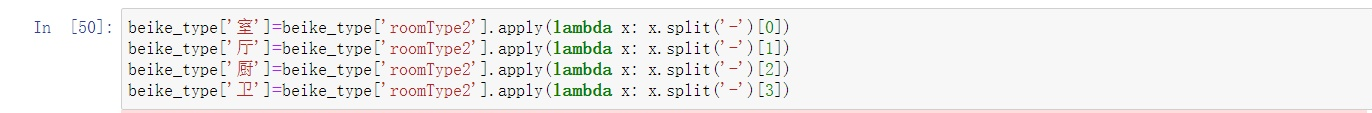
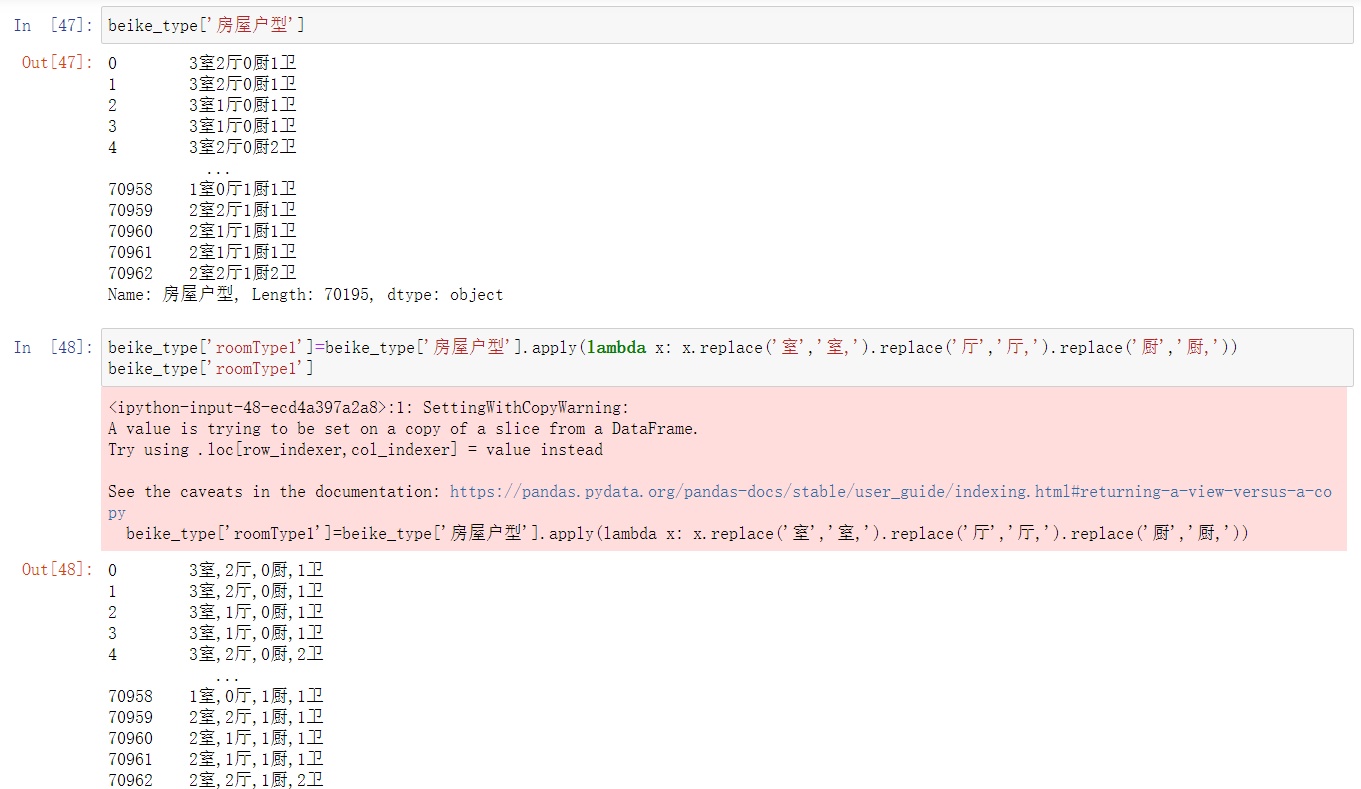


Since the number of bedrooms, living rooms, kitchens and bathrooms might influence people's choice of buying a house, we used Scikit-Learn to write a program in order to give them references based on the information about those sold second-hand houses in Guangzhou.

