**471 Midterm**

**Zhoumengdi Wang zxw534**

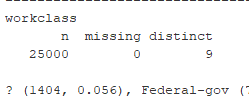
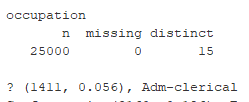
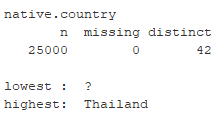
**1.Preprocessing data/Perform exploratory analyses**

(1) What is missing data?

The first thing we need to do is preprocessing the data. Now we need to take care of the missing data. Hmisc::describe(census\_train),this function can help us find the missing data.

I use Hmisc::describe to view the whole data.

According to the form, there is no missing data, but I found:



We can see that there are three variables has the value”?”. Although ”?” isn’t regarded as missing data, but ”?” means nothing in real. In fact, ”?” is the missing data.

1. How to deal with the missing data

With the missing data, there are some ways to deal with:

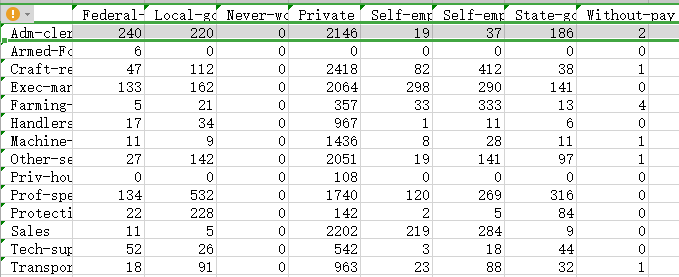
1. delete

2. use the highest frequency value to replace the missing data

3. use the relationships between variables

I will combine 2 and 3 to deal with the missing data. The first class which has the missing data is workclass. I think the workclass is influenced by occupation. Because different workclass is based on what kind of job you have. So my first step is dealing with the missing data in occupation. For the missing data in occupation, I would like to use highest frequency value to replace the missing data. For occupation, the highest frequency value is “Prof-specialty” so I use “Prof-specialty” to replace the missing data in occupation.

After dealing with the missing data in occupation, I need to find the relationship between the occupation and workclass. Show the table of these two vatiables:



We can find that when the occupation is “Prof-specialty”, the highest frequency value of workclass is “Private”. I use private to replace the missing data in workclass.

There is one variable which has missing date, the native\_country. Because it is a US census, the native country of most people in this census is US, so the highest frequency value of native country is United\_States.

After dealing with the missing data and checking there is no left missing data, group two part data.

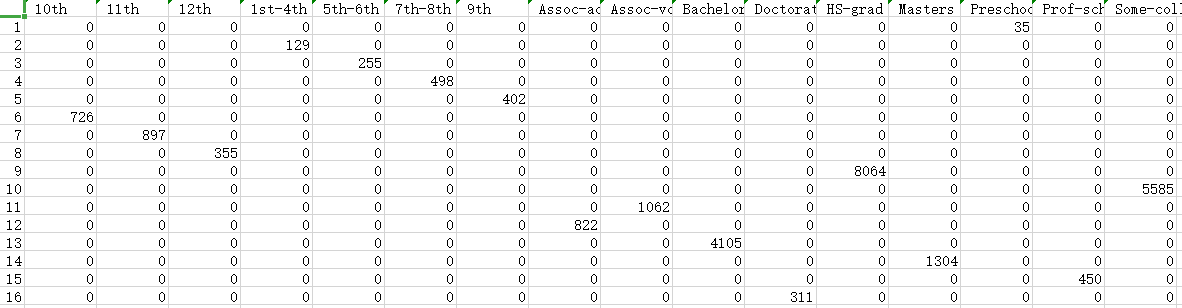
1551910736(1)

1. Check the variables

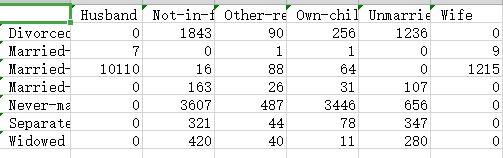
Close definition variables

There are two pairs variables, which has close definition, education and education.num, marital.status and relationship. Here are the tables of these two pairs:

