Sean Deery HW 3

# Introduction

The cost of marketing products and services is significant for companies. Banks are not excluded from this cost, as US banks spend billions of dollars on digital ads every year. These marketing campaigns have a significant influence on the performance of the company whether they reach the right market or not. That is why managers do not just rely on their instinct when developing their target audience. Managers have access to more and more data that they can use to gain a better understanding of the customers that end up buying their products.

Banks sell financial products and profit when people decide to invest their money in the bank’s products. To accomplish this, the bank needs to have the right products and market them to people who are in the right stage of life. The ability to define which customers are in that stage of life allows the bank to target marketing dollars at profit-generating consumers.

The following analysis examines data a bank has collected on its customers from a previous marketing campaign. The data will be used to learn about the customer base and what variables were correlated with a purchase. The results will allow the manager to gain insights into who they want to target in future marketing campaigns.

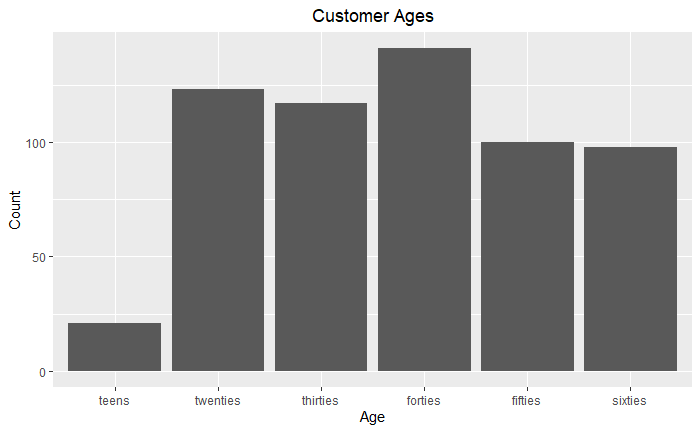
# Analysis and Models

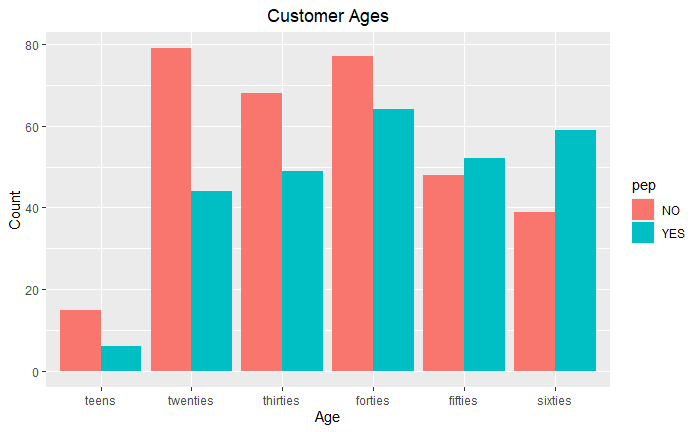
## About the Data

The dataset contains records on customers of a financial institution and whether they responded to a direct mail piece advertising a “Personal Equity Plan” (PEP) and bought the product. There are 12 variables including id, age, sex, region, income, married, children, car, save\_act, current\_acct, mortgage, and pep. The id variable is a unique identifier number for each customer. There are no missing values in the dataset.

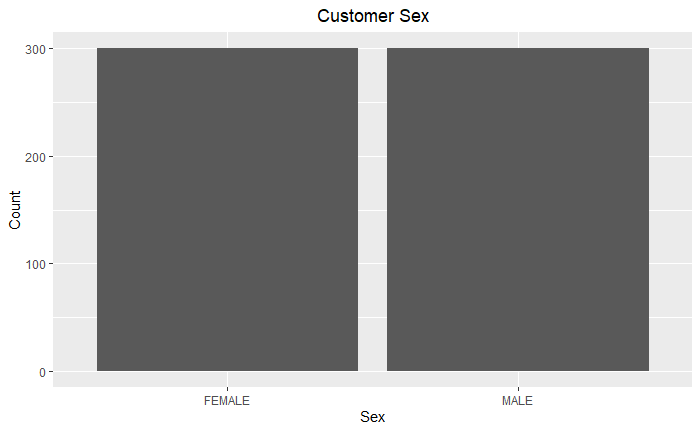
The id variable is removed because it does not hold any meaningful value. The children variable is read in as an integer, so it is converted to an ordinal type. The age variable is read in as an integer, so it is discretized into categories "teens", "twenties", "thirties", "forties", "fifties", and "sixties" and stored as a factor. The income variable is also read in as an integer, so it is discretized into categories "0-9,999", "10,000-19,999", "20,000-29,999", "30,000-39,999", "40,000-49,999", "50,000-59,999", "60,000-69,999". The rest of the variables, including sex, region, married, car, save\_act, current\_act, mortgage, and pep, are read in as characters so they are converted to factors.

