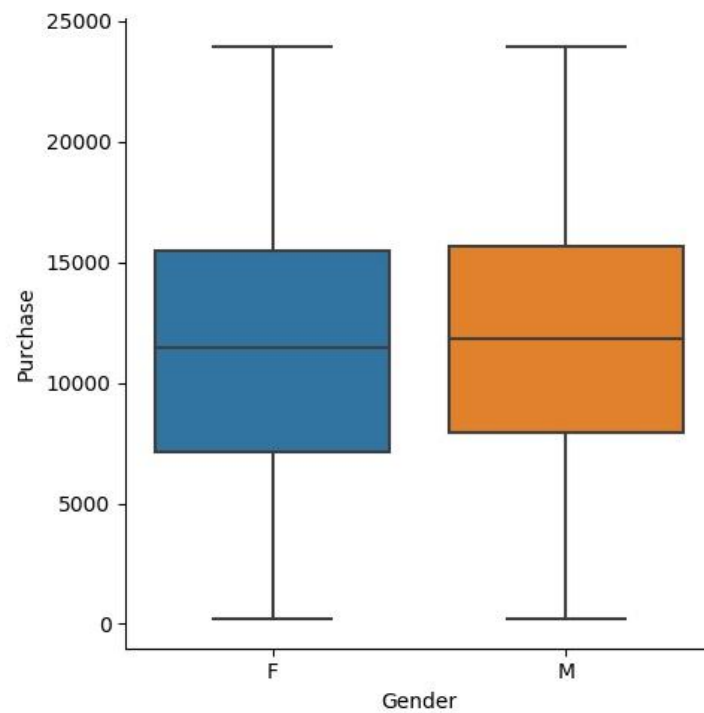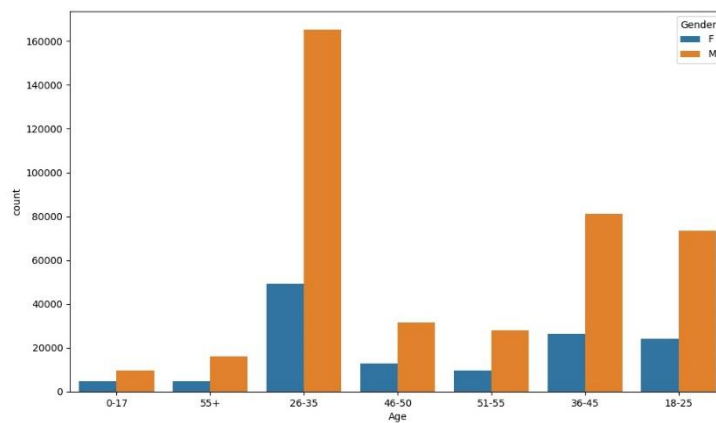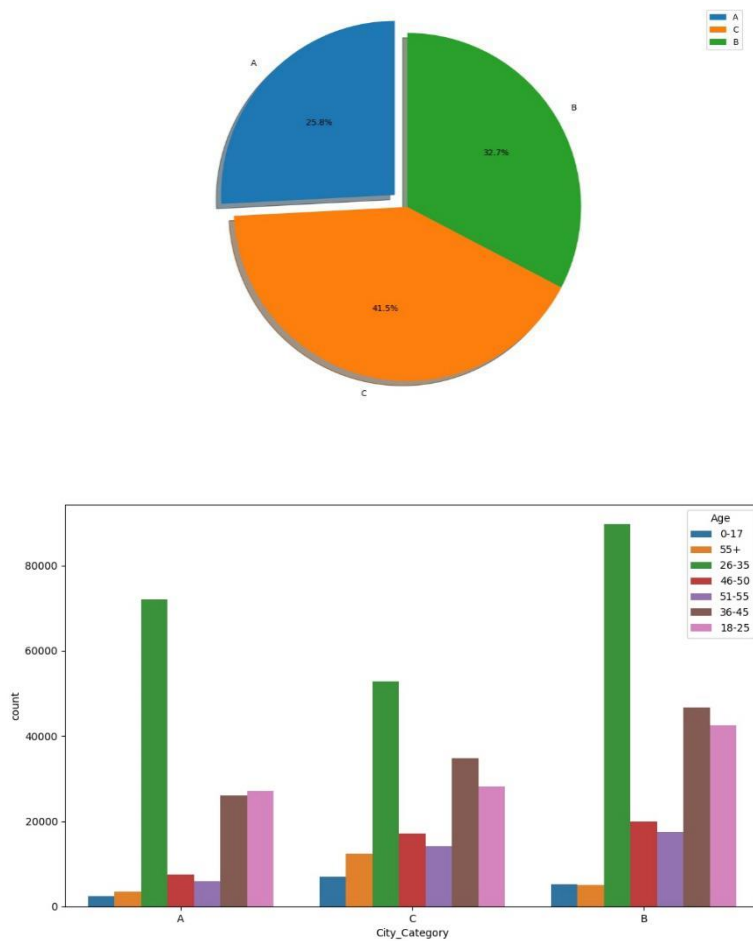1. Gender

