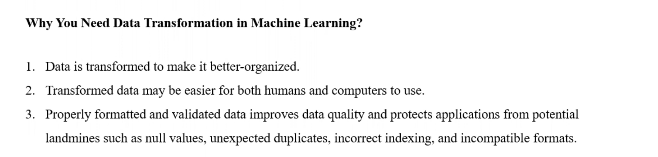
**Data Transformation:** Data transformation is the process of changing the format, structure, or values of data.

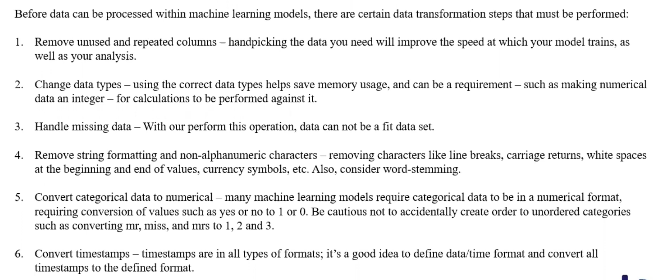


:

**Data transformation process:** Data transformation is basically done following these 4 simple steps.

* Data interpretation
* Pre-translation data quality check
* Data translation
* Post-translation data quality check

In details, These are the steps that are followed for data transformation.

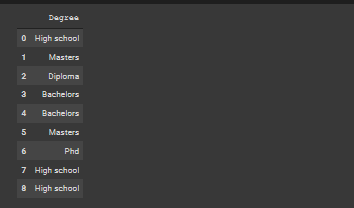


**Encoding** is the process of converting the data or a given sequence of characters, symbols, alphabets etc., into a specified format, for the secured transmission of data, and data encoding is the process of transforming data into feasible formats for work purposes.  
This is mostly done representing ‘strings’ or any sort of categorical data in finite numbers.  
  
We have two kinds of categorical data.

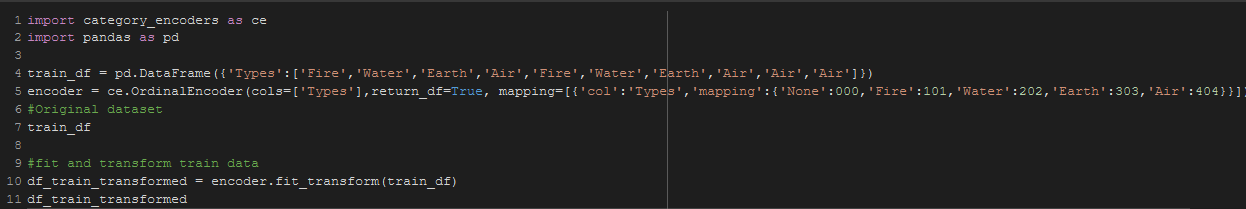
* **Ordinal Data:** The categories do not have an inherent order.  
  In Ordinal data, while encoding, one should retain the information regarding the order in which the category is provided.   
  For example, The highest degree a person possesses gives vital information about his qualification. The degree is an important feature to decide whether a person is suitable for a post or not.
* **Nominal Data:** The categories do not have an inherent order.  
  While encoding Nominal data, we have to consider the presence or absence of a feature. In such a case, no notion of order is present.   
  For example, the city a person lives in. For the data, it is important to retain where a person lives. Here, We do not have any order or sequence. It is equal if a person lives in Delhi or Bangalore.

There are different types of data encoding techniques available. Some of them are,

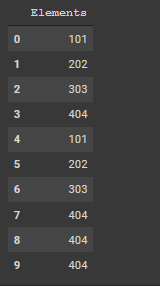
* **Label encoding or Ordinal encoding:** Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form.   
  Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.  
   