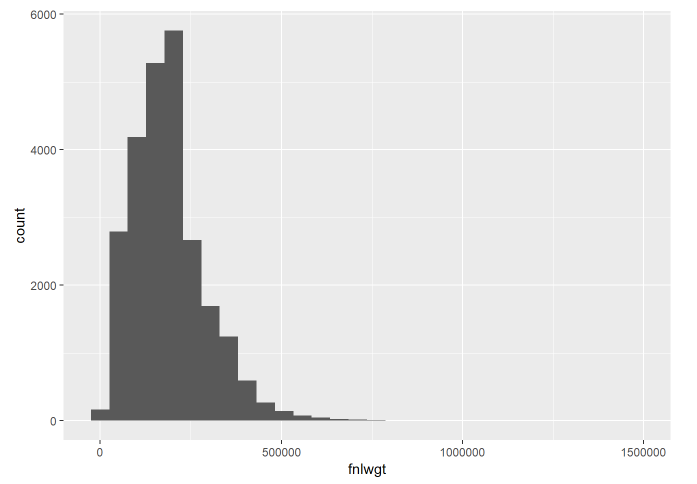
education and education.num,

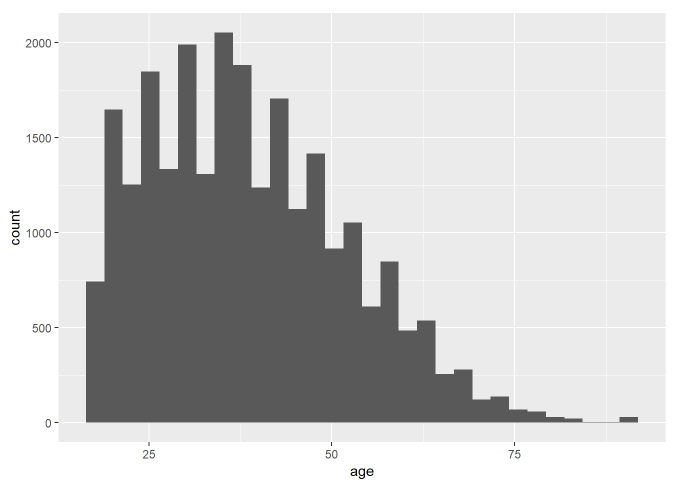
The table shows that each class of education.num matches a education level, so I think the education should be dropped.

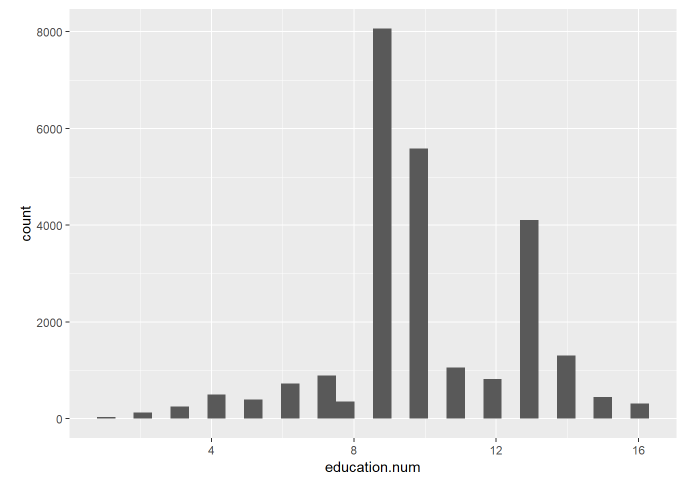
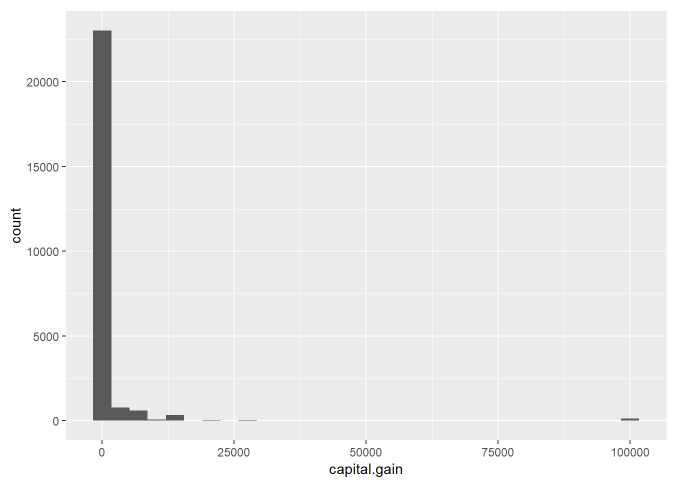
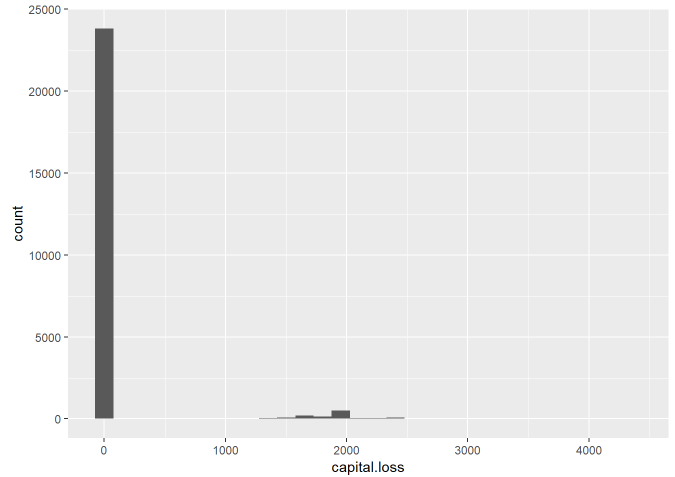
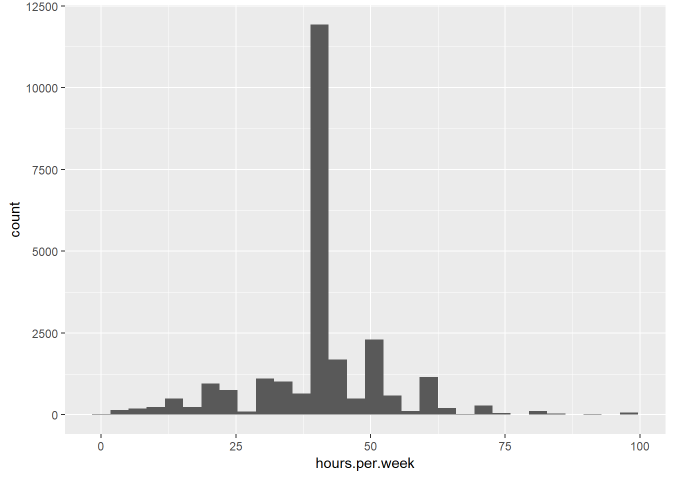
marital.status and relationship



