Daniel Hanks Jr
IST565 – Assignment Four

This reports purpose is to use available algorithms to accomplish a classification task. The data is in the form of a csv file and contains attributes on people's demographics and banking information on if they participate in a Personal Equity Plan (PEP).

The csv file Id, age, sex, region, income, married, children, car, save_act, current_act, mortgage and pep variables. Variable descriptions are as follows:

Id	ID number
age	Age (Years)
sex	Male/Female
region	Inner City/Town/Rural/Suburban
income	Income (yearly salary)
married	YES/NO
children	Amount of children (numerical)
car	YES/NO
save_act	YES/NO
current_act	YES/NO
mortgage	YES/NO
pep	YES/NO (Personal Equity Plan)

This report will discuss the meaning of support, confidence and lift values while explaining the rules discovered that will help the bank understand its customers better and also to develop new business opportunities. I am going to use Association Rule Learning to do this. The Apriori Algorithm is an algorithm that will do this, and is often used for mining frequent itemsets.

We hope to learn "interesting" rules from this process. An interesting rule is considered a more useful rule because it's unexpected. The lift, support and confidence are all ways to measure how interesting the rule is. The difference is the lift cannot be defined in the algorithm like support and confidence can. Lift is the target response divided by the average response. This will be demonstrated during this process below.

We'll start this process by importing the csv file mentioned above into R, which is a software program that will allow me to run the Apriori Algorithm on the data. It should be noted that I deleted the Id column from the csv because it's not needed for this. I also needed to discretize or approximate the columns for this process. Using the 'arules' package in R, I am able to run the Apriori Algorithm on the dataset I've created. I chose a support of .005 and confidence of .8 to create my first set of rules. The results I got look like this:

```
> rules <- apriori(factor_rules,parameter=list(support=.005, confidence=.8))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5   0.005      1
 maxlen target   ext
     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 2

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[571 item(s), 500 transaction(s)] done [0.00s].
sorting and recoding items ... [71 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 done [0.00s].
writing ... [33718 rule(s)] done [0.01s].
creating S4 object  ... done [0.02s].
```

As you can see towards the bottom of the image, it created 33,718 rules which would take a considerable amount of time to sort through, so I can sort the rules by lift and list the top 20. This gives me a list that looks like this example:

```
> inspect(head(sort(rules,by="lift"),n=20))
     lhs                    rhs          support confidence    lift
[1]  {age=27,
      region=TOWN,
      car=NO}           => {children=3}   0.006          1 8.62069
[2]  {age=27,
      region=TOWN,
      mortgage=NO}      => {children=3}   0.006          1 8.62069
[3]  {age=27,
      sex=FEMALE,
      car=NO}           => {children=3}   0.006          1 8.62069
[4]  {age=27,
      sex=FEMALE,
      region=TOWN,
      car=NO}           => {children=3}   0.006          1 8.62069
[5]  {age=27,
      sex=FEMALE,
      region=TOWN,
      mortgage=NO}      => {children=3}   0.006          1 8.62069
```

The higher the lift, the more interesting the rule and this example gives us the top 5. While we have potentially interesting data here, we want the rules to focus on whether or not people