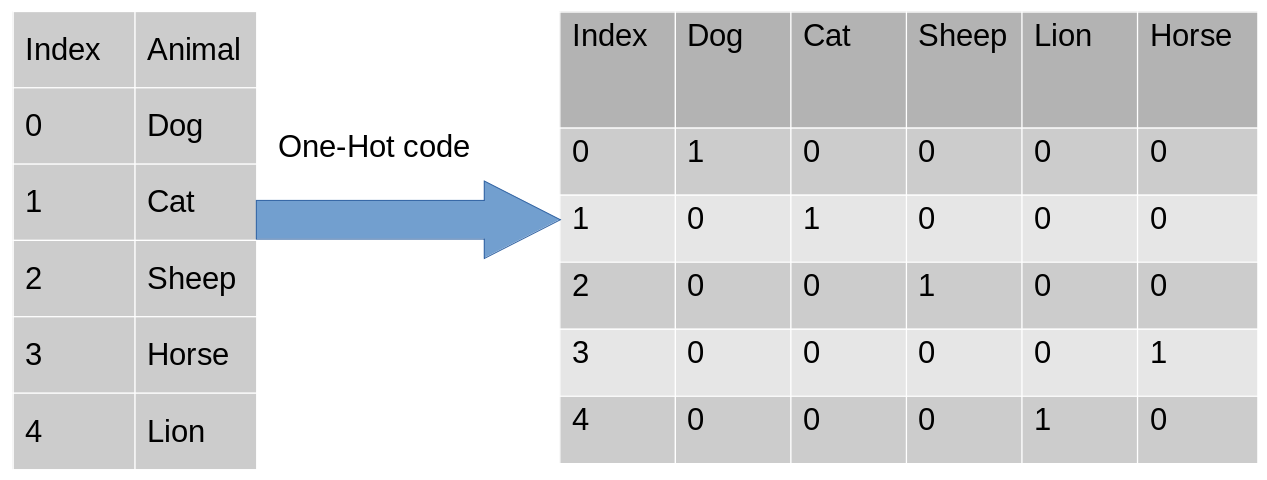


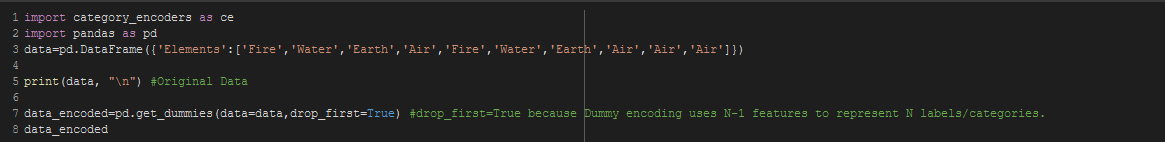
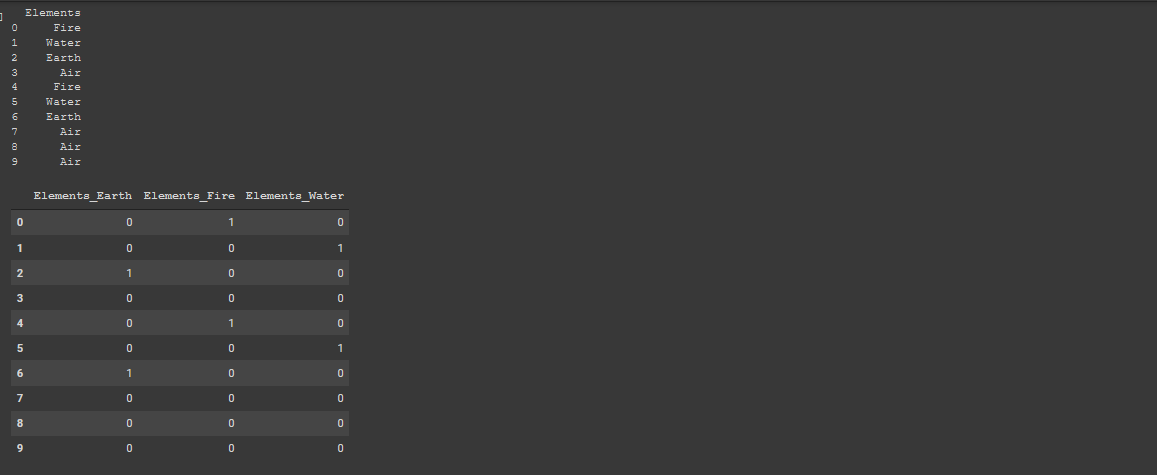
If we want to transform this dataset’s categorical values into numerical values, We use the label encoding technique.  
**Note:** We need to use the ***“category\_encoding”*** library for that.