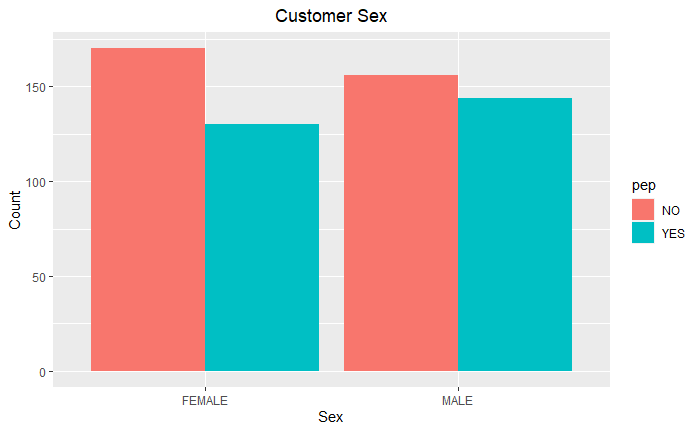
The age variable represents the age of the customer in years. The bar graph shows that there is a relatively small group of teen customers that take up only 3.5% of the customer base. The remaining age categories have a similar number of customers to each other, each having a percentage between 16.3% and 23.5%. The most common age category is a customer in their forties with a total of 141 customers (23.5%). The bar graph separated by whether or not the customers bought the PEP shows that the percentage of customers who bought the PEP increases as the age gets higher. This could be because people are naturally more receptive to the product when they are older, or it could be that younger people don’t respond to mail advertising as much as older people. It makes sense that customers in their twenties who are moving for college or their first job are less likely to be purchasing financial products from a mail advertisement.



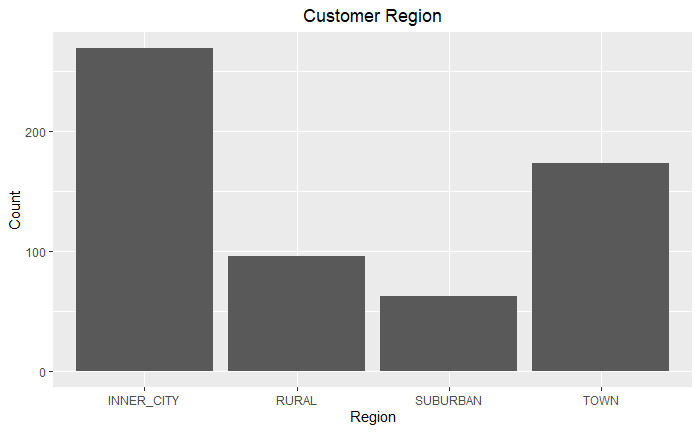


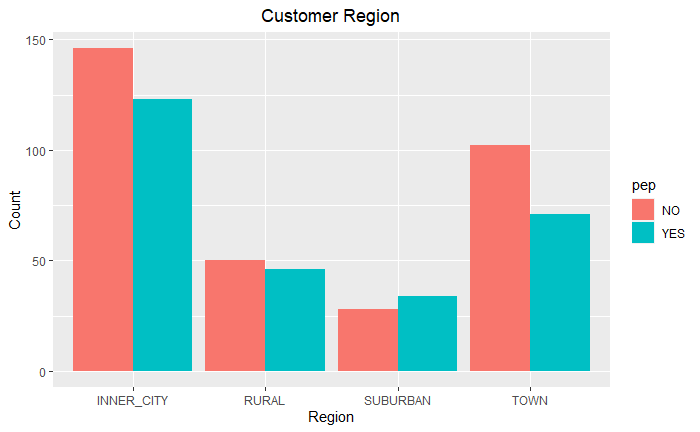
The sex variable represents the customer’s gender, “MALE” or “FEMALE”. The bar graph shows there are an equal amount of male and female customers. The bar graph separated by whether or not the customers bought the PEP shows that a greater percentage of the male customers bought the PEP.

