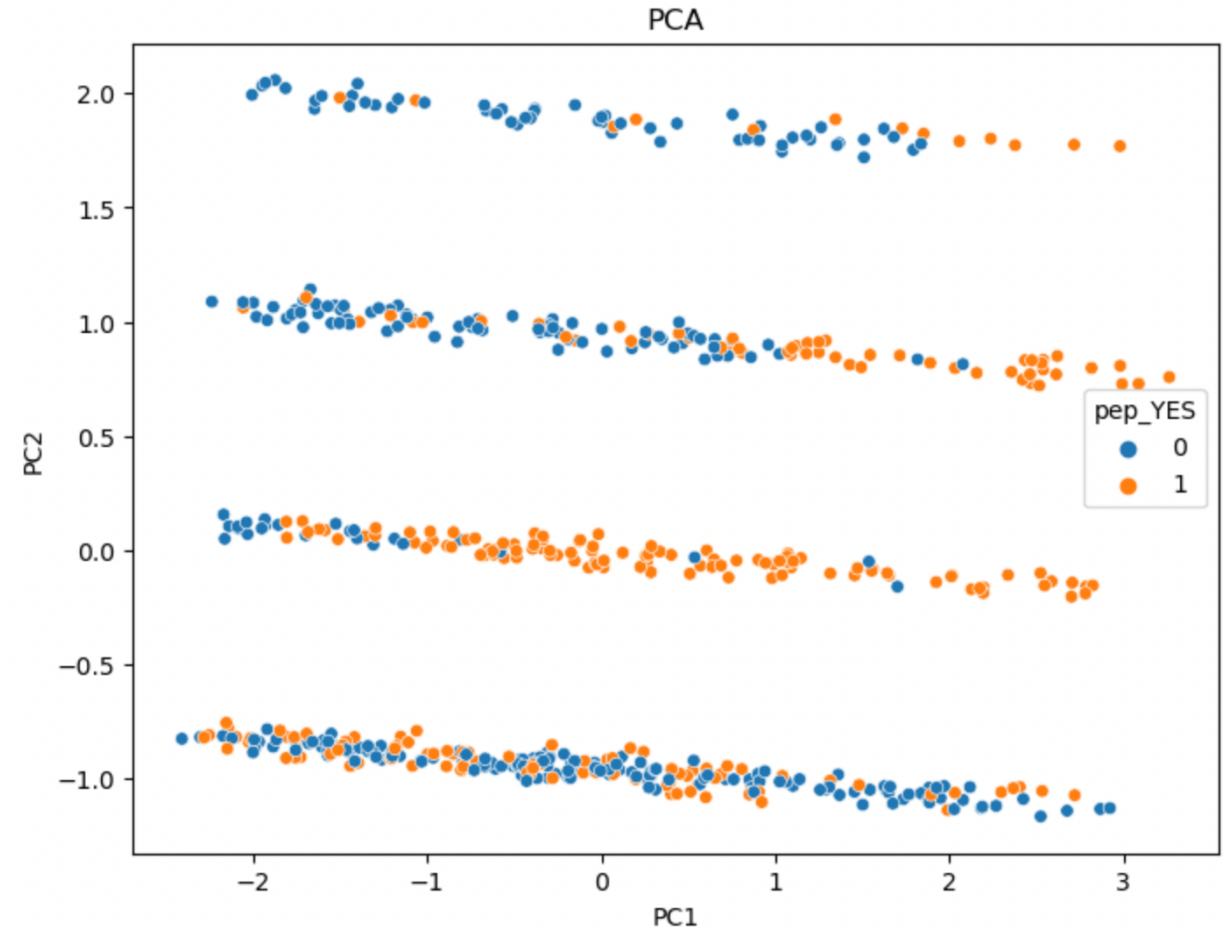
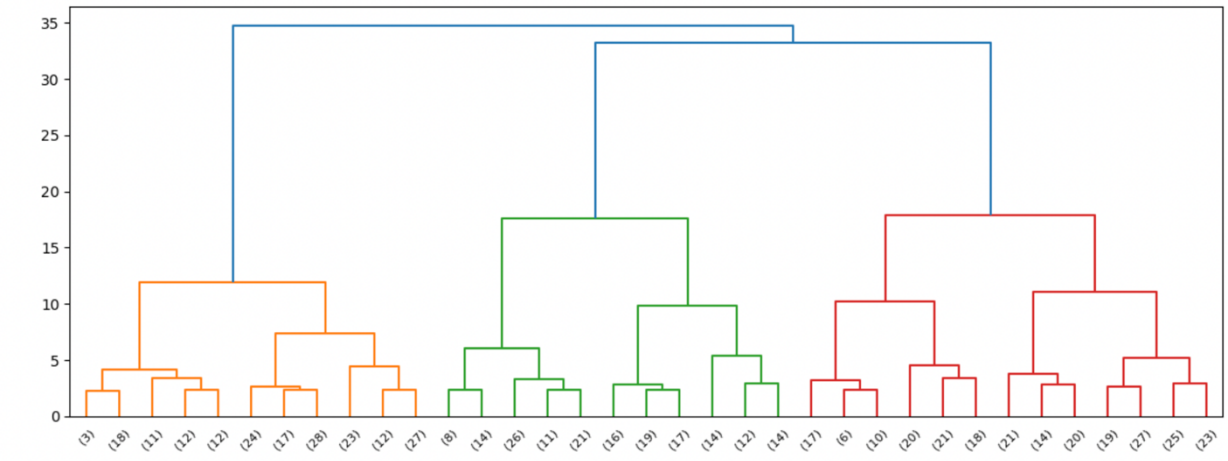
Bank dataset cluster