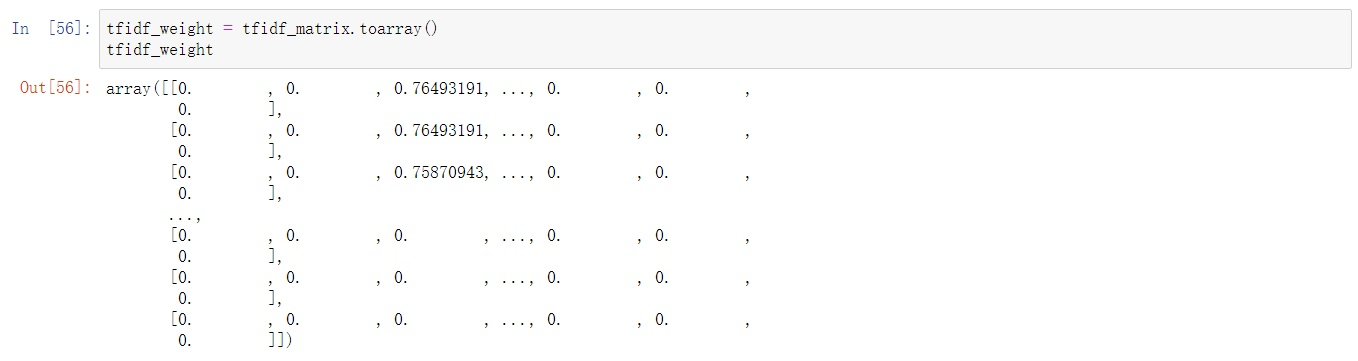
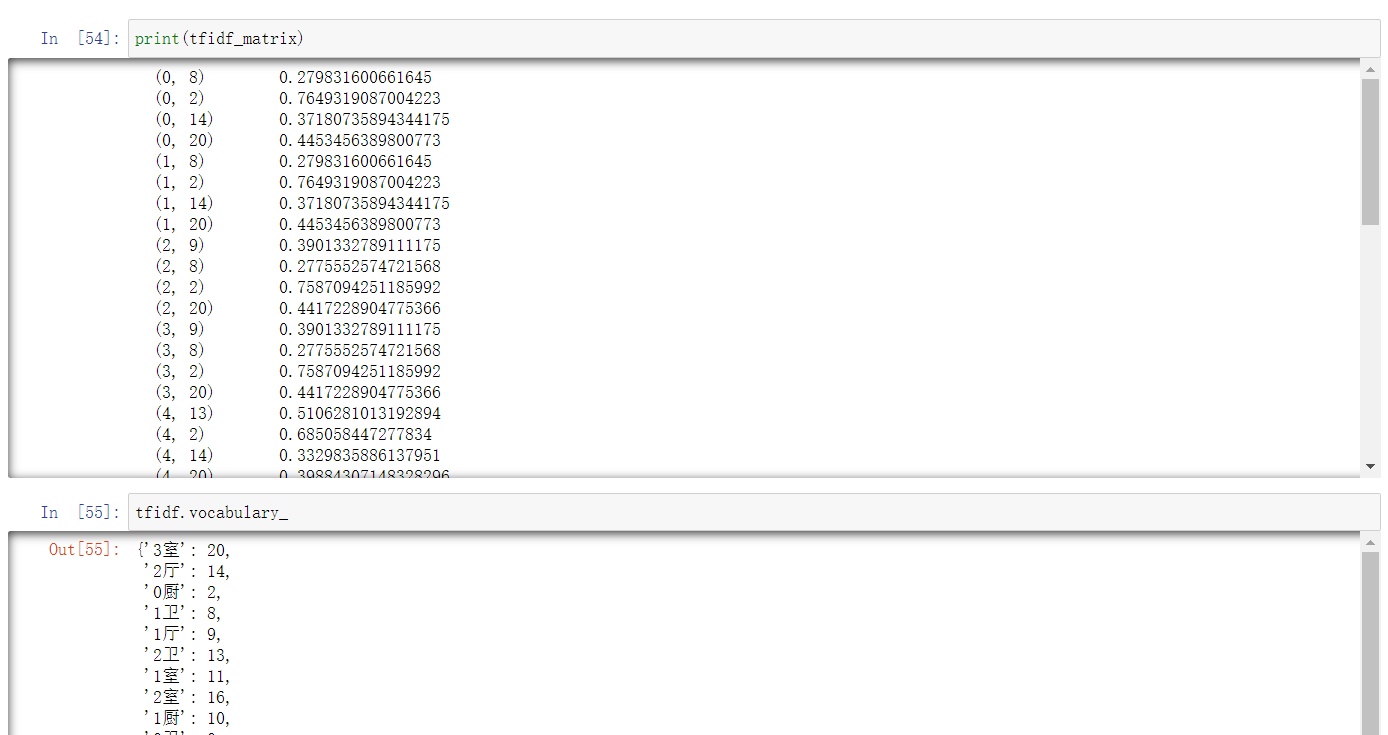
The first thing we should do is Data Cleaning. In data.csv, we can see that we also get the information about garages that have been sold, so we use beike[beike['房屋户型']!='车位'] to drop those datas. And we store them as beike\_type.



