

HW 03 – Association Rules

Introduction

Businesses are always looking to grow. Rather they're trying to increase customers, revenue, or another lucrative metric a business' main goal is to spread and prosper. However, the main issue with continuous growth is that it requires resources. A business must balance how much to sacrifice in order to increase, and how long it will take before the advantages of this investment outweigh the disadvantages (ROI). With a myriad of different factors contributing to a successful investment, how can a business be sure they are making the best decision? Luckily, in this present day and age there is data analytics that can help make informed decisions.

In the banking sector, which deals with a wealth of assets, the allocation of funds into fruitful investments is their bread and butter. Data analytics can help make informed decisions on customer support, risk assessment, and uncovering new profit opportunities like diversification and lowering time-to-market.

Banks have direct access to a wealth of historical data regarding customer spending patterns. They know how much money was deposited as salary any given month for their customers, how much went to savings, how much went to utilities, etc. This provides an excellent groundwork for more sophisticated analysis.

Once the initial analysis of customer spending patterns and preferred transaction channels is complete, the customer base can be segmented according to several appropriate profiles such as: easy spenders, cautious investors, deadline rush returners, etc. Knowing the financial profiles of all customers and even noncustomers helps the bank evaluate the expected spending and income next month and make detailed plans to secure the bottom line and maximize income. Data analytics is a tool that when used properly can provide massive dividends to any corporation it is involved with.

Analysis and Models

About the Data

Bank data comprised of 600 observations of 12 variables was analyzed to determine if future customers would want to obtain the new PEP (Personal Equity Plan) based on their demographics and banking information. The comma separated values (CSV) text document contained 7 demographic attributes:

1. Age
2. Sex
3. Region where they live (i.e. Inner City, Rural, Suburban, and Town)
4. Income

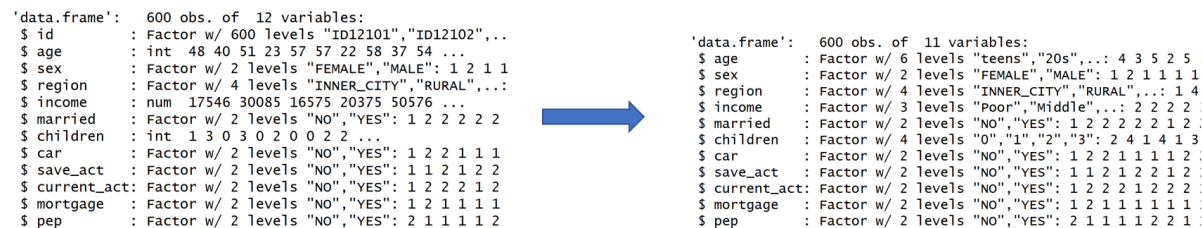
5. Marital Status
6. Number of Children
7. Car Ownership

It also contained the bank's personal information on the customer:

1. Customer ID
2. Do they have a savings account?
3. Do they have a current account?
4. Do they have a mortgage?
5. Have they purchased a PEP?

These 12 attributes allow for a wide scope of Association Rule discovery that will act as a strong guide in predicting if future customers would be interested in the equity plan.