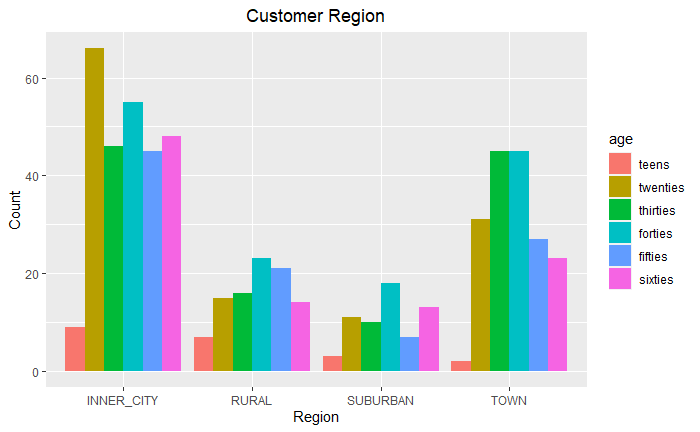




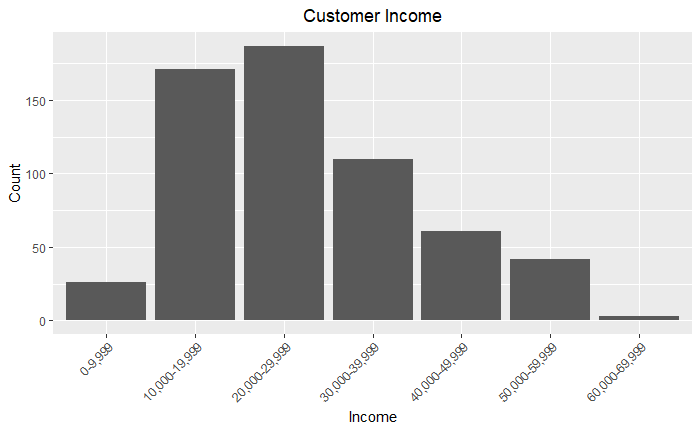
The region variable represents where the customer lives, either “inner\_city”, “rural”, “suburban”, or “town”. The bar graph shows that most of the customers live in the inner-city or a town, representing 73.6% of the customer base together. This makes sense as inner cities and towns are more populated than suburbs and rural areas. The bar graph separated by whether or not the customers bought the PEP shows that a greater percentage of the rural and suburban customers bought the PEP than the inner-city and town customers. This could be because people in suburban and rural areas are more settled and ready to buy financial products, but it could also have something to do with the age groups living in those areas. The bar graph separated by the age groups shows the inner-city has a high amount of customers in their twenties compared to the other areas.