I think marital.status is more clear than the relationship, so I decide to drop relationship. And I will combine “Married-AF-spouse” and “Married-civ-spouse” which seem like similar into a single category called “Married”.

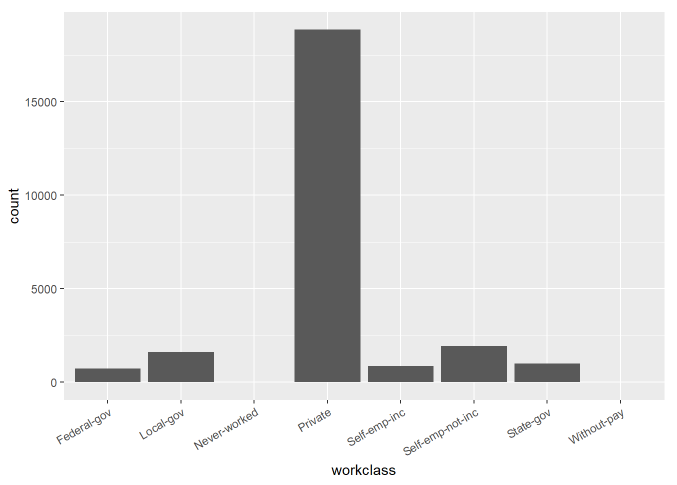
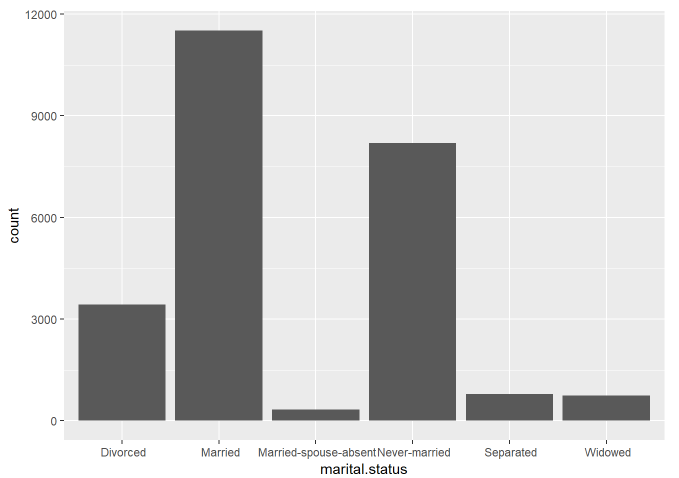
Check quantitative data