In this report, a bank dataset is used to cluster the data and find out what kind of people are most likely to buy PEP. To determine those, the apriori algorithm is used to find the association rules. The bank dataset contains numerical and categorical variables. The provided bank dataset includes customer characteristics, such as age, gender, residence region, income level, marital status, and mortgage status.

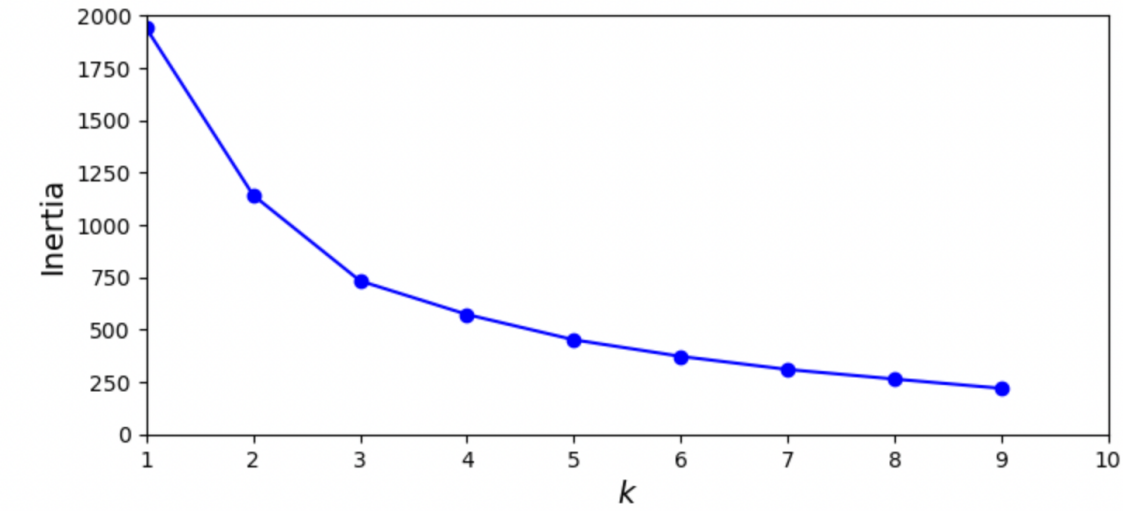
There are multiple features in this dataset so before performing the clustering, we will reduce the dimension. To do so, PCA is performed. The below graph shows the PCA plot.



