**Introduction**

Effective marketing for the firm is an ongoing sport, it would seem. Ensuring the proper group of existing and prospective clients receives marketing material, and in a way that is conducive to their reception and consideration is an evolving process. For example, sending an email to someone over 80, or recommending a home loan to a freshman in college are both likely to be unsuccessful campaigns. And the firm can obviously not continue without effective methods to draw new business, from both new clients and existing clients.

Recently, the firm began offering a personal equity plan (PEP going forward) similar to that introduced in the UK in 1987. In order to inform current and prospective buyers, the marketing department sent 600 mailers to various people throughout the city promoting the new product. This was done not just to drive business, but as a marketing exercise to determine guidelines regarding who would be a good candidate for future purchases of the firm’s new PEP.

This report is an examination of the demographic data collected by the marketing department. The goal is to determine which demographic factors contribute to the likelihood of a client purchasing the new PEP, so that the guidelines described above can be established and more effective marketing campaigns can be initiated.

**Section 1: Analysis and Models**

**Section 1.1: About the Data**

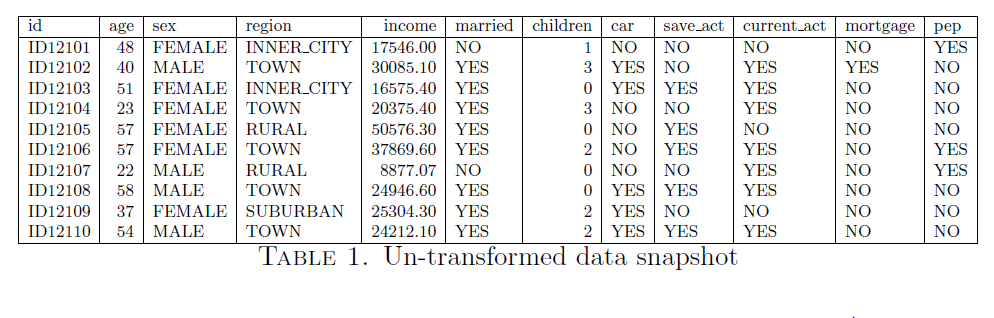
As mentioned, the marketing department has provided demographic data of everyone that received a PEP mailer last year. The dataset contains the following information for each of the 600 individuals:

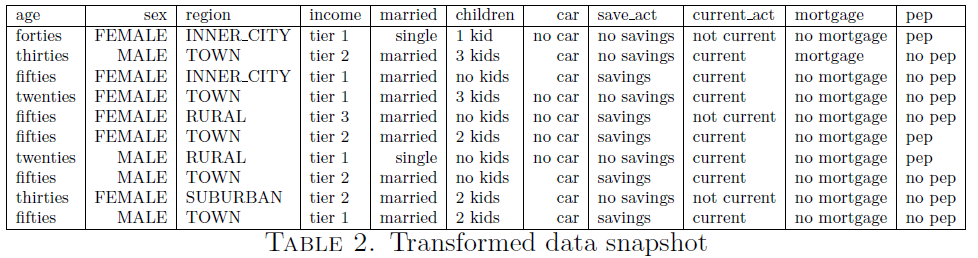
1. ID – reference id number
2. Age – customer age measured in years
3. Sex – male/female
4. Region – inner city/rural/suburban/town
5. Income – customer income
6. Married – is the customer married (yes/no)
7. Children – number of children
8. Car – does the customer own a car (yes/no)
9. Save\_act – does the customer have a savings account (yes/no)
10. Current\_act – does the customer have a current account (yes/no)
11. Mortgage – does the customer have a mortgage (yes/no)
12. Pep – did the customer buy a PEP after the last mailing (yes/no)

Since the goal of the investigation is to see which combination of the above characteristics contributes to the purchase of a PEP, a technique called association rule mining was implemented (more on this is Section 1.2). To do this, certain adjustments to the data had to be made.