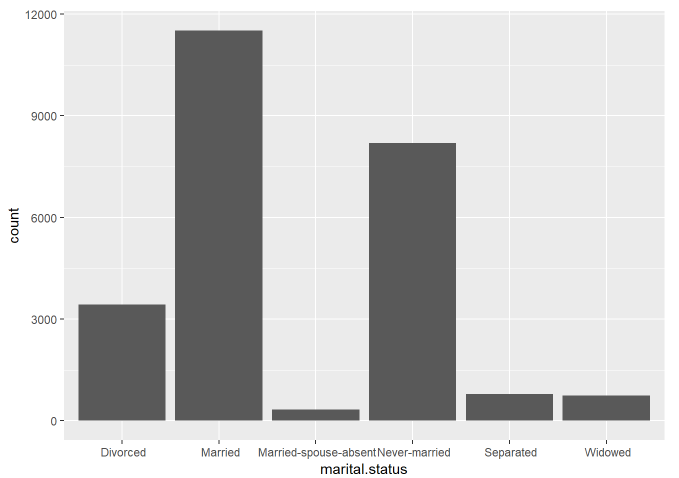


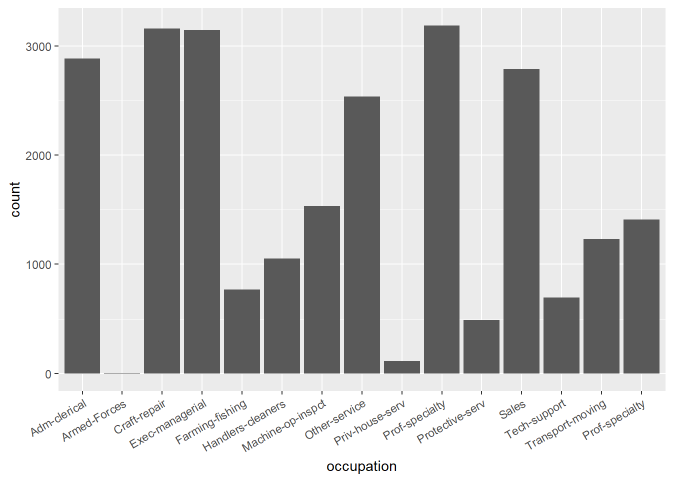
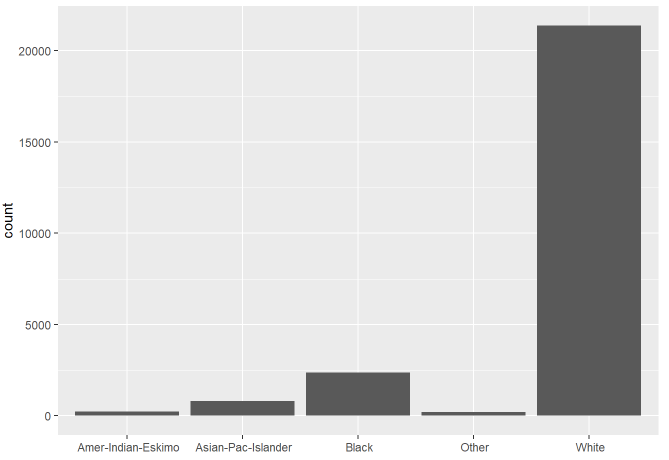
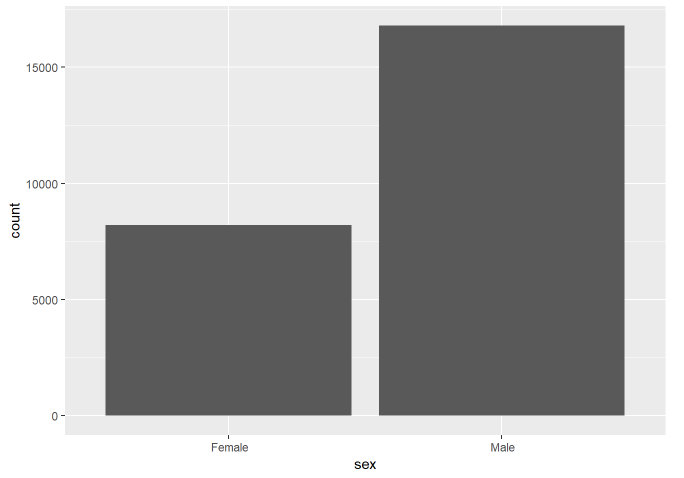
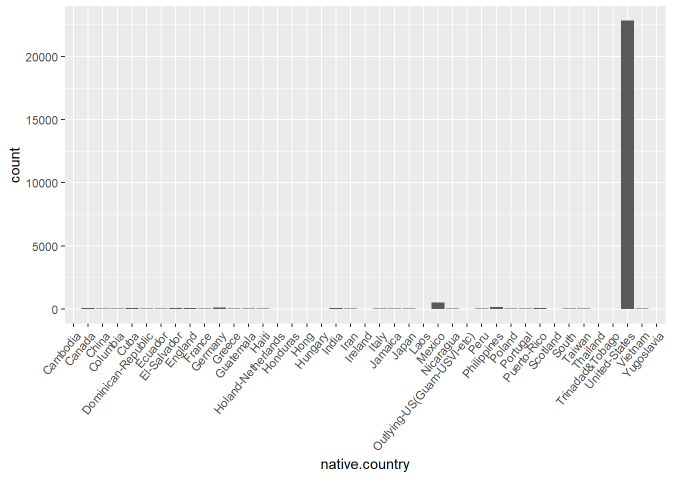
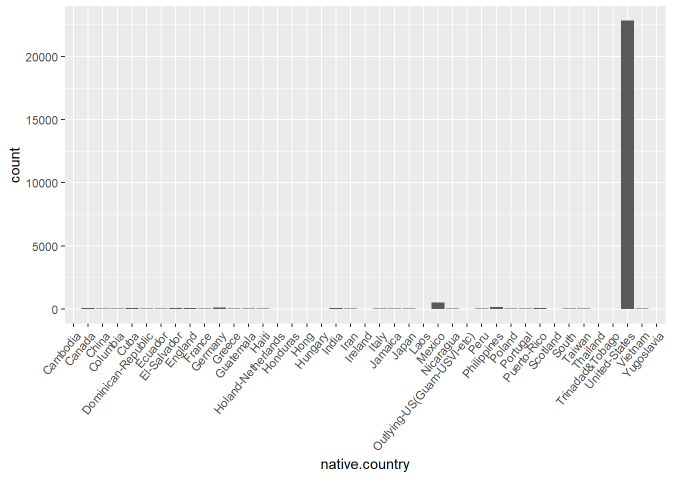


Distribution of most quantitative data is fine, however, capital.gain and capital.loss of most people is zero. I may delete them in adjusting models.