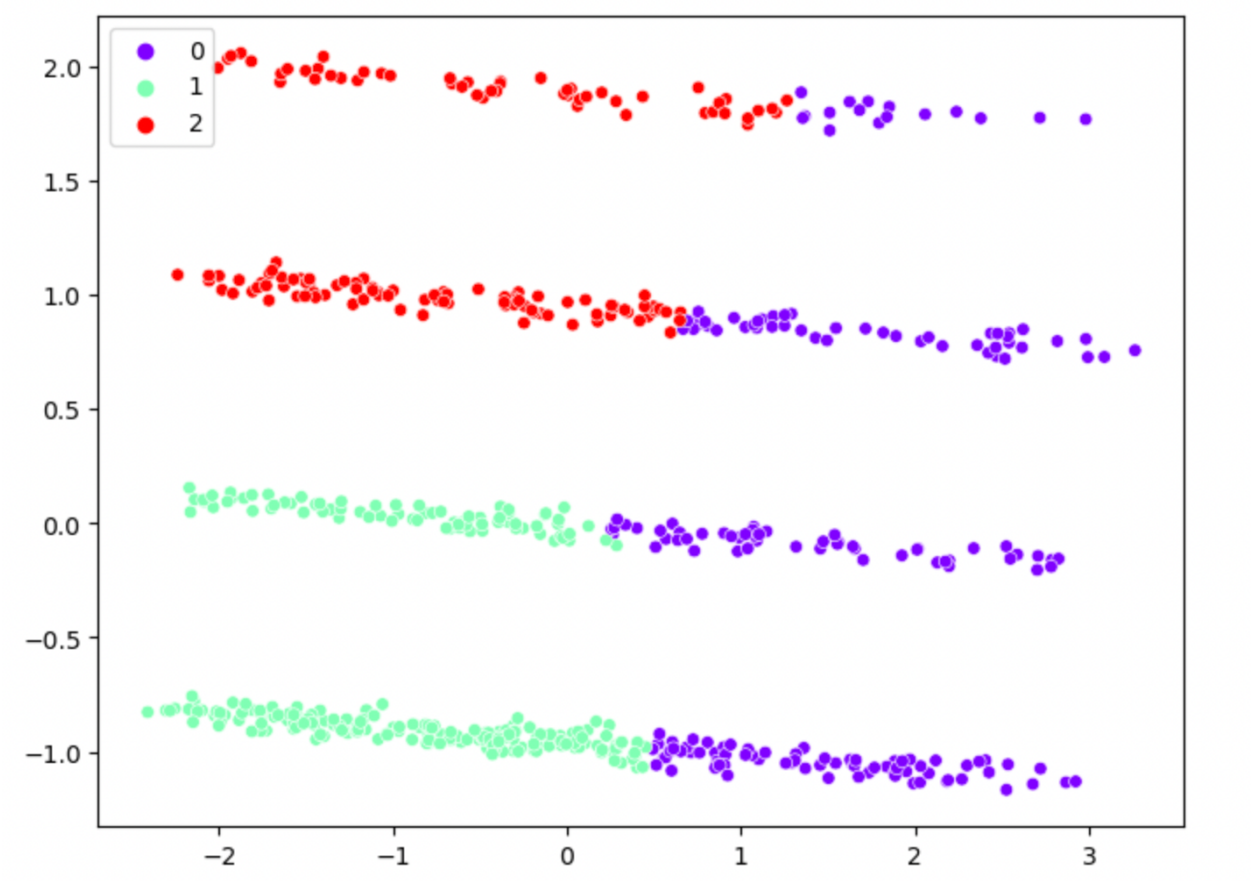
Now, let's perform HAC. The below dendrogram only shows the last 35 merges. Now, flcuster is used to label these. Although, based on the below HAC, a lot of interpretations cannot be made. There are 3 clusters and it doesn't have a lot of association with pep.



So let’s perform clustering the K-means clustering is used. To find the number of clusters, the method is used. From the image below, it is visible that there is an elbow at 3 so let's choose n as 3 for clustering.