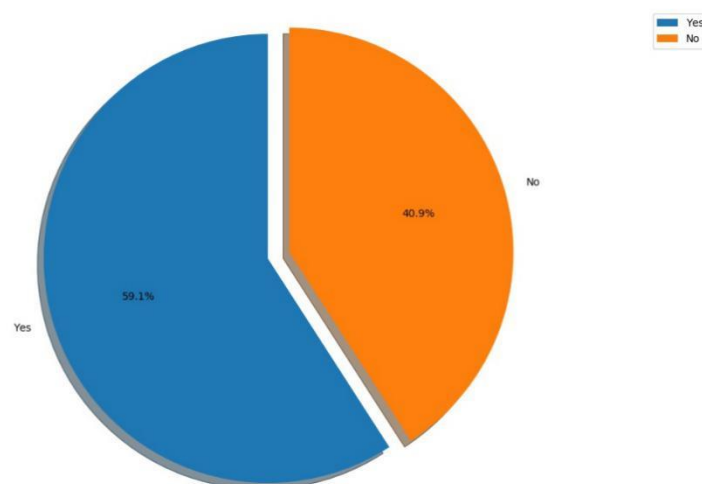Most of the customers are female rather than male.

2. age

Obviously, we can consider that the target age group of our stores is the age group of 26-35 years.

3. City

People living at City Category C have a greater purchasing power than others.

4. Marital

More than 60% customers are married, so the strategy of targeting families to ensure more clients succeed.

5. Evaluation: 10-fold cross validation

Evaluated classifications: Decision tree induction and Naïve Bayes

Using Classifier, predicting the gender base on Product_Category_1, Product_Category_2, Product_Category_3, Purchase. And we conclude from the result that the Product_Category_1/2/3 have no strong relationship with the customers' gender, and we cannot simply predict the customers' gender base on Product Categories.

Using 10-fold cross validation evaluate two classifiers, and the result of ROC curve graph show that the performance of Decision tree is better than Naïve Bayes.
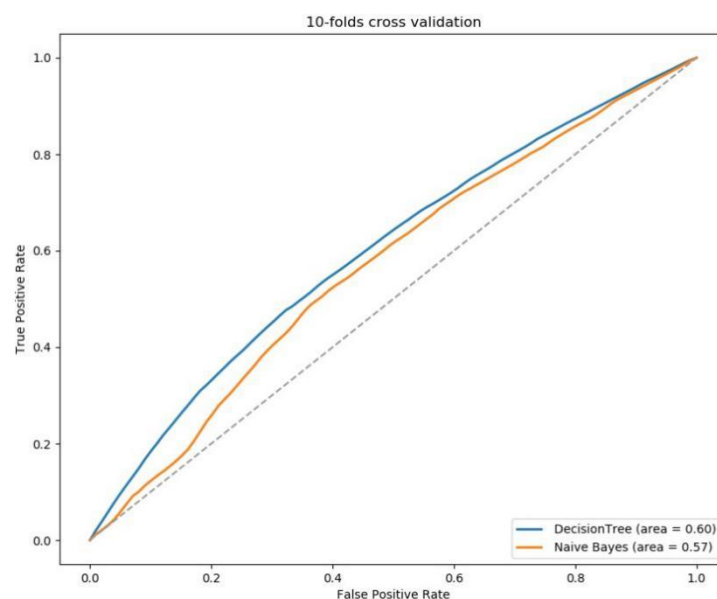
6. Frequent pattern:

Using User_ID and Product_ID, we want to find which commodities will be bought together. We took the first 20,000 transactions and used apriori algorithm to get the frequent patterns saved in the rules.csv file.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	frozenset(['P00110742'])	frozenset(['P00025442'])	0.271131561	0.270279482	0.118098	0.435575	1.611573	0.044817	1.292857
1	frozenset(['P00025442'])	frozenset(['P00110742'])	0.270279482	0.271131561	0.118098	0.436948	1.611573	0.044817	1.294496
2	frozenset(['P00110742'])	frozenset(['P00112142'])	0.271131561	0.262269939	0.110429	0.407291	1.552946	0.03932	1.244675
3	frozenset(['P00112142'])	frozenset(['P00110742'])	0.262269939	0.271131561	0.110429	0.421053	1.552946	0.03932	1.258955
4	frozenset(['P00057642'])	frozenset(['P00237542'])	0.243694615	0.234151329	0.100545	0.412587	1.762055	0.043484	1.303766
5	frozenset(['P00237542'])	frozenset(['P00057642'])	0.234151329	0.243694615	0.100545	0.429403	1.762055	0.043484	1.325464
6	frozenset(['P00112142'])	frozenset(['P00025442'])	0.262269939	0.270279482	0.104294	0.397661	1.471295	0.033408	1.211478
7	frozenset(['P00025442'])	frozenset(['P00112142'])	0.270279482	0.262269939	0.104294	0.385876	1.471295	0.033408	1.201273

7. Cluster analysis:

We used K-means method to classify the age as 7 clusters and used T-SNE method to present the clusters as two-dimensional graph. First, we tried the first 1000 transactions, and we found that the data can be separated as 7 clusters, then we tried the first 200,000 transactions then we got the well distribution of ages as 7 clusters.