We declared the column name “Types” in 4th line and set the values or categorical and then we used the “category\_encoder” library’s “fit\_transform” method to transform or convert the categorical data into numerical data. The output becomes something like this.

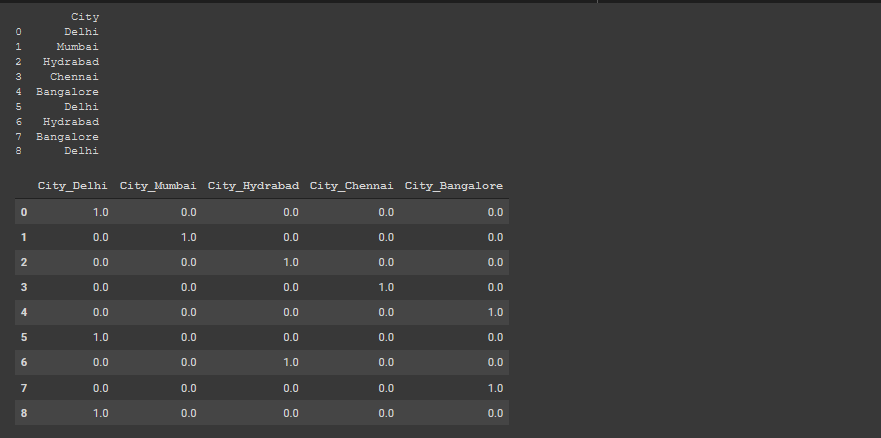


* **One hot encoding:** In digital circuits and machine learning, a one-hot is a group of bits among which the legal combinations of values are only those with a single high bit and all the others low.   
  *A similar implementation in which all bits are '1' except one '0' is sometimes called one-cold*.  
  **Example:** Suppose we have a dataset with a category animal, having different animals like Dog, Cat, Sheep, Cow, Lion. Now we have to one-hot encode this data.  
     
  After encoding, in the second table, we have dummy variables each representing a category in the feature Animal. Now for each category that is present, we have 1 in the column of that category and 0 for the others.
* **Dummy Encoding:** Dummy coding scheme is similar to one-hot encoding. This categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). In the case of one-hot encoding, for N categories in a variable, it uses N binary variables. The dummy encoding is a small improvement over one-hot-encoding.   
  **Note:** Dummy encoding uses N-1 features to represent N labels/categories.  
  **Example:** Here we are coding the same data using both one-hot encoding and dummy encoding techniques. While one-hot uses 3 variables to represent the data whereas dummy encoding uses 2 variables to code 3 categories.