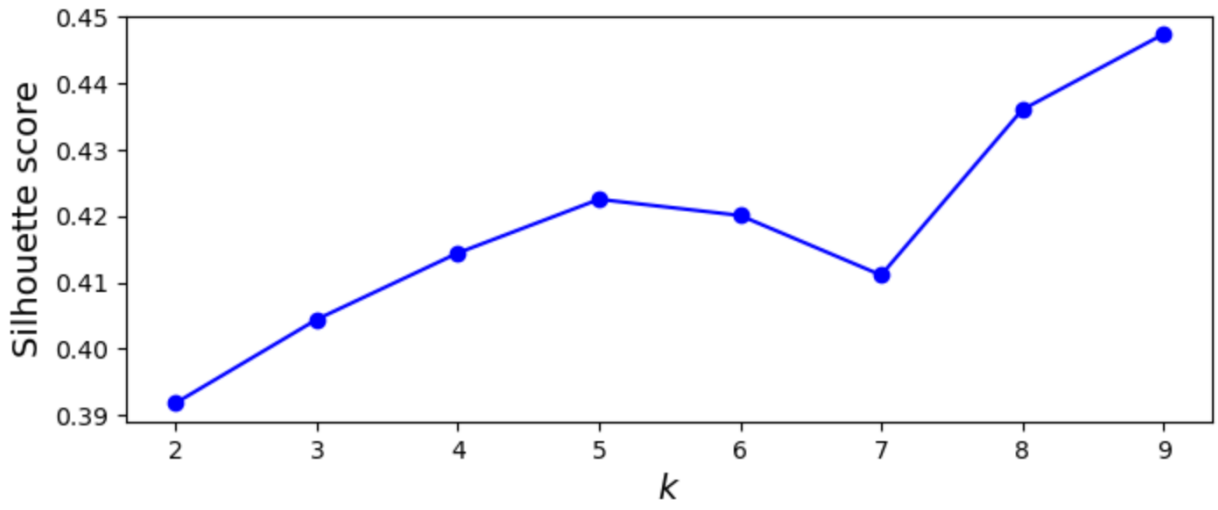


Using the K-means method for clustering and n as 3 and random state as 42 so the plot remains same each time, the below graph with clusters is produced.

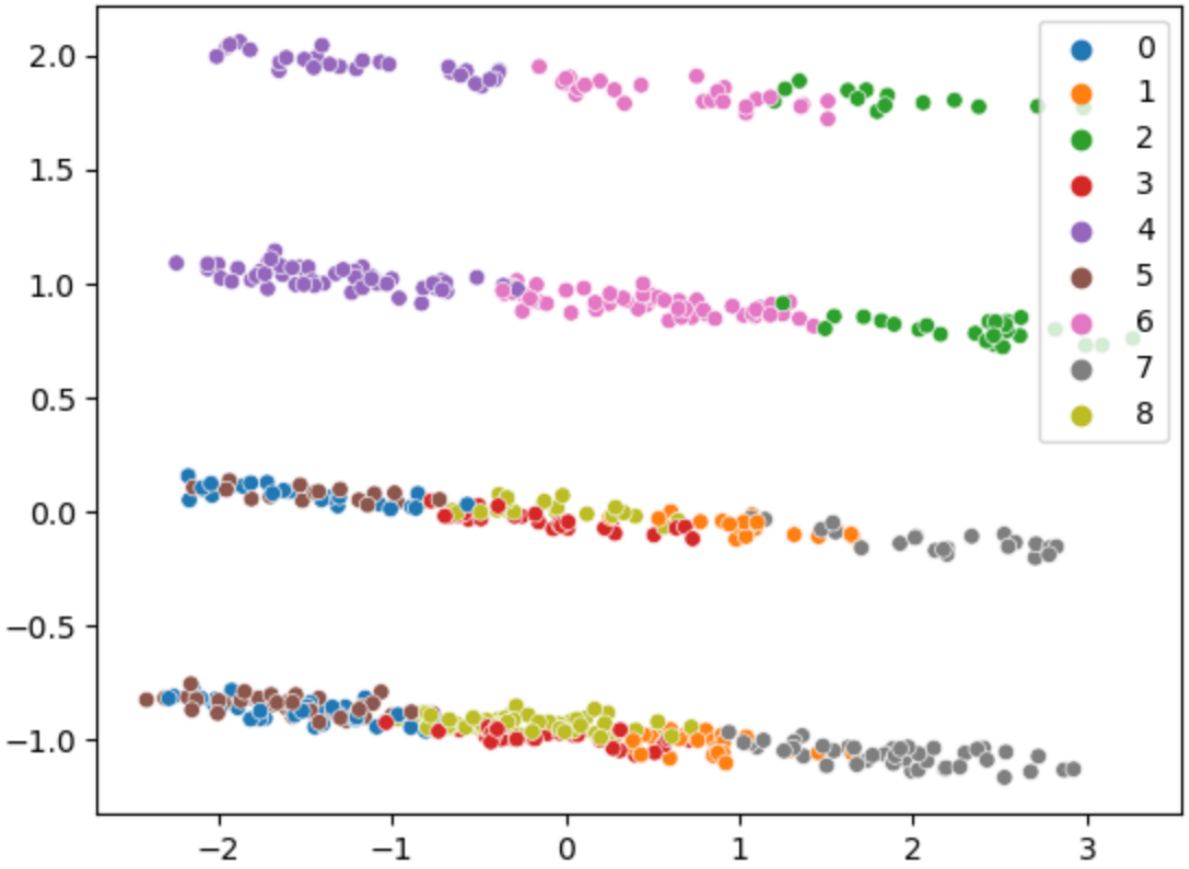


We can compare the normal PCA graph and the clustered PCA graph and see that there is no definite relation between pep and clusters. Now, lets choose the cluster using the Silhouette score. The below graph is the output.

```
plt.show()
```



It is visible that the Silhouette score is the highest when n is 9. So let's try with n as 9 this time. The below graph is the output.