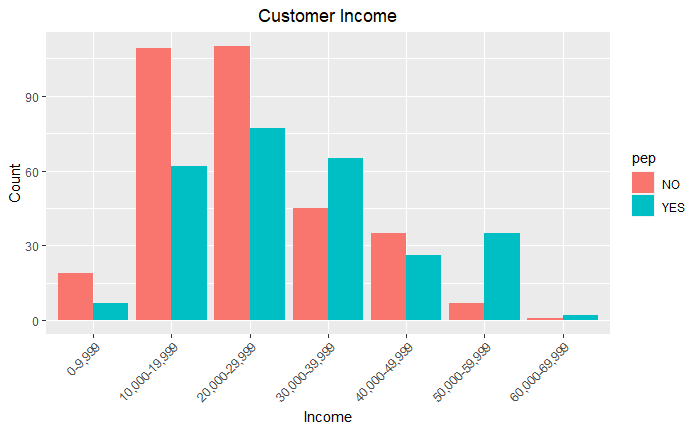




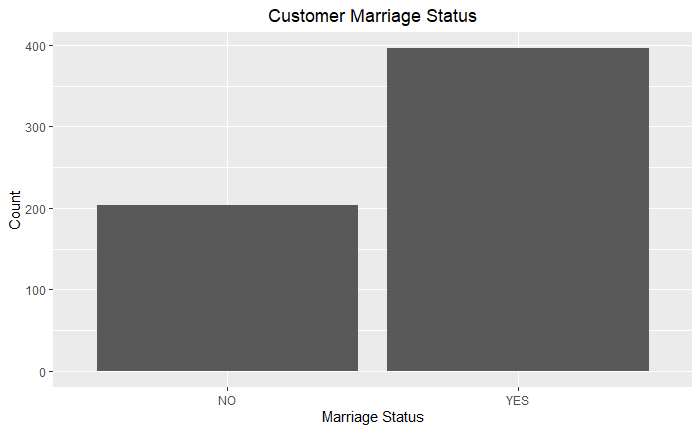


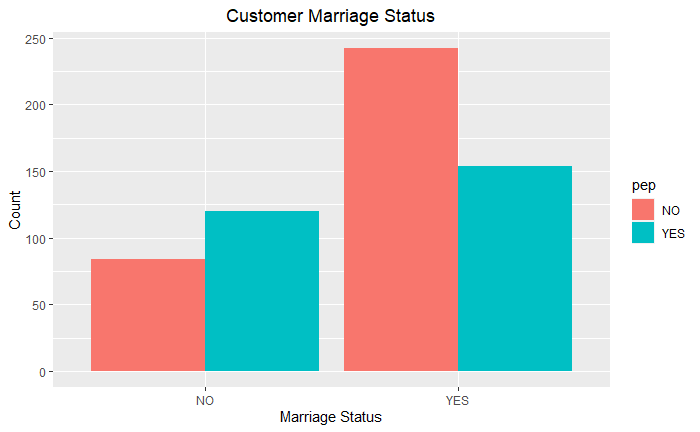
The income variable represents the income of the customer. The bar graph shows that the income is normally distributed around 20,000-29,000, skewed to the right due to some outliers making 50,000 or more. The bar graph separated by whether or not the customers bought the PEP shows that customers who make more money were more likely to buy the PEP. This makes sense because customers making less may not be financially ready to buy the product. It is also worth noting that there is a significant amount of customers making between 10,000 and 19,000 who bought the PEP, so it may still be worth it to market new products to customers even if they don’t have a high income.



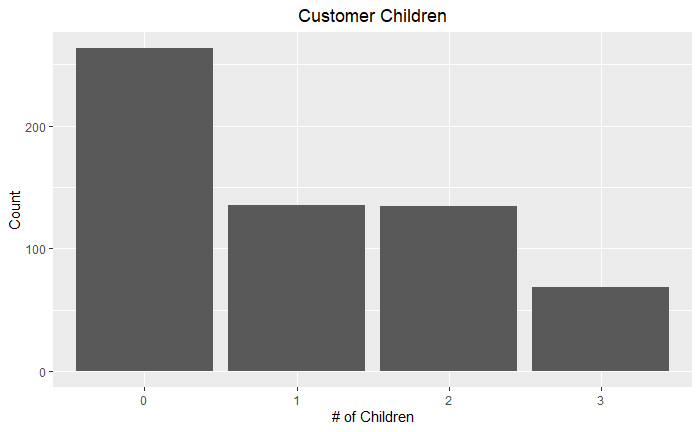


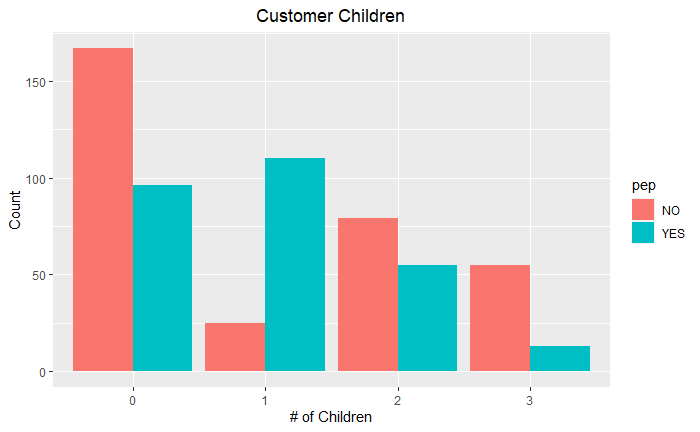
The married variable represents if the customer is married, “YES” or “NO”. The bar graph shows that there are about double the number of married customers as there are unmarried customers. The bar graph separated by whether or not the customers bought the PEP shows that a greater percentage of unmarried customers bought the PEP than married customers. Since there are more married customers, the total amount of married customers who bought the PEP is still greater than the total amount of unmarried customers who bought the product.