First, the entire dataset was screened for missing values, of which there were none. Next, the variable ID was removed as it is purely an identifier and a client’s ID number has no impact on whether they purchased a PEP, or any product for that matter. The age variable was broken into categories that reflect the client’s decade of life, e.g. a client that was 27 years old would be identified as “twenties”, a 43-year-old client would be “forties”, etc. Similarly, the income variable was broken into three groups based on the range of values of the variable. More precisely, if *m* = min(income) = $5,014.21 and *M* = max(income) = $63,130.10, then the income variable has a range of *M* – *m*= $58,115.89, and the resulting interval size by which the income values are subdivided is $58,115.89/3 = $19,371.96. Any client whose income was less than $19,371.96 was identified as “tier 1” income, likewise any client whose income was between $19,371.97 and $38,743.96 was “tier 2” and anyone whose income was above $38,743.97 was “tier 3”.

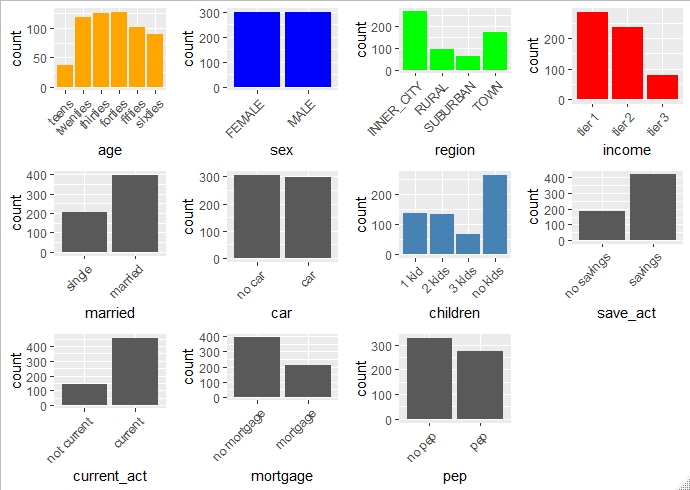
The children variable needed to be converted to an ordinal variable to a nominal variable, in other words the numbers needed to be… not numbers insofar as no calculation was to be done with the values, so the number was merely a label for a category. A value of 0 in the children column was relabeled as “No kids”, a 1 as “1 kid”, etc. Finally, each of the remaining variables with yes/no response options were recoded to reflect what the yeses and noes were indicating. For example, for the save\_act variable “yes” became “savings” and “no” became “no savings”. Here is a small snapshot of the data before and after these transformations:





In this form, the data can be thought of as transaction data. Transaction data is exactly what it sounds like, a collection of transactions where each instance is a separate collection of items, but in this case instead of milk, eggs, and toilet paper the items in a transaction are demographics. For example, the first “transaction” in the dataset is for a single mother of one in her forties who lives in the inner city, is in the low-income bracket, with no car, savings, mortgage, or current account, but did buy a PEP. The goal of association rule analysis (again, more on this in section 1.2) is to find which sub-collections of items or in this case characteristics lead to others, and having the data configured in this way allows for this to be done without losing context.

While categorical variables of this type do not necessarily have summary statistics, it is worthwhile to see the frequency of each characteristic, as seen in the plots below. While some interesting observations can be made (say, for example, that there was a roughly equal number of men and women in the dataset, or that there were more inner-city residents than suburbanites), it is which of these or combination of these most frequently led to PEP purchase that is of interest.

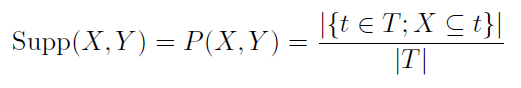


**Section 1.2: Analysis**

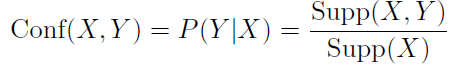
Before getting to the meat of the analysis, a brief discussion of association rule mining is necessary for the benefit of those that are not familiar. As mentioned in section 1.1, association rule analysis was used to determine which demographic information led to PEP purchases.

