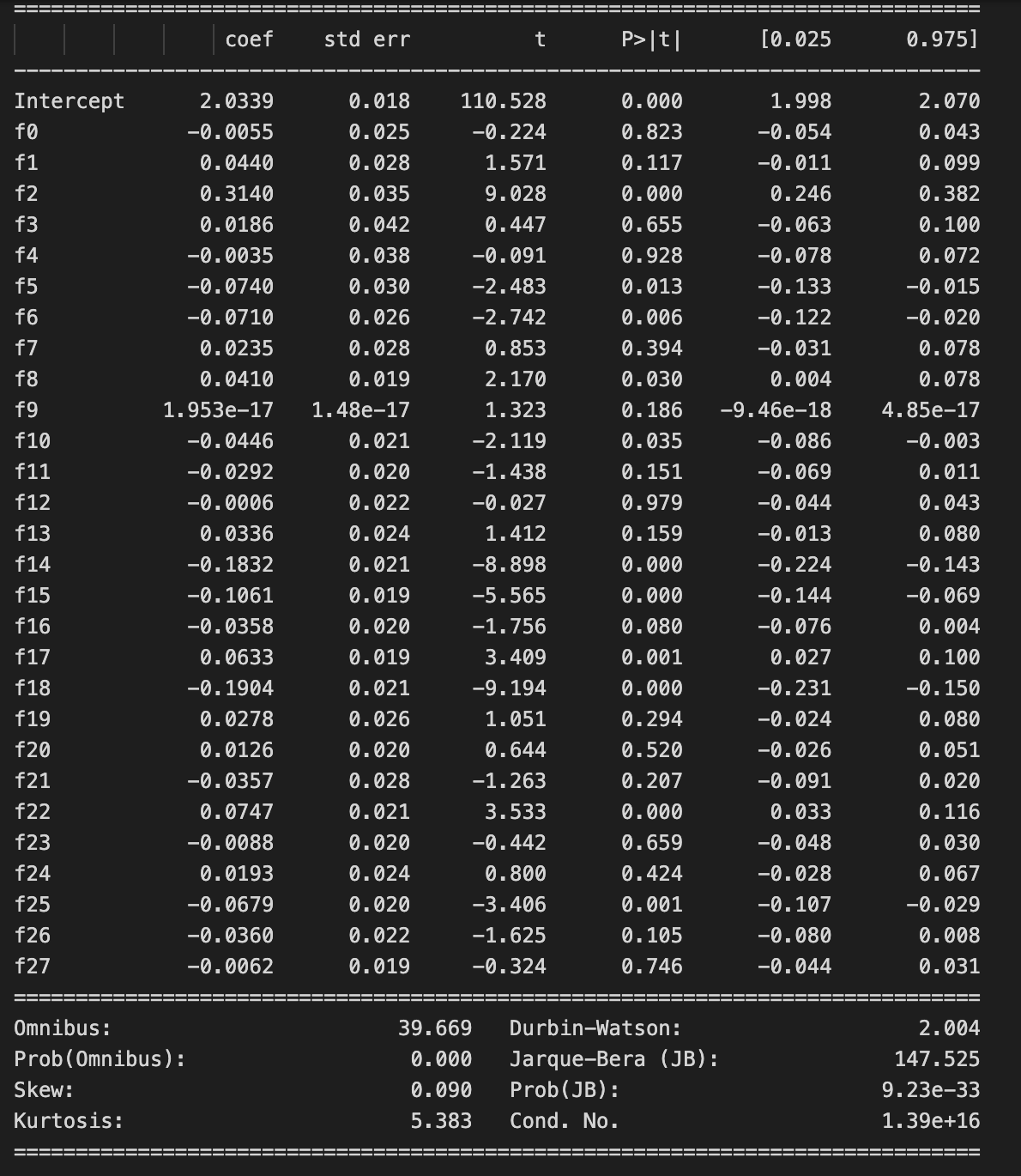
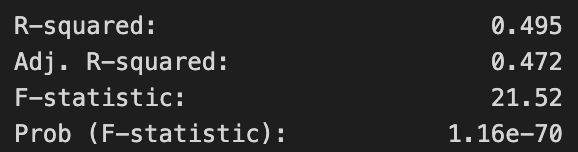
資管碩一 R10725026黃奕滔

1.

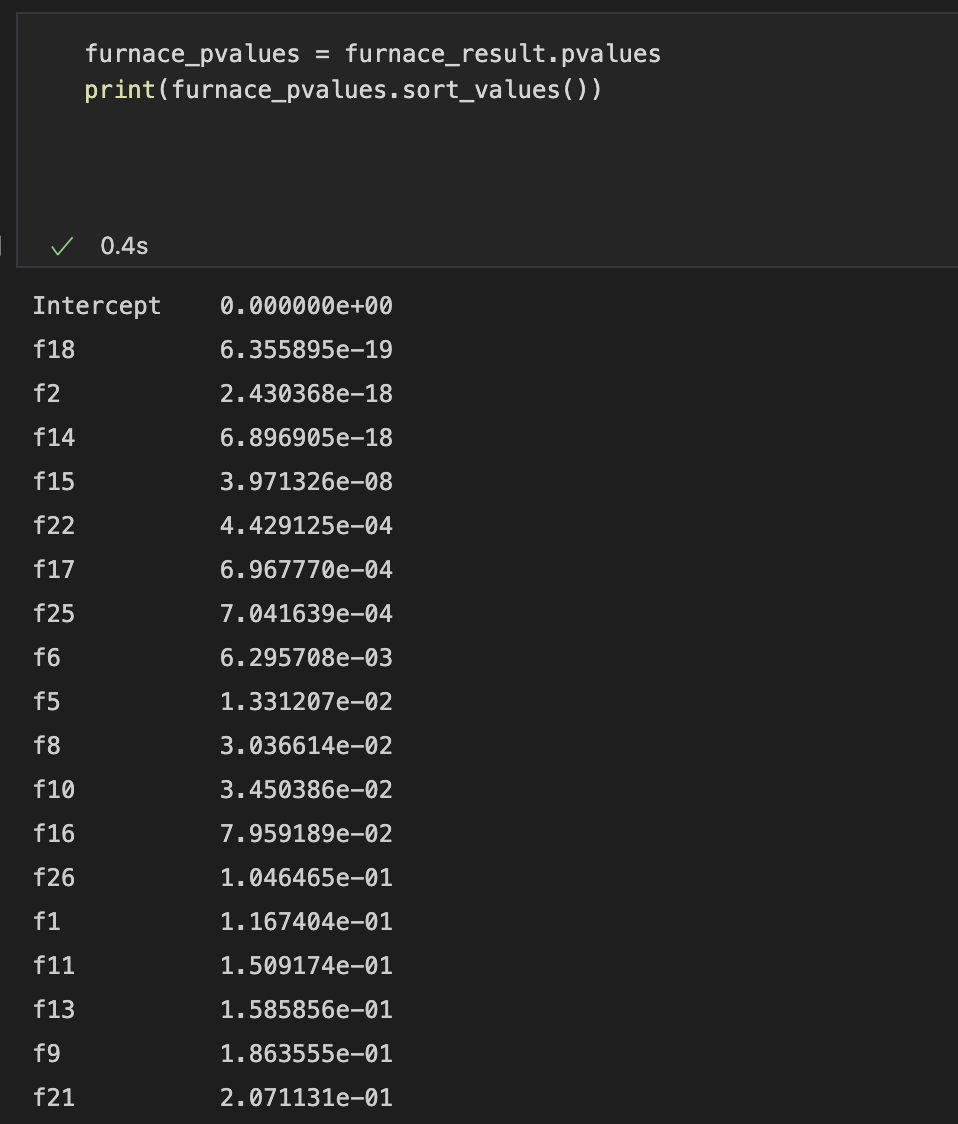
(a).