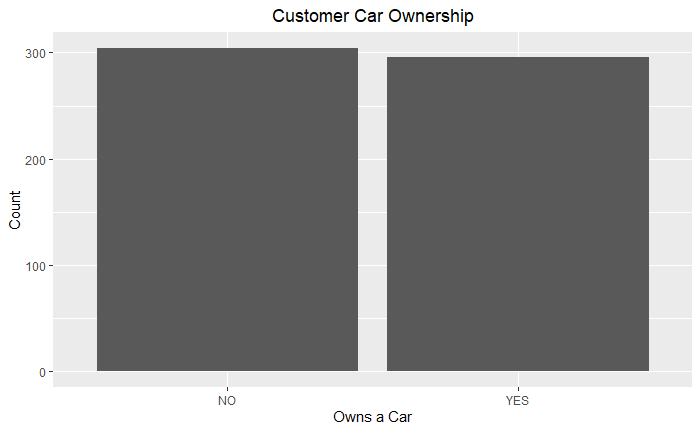


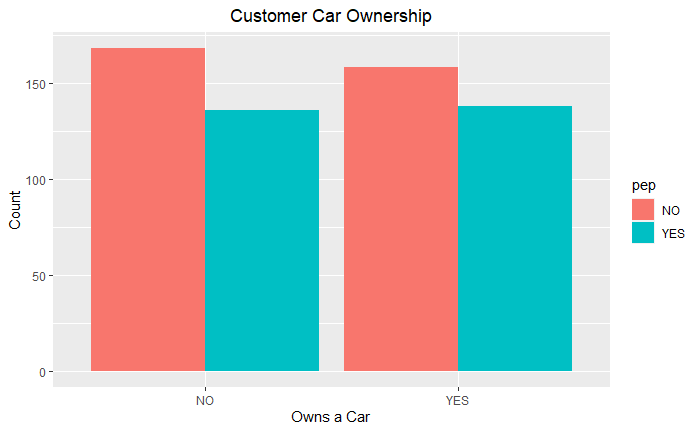
The children variable represents the number of children the customer has. The bar graph shows that almost half of the customers don’t have any children, with the other half having one, two, or three kids. The smallest group is the customers with three children. The bar graph separated by whether or not the customers bought the PEP shows that customers with one child have the greatest response rates to the advertisement. This may be because people who have had their first child are more actively planning their financial future than anyone else.



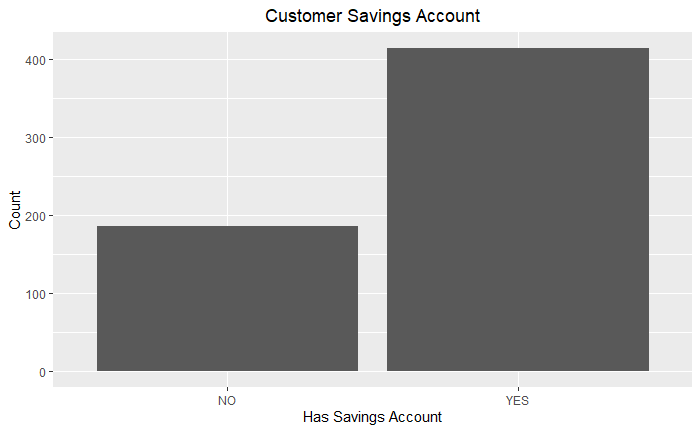


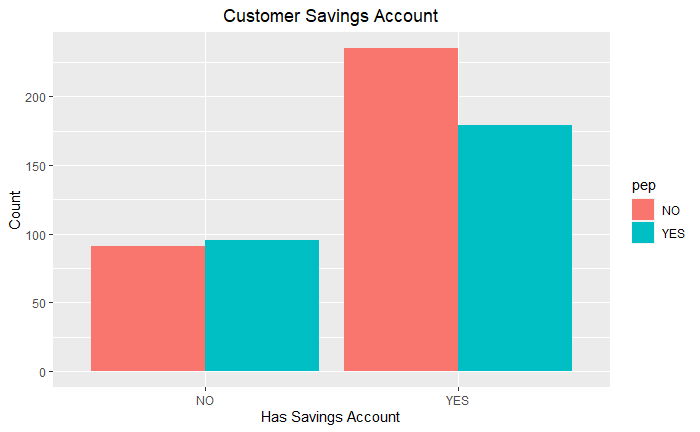
The car variable represents if the customer owns a car, “YES” or “NO. The bar graph shows there are about the same number of customers who own a car and customers who do not own a car. The bar graph separated by whether or not the customers bought the PEP shows that there are about the same response rates from customers who own a car and customers who do not own a car. The car variable will not be very helpful on its own in identifying who the company should target for marketing efforts.