Now, if the original PCA graph and clustered graph are compared, cluster 1 covers the most of the pep_YES that is 1. Cluster 2 also does a good job at covering some of the pep_YES that is orange in the original PCA. Cluster 5 also covers some of the orange points. Now, let's check their characteristics.

Most people in cluster 1 are above 40 and the mean salary is almost 30,000. They also happen to have either 1 or 0 children. In cluster 2, most people have either 2 or 3 children. All of them also have a savings account and their mean salary is 51414.26.They are also above 40 years old. Cluster 5 also has people that have 0 or 1 children. The average salary is 17533.83. Additionally, all of them have a savings account. Most of them also have a current account. Hence, we can target these people since they are most likely to apply for PEP.

PART III:

Association Rule

In this report, a bank dataset is used to find out what kind of people are most likely to buy PEP. To determine those, the apriori algorithm is used to find the association rules. The bank dataset contains numerical and categorical variables.

Before getting into the association rule, the data needs to be discretized. The age values are categorized into 3 bins, “10 to 30”, “30 to 60”, and “60+”. The income is also categorized into 3 categories, “lower class income”, “middle class income”, and “upper class income”. All the categorical values in the dataset are then converted into dummy variables.

Now, let’s get into the association rule. Using the apriori algorithm, the frequency of item sets is found, and the minimum support value is used as 0.2. Now, the association rule is used in the frequency itemset and the metric used is lift because it measures the strength of a relationship between items in a dataset. The minimum threshold is set to 0.7.

Once, we get the rules, the rules are filtered such that the consequent is ‘pep_YES’ since we need to find the variables that will most likely lead to a person applying for PEP. Once that’s done, the data is then sorted by confidence, lift, and support to find the top 5 interesting rules.

The rules that have been sorted by confidence, lift, and support are shown below.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
206	(current_act_YES, age category_30 to 60)	(pep_YES)	0.440000	0.456667	0.215000	0.488636	1.070007	0.014067	1.062519	0.116833
56	(sex_MALE)	(pep_YES)	0.500000	0.456667	0.240000	0.480000	1.051095	0.011667	1.044872	0.097222
58	(age category_30 to 60)	(pep_YES)	0.590000	0.456667	0.281667	0.477401	1.045404	0.012233	1.039676	0.105932
230	(age category_30 to 60, income category_middle...	(pep_YES)	0.433333	0.456667	0.203333	0.469231	1.027513	0.005444	1.023671	0.047252
20	(car_YES)	(pep_YES)	0.493333	0.456667	0.230000	0.466216	1.020911	0.004711	1.017890	0.040427
44	(current_act_YES)	(pep_YES)	0.758333	0.456667	0.351667	0.463736	1.015481	0.005361	1.013183	0.063082
60	(income category_middle class income)	(pep_YES)	0.548333	0.456667	0.253333	0.462006	1.011692	0.002928	1.009925	0.025587
170	(current_act_YES, save_act_YES)	(pep_YES)	0.531667	0.456667	0.233333	0.438871	0.961032	-0.009461	0.968287	-0.079680
32	(save_act_YES)	(pep_YES)	0.690000	0.456667	0.298333	0.432367	0.946789	-0.016767	0.957191	-0.153471
8	(married_YES)	(pep_YES)	0.660000	0.456667	0.256667	0.388889	0.851582	-0.044733	0.889091	-0.338889

Fig 1. Sorted by Confidence

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
206	(current_act_YES, age category_30 to 60)	(pep_YES)	0.440000	0.456667	0.215000	0.488636	1.070007	0.014067	1.062519	0.116833
56	(sex_MALE)	(pep_YES)	0.500000	0.456667	0.240000	0.480000	1.051095	0.011667	1.044872	0.097222
58	(age category_30 to 60)	(pep_YES)	0.590000	0.456667	0.281667	0.477401	1.045404	0.012233	1.039676	0.105932
230	(age category_30 to 60, income category_middle...	(pep_YES)	0.433333	0.456667	0.203333	0.469231	1.027513	0.005444	1.023671	0.047252
20	(car_YES)	(pep_YES)	0.493333	0.456667	0.230000	0.466216	1.020911	0.004711	1.017890	0.040427
44	(current_act_YES)	(pep_YES)	0.758333	0.456667	0.351667	0.463736	1.015481	0.005361	1.013183	0.063082
60	(income category_middle class income)	(pep_YES)	0.548333	0.456667	0.253333	0.462006	1.011692	0.002928	1.009925	0.025587
170	(current_act_YES, save_act_YES)	(pep_YES)	0.531667	0.456667	0.233333	0.438871	0.961032	-0.009461	0.968287	-0.079680
32	(save_act_YES)	(pep_YES)	0.690000	0.456667	0.298333	0.432367	0.946789	-0.016767	0.957191	-0.153471
8	(married_YES)	(pep_YES)	0.660000	0.456667	0.256667	0.388889	0.851582	-0.044733	0.889091	-0.338889