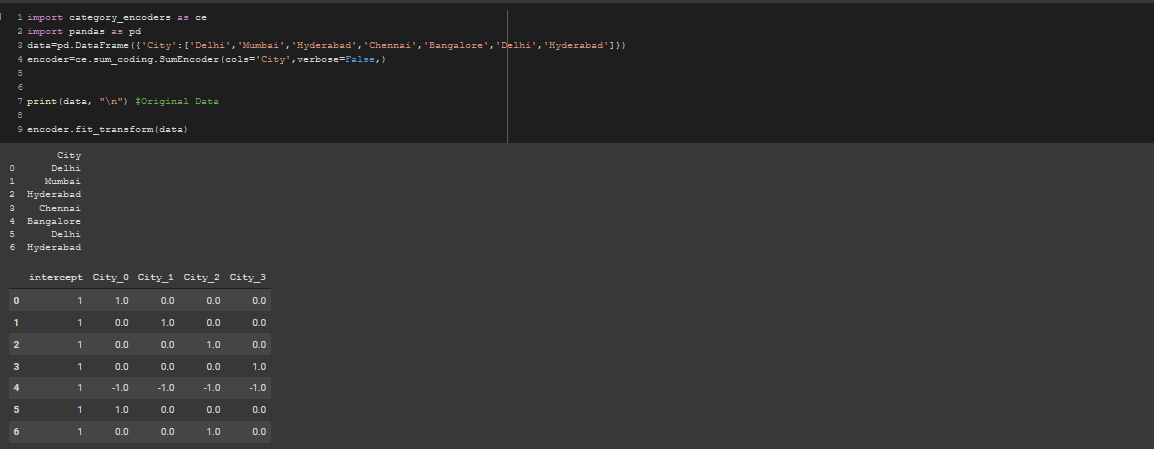
  


It’s featuring n-1 or 4-1 =3 features. All 4 features can be shown by setting the ***“drop\_first = False”***

* **Effect Coding:** This encoding technique is also known as Deviation Encoding or Sum Encoding. Effect encoding is almost similar to dummy encoding, with a little difference. In dummy coding, we use 0 and 1 to represent the data but in effect encoding, we use three values i.e. 1,0, and -1.

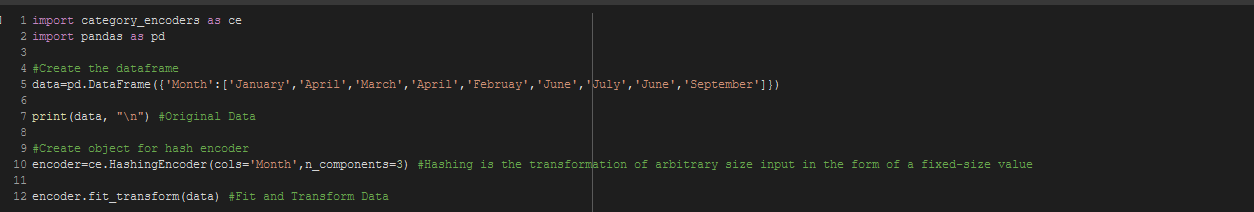


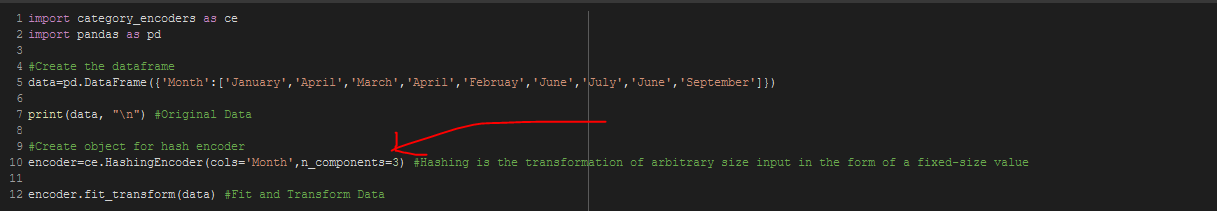
The row containing only 0s in dummy encoding is encoded as -1 in effect encoding. In the dummy encoding example, the city Bangalore at index 4 was encoded as 0000. Whereas in effect encoding it is represented by -1-1-1-1.



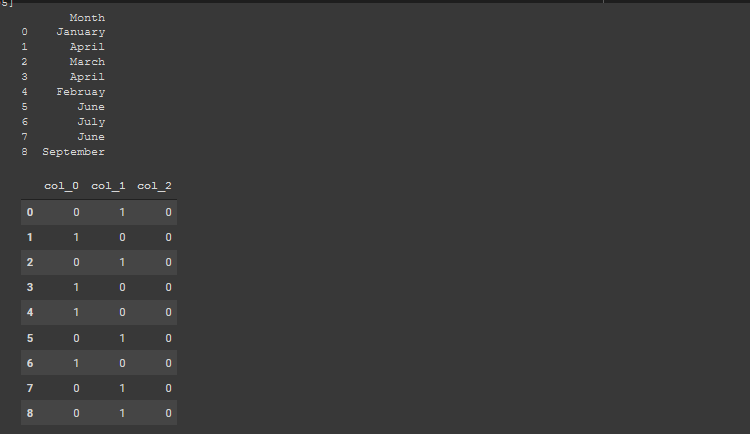
* **Hash Encoding:** To understand Hash encoding it is necessary to know about hashing. Hashing is the transformation of arbitrary size input in the form of a fixed-size value. We use hashing algorithms to perform hashing operations i.e to generate the hash value of an input.