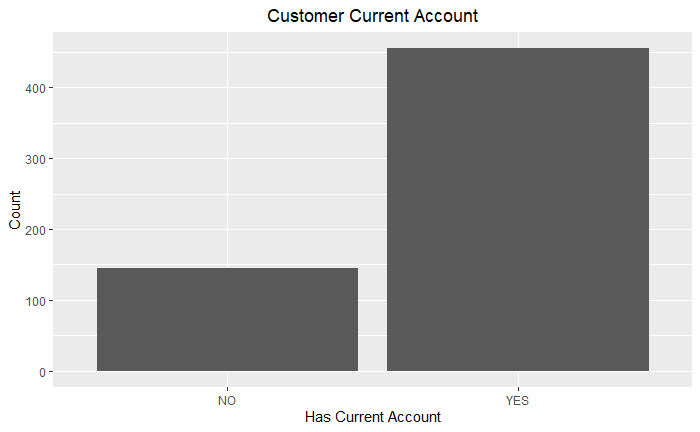


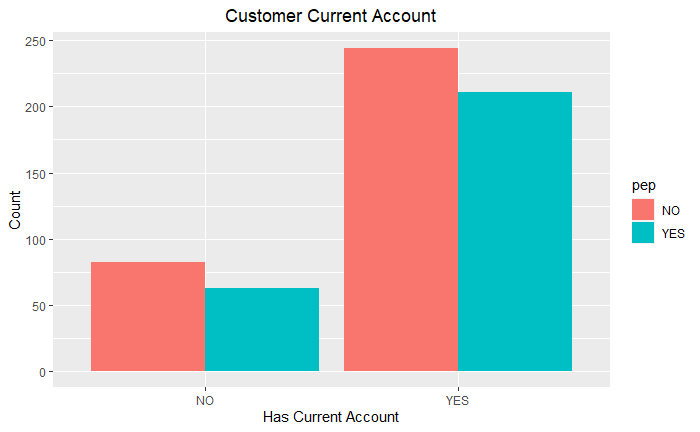
The save\_act variable represents if the customer has a savings account, “YES” or “NO”. The bar graph shows that about 2/3 of the customers have savings accounts. The bar graph separated by whether or not the customers bought the PEP shows that customers who do not have a savings account bought the product at a greater rate than customers who have a savings account. It is worth noting that the total number of customers who have a savings account and bought the product is still greater than the total for customers who do not have a savings account.



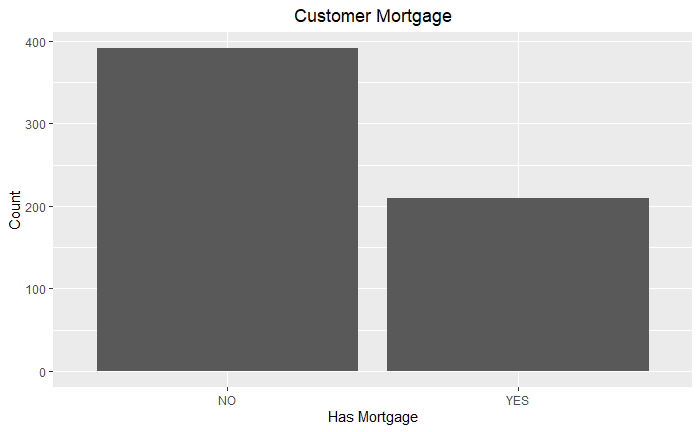


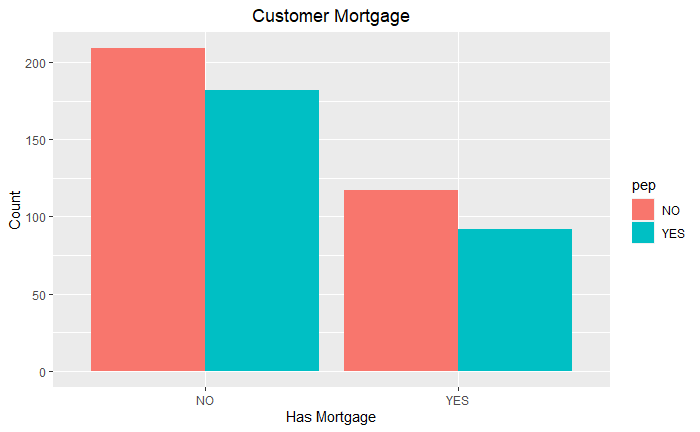
The current\_act variable represents if the customer has a current account, “YES” or “NO”. The bar graph shows that about 75% of the customers have a current account. The bar graph separated by whether or not the customers bought the PEP shows both groups had similar response rates to the advertisement. Current\_act will not be very helpful on its own in identifying who the company should target for marketing efforts.