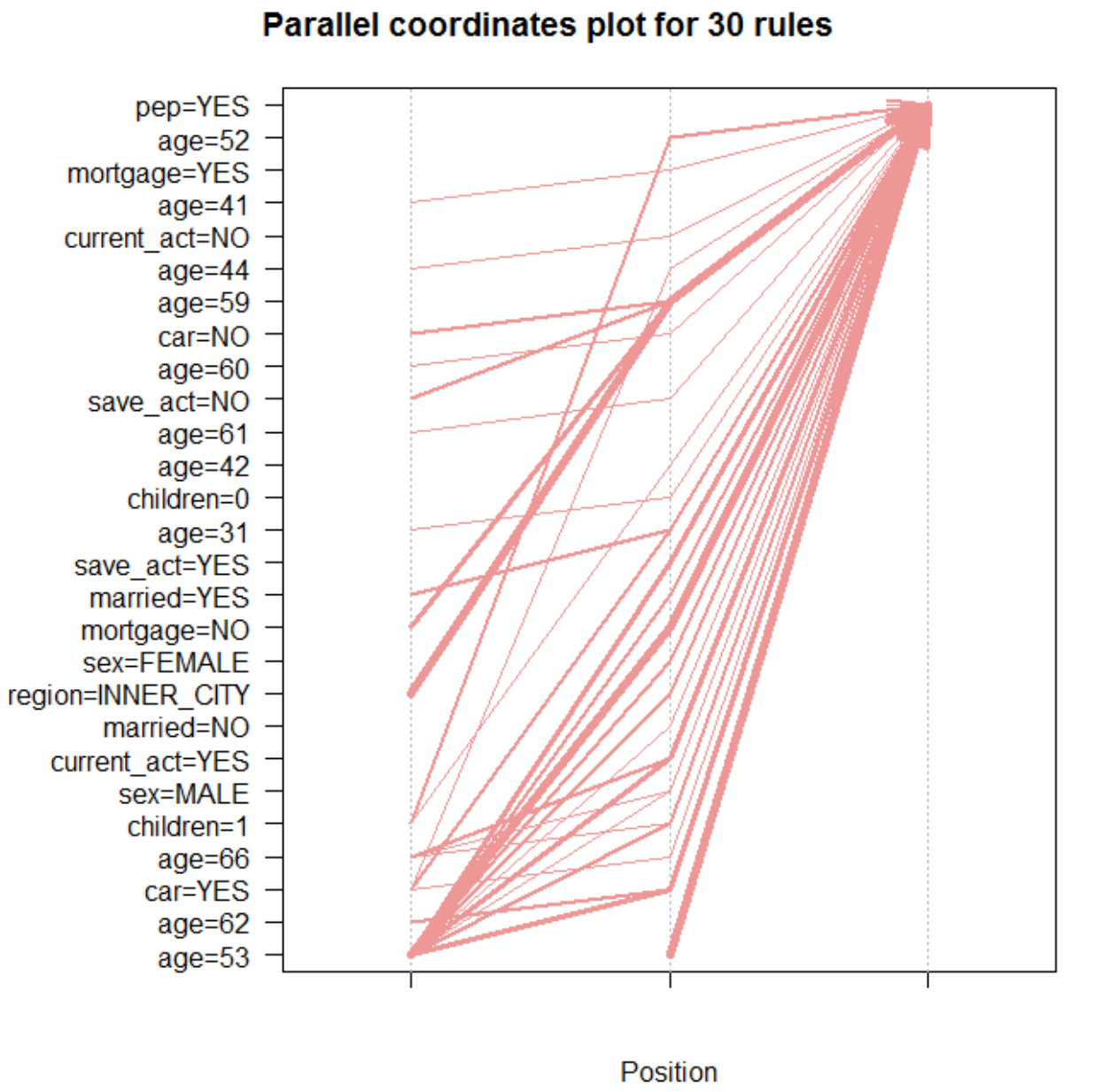
participle in the Personal Equity Plan (PEP). We'll set the rhs column to only look at the pep column. We come up with the following results:

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5   0.005      1
 maxlen target   ext
     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 2

set item appearances ...[2 item(s)] done [0.00s].
set transactions ...[571 item(s), 500 transaction(s)] done [0.00s].
sorting and recoding items ... [71 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 done [0.00s].
writing ... [5315 rule(s)] done [0.00s].
creating S4 object  ... done [0.02s].
```

This is a much more manageable set of rules at 5315. We can inspect the 50 lift rates (top 10 shown for the example):

```
     lhs                                rhs       support confidence lift
[1]  {age=53}                        => {pep=YES} 0.014   1          2.145923
[2]  {age=62,car=YES}                => {pep=YES} 0.008   1          2.145923
[3]  {age=66,children=1}             => {pep=YES} 0.006   1          2.145923
[4]  {age=66,sex=MALE}               => {pep=YES} 0.006   1          2.145923
[5]  {age=66,car=YES}                => {pep=YES} 0.006   1          2.145923
[6]  {age=66,current_act=YES}        => {pep=YES} 0.008   1          2.145923
[7]  {age=53,children=1}             => {pep=YES} 0.008   1          2.145923
[8]  {age=53,married=NO}             => {pep=YES} 0.006   1          2.145923
[9]  {age=53,region=INNER_CITY}      => {pep=YES} 0.008   1          2.145923
[10] {age=53,sex=FEMALE}             => {pep=YES} 0.008   1          2.145923
```

Even off these 10 examples I can tell you male or females over 50 are likely to be enrolled in a PEP. If we want 30 strong rules I could plot them as such:



This re-enforces my belief that anyone over the age of 50 is highly likely to be enrolled in the PEP. Based on this you could go a couple routes with the recommendations to the bank. First, marketing to the older crowd who are more experienced with investing in such financial plans, or may be worried about retiring could be beneficial since they are more apt to looking into such things. Second, maybe you make a conscious effort to reach the younger crowd and tell them the benefits on investing in PEP. This may help increase more PEP investments.