Watch qualitative data



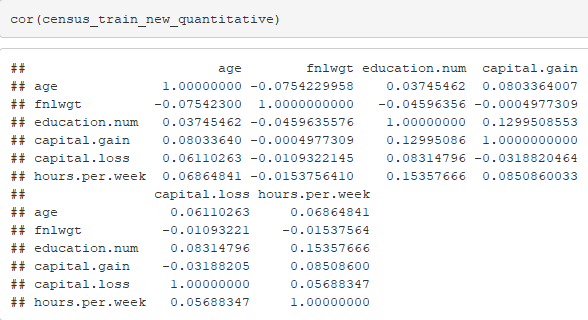




For native.country, almost all people’s native.country is United-Staste. So I will combine other countries into a new label“other country”

1551912419(1)

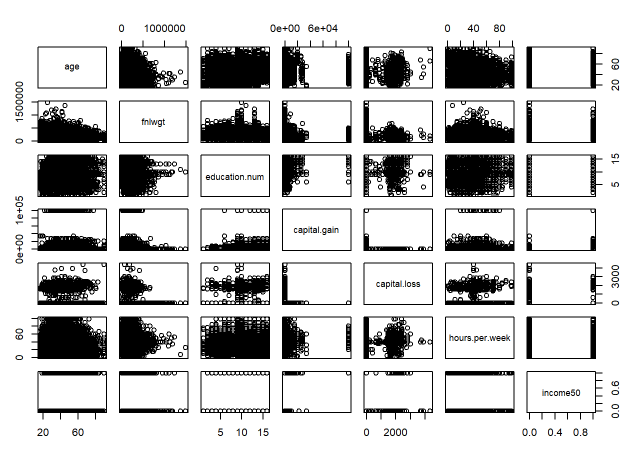
Check collinearity to find if there is some relationships between quantitative variables.



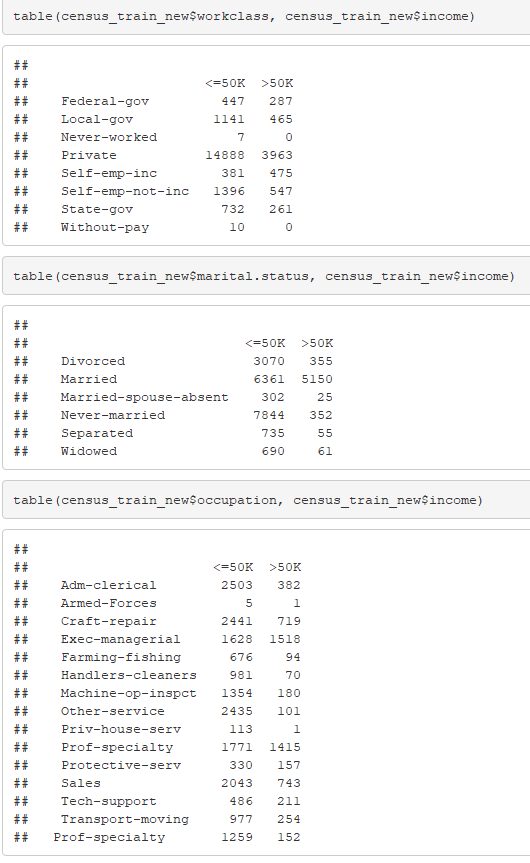
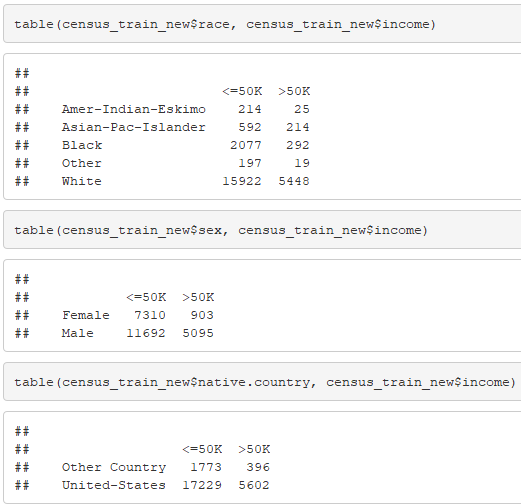
Most variables do not have strong collinearity, but the collinearity values of education.num-capital.gain and education.num-hours.per.week is higher than 0.1. According to ISLR, I can delete some variables which have collinearity. But,I don’t want to delete education.num-hours.per.week. Because I just choose education.num rather than education, and hours.per.week is obviously related to income. How much time you work, how much money you can get.

check variables and income

Find whether there is some variable can determinate income directly or has some determinative role.



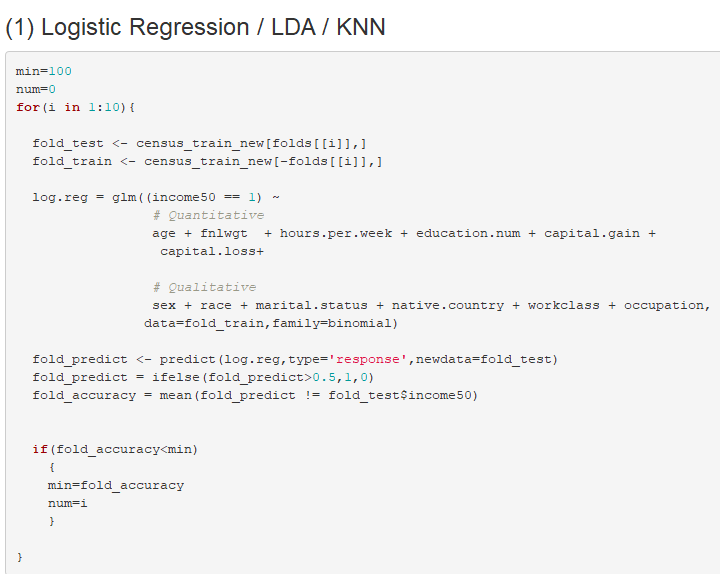
No quantitative variable is the determinative role of income.