An association rule is a relationship between one group of items (or in this case characteristics) to another. For example, there could be a relationship between the set {no kids, mortgage, car} and {married, PEP}. Each association rule, and each itemset, has three levels of measure that can help determine its quality. Suppose that *I* is the set of all items, *T* is the set of all transactions, and *X* and *Y* are subsets of *I*. Then define

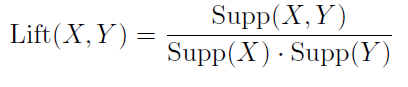
1. Support – this is the probability that the rule occurs



1. Confidence – this is the probability of the resulting set being in the transaction given that the original set is

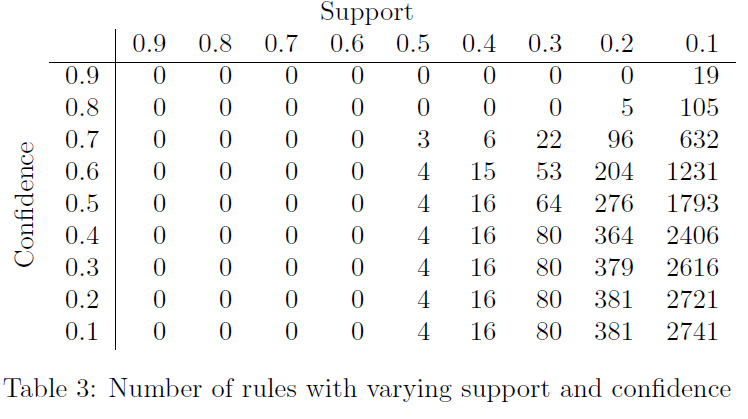


1. Lift – this compares the probability of them occurring together to the probability of them both occurring but separately



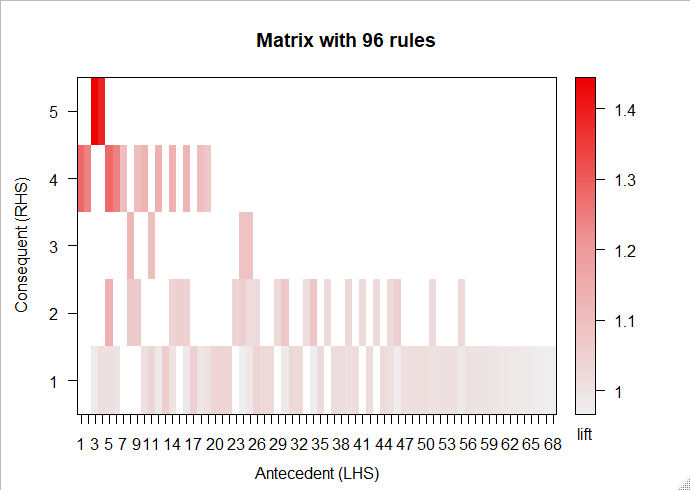
The main technique is to look for rules that have high levels of these measures. However, there is no rule for what is “high” in this case, so different thresholds are set and used to see what interesting (translated: sufficiently high support, confidence, and/or lift) are determined.

To get an idea of where to start the search for good support and confidence thresholds, a number of rule sets were generated with varying levels of support and confidence. Table 3[[1]](#footnote-1) shows how many rules were generated when support and confidence ranged from 0.9 to 0.1.