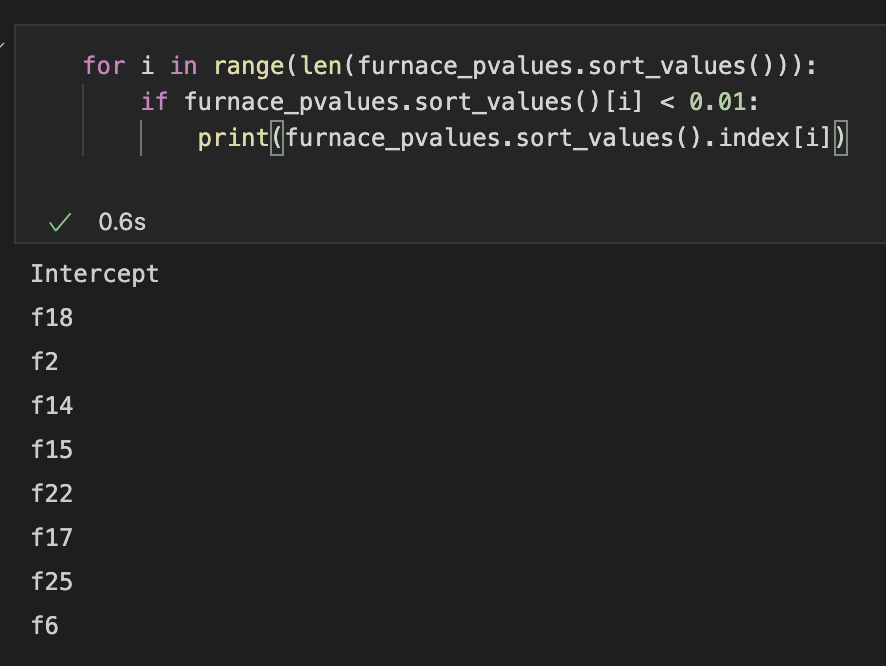




(b). Before adapting linear regression, we should check whether there truly exist linear relationships between features and labels. Maybe there are some interactions among the features or higher degree polynomial terms.

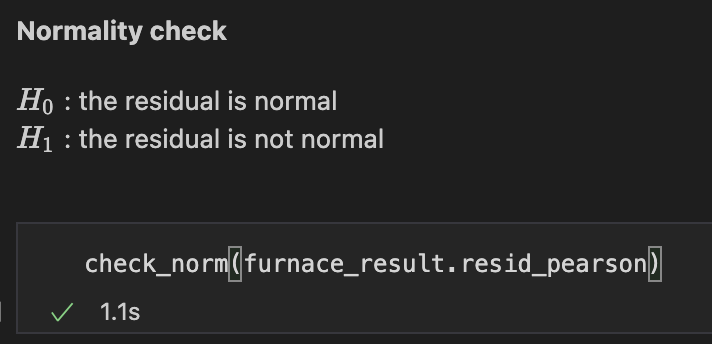
(c). Sorted p-values and significant features

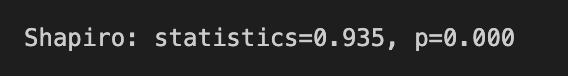




(d).

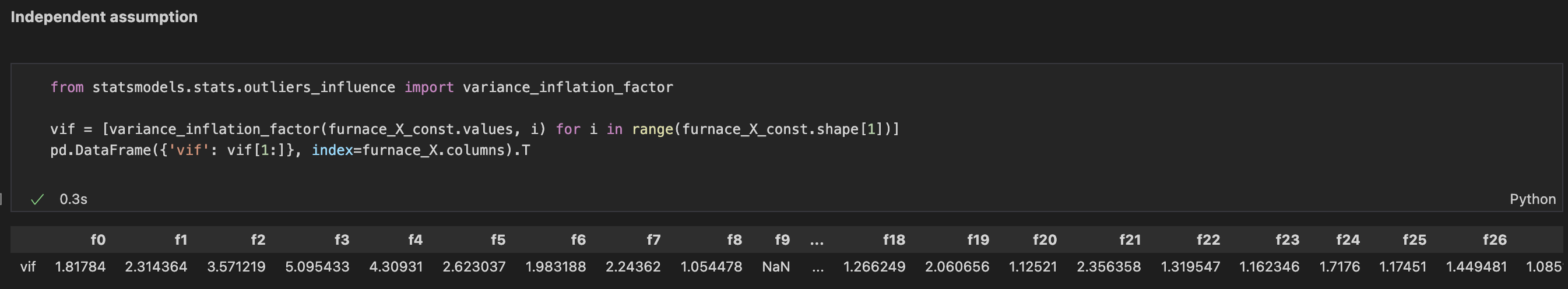
(1). Normality test





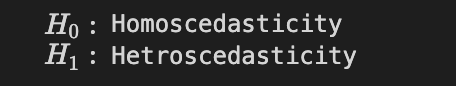
P-value of the Shapiro Normality test < 0.05, we have 95% confident to reject null hypothesis. Thus, residual distribution isn’t normal.

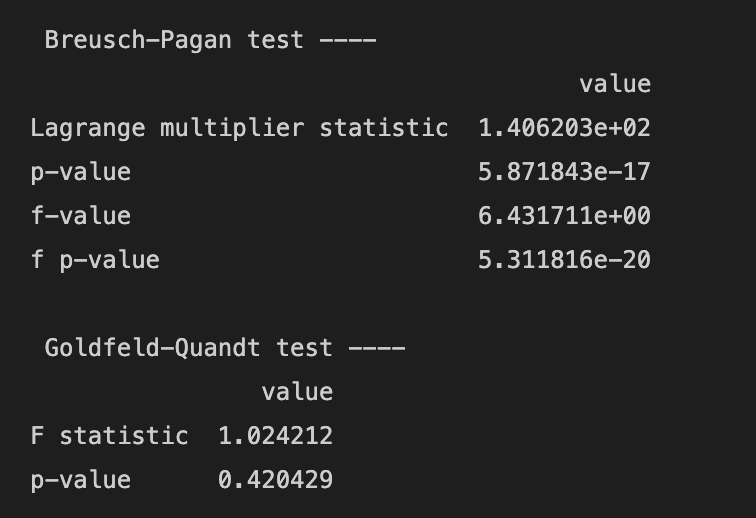
(2). Independence (aka check multicollinearity)



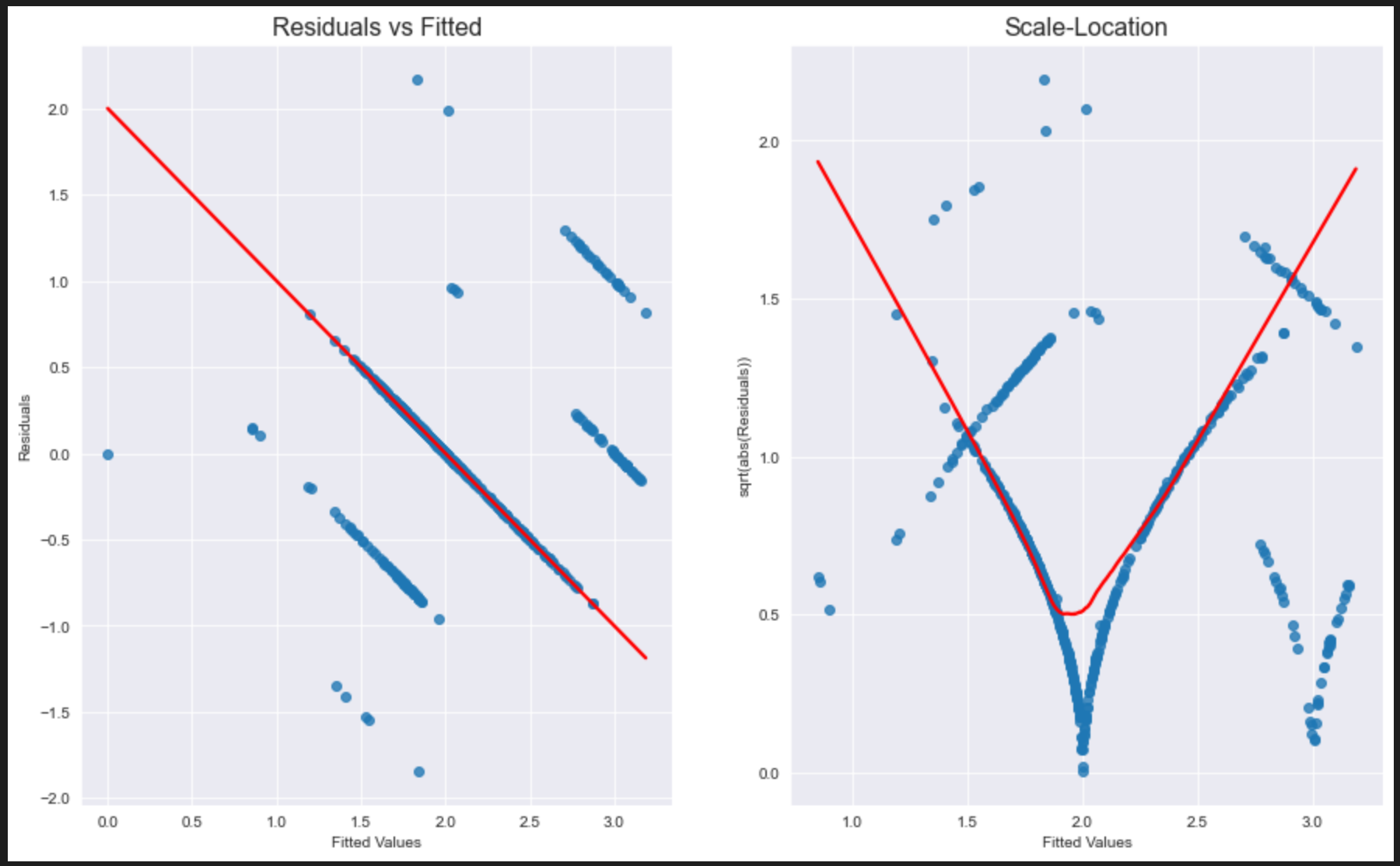
Through checking VIF (Variance Inflation Factor), there’s no strong multicollinearity features (VIF > 10) that must be removed.

(3). Homogeneity of Variance (aka Homoscedasticity)



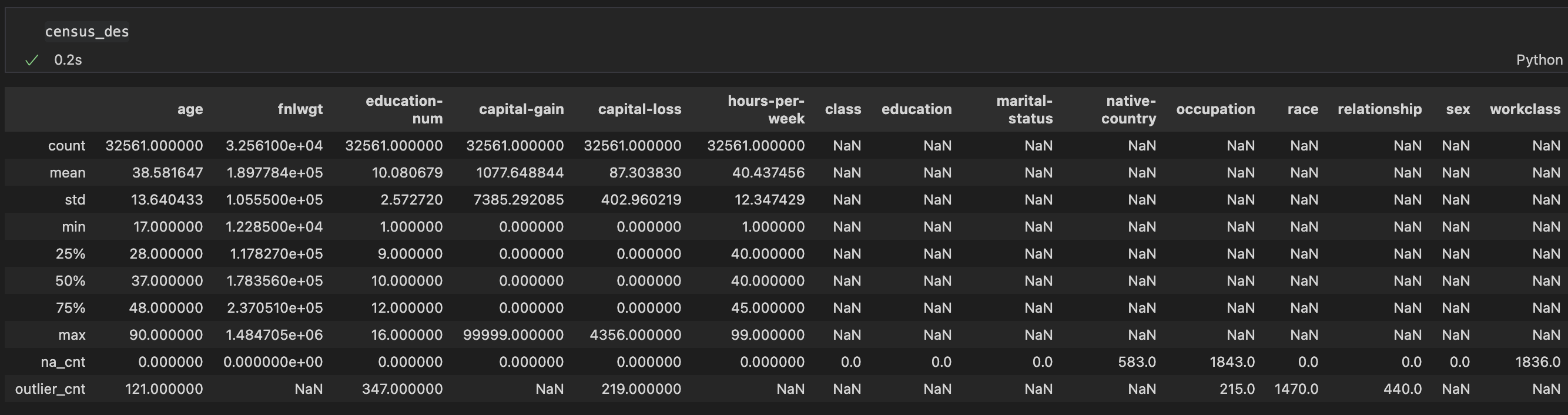


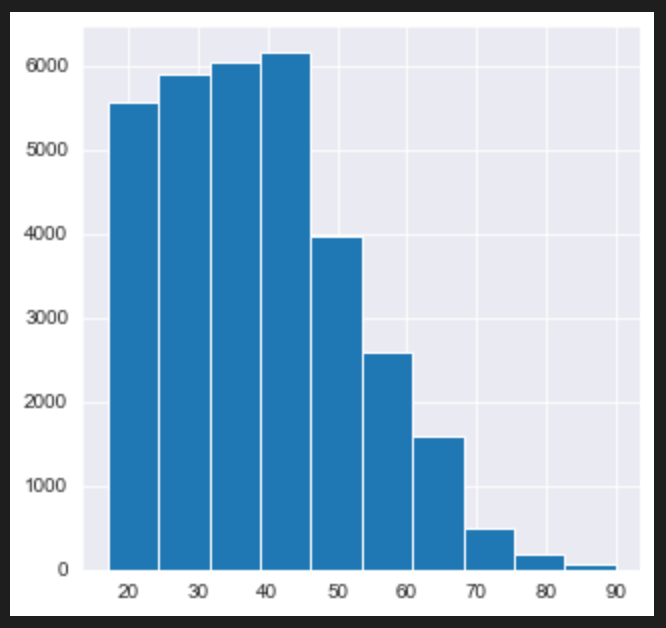
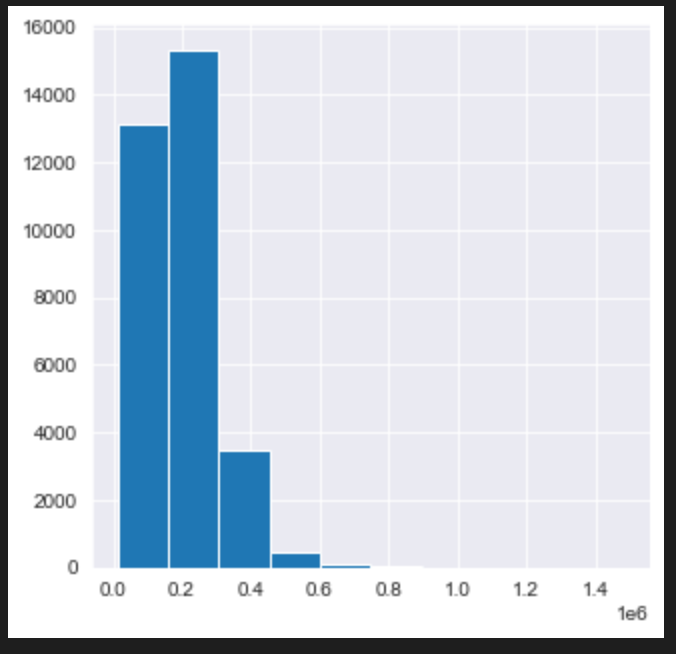
B-P test reject null hypothesis, while G-Q test doesn’t. We can’t surely infer whether Homogeneity of Variance exist or not.

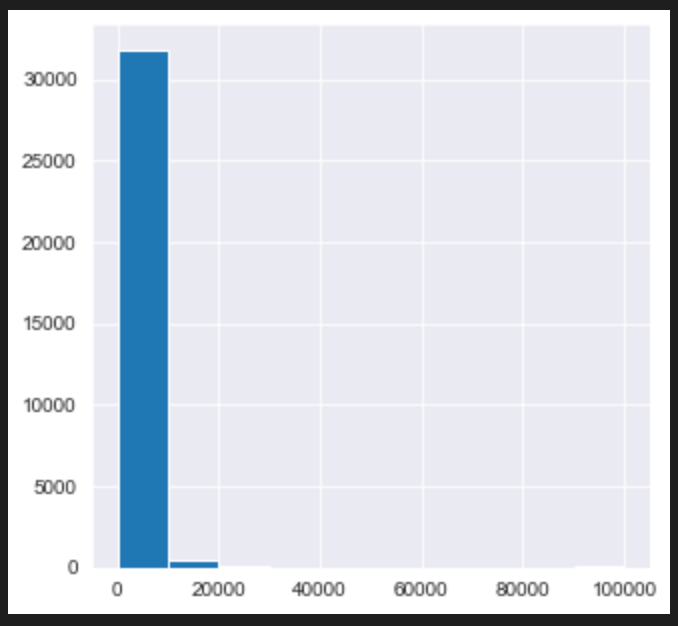
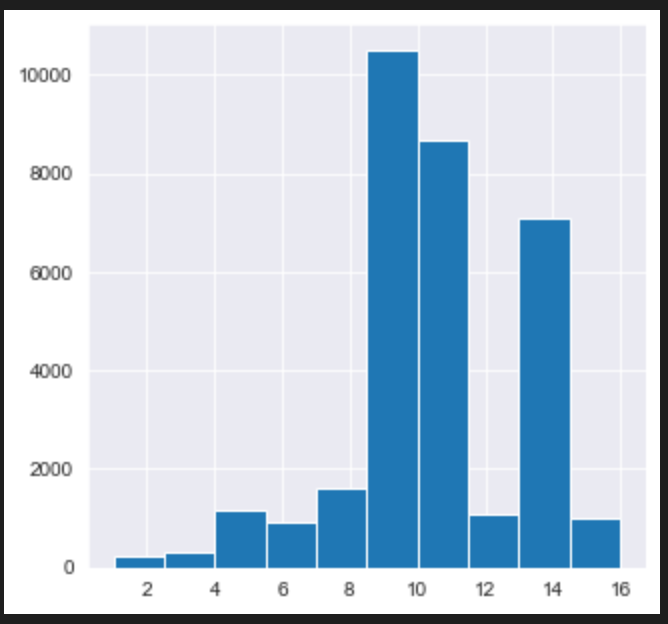


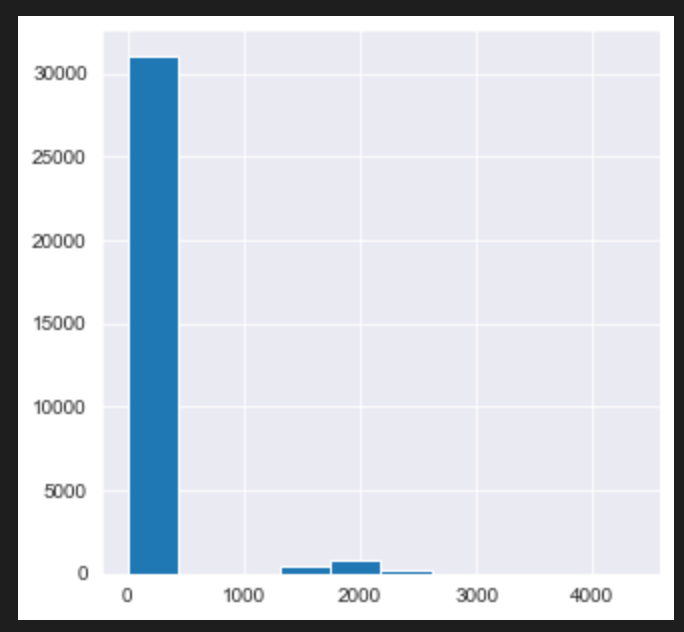
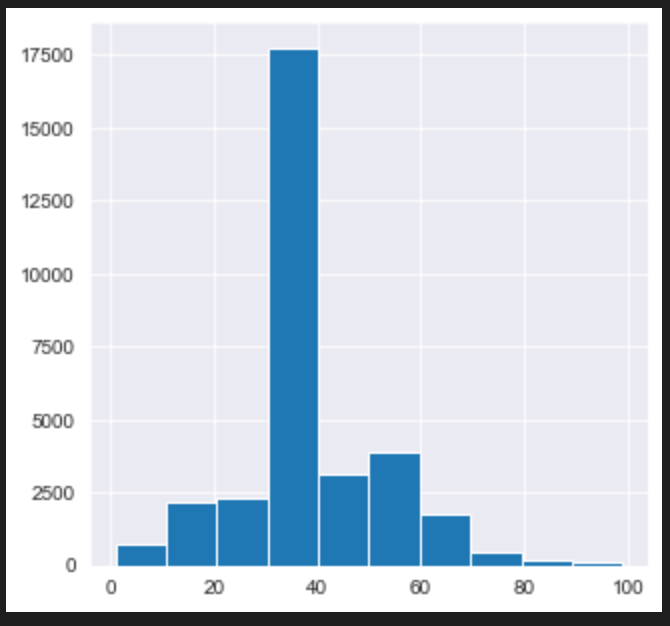
2.

(1). Distribution plot is posed by column number order.

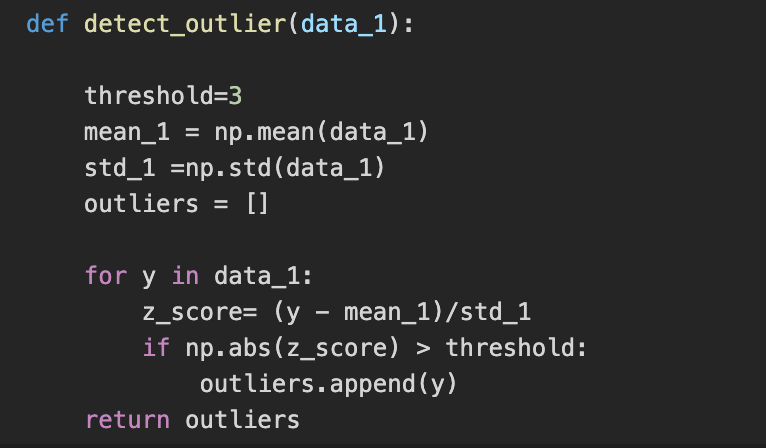


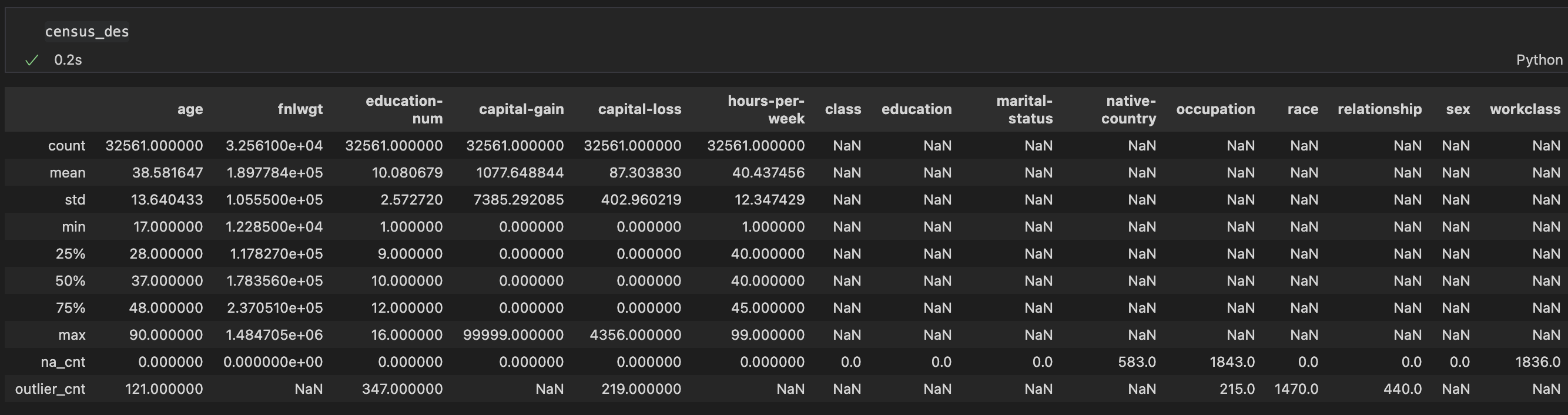
 



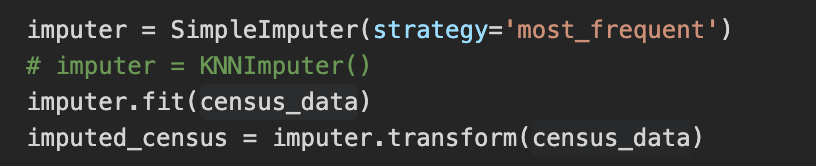
 

(2). Using z-score : If |z-score| > threshold, the value is outlier. (thres. is set to 3)



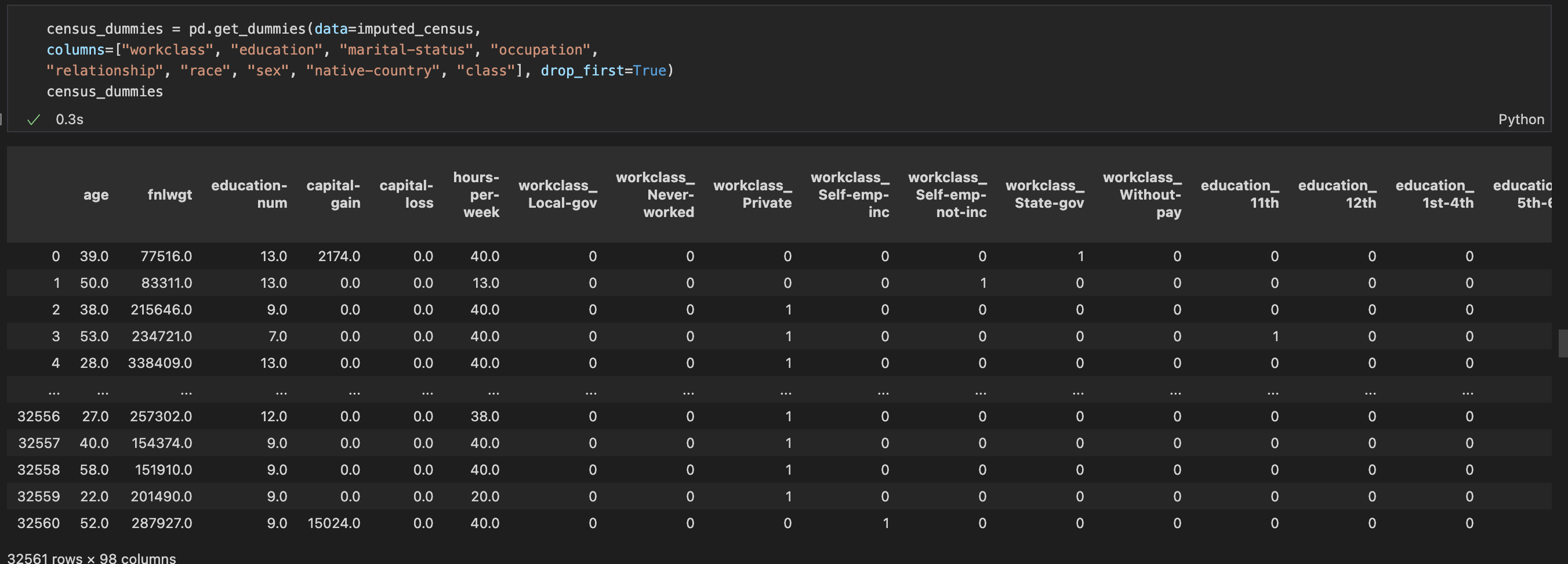


Take most-frequent value to replace nan (impute the missing values)

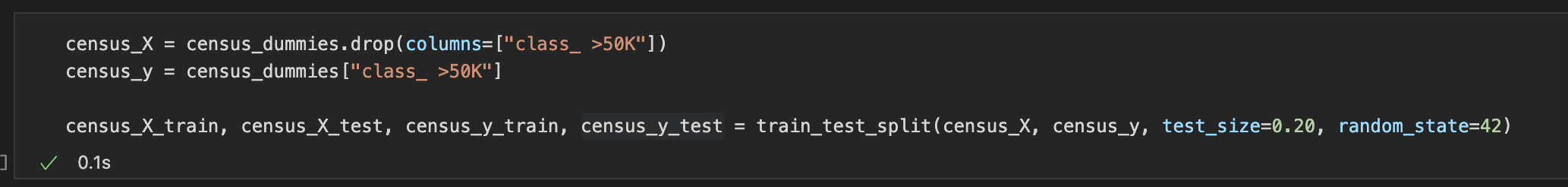


(3). We use the pandas.get\_dummies() method to transform the categorical columns into dummy one. Because the category columns in the dataset is mostly binary, we can use drop\_first=True option to make appropriate dummies.

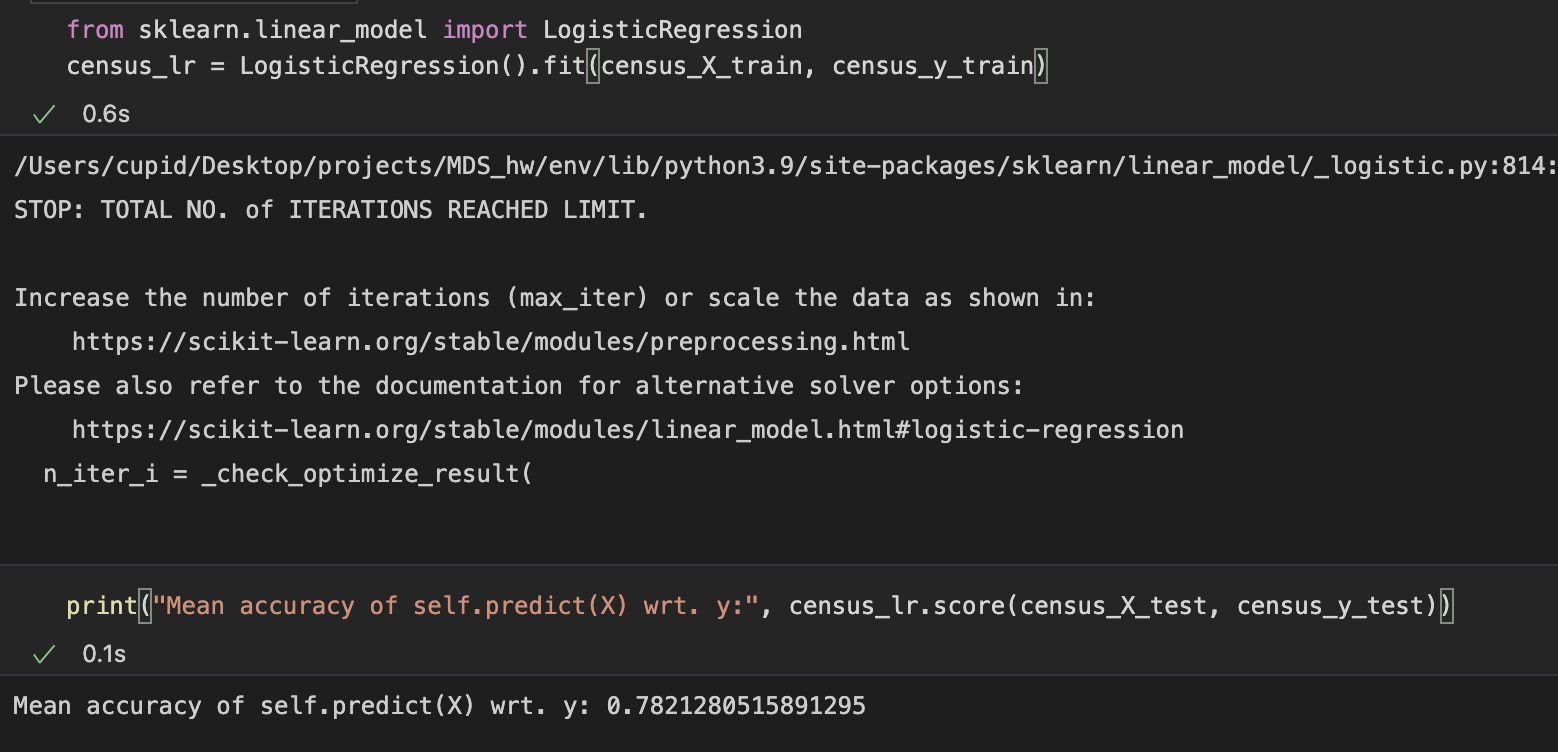
Ex.



(4). We can use the sklean.model\_selection.train\_test\_split() method. With chosen random\_state option, we can perform randomly split. Meanwhile, set the proportion of train/test dataset by test\_size option.