Just like one-hot encoding, the Hash encoder represents categorical features using the new dimensions. Here, the user can fix the number of dimensions after transformation using***the n\_component*** argument. Here is what I mean – A feature with 5 categories can be represented using N new features similarly, a feature with 100 categories can also be transformed using N new features.



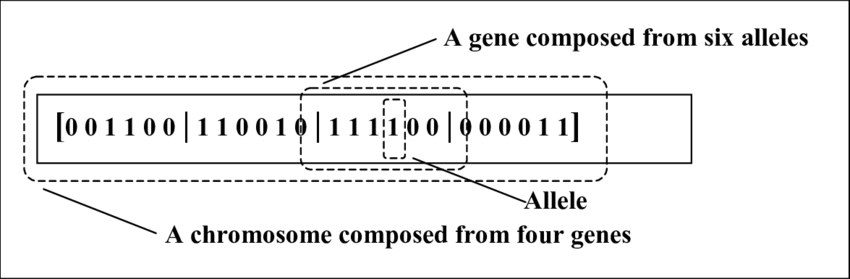


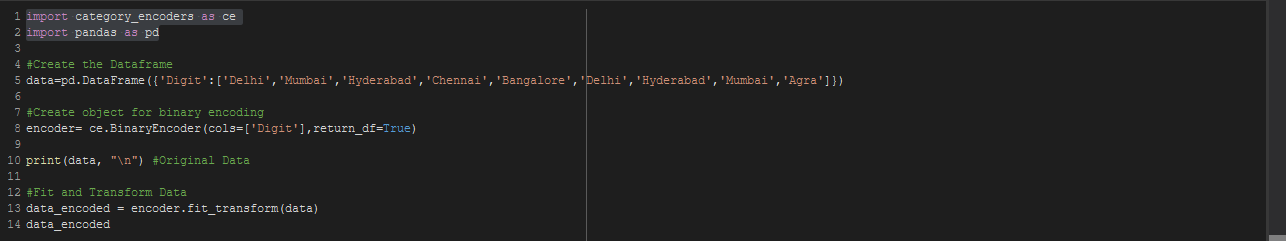
We can set the column size by changing the ***“n\_component”*** ‘s values.



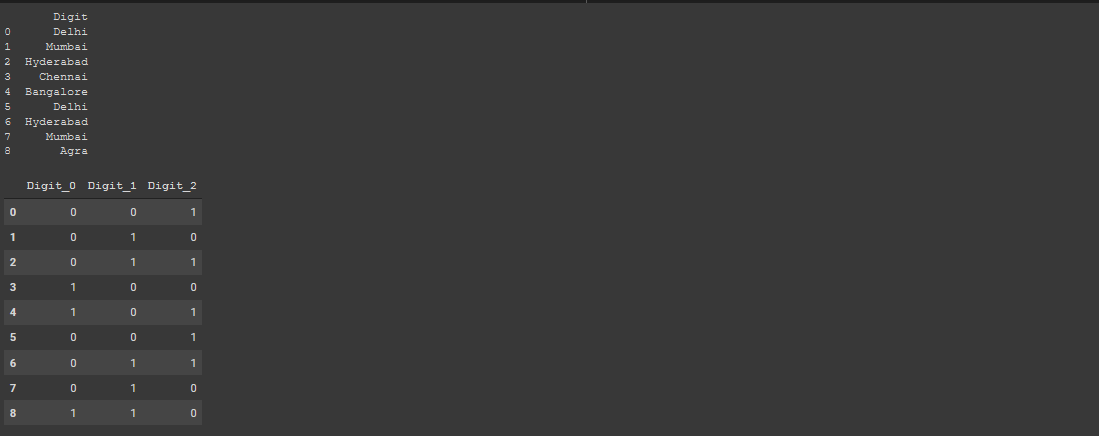
* **Binary Encoding:** Binary encoding is a combination of Hash encoding and one-hot encoding. In this encoding scheme, the categorical feature is first converted into numerical using an ordinal encoder.

