unsystematic risks.

#####################################################################################

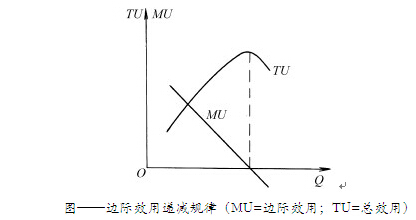
From the figure, we can gain a growth curve that describes the increasing of universities or colleges’ total quantity by the increasing of risk-factor index. Our figure is similar with the standard risk growth curve in most of the models.





# Marginal efficiency of investment

The marginal efficiency of investment is that [rate of discount](https://en.wikipedia.org/wiki/Discounting) which would equate the price of a [fixed](https://en.wikipedia.org/wiki/Fixed_asset) investment [asset](https://en.wikipedia.org/wiki/Asset) with its [present discounted value of expected income](https://en.wikipedia.org/wiki/Discounted_cash_flow). We must consider the performance of our investment, so we need to introduce the marginal efficiency to our model.



MU= marginal utility, TU=total utility, AU=average utility.

As a simple example, we will face a law of diminishing marginal utility if we invest a small amounts. On the other hand, we must think over the law of increasing marginal utility within an excessive investment. We consider the investment needs to be an appropriate amount. This is because the MU will returns to negative with increasing of Q (quantity). These variables can be expressed as follows: