IDS Final Report

산업경영공학과 201801158 김 찬

**INDEX**

1. Objective

2. Data Introduction & Processing

- Introduction

- Data Processing

- Data Visualizing

3. Analysis Techniques

- Clustering

- Association Rule

- Classification Decision Tree

- Regression Decision Tree

- Multiple linear Regression

4. The explanations about Feedback

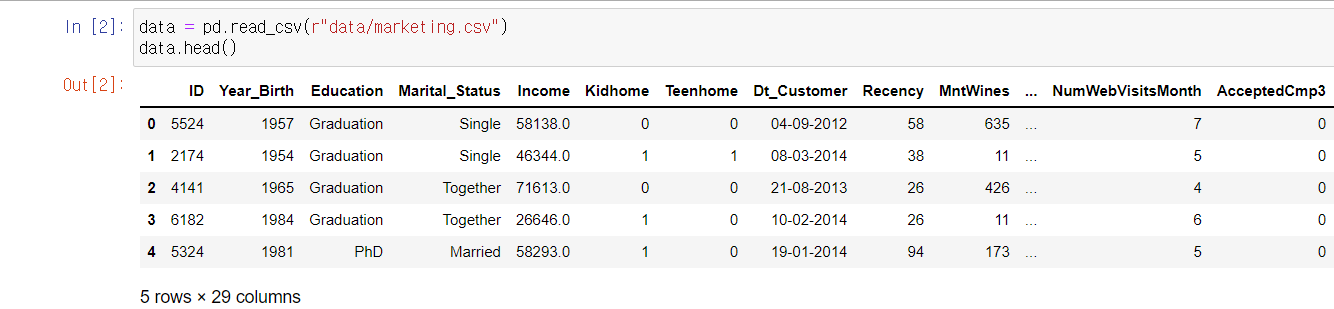
5. Conclusion

6. Reference

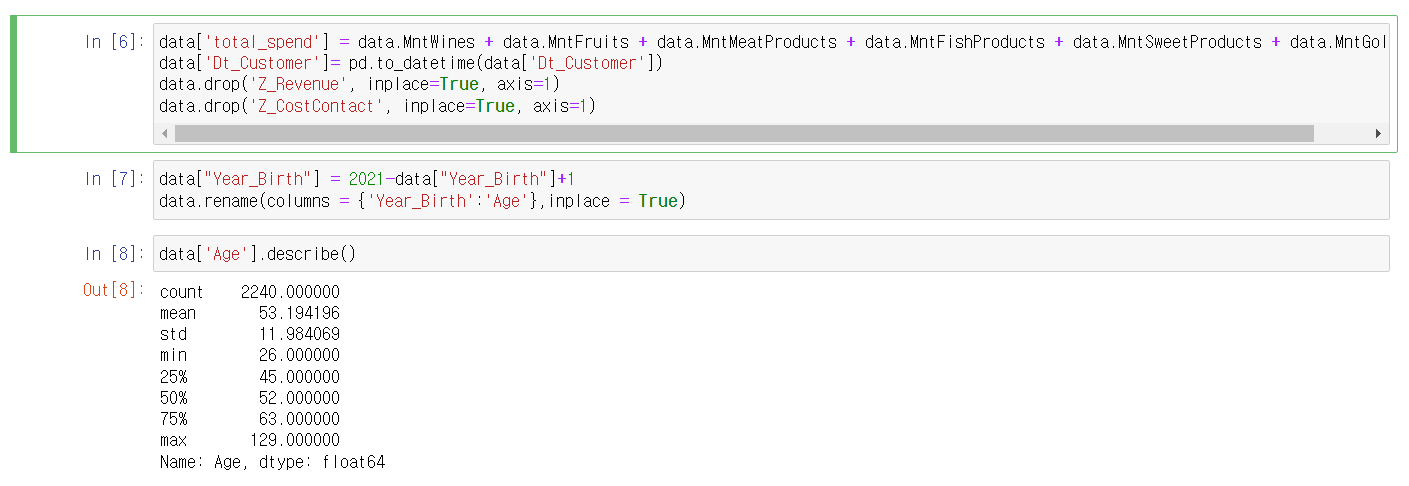
1. **Objective**

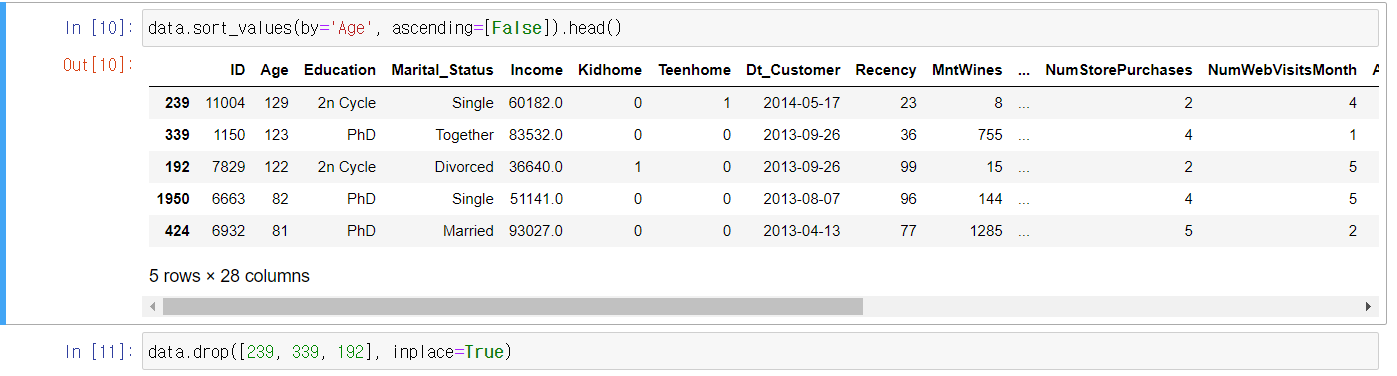
We wanted to analyze the product purchase and group by customer consumption pattern, predict the consumption, establish marketing strategies from this data set. this data set is told us “The information about many customer”, included “the product that customers purchased”, “education”, “age” etc

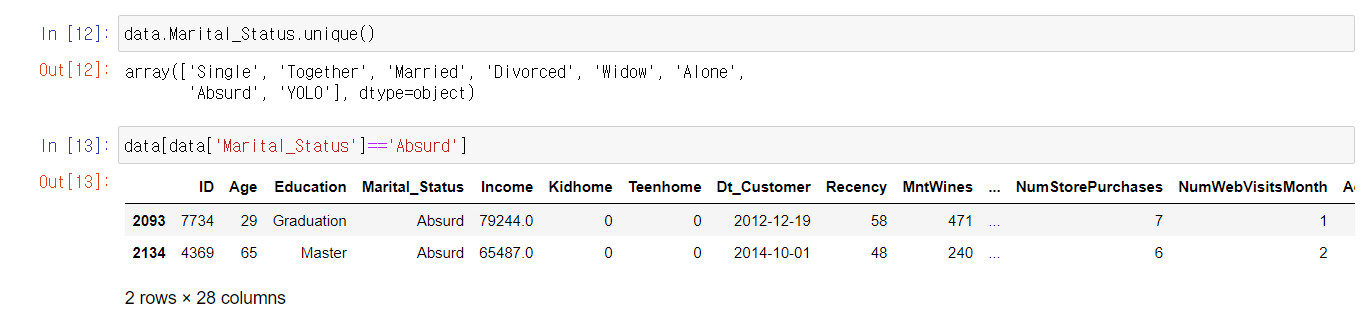
텍스트, 모니터, 스크린샷, 화면이(가) 표시된 사진

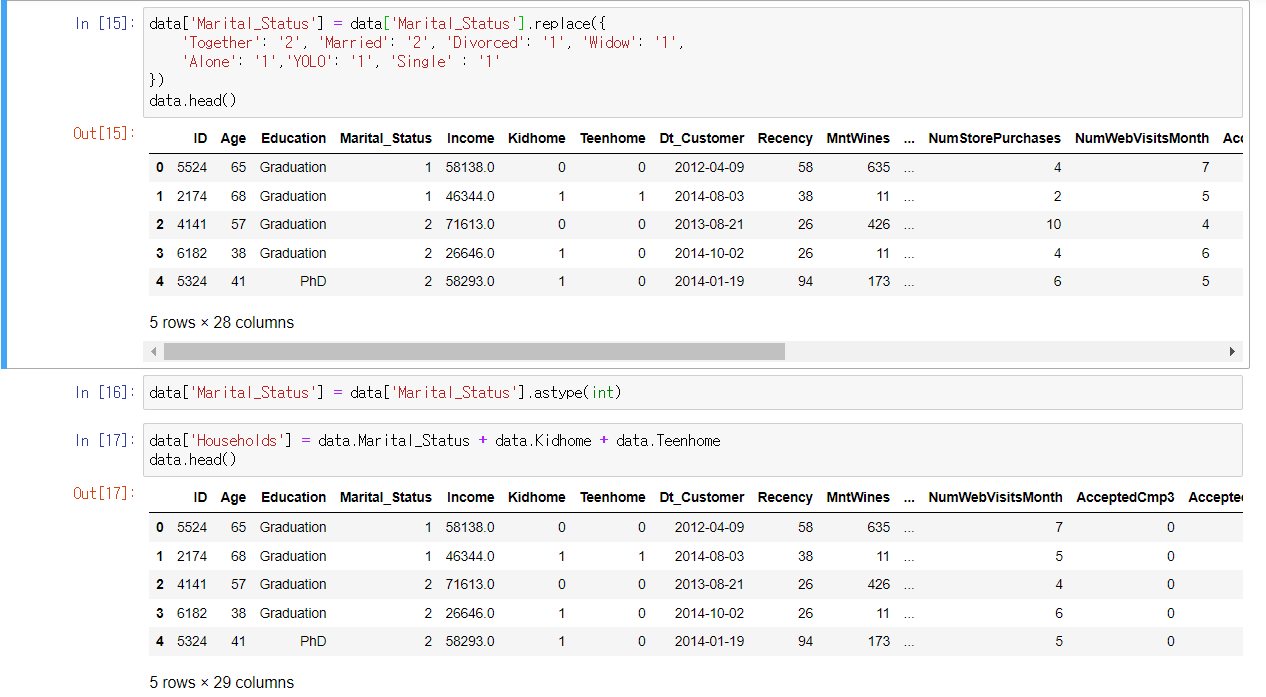
자동 생성된 설명**2. Data Introduction and Processing**

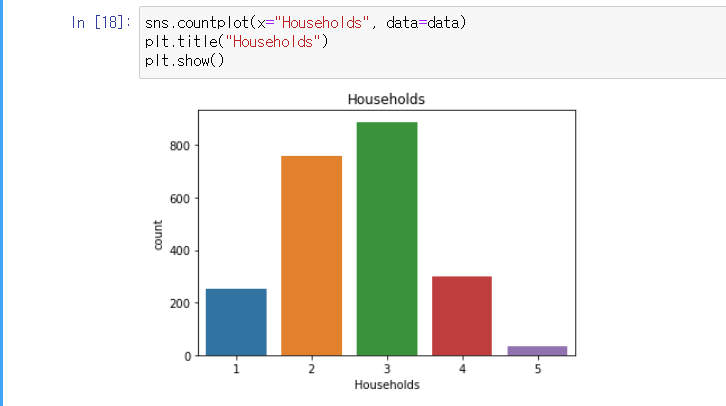
The picture is described about the data set, We had to delete the missing values or outlier values. see the next picture. in this set, Z\_CostContact and Z\_Revenue are the unique values, so we needed to remove these columns. to easily represent consumption, we made a column, “total\_spend”, combine these columns related to this. also, The column, “Age”, is used to delete the outlier more than 100 in consumer age.

 The maximum age is 129. we finded out the values of age more than 100. The tree outliers is needed to remove. as this result, the row, ‘239’, ‘339’,‘192’, have been deleted.

 Next, the missing row, ‘absurd’, in ‘Marital\_Status’ was deleted. we don’t know what this word means.

 We added column “Households” and each “Marital\_Status” was transformed into the number of a family, “Households” is added by “the number of family” and “the number of children(Teen, Kid)”. and this result is follwing as the same :





After we finished calculating and cleaning these outlier vales, we analyzed the correlationship between each variables. finally, we discovered the highest correlationship(0.66) at income. this fact influenced to easily fill the missing values. we could estabilsh the stardard filled these values in total\_spend and decide “Spending\_section labels=['Low\_spend','mid\_spend','high\_Spend']. also, the mean of each interval in Spending\_section was filled in missing values in income. at this result, In data set, we could not find out the missing values. also, we made the column the number of date visiting customer.

텍스트이(가) 표시된 사진

자동 생성된 설명

- Scaling and encoding to make the range (or distribution) of all characteristics the same.

텍스트이(가) 표시된 사진

자동 생성된 설명

**3-1. Clustering**

Before the association Rule, we divided two cases, one is with PCA, other is not. In calculating result, it's good when k = 4, This result and strategy is following :

Cluster 0 -> less consume, they purchased a lot of fish, but less meat. In discount product, has a big interest. also, they used both oneline purchase and offline. the family number almost 2.5 and education is high 2nd. they less joined in campain

Bronze consumer (we need to promote a discount product in every aspect.)

Cluster 1 -> high consume, they purchased a lot of meat, but they're interested in gold product. the less interested in discount product. purchase way is most catalog in online.

Gold consumer (we need marketing stratage to sell meat product for 1 person, and promote gold product at online)

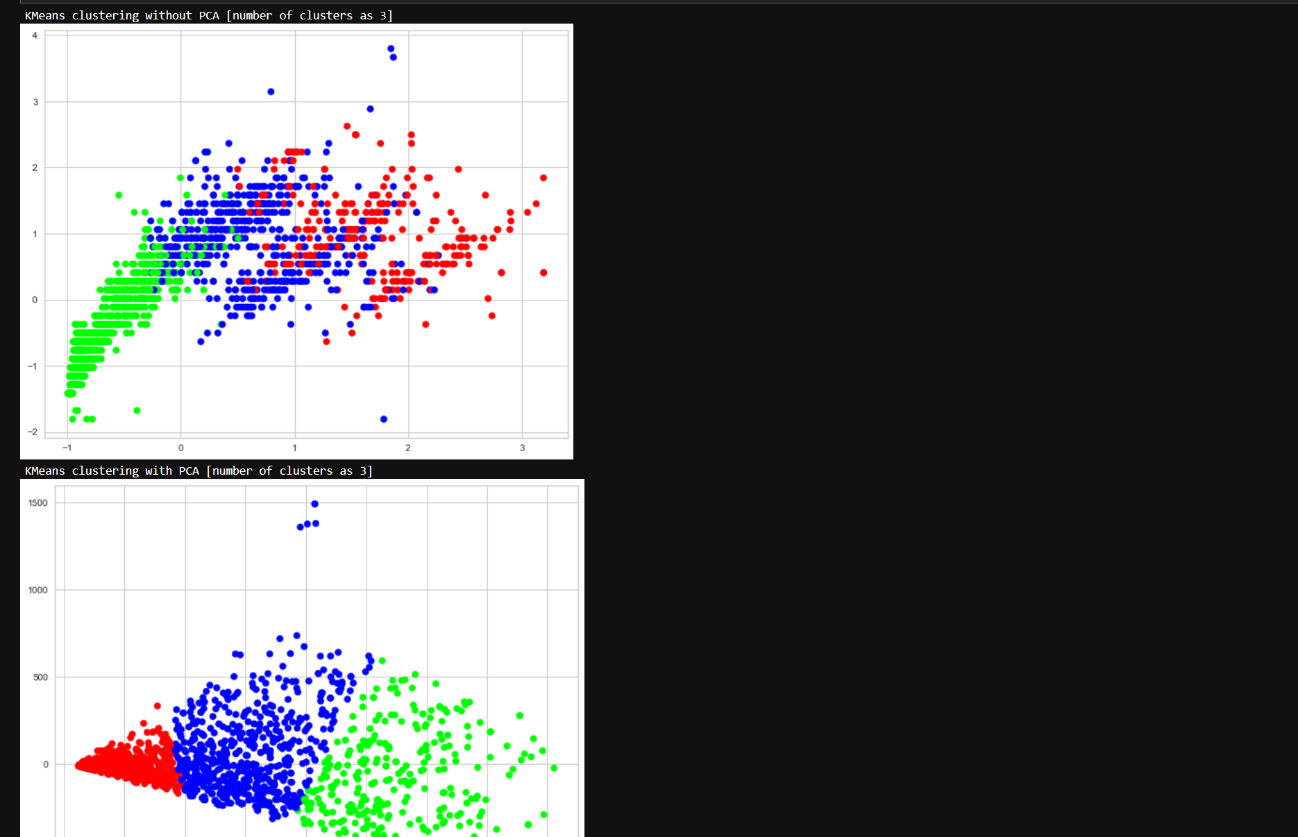
Cluster 2 -> less consume, they're not interested in meat product.

Silver consumer (we need to promote about meat product.)

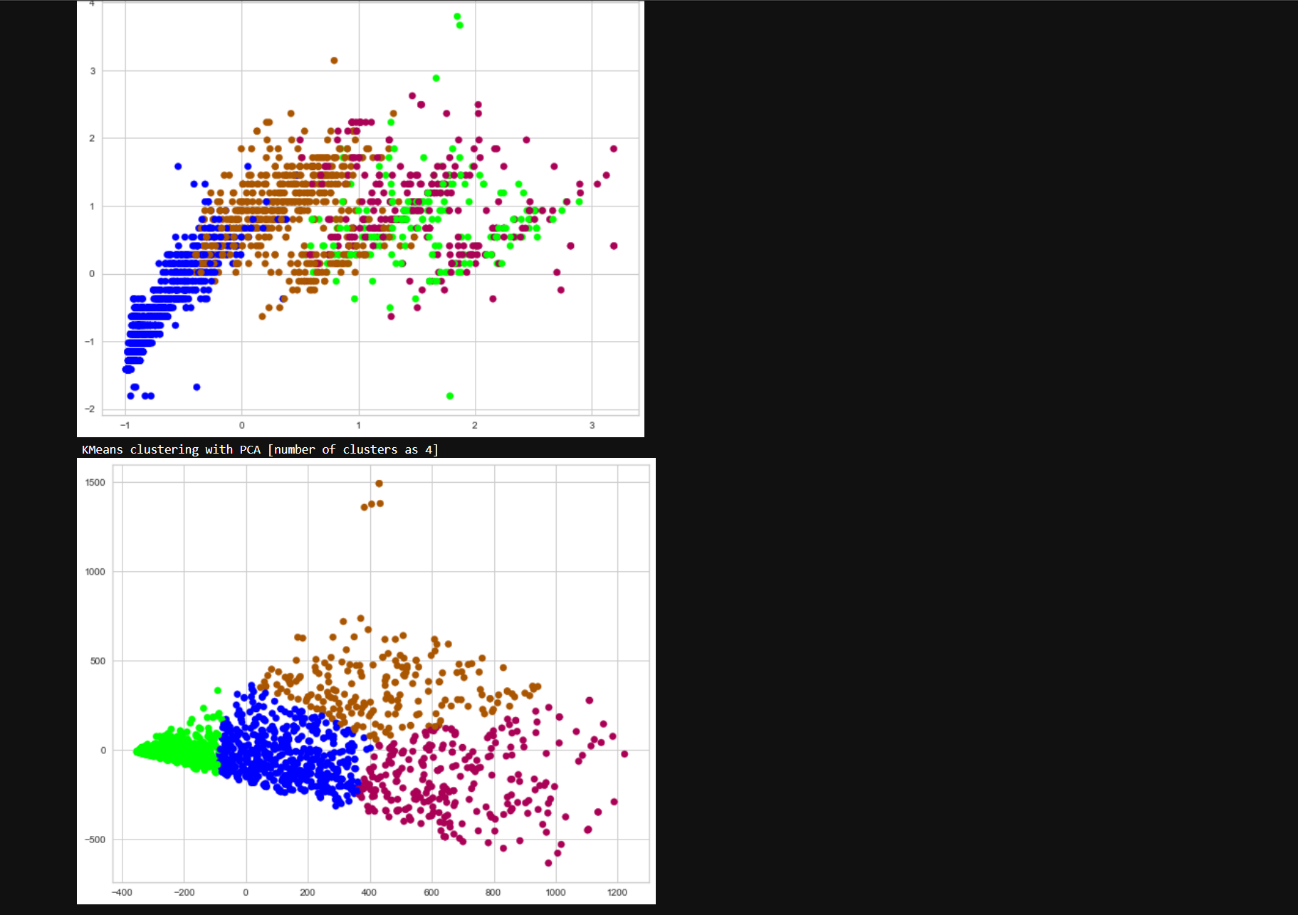
Cluster 3 -> very high consume, they purchase Wine and their purchase way has both online and offline.

Platinum consumer (we need to promote wine product)

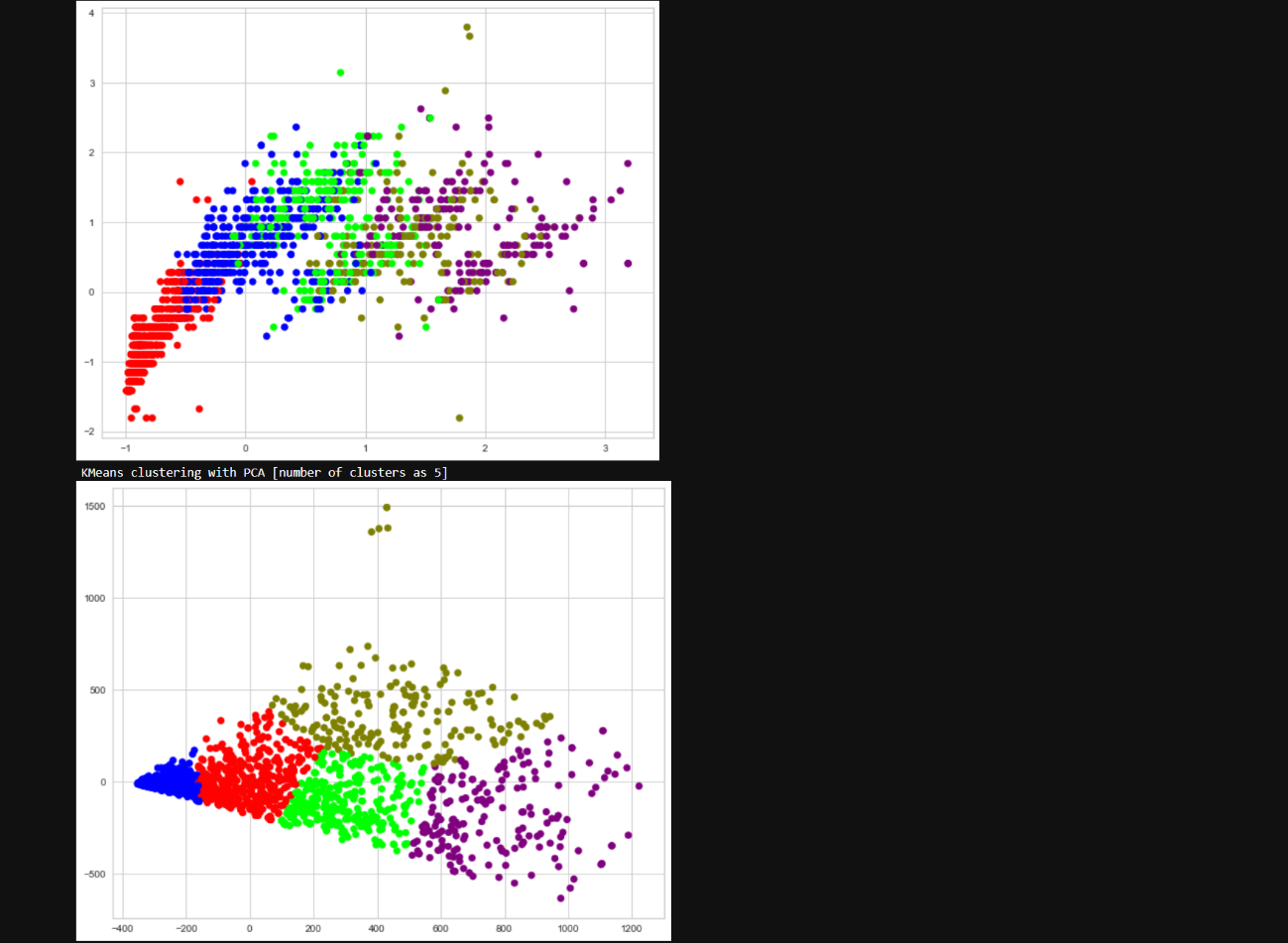
(Without PCA and with PCA cluster as 3)



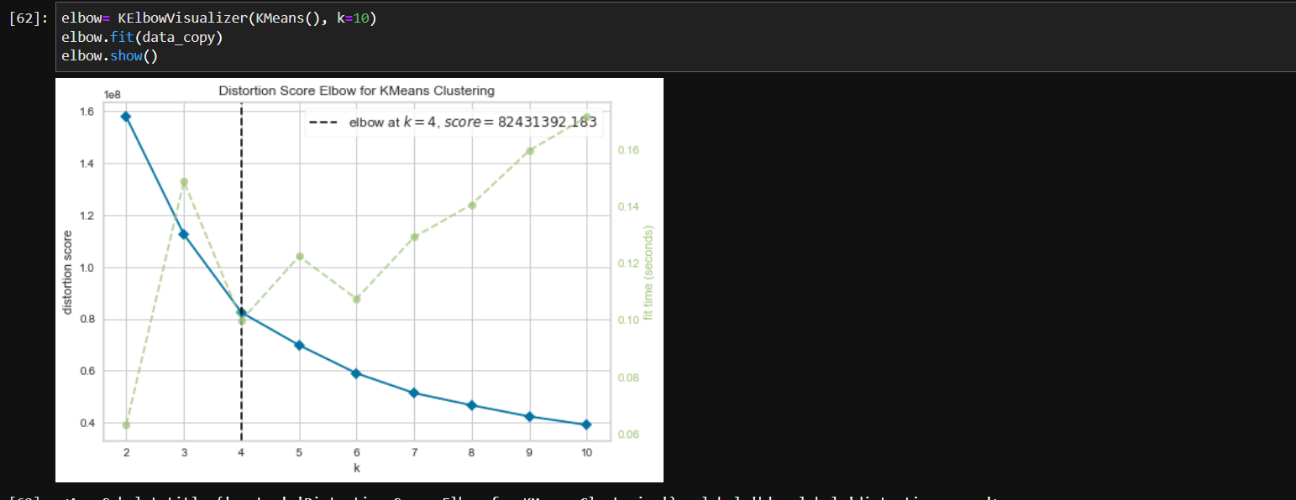
(Without PCA and with PCA cluster as 4)



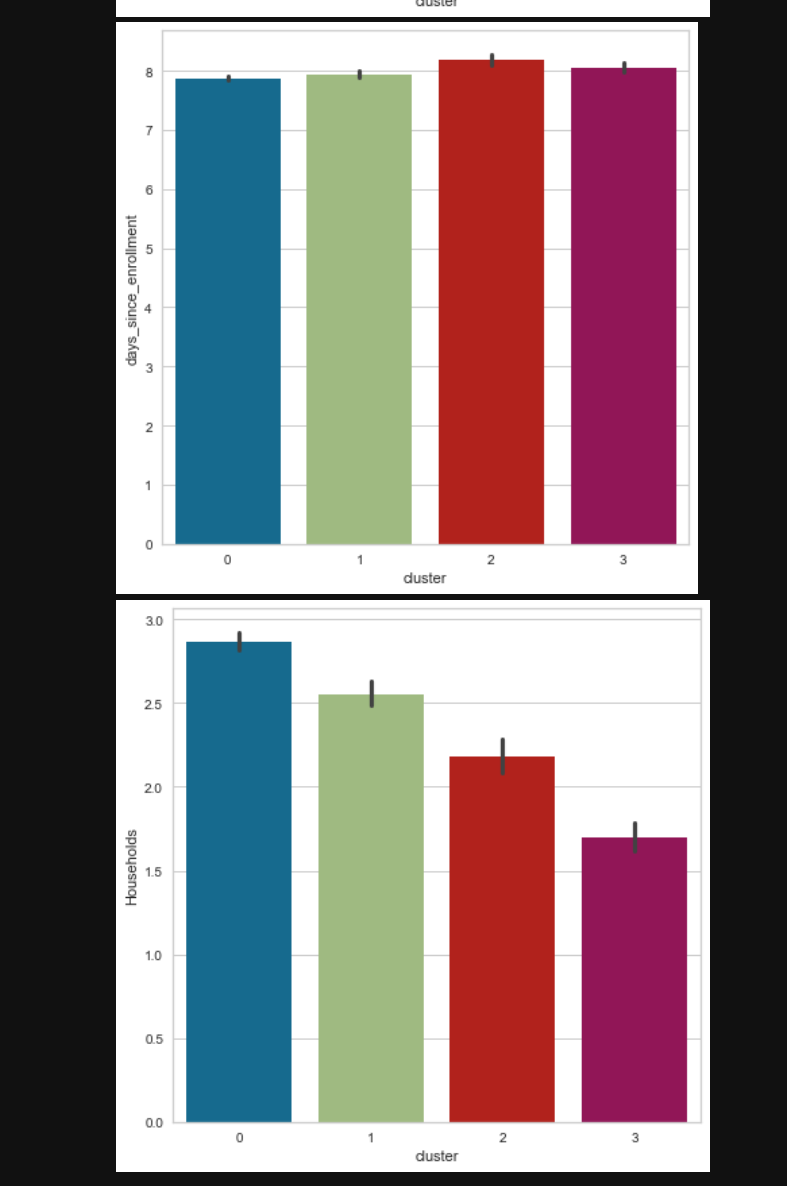
(Without PCA and with PCA cluster as 4)



The optimal K is 4.

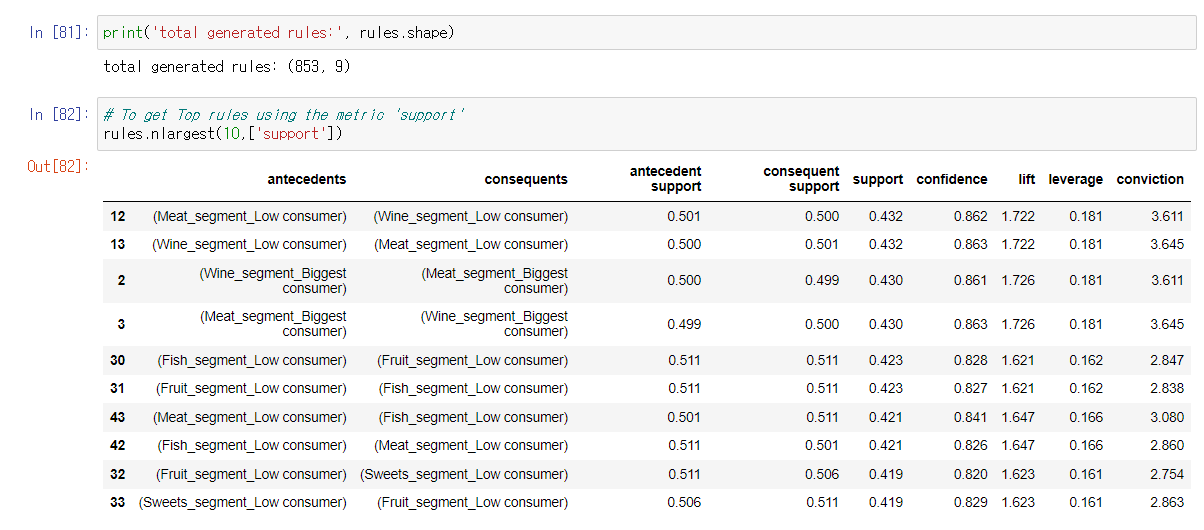


The results of various clustering.



**3-2. Association Rule**

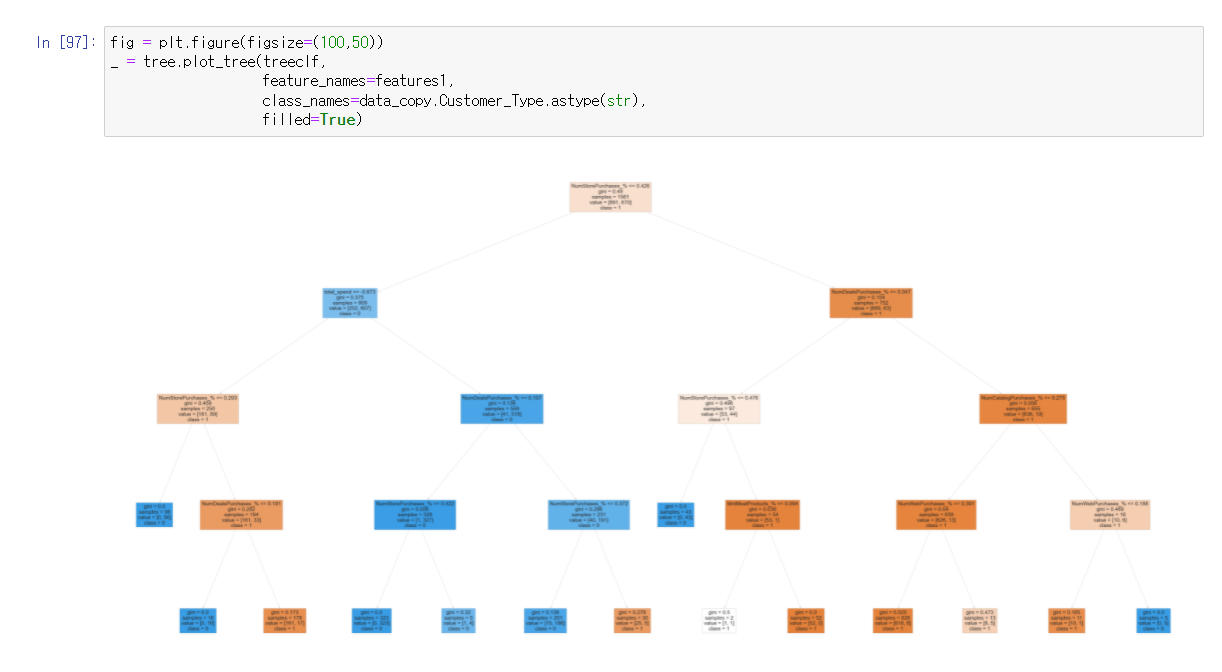
We cleaned and transformed any columns into new colums to easily understand. and divided low consume section, high consume section. In this result, we get the high support value between wine and meat(0.432), we need to set marketing strategy to sell the product with meat and wine. if we didn't divide two section, then we get the higher support value.



**3-3. Classification Decision Tree**

we testified the maximum depth in decision tree, when it has the highest accurcy, we choose it. in max\_depth = 4, the decision tree is following :





**3-4. Regression Decision Tree**

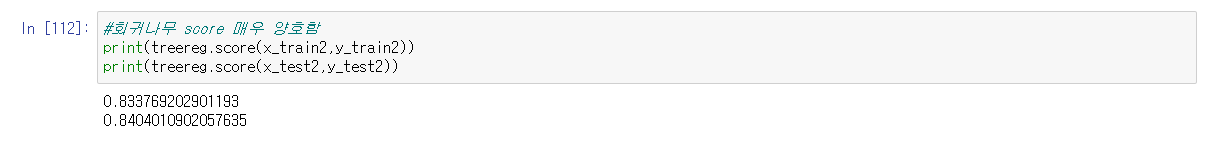
The personal information(feature\_cols)of new consumer without data, learning and training original data,

We predicted purchase times, how much purchase them though the regression tree.

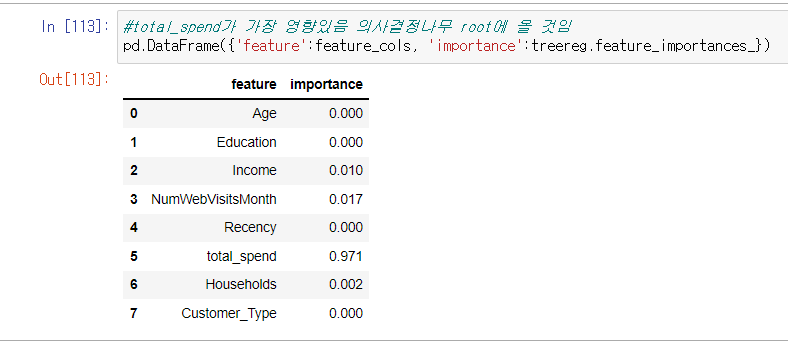
However, in regression, it is a limitation when happended error at the division point

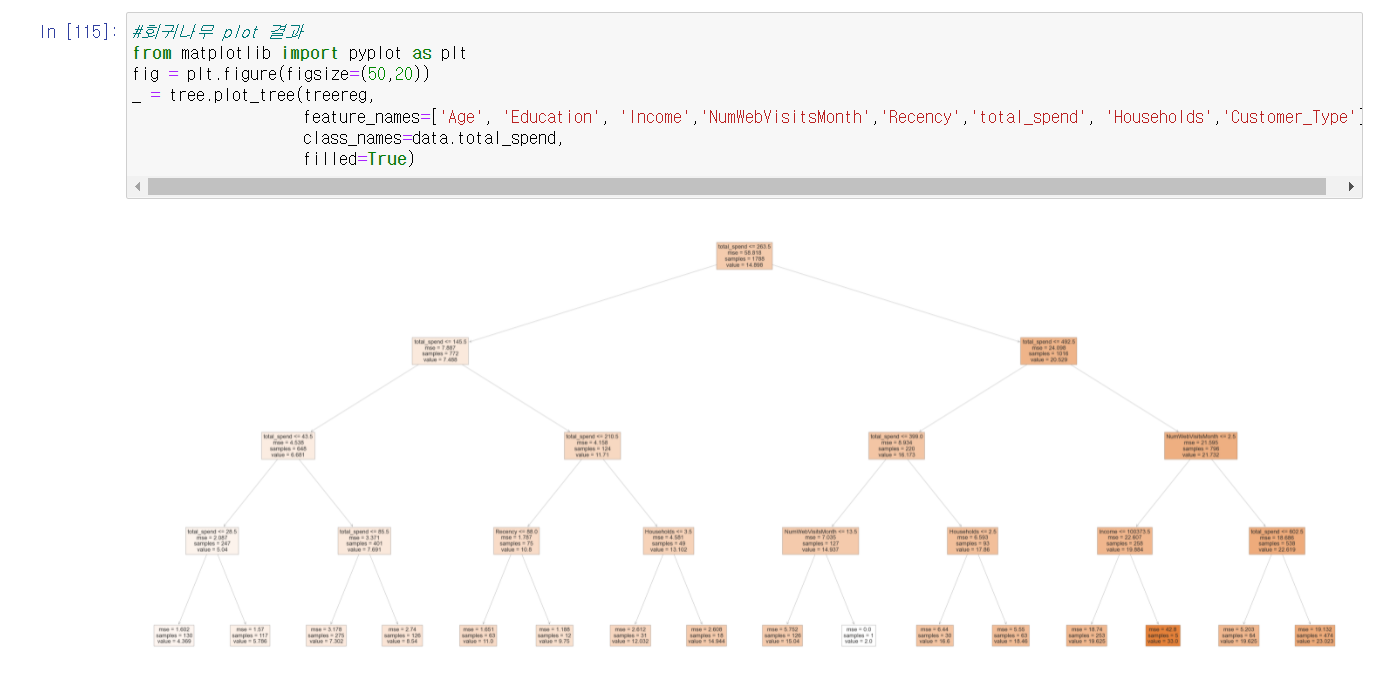
because the continuous variable is set the discontinuous variables,

Therefore, we predicted the purchase times about new consuemr as multi-linear regression again.



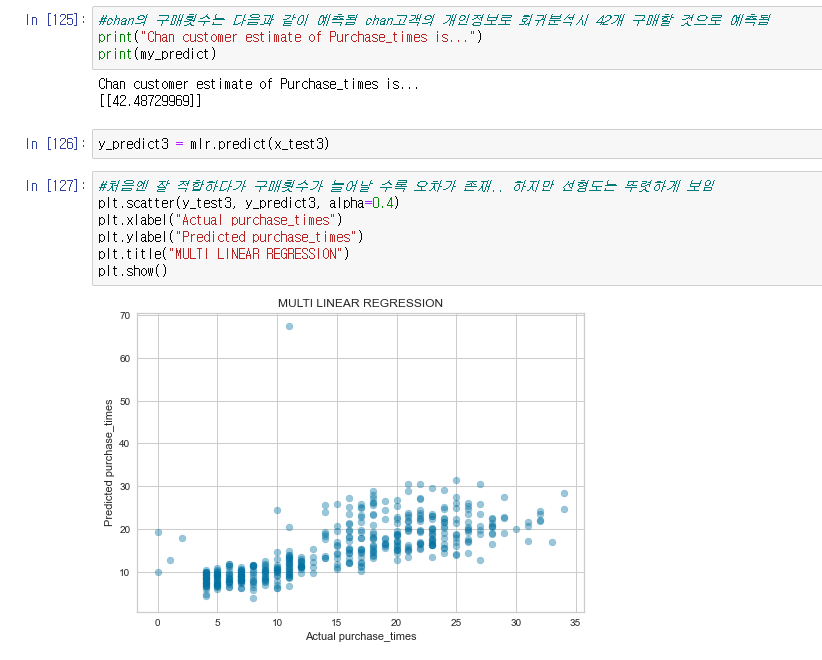
It's good score, moreover, we could find out the importance of the columns affect at the total\_spend.



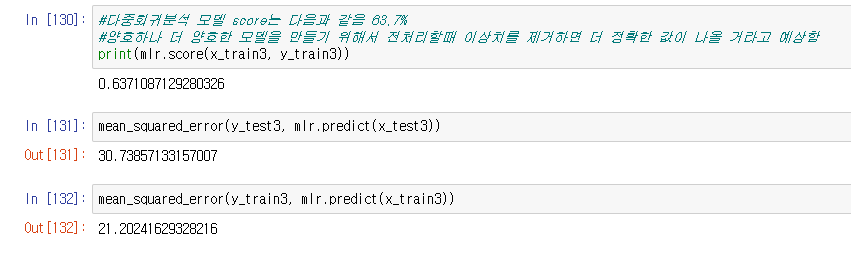


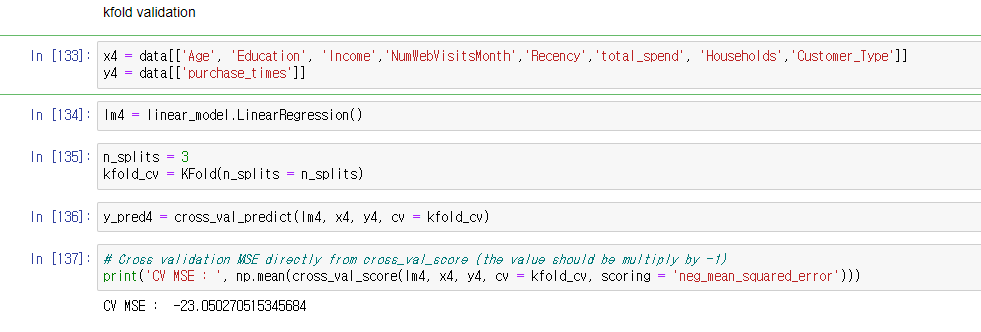
**3-5. Regression Decision Tree**

As we told a limitation about regression decision tree, we tried multi-linear regression. the new information with new consumer was anlyzed by the original data. in beginning, it's suitable to predict a new data, but when the purchase times are increasing, it happened the error and found out the linear relationship.



The score is 0.64. it's suitable, but when we more remove missing values, could get the correct value to make more suitable model, form. moreover, we did a k-fold to achieve this objective.





**4. The explanations about Feedback**

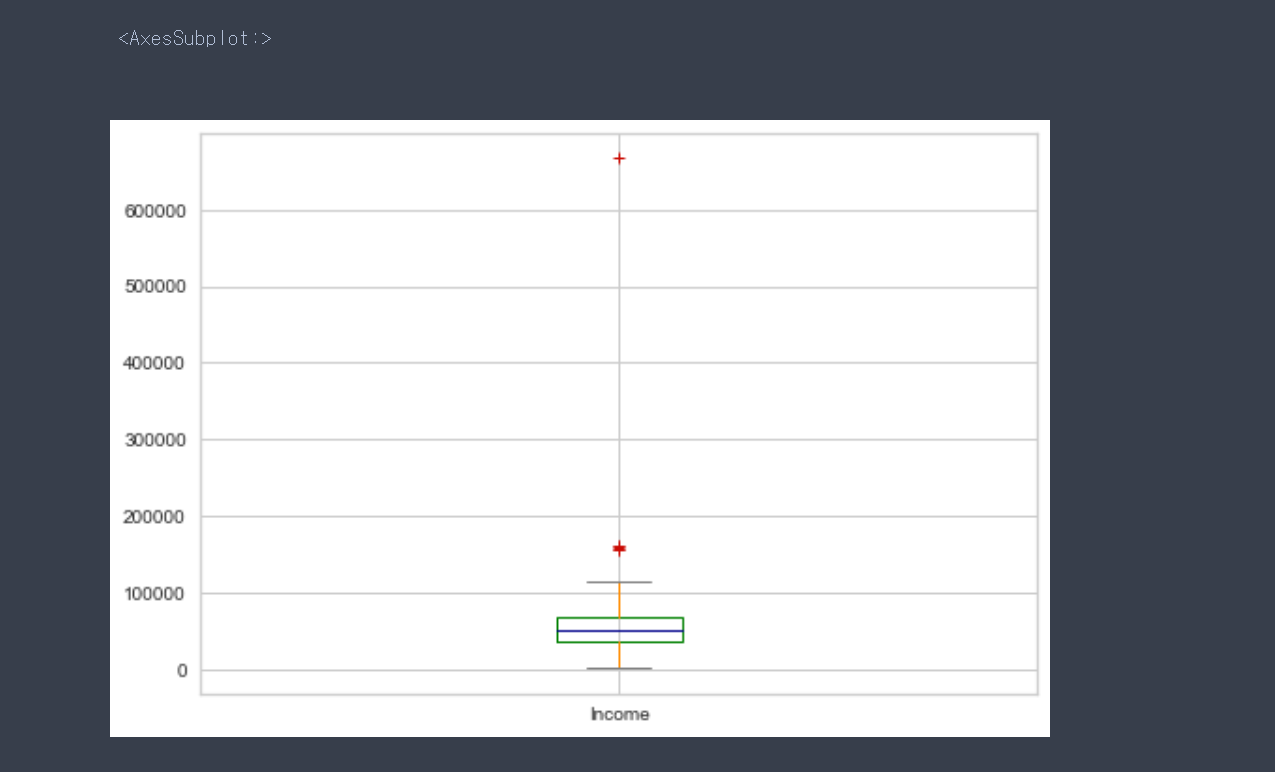
we was analyzed five Analysis technique. Association Rule, Clustering, Classification Tree, Regression Tree, Multiple Regression

Among them, the accuracy of classification tree and regression tree and MLR was not high.

There are Six opinions

The first one. <Train Data Set Quality>

we did preprocess and aggregation for all Analysis. therefore, training data for only decision trees could not be provided. Also, we think that the outlier values that we couldn't check had an effect on the results. likes in Fig1



[Fig1]

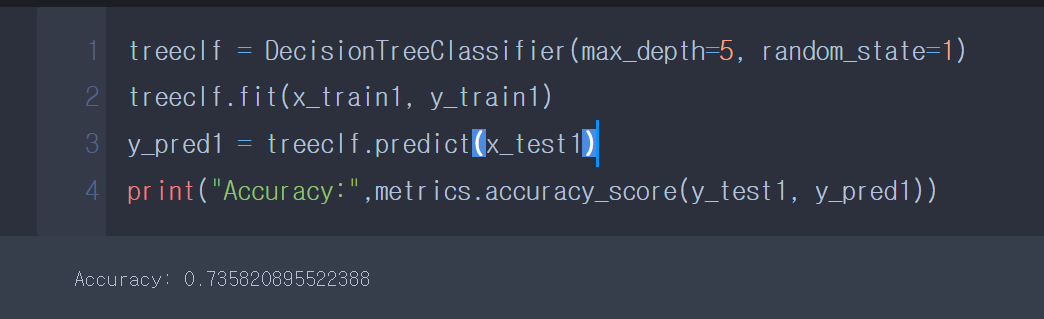
Second, Feature selection

Target is Customer\_Type (online or offline purchase),

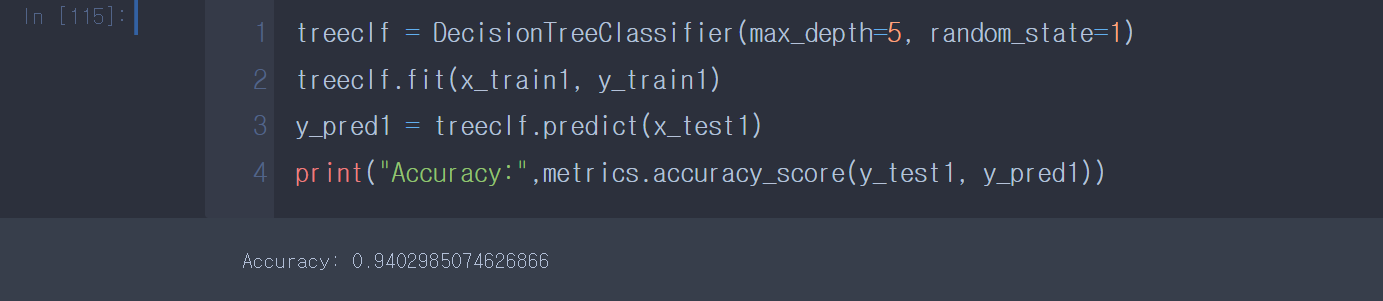
Features are [Income, Total Expenditure, Purchase Multiple, Household Education Age]

we had to choose features that is related to the target variable. However, I didn't use all the columns because, Our purpose is to predict the customer type as the most basic information for new customers. The Feature column consists of only basic information for new customers.

If features are included in all columns, better accuracy is required as shown in Fig3. But I don't want this result.



[Fig2] Analysis that not include all columns

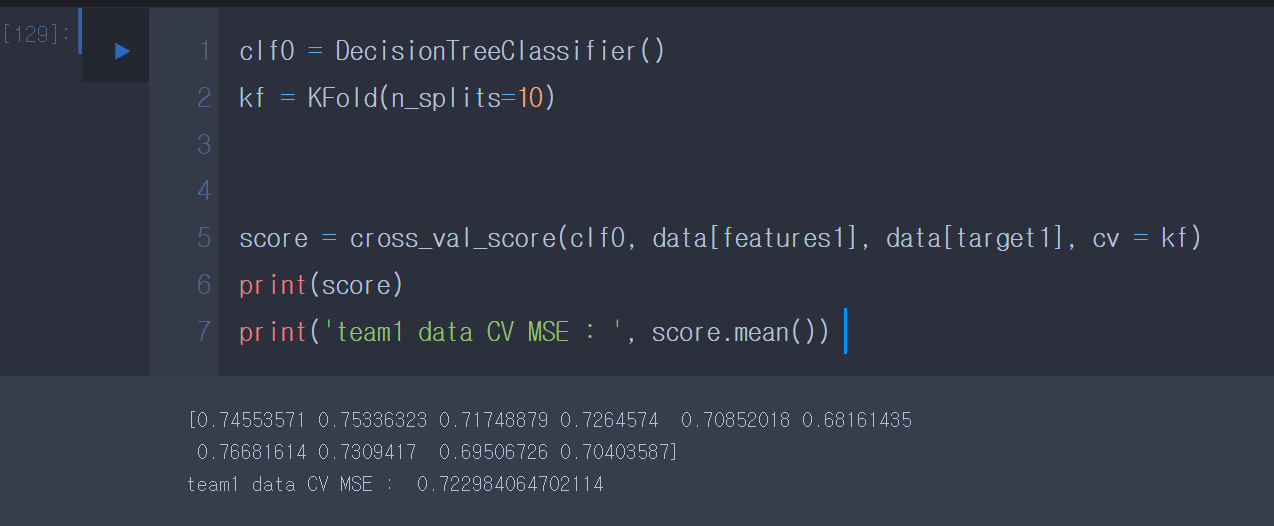


[Fig3] Analysis that include all columns

So we found out that there was a Naïve Bayes Classification, a method of removing noise instances from training data. we're going to study more about this.

Third, <cross\_validation>

No cross\_validation was performed on the classification tree. As a result, it can be seen that the accuracy is improved at k=7, as shown in Fig. 4.



[Fig 4]

Fourth, there are columns that do not have normalize and scale in the data used in the some feature.

Fifth, it will be more accurate if the weight is appropriate for columns that seem important.

Sixth, we analyzed using regression tree to predict the number of purchases by customers. It also includes only basic customer data, not all columns.

**6. Reference**

-regression

<https://machinelearningmastery.com/repeated-k-fold-cross-validation-with-python/>

<https://scikit-learn.org/stable/modules/cross_validation.html>

- Asscociation Rule

<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1057.9306&rep=rep1&type=pdf>

- Classification

<http://ceur-ws.org/Vol-2824/paper16.pdf>

- Clustering

<https://www.sciencedirect.com/science/article/pii/S2352711020303551>