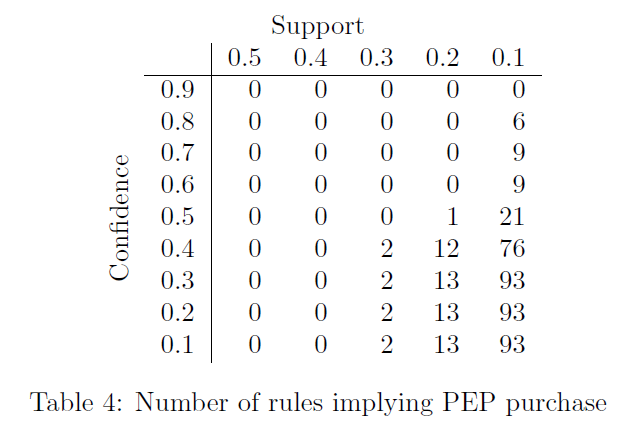


From these values, it is clear that the search can be limited to support values below 0.5. However, to get a decent collection the support should be set closer to 0.2 to allow for higher confidence levels. As an example of what can be learned from this, there is a rule with 0.21 support and 0.74 confidence indicating that living in town is correlated to a savings account, with lift greater than 1. However, there is another rule with identical support and confidence relating town to current account, but with lift less than 1, which indicates that living in town actually reduces the likelihood that a client has a savings account. This is why lift is important, so that seemingly good rules are not relied on to determine implications.

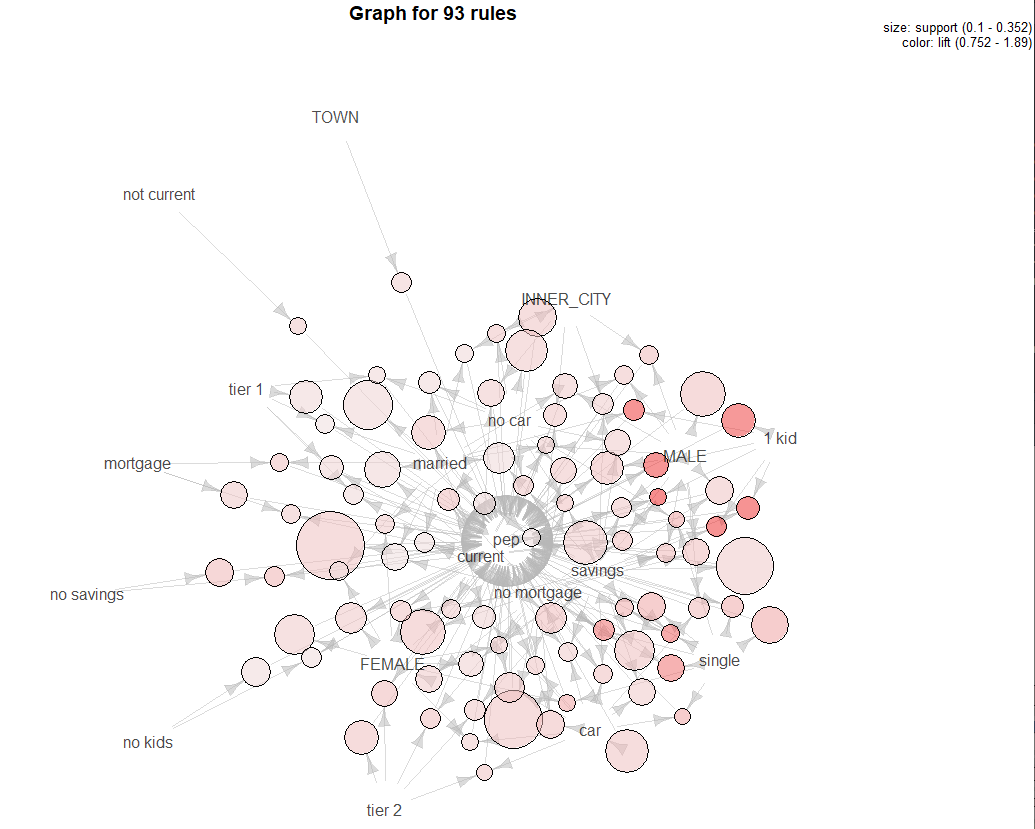
Consider further the 96 rules with support greater than 0.2 and confidence greater than 0.7. The following plot groups the rules by which item appears as the consequent (right hand side, or the *Y* in the rule *X* -> *Y*) for each rule. The five possible consequents are Current account, Savings account, No Mortgage, Married, and no PEP, in that order. There are no indicators, but that’s ok, as the interest here is in which items appeared as consequents. Notice that No PEP only appeared twice, but with very high lift in each instance.