After that, we need to format '房屋户型' column by applying lambda x to add a column named 'roomType1'.

'roomType2' is another form of '房屋户型'.

Through these commands, we'll get a clearer dataframe.

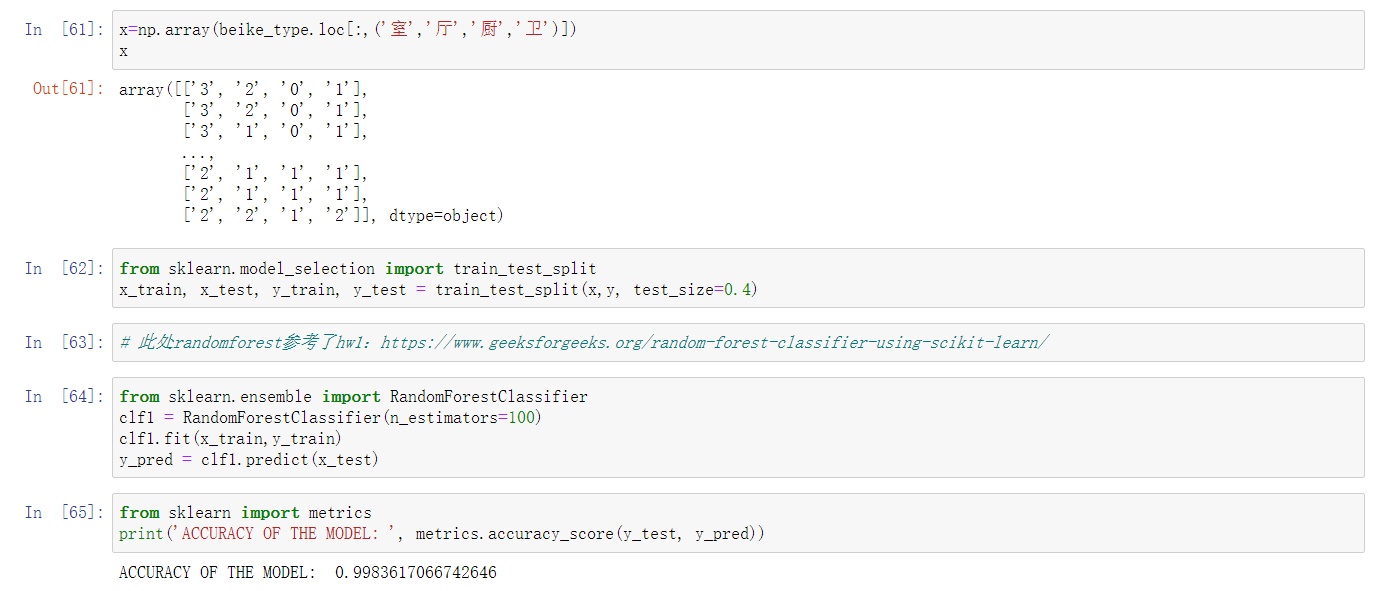


Before using RandomForestClassifier, we used TfidfVectorizer and K-Means to give each transaction a label.