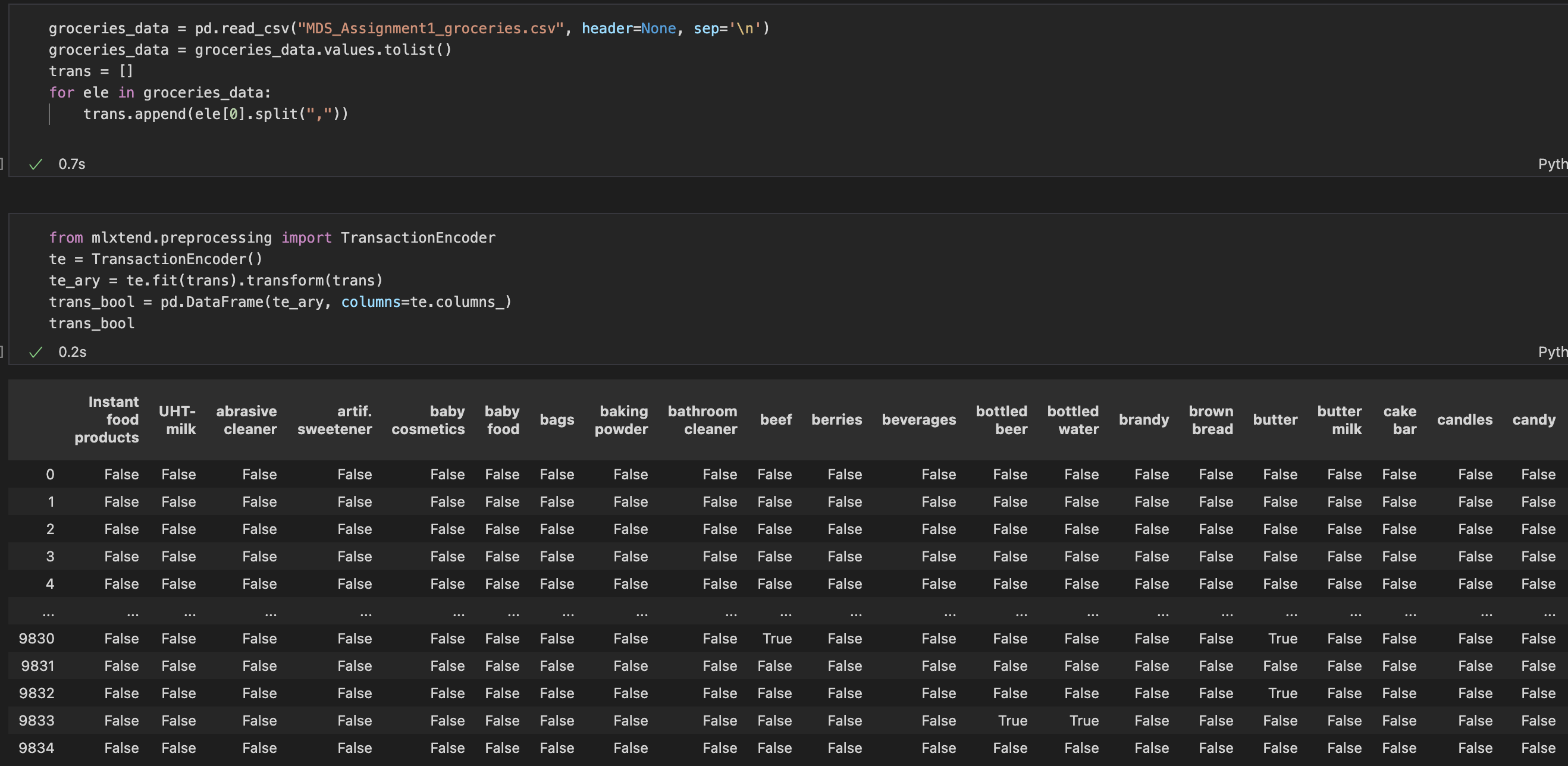


(5). Fit the logistic regression model with Xy\_train and evaluate model accuracy with score method using Xy\_test.

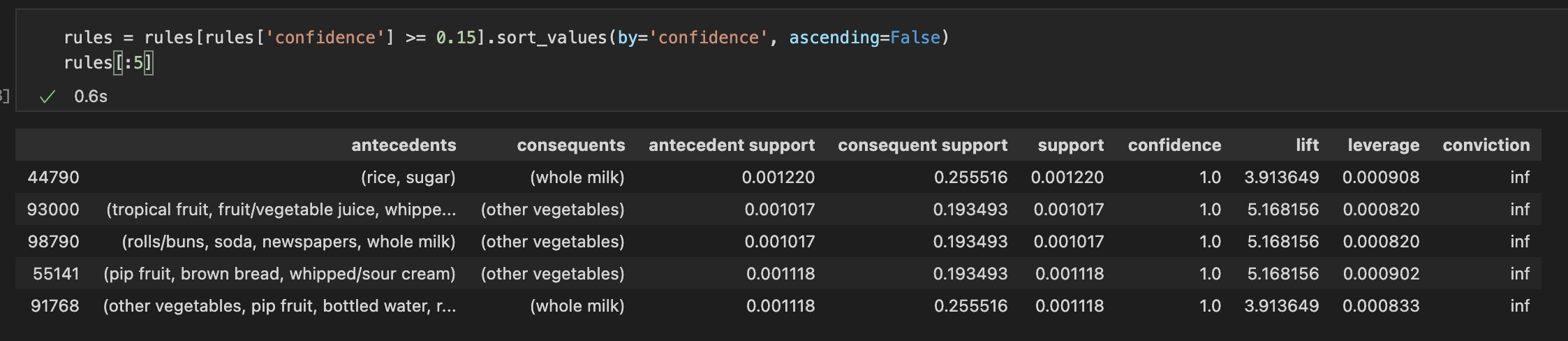


3.

(1). There’s a package named mlxtend with a method called TransactionEncoder(), which can help us transform the transaction data into Boolean table form.