Then the numbers are transformed into binary numbers and then binary values are split into different columns.  
Binary encoding works really well when there are a high number of categories. For example the cities in a country where a company supplies its products.

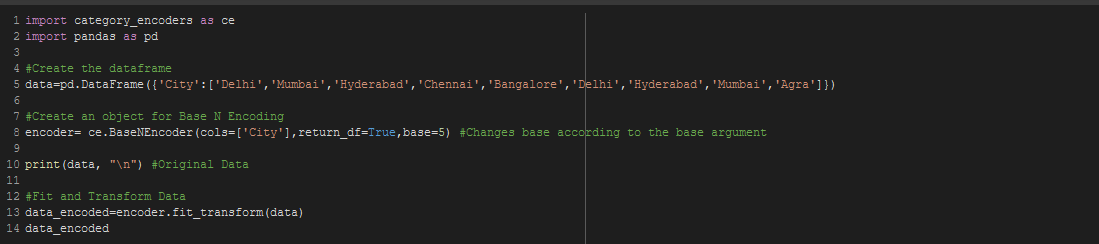




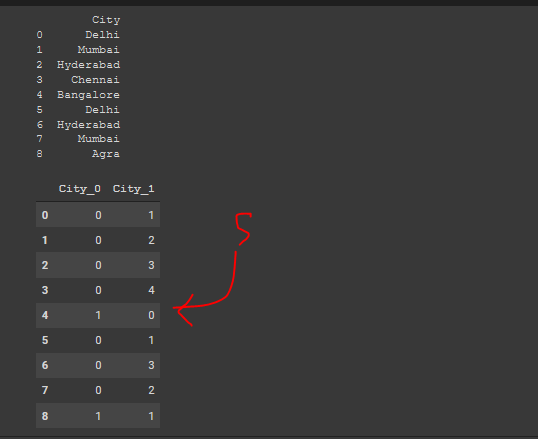
Binary encoding is done in a random manner and each and the same data has the same encoding.



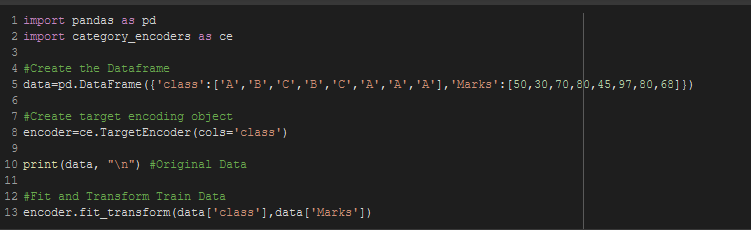
* **Base N coding:** Base-N encoder encodes the categories into arrays of their base-N representation. A base of 1 is equivalent to one-hot encoding (not really base-1, but useful), a base of 2 is equivalent to binary encoding. N=number of actual categories is equivalent to vanilla ordinal encoding.

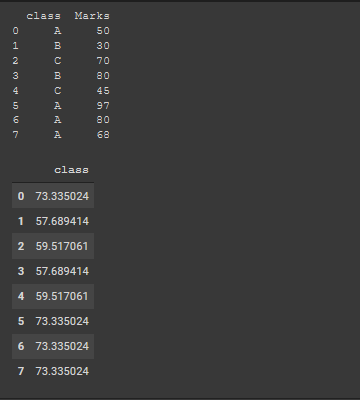


The encoding base can be changed by changing the base argument.



* **Target coding:** In target encoding, we calculate the mean of the target variable for each category and replace the category variable with the mean value. In the case of the categorical target variables, the posterior probability of the target replaces each category.





It calculates the means of each element by their set weights.