From the outset, the goal of this study was to determine the demographic characteristics that were indicative of whether a client would purchase a PEP. To that end, the next step in the analysis was to focus the association rule mining to only those rules where PEP was the consequent or right-hand side. A similar approach was taken to ascertain good starting points for support and confidence, but since there were no rules whatsoever with support greater than 0.6, the search was limited to support values below 0.6. The number of rules generated for varying support and confidence levels with PEP as the focused consequent are in table 4.



To get an idea of what the rules consisted of, consider the following plot showing each rule. The nodes are the characteristics themselves, and the arrows show the direction of the relationship.



This shows not only the complexity of how the demographic characteristics influence PEP purchases individually and as groups, but also quickly identifies the rules with high lift. Notice the cluster of high lift rules stemming from the 1 kid characteristic. Indeed 6 of the highest lift rules have 1 kid in the antecedent (left-hand side). This begs further investigation, which is discussed in Section 2.

**Section 2: Results**

As shown in section 1.2, there were 93 rules with PEP as the consequent. The most common rule was {Current} -> {PEP}, occurring 211 times. This rule had support 0.35 and confidence 0.46, however its lift was only 1.02, indicating that while 211 of the 274 clients who did purchase a PEP have a current account, this is not a strong enough indicator on its own. A better rule was {1 kid, current, savings} -> {PEP}. This rule had support 0.11, confidence 0.86, and lift 1.89, and occurred 63 times. This is a much better indicator of PEP purchasing as it allows for better targeted marketing than just blasting all current account holders (though there is no harm in doing so). Somewhat surprisingly, the rule {married, no mortgage, savings} -> {PEP} had support 0.11, confidence 0.35, and lift 0.76. A lift value less than 1 indicates effectively a negative correlation, so being married with a savings account and no mortgage actually has a negative effect on the likelihood of purchasing a PEP, thus it is recommended not to focus on this group. In fact, {married} appears in 11 of the 20 rules with lowest lift, so it is additionally recommended to not focus on married individuals.

Two complimentary rules arose, {single, no mortgage, savings} and {single, no mortgage, current}. These rules had support 0.11 and 0.12, confidence 0.74 and 0.72, and lift 1.63 and 1.57, respectively. Moreover, even focusing just on single people with no mortgage produced support 0.15, confidence, 0.71, and lift 1.55. Clearly, unmarried people with no mortgage would be a good group to focus on. Indeed {single} was part of 10 of the 20 rules with the highest lift, so it very highly recommended that those clients who are not married are good target candidates for PEP purchases.

**Section 3: Conclusions**

Obviously, not every one of the firm’s clients is a good candidate for PEP marketing. For example, it was shown in section 2 that married individuals are quite unlikely to purchase the PEP, whereas single individuals did have a high likelihood to purchase. And while predicting human behavior always has uncertainty there are certain guidelines that this brief study has produced, summarized here:

1. Individuals with 1 child are very likely to purchase the PEP, especially those with current accounts and savings accounts
2. Married individuals are very unlikely to purchase the PEP
3. Single individuals, especially those with no mortgage should be targeted
4. All current account holders should receive future PEP marketing

Now, it is important to note that this list is by no mean comprehensive. It would be interesting to see which clients responded to the mailer but ultimately did not purchase the PEP. This could provide further insight into which clients would purchase the PEP in the future. Also, the guidelines are not static, meaning they can be adapted as time goes on and additional campaigns go out.

1. The values were started high because, by their nature, higher support and confidence values are a good thing. [↑](#footnote-ref-1)