Once the data was ported into the R studio environment, all features were checked for proper character assignment and utility (*Figure 1.*). The first feature 'id' is the bank's internal primary key that is assigned to each individual customer. This feature provides no useful information to the analysis and is therefore removed from the data-set.



```
'data.frame': 600 obs. of 12 variables:
 $ id      : Factor w/ 600 levels "ID12101","ID12102",...
 $ age     : int  48 40 51 23 57 57 22 58 37 54 ...
 $ sex     : Factor w/ 2 levels "FEMALE","MALE": 1 2 1 1
 $ region  : Factor w/ 4 levels "INNER_CITY","RURAL",...:
 $ income  : num  17546 30085 16575 20375 50576 ...
 $ married : Factor w/ 2 levels "NO","YES": 1 2 2 2 2
 $ children: int   1 3 0 3 0 2 0 0 2 ...
 $ car     : Factor w/ 2 levels "NO","YES": 1 2 2 1 1 1
 $ save_act: Factor w/ 2 levels "NO","YES": 1 1 2 1 2 2
 $ current_act: Factor w/ 2 levels "NO","YES": 1 2 2 1 2 2
 $ mortgage: Factor w/ 2 levels "NO","YES": 1 2 1 1 1 1
 $ pep     : Factor w/ 2 levels "NO","YES": 2 1 1 1 1 2

'data.frame': 600 obs. of 11 variables:
 $ age     : Factor w/ 6 levels "teens", "20s",...: 4 3 5 2 5
 $ sex     : Factor w/ 2 levels "FEMALE","MALE": 1 2 1 1 1 1
 $ region  : Factor w/ 4 levels "INNER_CITY","RURAL",...: 1 4
 $ income  : Factor w/ 3 levels "Poor","Middle",...: 2 2 2
 $ married : Factor w/ 2 levels "NO","YES": 1 2 2 2 2 1 2
 $ children: Factor w/ 4 levels "0","1","2","3": 2 4 1 4 1 3
 $ car     : Factor w/ 2 levels "NO","YES": 1 2 2 1 1 1 2
 $ save_act: Factor w/ 2 levels "NO","YES": 1 1 2 1 2 2 1
 $ current_act: Factor w/ 2 levels "NO","YES": 1 2 2 1 2 2 2
 $ mortgage: Factor w/ 2 levels "NO","YES": 1 2 1 1 1 1 1
 $ pep     : Factor w/ 2 levels "NO","YES": 2 1 1 1 1 2 1
```

Figure 1. Data Cleaning

Next issue is the non-nominal feature types in the data-set. To run the AR mining model "apriori" in R all inputs must be converted to nominal features (whether an item occurs in a transaction or not). The "age" feature is listed above as a character type *int* and "income" is a character type *num*. Both of these parameters can be discretized by setting up business relevant classes/groupings for different ranges within the feature. Looking at age:

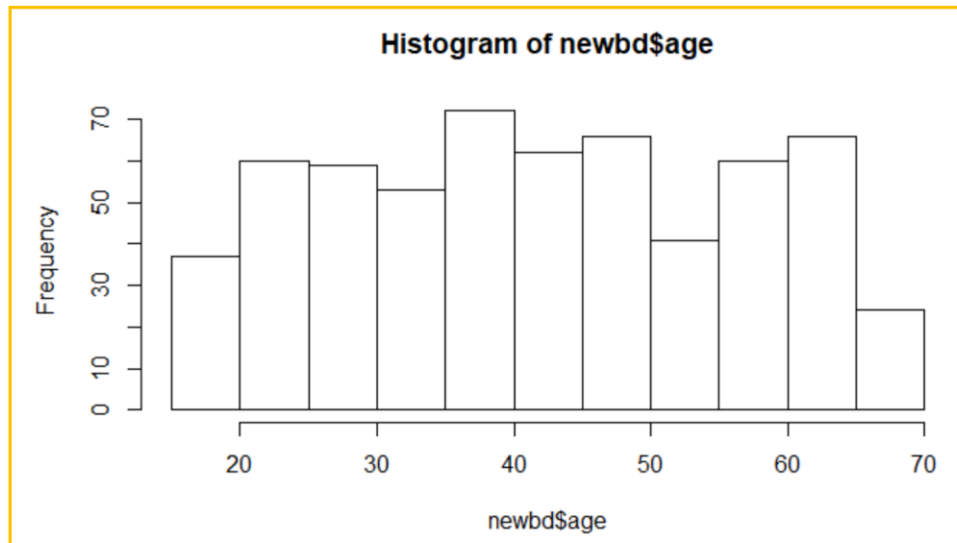


Figure 2. Age Histogram

The histogram partitions the age feature into 6 distinct groups: adolescents (0-20), young adults (20-30), thirties, forties, fifties, and sixties. While this is a good initial guess; does it provide enough granularity into the younger sector? For instance, did parents open PEPs for their children? Taking a deeper dive into this feature reveals that there are no customers under 18 years old.

```
> newbd[newbd$age < 18 , ]
<0 rows> (or 0-length row.names)
```

Figure 3. Age Investigation

This is because PEPs are investment plans introduced in the U.K. that only allowed customers over the age of 18 to participate¹. Age groupings were then formalized as teens, twenties, thirties, forties, fifties, and sixties.

Income plotted as a histogram (Figure 4.) reveals the income range to be between £5,000 and £65,000.