No qualitative variable has the determinative role of income.

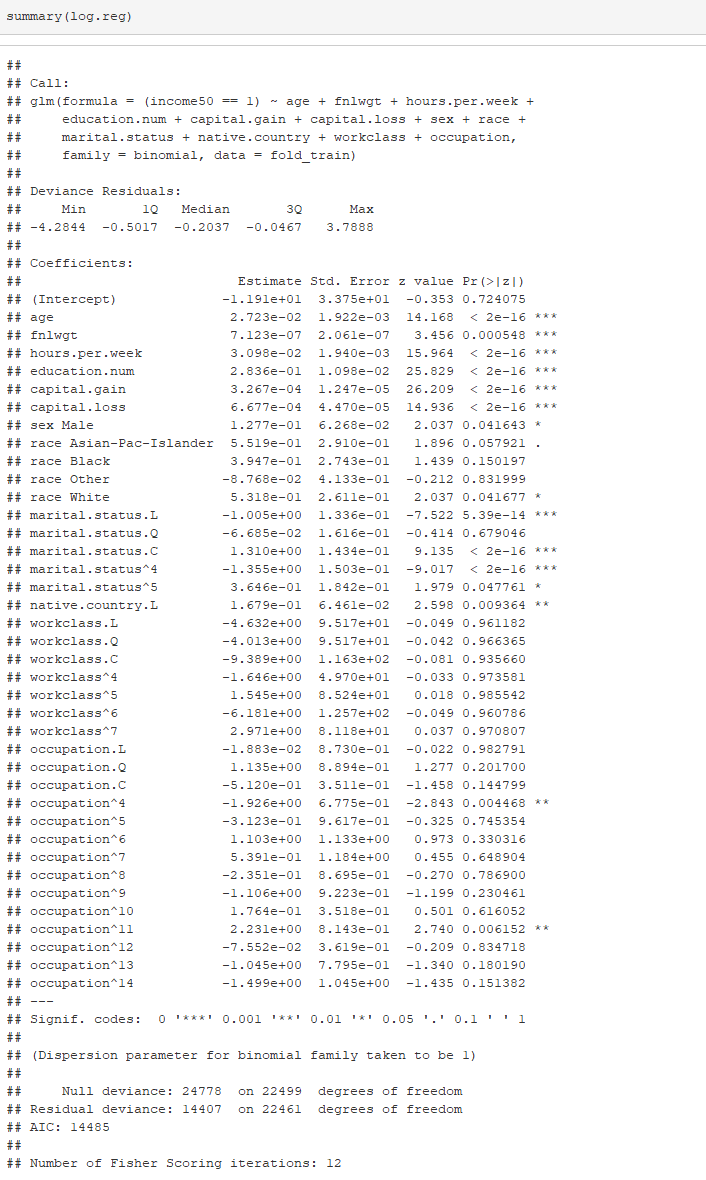
1. **Creating models**

1551913217(1)In this part, I will use 10 fold cross-validation to find a better model.



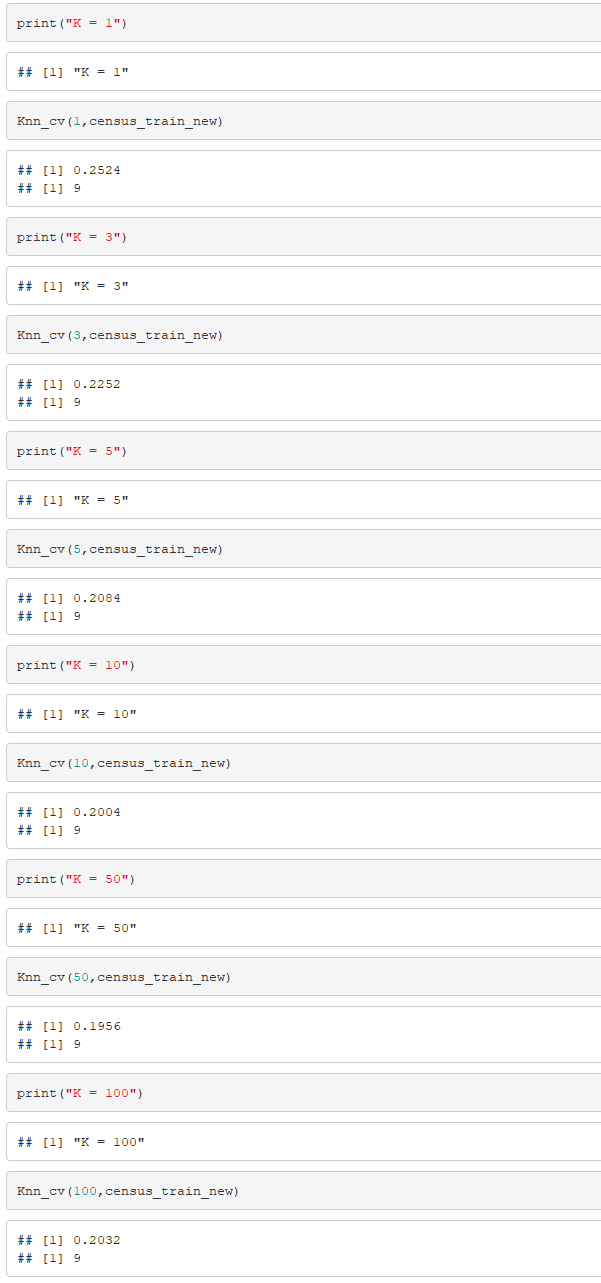
1551913320(1)

