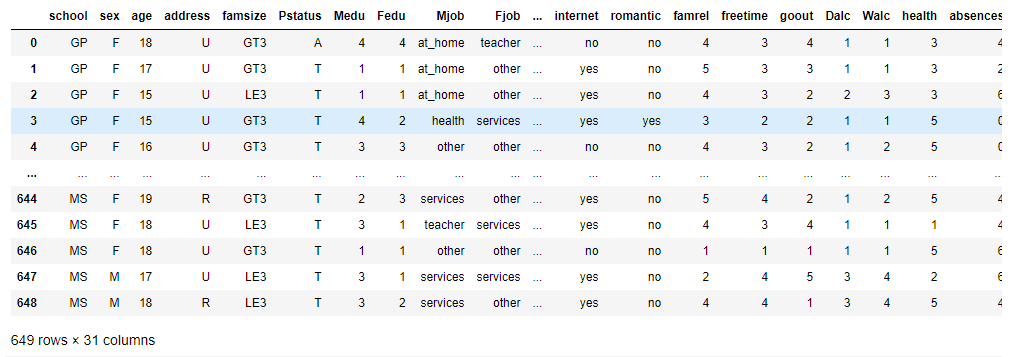
## 1、 Data preparation

Use the Pandas module to obtain the read\_ csv method reads the file. Through observation, it can be found that the file data is not in the standard CSV format.

图形用户界面

中度可信度描述已自动生成

Use the above code to process the data format and save it to the DataFrame data table. The processed data is shownbelow:



### 1.1 Viewing Empty Data

### 1.2 Characteristics of observed data

By observing the data type, we can find that in addition to numerical data, the data also contains string data. There are two types of classification, such as school, multi classification type Mjob, etc. In order to build a linear model later, these data need to be converted into numerical types in advance.

## 2、 Data analysis

### 2.1 Data cleaning and preprocessing

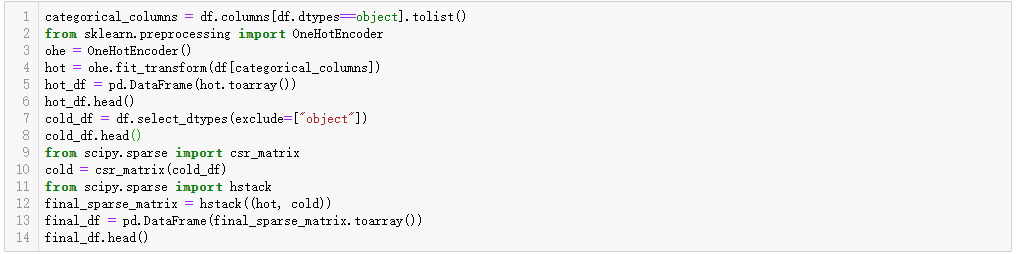
As with pd.get\_ dummies, the one pot code converts all classification values into 0 and 1, where 0 means no existence, 1 means existence, but it is the same as pd.get\_ The difference with dummies is that its computing cost is not high. OneHotEncoder uses a sparse matrix instead of a dense matrix to save space and time.

Sparse matrices save space by storing only data whose value does not include zero. Save the same amount of information by using fewer bits.

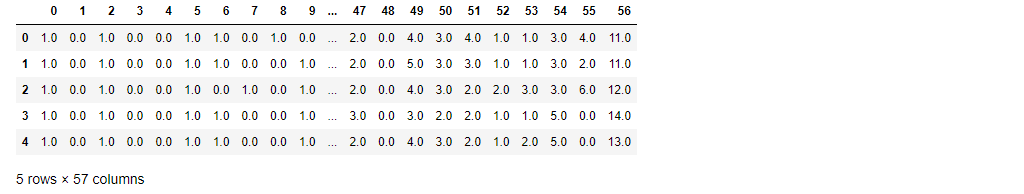
In addition, OneHotEncoder is a scikit born converter, which means it is specifically designed for machine learning pipelines.

In previous scikit-learn versions, OneHotEncoder only accepted digital input. In this case, use LabelEncoder to take the intermediate step of first converting all classification columns to numerical columns.