¹ <https://www.investopedia.com/terms/p/pep.asp>

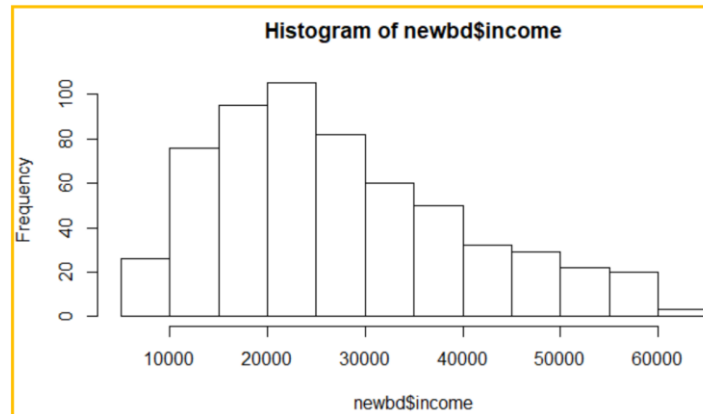


Figure 4. Income Histogram

The income tax rates on the United Kingdom government website as of April 2019 reveals three different tax brackets this data-set falls into which has been prescribed as the "Poor", "Middle", and "Affluent" classes.²

Band	Taxable income	Tax rate
Personal Allowance	Up to £12,500	0%
Basic rate	£12,501 to £50,000	20%
Higher rate	£50,001 to £150,000	40%

Figure 5. British tax brackets

The last feature to address is the "children" attribute which was set to nominal values ranging from 0-3. After all features were converted into nominal values, all binary outputs were converted from "YES" to "[variable_name]=YES" to improve readability of the AR mining output. Data-set was checked for randomness by plotting the "pep" feature (Figure 6.). In a balanced randomized data-set there should be roughly equal amounts of good and bad outcomes.

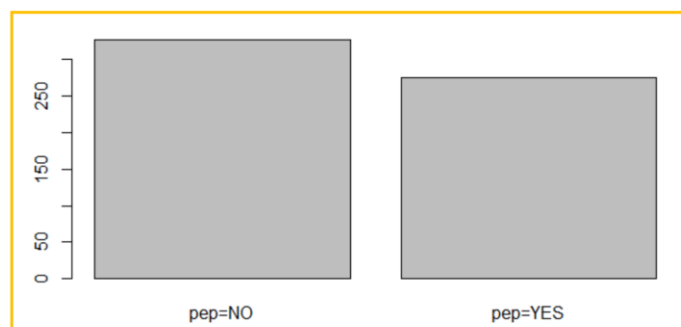


Figure 6. Output Distribution

² <https://www.gov.uk/income-tax-rates>

Apriori Model

Association Rules (AR) mining is a form of unsupervised learning that can be modeled in R with the apriori algorithm. It takes in transactional data and outputs rules that are correlated based on distinct features. In order to extract value from this model there are 4 key features that must be understood and specified:

1. Support \rightarrow the fraction of which our item set occurs in our dataset over total number of transaction
2. Confidence \rightarrow probability that a rule is correct for a new transaction if only given the input features
3. Lift \rightarrow the ratio by which the confidence of a rule exceeds the expected confidence
4. Length of rule \rightarrow Rules of varying itemset length can be set by either a “maxlen” or “minlen” value

Based off the above definitions, it is assumed that the rarest and most valuable rules will be ones with a Lift greater than 1. A Lift of this magnitude represents a strong overlap/association between the inputs on the left-hand side (LHS) and the outputs on the right-hand side (RHS) of the item-set. Meaning that when the LHS of the itemset is present it is more than likely that the RHS will be as well.