TF-IDF is the abbreviation for 'Term Frequency-Inverse Document Frequency'. We can get the weight of each word by transforming 'roomType1'.

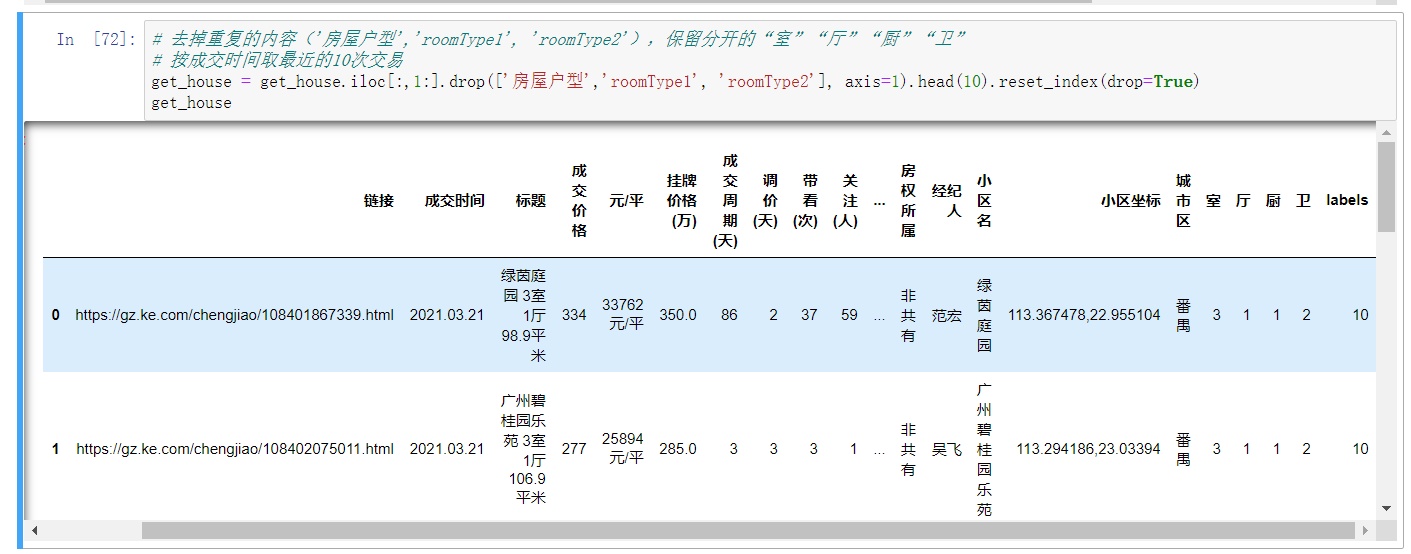
We can see the number that represents each word by using vocabulary\_.

Then turn the matrix to array to put into K-Means to fit a model.



The command '.intertia\_' is the way to let us know whether the model is proper. The lower the value, the better the fitting.

Add the label we got to the dataframe, and extract the values in 'labels' and change them to array 'y'.



Get the amount of each room in an array and assign it to x. After spliting x and y into x\_train, x\_test, y\_train and y\_test, we can use RandomForestClassifier to fit another model, then check the accuracy of the model.

Finally, we, just pretend we are those who want to buy a second-hand house in Guangzhou, can input a (the number of bedrooms), b (the number of living rooms), c (the number of kitchens) and d (the number of bathrooms), can get the information of those houses that have traded in the past.

Sort by the deal time, drop the raws that don't contain the

url of the house sale and sort the index of the dataframe, we can get last ten transactions of second-hand house sale in Guangzhou. With that, we can find out how much money is needed to buy the second-hand house with the house types that meet our expectations, and what the configuration of those houses might be in order to make a better choice of future house.