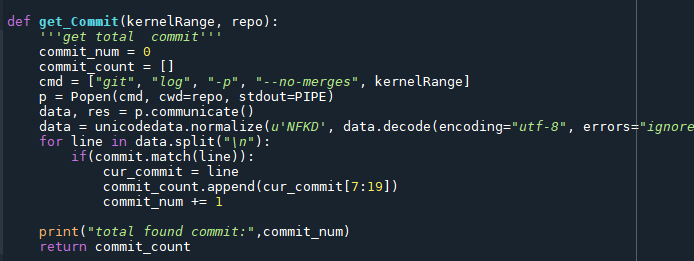
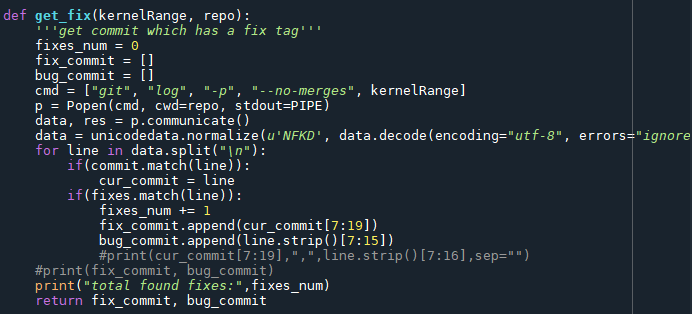
Assumption: The quality of the code may be related to the location

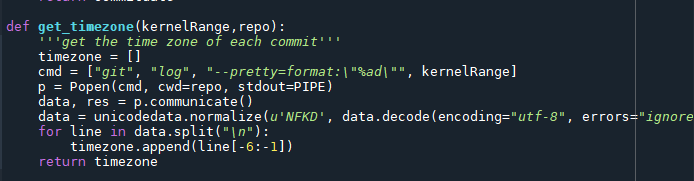
There are many factors that affect the quality of the code. Here we consider that the quality of the code may be related to the location. The standard for measuring the quality of the code is whether the content of the commit contains a fix tag. If it contains a fix tag, it means that it is not a good code; if it does not contain a fix tag, it means it is a good code. The area you are in is represented by a time zone, and different time zones represent different regions

First, we first extract the data we want. The version we studied here is version 4.7. First extract the hashes of all commits in all git logs, and then extract the hashes of commits containing fix tags. After this, we also use the same method to extract the author's name, email, commit content, etc. from the log for further analysis。Create an empty list, if the hash value of a certain commit appears in the list of commits containing fix tags, add 0 to the list just created, otherwise add 1 to the list just created

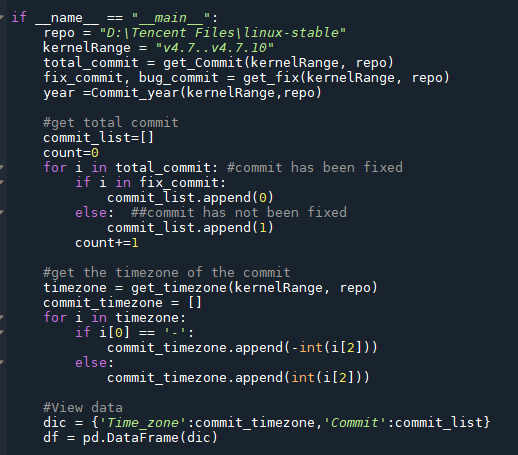


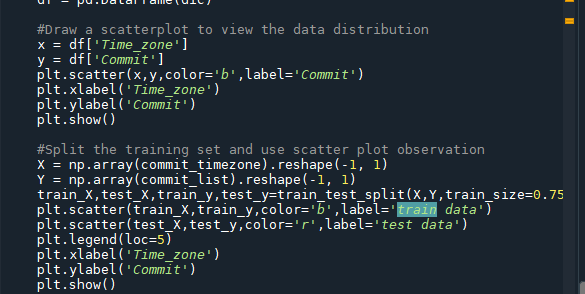


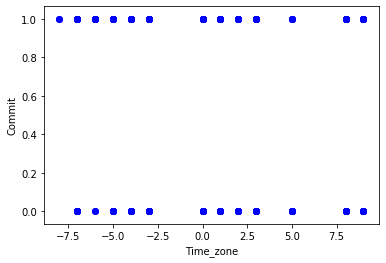
Then, we extract the time zone corresponding to each commit. This can be extracted from the commit time of each commit. Here we also have to do some processing to convert the time zone into a digital form. For example, convert "-0100" to -1, "+0600" to 6, and so on.



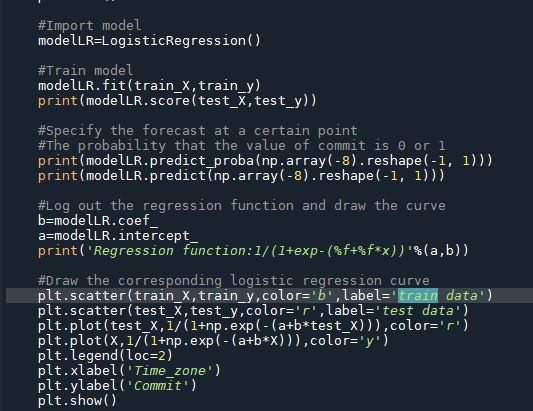
Use the pandas dataframe method to integrate the data just now, and use the matplotlib module to draw the corresponding scatter plot in order to observe the relationship between them.

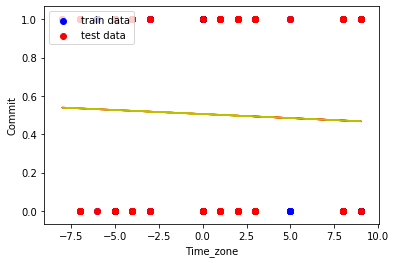


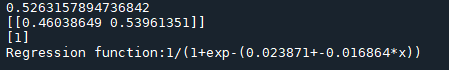




Because the value of the dependent variable is only 0 or 1, we chose to use logistic regression to evaluate the relationship between them. First divide the obtained data into training set and test set. Use the LogisticRegression method in the sklearn library to build a logistic regression model. Use this model to predict the probability that the predicted value of different points is 0 or 1. After this, add the regression curve corresponding to the model to the scatterplot and find the regression equation.







Finally, the correlation coefficient of the model is calculated to be 0.5263157894736842. This can show that there is a certain relationship between the quality of the code and the different time zones, but the correlation is not very strong. The degree to which the model can explain is average. The possible reason for such a result is that there are too few types of independent variables, which cannot be explained well, and other variables need to be introduced to better optimize the model.