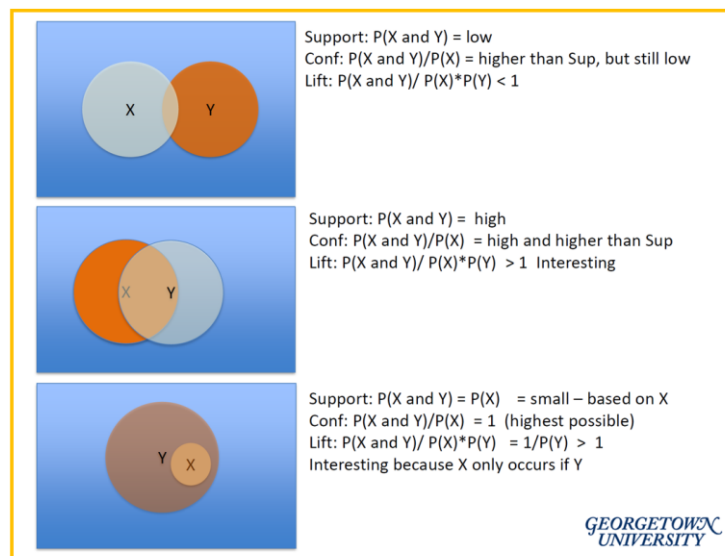


Figure 7. Correlation Parameters

3

Ideal rule query would search the data-set for strongly correlated rules (Lift), then filter for most Confident rules, and lastly pick out the ones with the highest Support. However, the algorithm does not allow for filtering by lift initially. To circumvent this the master rule-set was built using a high Confidence (> 0.9), Low support (> 0.001), and a low minimum set size (> 3) to enable a large sum of rules to be created. This broad approach created 843,180 rules that can now be filtered based on the business objective.

³ Association Rule Mining with Tweets: Thinking Outside the Basket (Gates, 2018)

Results

Based off the reasoning in the previous section the first iteration of filtering the rules was based on maximizing Lift then Confidence. (Figure 8.)

```
> inspect(sorted_rules[1:30])
```

	lhs	rhs	support	confidence	lift	count
[1]	{region=SUBURBAN, income=poor, children=1}	=> {age=teens}	0.00167	1	16.2	1
[2]	{region=SUBURBAN, income=poor, save_act=save_act=NO}	=> {age=teens}	0.00167	1	16.2	1
[3]	{region=TOWN, income=poor, children=3}	=> {age=teens}	0.00167	1	16.2	1
[4]	{income=poor, mortgage=mortgage=YES, pep=pep=YES}	=> {age=teens}	0.00333	1	16.2	2
[5]	{region=SUBURBAN, income=poor, children=2, pep=pep=YES}	=> {age=teens}	0.00167	1	16.2	1

Figure 8. 1st Filter Iteration (Lift, Confidence)

While these rules have extremely high Lift; on closer inspection the support on these rules is exceptionally small. Of the 30 rules inspected the "count" of transactions that support each rule is never higher than 2. In a data-set of 600 transactions, a rule based off 2 transactions is not significant enough to rely upon. Replacing Confidence with Support (Figure 9.) only yielded a "Count" between 2 and 4.

```
> inspect(sorted_rules2[1:30])
```

	lhs	rhs	support	confidence	lift	count
[1]	{sex=MALE, income=poor, married=married=NO, car=car=NO, save_act=save_act=YES}	=> {age=teens}	0.00667	1	16.2	4
[2]	{sex=MALE, income=poor, married=married=NO, car=car=NO, save_act=save_act=YES, current_act=current_act=YES}	=> {age=teens}	0.00667	1	16.2	4
[3]	{income=poor, children=3, current_act=current_act=YES, mortgage=mortgage=NO}	=> {age=teens}	0.00500	1	16.2	3
[4]	{region=TOWN, income=poor, save_act=save_act=YES, mortgage=mortgage=NO}	=> {age=teens}	0.00500	1	16.2	3
[5]	{income=poor, car=car=NO, save_act=save_act=YES, pep=pep=YES}	=> {age=teens}	0.00500	1	16.2	3

Figure 9. 2nd Filter Iteration (Lift, Support)

After these first two iterations it become clear that Lift cannot be the initial parameter maximized. Lift is the ratio of Confidence over the Expected Confidence; it describes the increase in probability that the RHS of the rule will appear given the LHS. While this parameter highlights the most significant rules, it falls prey to uniqueness. Rules that are so unique that there are only 1 or 2 transactions in the data-set. These rules do not hold enough weight (Support) to be deemed valuable to the business.