Well, the original logistic regression model ‘s performance is not bad.



 Little worse than logistic regression

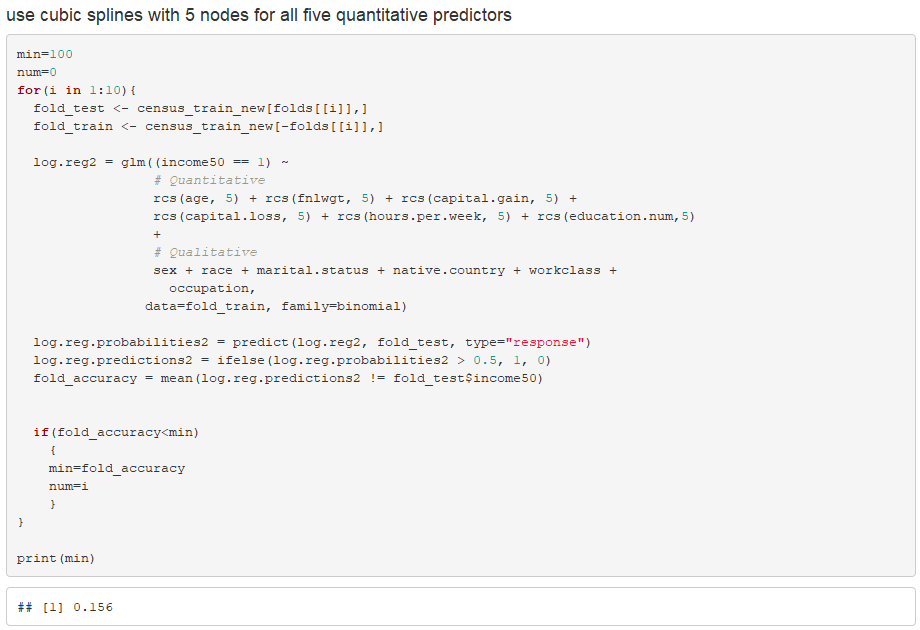


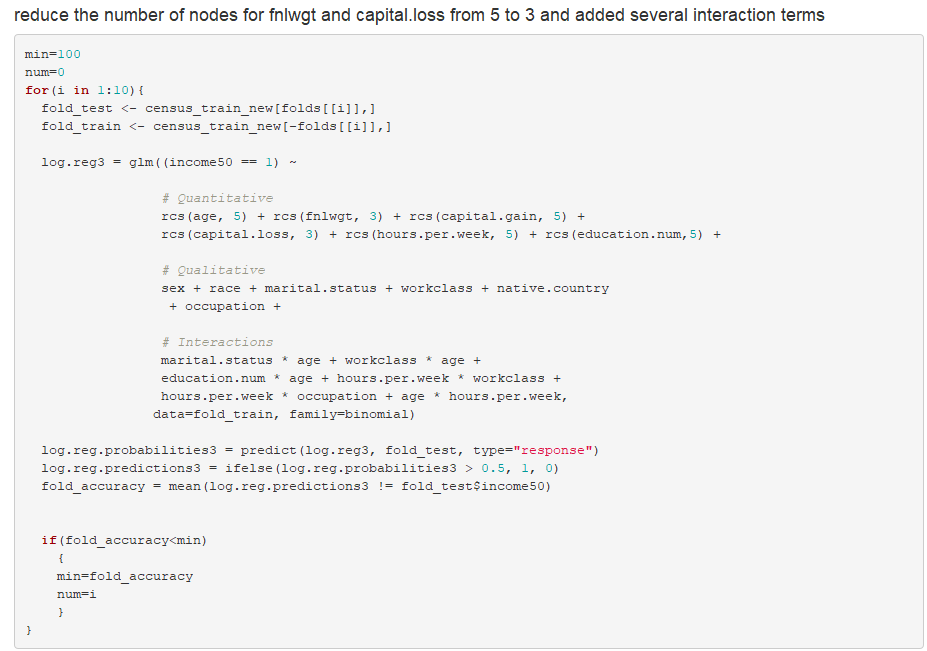


With K increase, the error rate is decrease, but if K keeps increasing, the error rate will increase. So I guess there always is a threshold value of K in KNN model.

The LDA and KNN both performance worse than logistic. We decided to go with the logistic regression and fine-tune it.