Here, we choose to convert the data to OneHotEncoder. The specific codes are as follows:



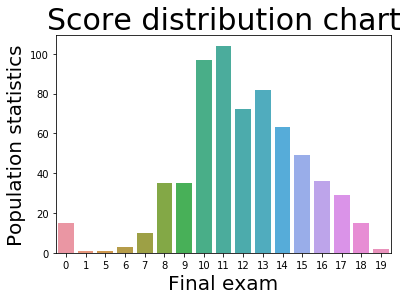
Results after conversion:



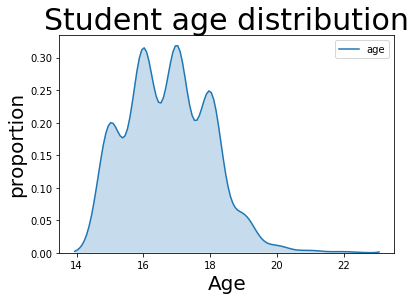
### 2.2Data analysis

Take the relationship between students' age, location and performance as an example to analyze.

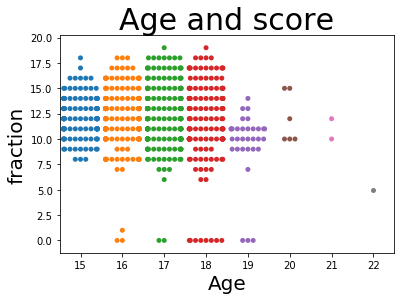
First, draw the score distribution histogram. It can be found that the histogram of students' scores and number of students roughly conforms to the normal distribution, mainly focusing on the average of 10-11 points, but there is a significant change at 0 points. It is reasonable to doubt whether 0 points include missing data or special circumstances such as absence from examinations.



Next, analyze the relationship between student achievement and student age. It can be found from the student age distribution map that the students' ages are mainly between 15 and 19 years old.



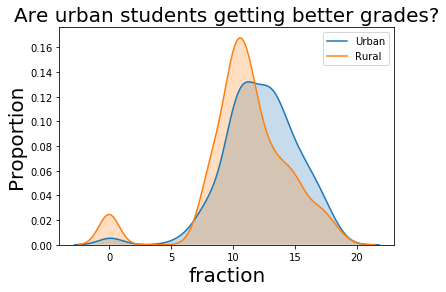
Next, discuss the distribution of grades in different age groups. By drawing the score distribution map, we can find that in the main age group of 15-18, the distribution of students' scores in each score segment is basically the same, but in the high segment, the age groups of 17 and 18 are more concentrated. From this, we can see that there is some relationship between age and scores.



Then analyze the relationship between residence and achievement. It can be found that most students live in cities by drawing the histogram of the number of people living in the city.



By drawing the distribution curve of students' grades in urban and rural areas, it can be found that the peak distribution of students' grades in rural areas is lower than that in urban areas. At the same time, in the extreme data of 0, rural students' performance is more obvious than that of urban students.



## 3、Data randomization

Select under the sklearn train\_ test\_ Split method, randomly divide the data set into training set and test set.



## 4、Build model

### 4.1Multiple linear regression model

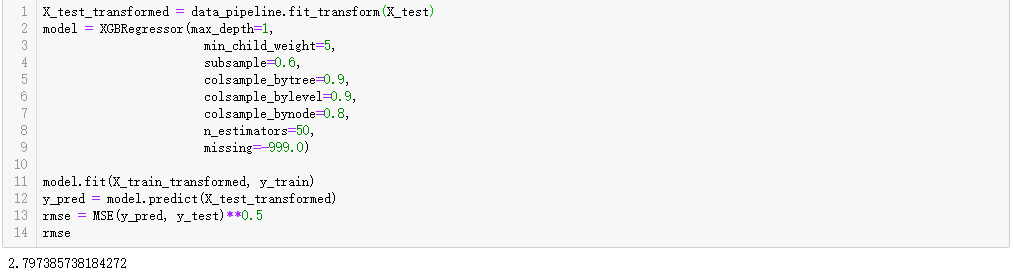
图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成**4.2XGBoost model**



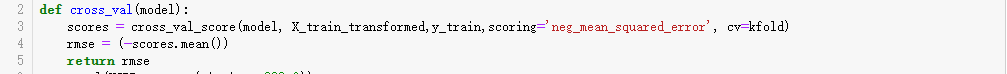


## 5、Model comparison

Compared with XGBoost model and Multiple linear regression model, the RMSE optimal result of XGBoost model is 2.79, which is higher than the optimal result of linear model 2.39; Through data comparison, we can find that the regression effect of XGBoost model is worse than that of linear model. To sum up, the linear model is simpler than other machine learning models for this dataset, with smaller RMSE and better regression effect.

### 5.1Calculate MSE and RMSE

By using cross\_val\_score method in the sklearn library calculates the RMSE.



## 6、Summarize the harvest

In data mining, the more complex the model, the better, but the model matching the data characteristics should be selected. In this case, the result of selecting the complex XGBoost model is not better than the simple linear model.