Due to this new learning all subsequent iterations of filtering ranked Support higher than Lift. After a few more trials of exploratory development the hierarchy of Support, Lift, and then Confidence was chosen (Figure 10.). High Support ensures that a large enough sample size has been considered for the rule to make business sense; high Lift ensures that there is a dependence between the LHS and RHS; high Confidence ensures that the rule suggested is the most likely outcome when given the same LHS.

```
> inspect(sorted_rules5[1:30])
```

	lhs	rhs	support	confidence	lift	count
[1]	{married=married=YES, children=0}	=> {income=Middle}	0.270	0.900	1.08	162
[2]	{children=0, car=car=YES}	=> {income=Middle}	0.187	0.903	1.08	112
[3]	{children=0, mortgage=mortgage=NO, pep=pep=NO}	=> {married=married=YES}	0.173	0.972	1.47	104
[4]	{age=40s, current_act=current_act=YES}	=> {income=Middle}	0.165	1.000	1.20	99
[5]	{children=1, pep=pep=YES}	=> {income=Middle}	0.165	0.900	1.08	99

Figure 10. Final Filter Iteration (Support, Lift, Confidence)

With the hierarchy set, the attention turned to creating rules that will predict if a customer will purchase a PEP given their demographic and customer information. This is done by only viewing rules that have “pep=YES” on the RHS (Figure 11.)

```
> inspect(yesPEPrules_sort[1:30])
```

	lhs	rhs	support	confidence	lift	count
[1]	{income=Middle, children=1, save_act=save_act=YES, current_act=current_act=YES}	=> {pep=pep=YES}	0.0917	0.902	1.97	55
[2]	{income=Middle, married=married=YES, children=1, current_act=current_act=YES}	=> {pep=pep=YES}	0.0833	0.909	1.99	50
[3]	{sex=FEMALE, income=Middle, children=1}	=> {pep=pep=YES}	0.0767	0.920	2.01	46
[4]	{married=married=NO, children=0, mortgage=mortgage=NO}	=> {pep=pep=YES}	0.0750	0.938	2.05	45
[5]	{married=married=YES, children=1, save_act=save_act=YES, current_act=current_act=YES}	=> {pep=pep=YES}	0.0733	0.917	2.01	44

Figure 11. Customers that purchased a PEP

In terms of the number of features in a rule; less is more. Even thou the algorithm has generated rules consisting of up to 10-items it appears that rules of approximately 4-6 items yielded the best results. Notably, parents in the middle-income bracket with 1 child are extremely likely to take out a PEP. The single child middle-income combo accounts for 113 out of 600 transactions (~18%) in the data-set which is a small enough subgroup to confidently say there is a correlation. Perhaps parents with only one child have enough dispensable income to make this type of investment. Or, maybe only having 1 child is an artifact of smart crafty investors realizing how expensive children are and settling for a minimal amount. It could also be that once an adult becomes a parent, they start taking their finances more seriously and want to create a safety net for the future. More psychoanalysis into the customers motivations are needed before coming to a concrete conclusion.

The top 5 most actionable rules from *Figure 11*. that predict a customer will purchase a PEP are shown below:

```
> inspect(yesPEPrules_sort[c(1,4,18,26,30)])
```

	lhs	rhs	support	confidence	lift	count
[1]	{income=Middle, children=1, save_act=save_act=YES, current_act=current_act=YES}	=> {pep=pep=YES}	0.0917	0.902	1.97	55
[2]	{married=married=NO, children=0, mortgage=mortgage=NO}	=> {pep=pep=YES}	0.0750	0.938	2.05	45
[3]	{region=TOWN, income=Middle, children=1}	=> {pep=pep=YES}	0.0517	0.939	2.06	31
[4]	{region=INNER_CITY, children=1, save_act=save_act=YES, current_act=current_act=YES}	=> {pep=pep=YES}	0.0450	0.900	1.97	27
[5]	{age=40s, income=Middle, children=1, save_act=save_act=YES}	=> {pep=pep=YES}	0.0433	1.000	2.19	26

Figure 11. Most Interesting Rules

These 5 rules all have a Lift of 1.97 or more, a Confidence of 0.9 or more, and apply to at least 26 transactions in the data-set. Given that there are only 274 transactions with customers who bought PEPs, that's at least a 9.5% inclusion of all transactions in each rule.