Fig 2: Sorted by Lift

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
44	(current_act_YES)	(pep_YES)	0.758333	0.456667	0.351667	0.463736	1.015481	0.005361	1.013183	0.063082
32	(save_act_YES)	(pep_YES)	0.690000	0.456667	0.298333	0.432367	0.946789	-0.016767	0.957191	-0.153471
58	(age category_30 to 60)	(pep_YES)	0.590000	0.456667	0.281667	0.477401	1.045404	0.012233	1.039676	0.105932
8	(married_YES)	(pep_YES)	0.660000	0.456667	0.256667	0.388889	0.851582	-0.044733	0.889091	-0.338889
60	(income category_middle class income)	(pep_YES)	0.548333	0.456667	0.253333	0.462006	1.011692	0.002928	1.009925	0.025587
56	(sex_MALE)	(pep_YES)	0.500000	0.456667	0.240000	0.480000	1.051095	0.011667	1.044872	0.097222
170	(current_act_YES, save_act_YES)	(pep_YES)	0.531667	0.456667	0.233333	0.438871	0.961032	-0.009461	0.968287	-0.079680
20	(car_YES)	(pep_YES)	0.493333	0.456667	0.230000	0.466216	1.020911	0.004711	1.017890	0.040427
206	(current_act_YES, age category_30 to 60)	(pep_YES)	0.440000	0.456667	0.215000	0.488636	1.070007	0.014067	1.062519	0.116833
230	(age category_30 to 60, income category_middle...	(pep_YES)	0.433333	0.456667	0.203333	0.469231	1.027513	0.005444	1.023671	0.047252

Fig 3: Sorted by Support

After looking at the above rules, the top 5 interesting rules would be the one sorted by the confidence level as shown below.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
206	(current_act_YES, age category_30 to 60)	(pep_YES)	0.440000	0.456667	0.215000	0.488636	1.070007	0.014067	1.062519	0.116833
56	(sex_MALE)	(pep_YES)	0.500000	0.456667	0.240000	0.480000	1.051095	0.011667	1.044872	0.097222
58	(age category_30 to 60)	(pep_YES)	0.590000	0.456667	0.281667	0.477401	1.045404	0.012233	1.039676	0.105932
230	(age category_30 to 60, income category_middle...	(pep_YES)	0.433333	0.456667	0.203333	0.469231	1.027513	0.005444	1.023671	0.047252
20	(car_YES)	(pep_YES)	0.493333	0.456667	0.230000	0.466216	1.020911	0.004711	1.017890	0.040427

These rules have the highest confidence and also the highest lift. The support for these is also above 0.2. Now, let's analyze the rule 1 above. In rule 1, the support is 0.215 which means 21.5% (0.215) of the dataset has both "current_act_YES" and "age category_30 to 60" as antecedents and "pep_YES" as the consequent. The confidence is approximately 0.4886, which means that when a person has a current account and their age is between 30 to 60, there's a 48.86% chance that they will apply for PEP. Lift is a measure of how much more likely the PEP is to occur when "current_act_YES" and "age category_30 to 60" are present compared to when they are not and since the value is greater than 1, it suggests that there is a positive influence.

Based on the above evidence, it can be concluded that the target audience for PEP would be people who have a current account, and their age is between 30 to 60. From our second rule, it is shown that men are more likely to apply for PEP. Our third rule states that anyone between the ages of 30 to 60 should be targeted. Moreover, people with an income wage between 20000 to 45000 in that age group would also be interested in applying for PEP. Additionally, the people who have cars are likely to apply for PEP. These people should be targeted since they are most likely to buy PEP. `

"I certify that this assignment represents my work. I have not used any unauthorized or unacknowledged assistance or sources in completing it, including free or commercial systems or services offered on the internet."