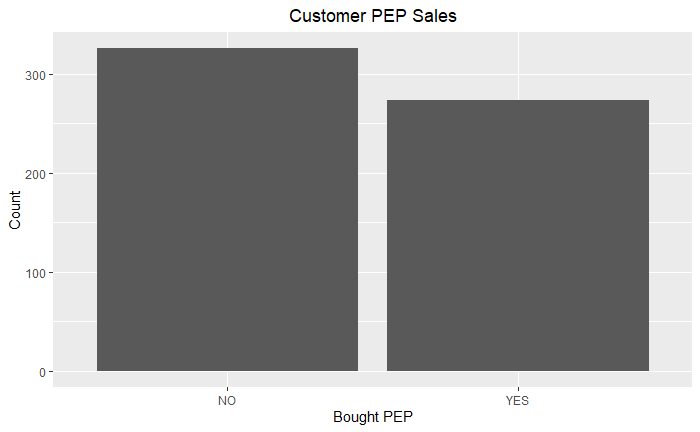


The mortgage variable represents if the customer has a mortgage, “YES” or “NO”. The bar graph shows the group of customers who does not have a mortgage is about twice the size of the group of customers who have a mortgage. The bar graph separated by whether or not the customers bought the PEP shows both groups had similar response rates to the advertisement. The mortgage variable will not be very helpful on its own in identifying who the company should target for marketing efforts.





The pep variable represents if the customer bought a PEP after the last mailing, “YES” or “NO”. The bar graph shows that 45.7% of customers responded to the mail advertisement and bought the PEP product. The fact that the target variable is split is beneficial for analysis.



## Association Model (Apriori)

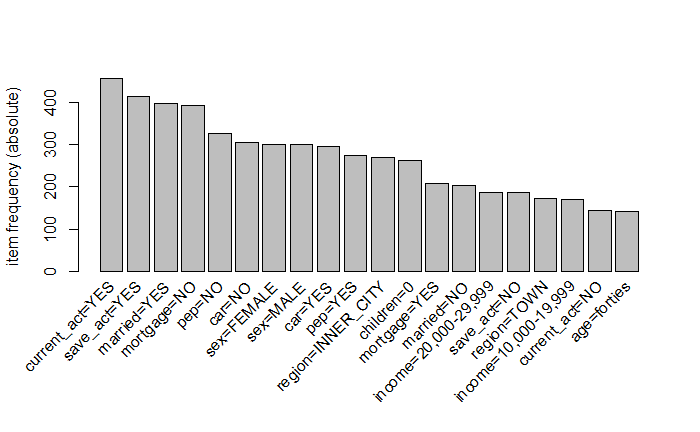
Association Rule Mining finds rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. The Apriori algorithm is a method to efficiently generate frequent itemsets and then prune rules with low confidence. To generate candidates efficiently the algorithm calculates the support for each itemset. The support is the fraction of transactions that contain the itemset. The algorithm discards the itemset if the support is below the minimum support threshold. If the support is above the threshold, the algorithm builds larger itemsets with it and repeats the process until it runs out of itemsets. To prune the generated rules down, the algorithm calculates the confidence of the rule. The confidence shows how frequently items in Y appear in transactions that contain X. The algorithm starts with rules with the highest number of items, and if the confidence does not meet the confidence threshold then it discards that rule and any rule with the same right-hand side item. This results in a list of rules that includes the rule’s support, confidence, and lift. The lift is a measure of dependent or correlated events. It is analyzed because support and confidence are limited and could sometimes be misleading.

# Results

The item frequency plot shows what items are most likely to be part of the rules that will be mined.