(2)Adjusting Logistic Regression model /improve models’ performance

The model performances worse before removing capital.loss and capital.gain, so I will keep them.



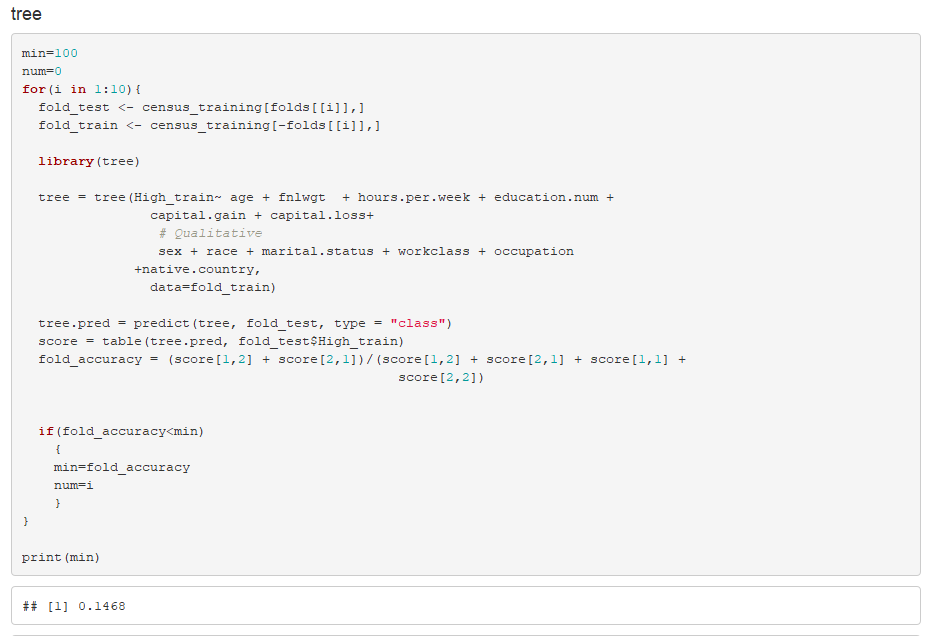
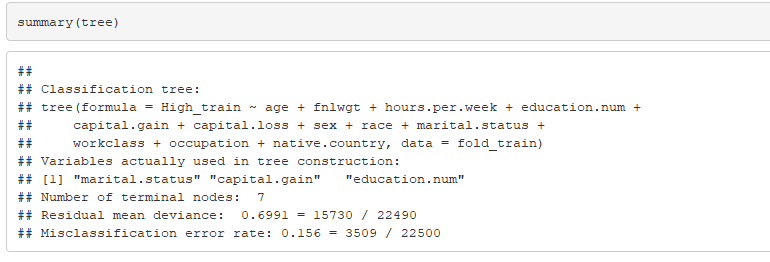




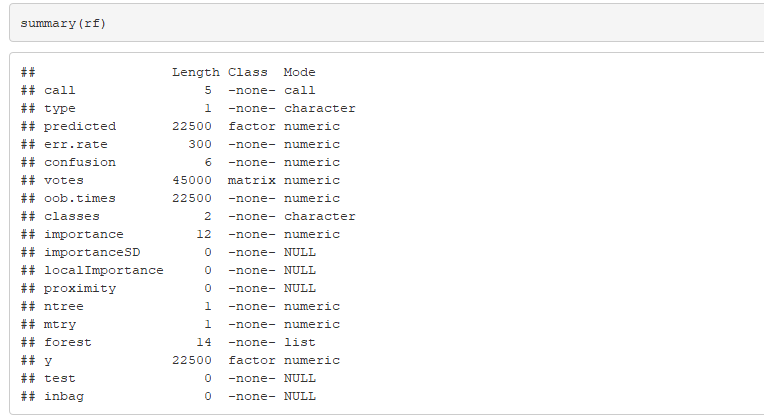
1551914250(1)

After trying these model, the performance still isn’t satisfactory. I’d like to try tree.

(3)Tree / RandomForest

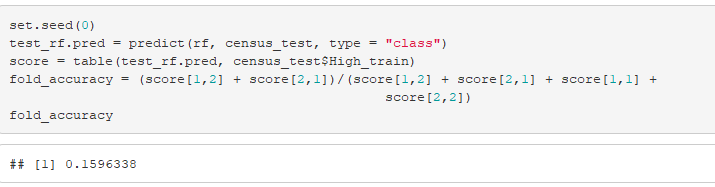
Decision Tree performances better than LDA and KNN, but still worse than Logistic Regression

1551914948(1)



According to all models, I think the best one is randomForest, I will choose this one to do test.

**3. Test Data**

Do the same reprocessing as training data, and use randomForest to do the test.

The final error rate is around 0.16. Not bad!

After doing this midterm project, I think preprocessing is very important. We have to view the data by eyes firstly. If we input the data directly, some missing data won’t show off. That may cause big problem in fitting data and testing data.

And I think I should do more data processing, I should continue merging some variables.

Although the best model in my project is randomForest, Logistic Regression also performances well. Logistic Regression is a basic classifier, but it is still very useful, and Logistic Regression can do many extensions.