Rule #1: When a customer is from the Middle Class, has 1 child, a savings account, and have an active account with the bank there is a strong possibility that he/she may take out an equity plan. Support (0.0917) is the highest for this rule and it is calculated as the ratio of the number of transactions that include all items in the rule over the total number of transactions in the database. This means that **Rule #1** had the most examples of the rule occurring in the data-set compared to the other four rules. The Confidence (0.902) is a conditional probability that a randomly selected transaction will include all the items on the RHS, given that the transaction includes all the items in the LHS. In **Rule #1** there is a 90.2% chance of that occurring. Lift (1.97) is the ratio of Confidence over the Expected Confidence; it describes the increase in probability that the RHS of the rule will appear given the LHS. A Lift higher than 1 signifies the Confidence of a rule is actually more than the Confidence of the RHS only.

The most useful part of this rule lies in its gender and marital neutrality which creates a much larger pool of potential customers to market the PEP to. This rule generally states that, if an average income parent is an active customer, they will be likely to buy the PEP regardless of where they live, how old they are, or if they're married or not. This pool of customers appears to be middle class workers who are financially conscious. They are active members in the bank and are already saving money for the future. Perhaps to increase their family size.

Rule #2: When a customer is not married, has 0 children, and isn't currently paying a mortgage there is a strong possibility that he/she may take out an equity plan. This rule has the fewest number of items on the LHS and still performs well. The biggest draw to implementing this rule is targeting the young adult pool. This rule dictates that regardless of age people who have less responsibilities (i.e. not married, no kids, and no house-note) can be easily persuaded into purchasing a PEP. Maximizing the

benefits of this rule would work best when customers of this segment are in the bank, as people with less responsibilities are more susceptible to impulse buys. Employees should be trained on how to specifically address the “Wild and Free” demographic when making their sales pitch.

Rule #3: When a customer lives in town, is part of the Middle class, and has 1 child there is a strong possibility that he/she may take out an equity plan. This rule does a beautiful job making the life of the marketing department easy. There is no distinction about male or female, marital status, or even being an active customer to the bank. This marketing strategy should mainly coincide with telephone calls and mail services to all residents in nearby towns since this demographic is so large.

Rule #4: When a customer lives in the inner city, has 1 child, a savings account, and is a current active member of the bank there is a strong possibility that he/she may take out an equity plan. As with **Rule #3** this rule allows the marketing department to do a shotgun approach when soliciting customers to buy the PEP. Instead of just focusing on the slower and smaller town sector, they can split their time with the “City Hustlers” who, like their small town counterparts, have children but also are more active in the banking scene. Telephone calls and mail services would still be the primary form of communication with less emphasis on telephones as busy people are more likely to hang up on telemarketers. Urban designed/themed postcards should be a staple in grabbing the attention of these busy people.

Rule #5: When a customer is in their 40s, middle class, has 1 child, and has a savings account there is a strong possibility that he/she may take out an equity plan. The most useful part of this rule lies in its gender and marital neutrality while still specifically pointing out customers in their forties. Perhaps this age group is worried about the future and how much they are going to be able to save for retirement. These “Late Starters” are dispositioned to saving as much as they can as fast as they can. The most effective sales pitch for this group would be face to face as they are older and need more assurance that the product being offered will be beneficial in the long run. Employees should be trained to act confidently and highly knowledgeable when making a sales pitch to this demographic.

Conclusion

Using AR mining in R with the apriori algorithm allowed the bank data-set to be segmented based on differences in demographics and banking information. The most viable asset to this method was being able to filter the plethora of association rules that were generated from the model using a few key features. 5 distinct rules of interest were found by reviewing the top 30 rules based on filtering by highest Support, then Lift, and finally Confidence. These 5 rules were chosen based on their statistical performance as well as their actionability in the business sector. For instance, all of these rules are sex agnostic and four of the five do not divide customers based on their marital status. With those two factors being two of the main groupings in society, circumventing them allows the marketing team to target a much broader population.

Classifying the customers into different segments will allow for more efficient marketing campaigns and provide a hot bed for deeper analysis. For instance, the Wild and Free, City Hustlers, and Late Starters are all sociodemographic segments that can easily be identified in the bank’s database. But, what’s their motivation behind seeking a PEP? Are they planning on growing their family? Are they worried about

their retirement? Do they just have a lot of excess money? There are so many more psychographic nuances that can be properly explored within each segment now that their broad behavior has been identified.

In conclusion, this analysis has laid the foundation for segmenting the customer base in a way that provides the most value back to the business. Rather it's deciding on the marketing budget or estimating how many people will buy the PEP next quarter or next year, this model/method can be utilized to make sure the business objectives align with the actual data and past performance of the business.