Text

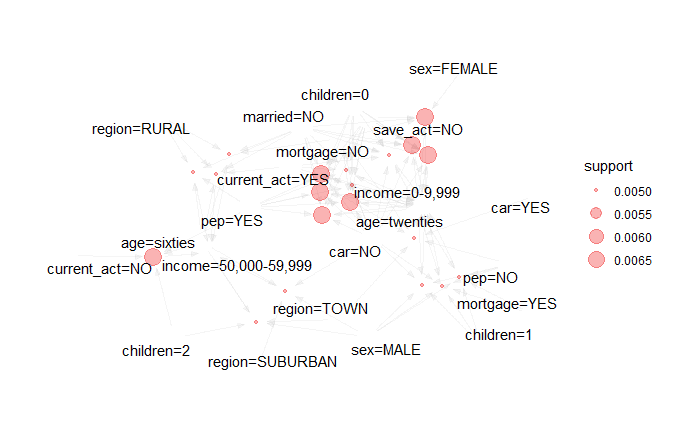
Description automatically generated



To find a list of rules with high lift and high confidence, the Apriori algorithm was first run with a support threshold of 0.001 and a confidence threshold of 0.8. It generated a list of rules where the top 20 had a confidence of 1 and a lift of 200. The problem is that all of the rules had “income=60,000-69,000” on the right-hand side, which is not very interesting. The algorithm was run again with a support threshold of 0.005 and a confidence threshold of 0.8. It generated a list of rules where the top 20 had a confidence of 0.8-1 and a lift of 14.3-23.1. This list also had income on the right-hand side but it was a mix of “income=0-9,999” and “income=50,000-59,000”.

Text

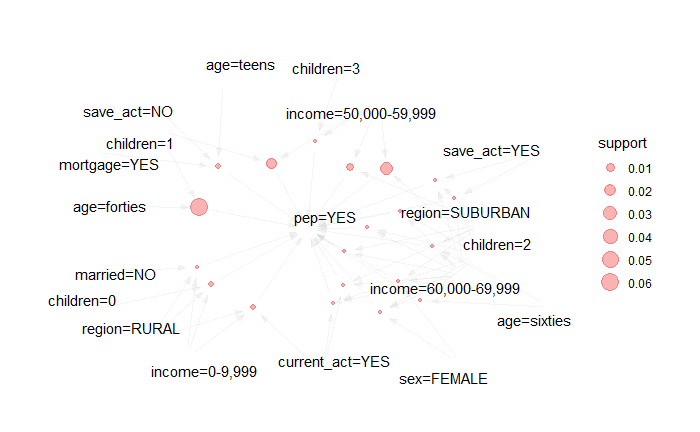
Description automatically generated



Setting the PEP variable as the right-hand side ensures the generated rules show what variable values generally correlate with buying the product. The support threshold was moved up to 0.3 when the lift started to decrease. The algorithm generated rules that happened frequently given the high support threshold. These included having a child in their forties, and having a child with some income. Since the goal is not to find the most obvious rules, the support is decreased back to 0.002 where the confidence and lift remained the same.

Text

Description automatically generated



The algorithm generated a list of rules where the top 20 had a lift of 2.19 and a confidence of 1. The most interesting rules that would help the business include:

1. {region=SUBURBAN, income=60,000-69,999} => {pep=YES}
   1. Support: 0.003
   2. Confidence: 1
   3. Lift: 2.19
2. {age=forties, children=1}
   1. Support: 0.060
   2. Confidence: 1
   3. Lift: 2.19
3. {sex=FEMALE, income=60,000-69,999, children=2} => {pep=YES}
   1. Support: 0.003
   2. Confidence: 1
   3. Lift: 2.19
4. {age=teens, save\_act=NO, mortgage=YES} => {pep=YES}
   1. Support: 0.005
   2. Confidence: 1
   3. Lift: 2.19
5. {region=RURAL, income=0-9,999, children=0} => {pep=YES}
   1. Support: 0.005
   2. Confidence: 1
   3. Lift: 2.19

# Conclusions

Analysis of the bank customer records produced some insights into who responded to the previous advertisement and bought the product. This should give the business a better idea about how their marketing efforts worked on different groups of people and what to keep in mind when developing a marketing campaign in the future. The initial data analysis looked at each variable on its own to gain an understanding of how it might correlate with PEP when combined with other variables. Running the Apriori algorithm for associative rule mining generated a list of these rules, which can be explained by looking at rule two in the list of the five most interesting rules.

Rule two in the list is one of the most interesting because it has a high support value. It states that when a customer is in their forties and has one child, they were very likely to buy the product. The support is 0.06 which means 6% of the customers are in this situation of being in their forties, having a child, and having bought the product.

The confidence is 1 which means that every customer in their forties who has a child bought the product. This is an important measure to take into consideration for marketing purposes. Each advertisement costs the company money, so it is best to send out advertisements to customers like this to maximize the return on the marketing campaign. The lift is 2.19 which further supports that there is a meaningful correlation between a customer being in their forties with a child and having bought the product from the advertisement.