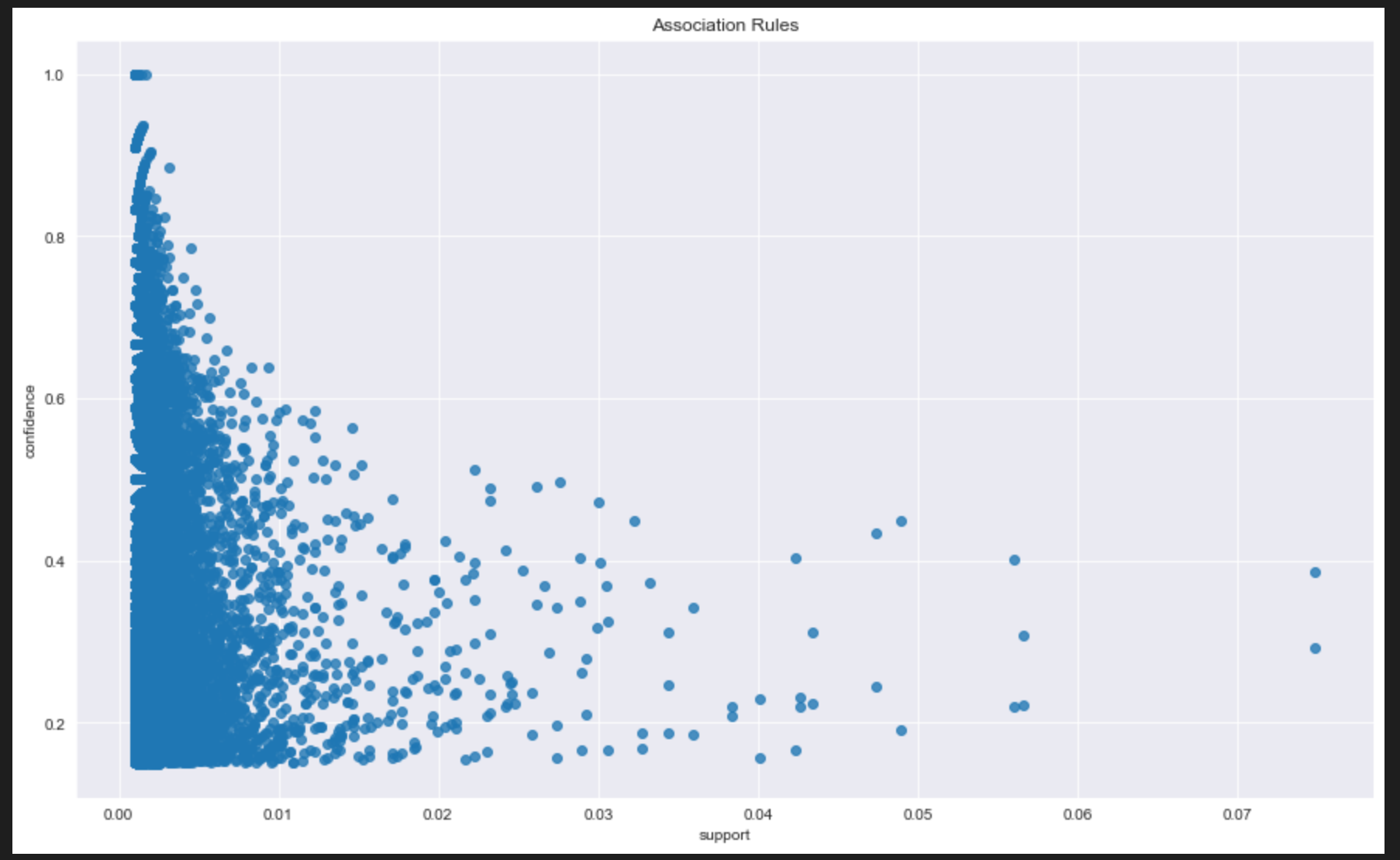
(2).



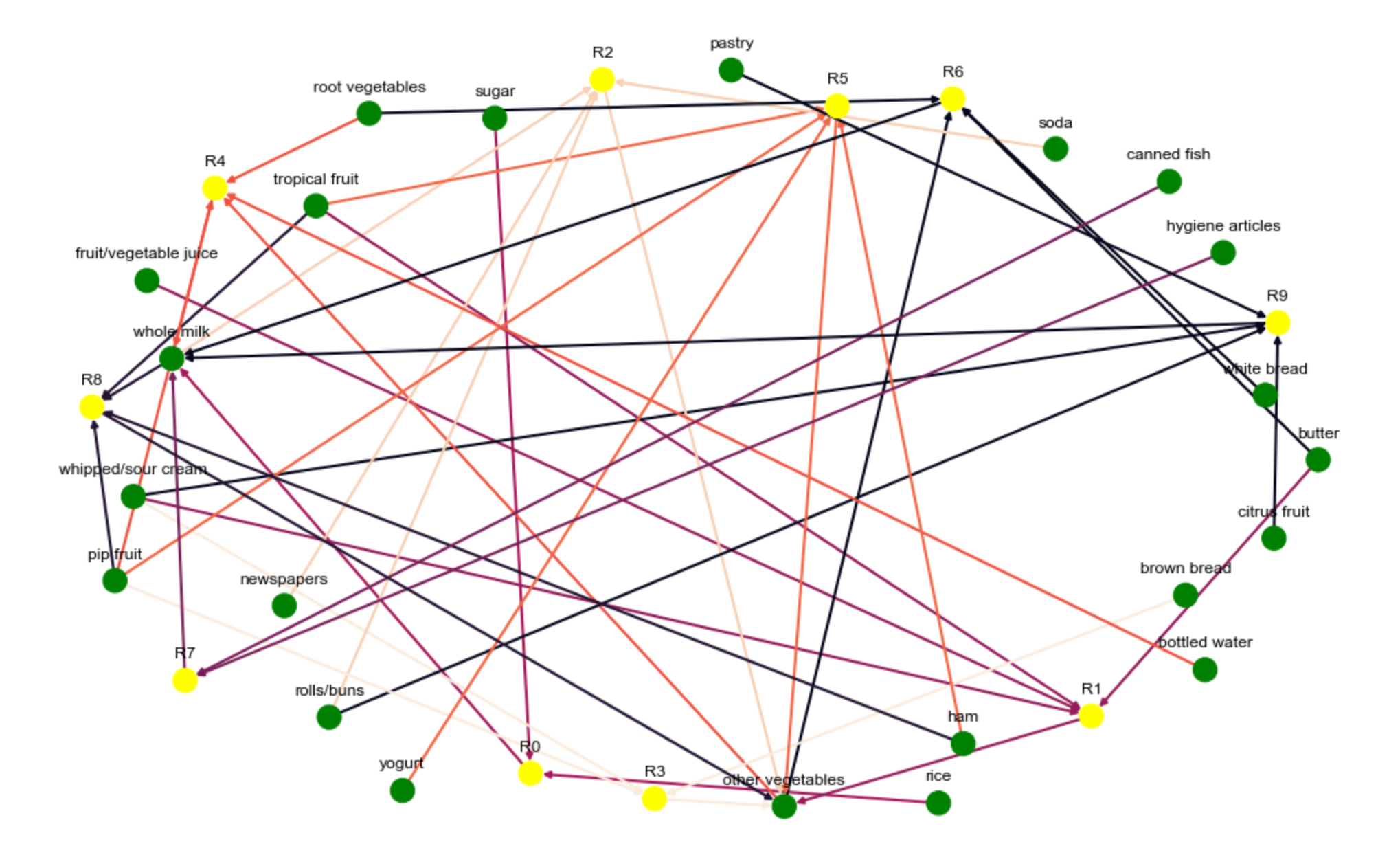
(3). As I see it, row 93000 is the most interpretable row in the first five rules.

It’s reasonable that people awarded of the important of health to consume balanced amount of vegetables, fruits, protein and fat. Due to the aforementioned reason, when someone buys fruit, juice, hopefully will also buy vegetables to fulfill daily demand of nutrients.

(4). Relationship between confidence and support:



10 of the rules connections:



Ref.

linear regression 回歸模型 and 檢測

<https://towardsdatascience.com/verifying-the-assumptions-of-linear-regression-in-python-and-r-f4cd2907d4c0>

outlier detection

<https://medium.com/datadriveninvestor/finding-outliers-in-dataset-using-python-efc3fce6ce32>

impute data (Compensate missing value)

<https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779>

association rules

<http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/>

<https://artsdatascience.wordpress.com/2019/12/10/python-%E5%AF%A6%E6%88%B0%E7%AF%87%EF%BC%9Aapriori-algorithm/>

<https://pbpython.com/market-basket-analysis.html>

(association rules visualization)

<https://intelligentonlinetools.com/blog/2018/02/10/how-to-create-data-visualization-for-association-rules-in-data-mining/>