Bank Data Analysis

Dataset

The marketing department of a financial firm keeps records on customers, including demographic information and, number of type of accounts. When launching a new product, such as a "Personal Equity Plan" (PEP), a direct mail piece or a targeted email, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product. Based on this database of prior cases, the managers decide to use data mining techniques to build customer profile models in order to predict the behavior of future customers.

Each record is a customer description where the "pep" field indicates whether that customer has purchased a PEP. For classification problems, this field is used as the target attribute (with "YES" and "NO") as class labels. The data contains the following fields:

id	a unique identification number (categorical, str)
age	age of customer in years (numeric, int)
income	income of customer (numeric, float)
children	number of children (numeric, int)
gender	MALE / FEMALE
region	INNER_CITY/RURAL/SUBURBAN/TOWN
married	Customer married (YES/NO)
car	Customer owns one or more cars (YES/NO)
save_acct	Customer has a savings account (YES/NO)
current_acct	Customer has a current checking account (YES/NO)
mortgage	Customer have a mortgage (YES/NO)
pep	Customer purchased a PEP, Personal Equity Plan (YES/NO)

Questions:

1. **Explore the general characteristics of the data as a whole: examine the means, standard deviations, and other statistics associated with the numerical attributes; show the distributions of values associated with categorical attributes.**

To explore the general characteristics of the data I performed a summary of each attribute to show the central tendency and dispersion of the data. This summary includes the five-number summary (the minimum and maximum values, 1st and 3rd quartiles, and median) and the mean values for the data set. To observe the mode values, I created a mode function and calculated the mode for each attribute in the dataset. See the figures 1 and 2 below for the output.

Figure 1. Basic Statistical Description

```
> #Basic Statistical Data Description. I will be measuring the dispersion of the data.
> summary(bank_data)
       id            age             income         children       gender
 ID12101:  1   Min.   :18.00   Min.   : 5014   Min.   :0.000   FEMALE:300
 ID12102:  1   1st Qu.:30.00   1st Qu.:17265   1st Qu.:0.000   MALE  :300
 ID12103:  1   Median :42.00   Median :24925   Median :1.000
 ID12104:  1   Mean   :42.40   Mean   :27524   Mean   :1.012
 ID12105:  1   3rd Qu.:55.25   3rd Qu.:36173   3rd Qu.:2.000
 ID12106:  1   Max.   :67.00   Max.   :63130   Max.   :3.000
 (Other):594
       region    married    car      savings_acct current_acct mortgage
 INNER_CITY:269   NO :204   NO :304   NO :186      NO :145      NO :391
 RURAL      : 96   YES:396   YES:296   YES:414      YES:455      YES:209
 SUBURBAN   : 62
 TOWN       :173


  pep
 NO :326
 YES:274
```

Figure 2. Modes of Attributes*

```
> #Create a function to calculate the modes of the numeric and categorical valaues
>
> modevalue<- function(x){
+   modeval<- unique(x)
+   modeval[which.max(tabulate(match(x, modeval)))]
+
+ }
>
> modevalue(bank_data$age)
[1] 40
> modevalue(bank_data$income)
[1] 38248.3
> modevalue(bank_data$children)
[1] 0
> modevalue(bank_data$gender)
[1] FEMALE
Levels: FEMALE MALE
> modevalue(bank_data$region)
[1] INNER_CITY
Levels: INNER_CITY RURAL SUBURBAN TOWN
> modevalue(bank_data$married)
[1] YES
Levels: NO YES
> modevalue(bank_data$car)
[1] NO
Levels: NO YES
> modevalue(bank_data$savings_acct)
[1] YES
Levels: NO YES
> modevalue(bank_data$current_acct)
[1] YES
Levels: NO YES
> modevalue(bank_data$mortgage)
[1] NO
Levels: NO YES
```

*Note that multiple modes exist for the data. The smallest value is reported.

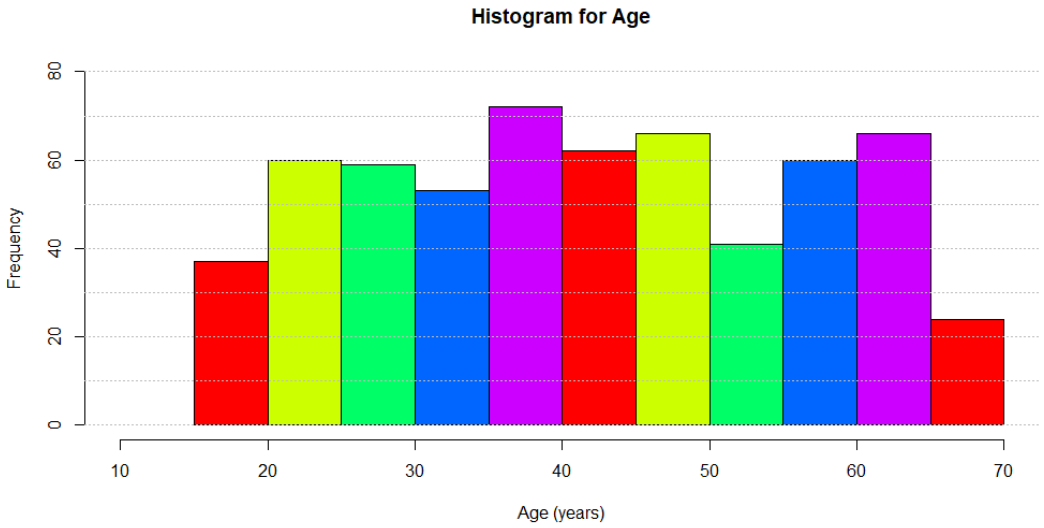
```
> modevalue(bank_data$pep)
[1] NO
Levels: NO YES
>
```

In figure 2 above the following mode values were observed:
*note the mode values are the smallest values observed

Age- 40
Income- $38,248.3
Gender- Female
Region- Inner City
Married- Yes
Owns a car- Yes
Has a savings account- Yes
Current checking account- Yes
Mortgage- No
Purchased a PEP- No

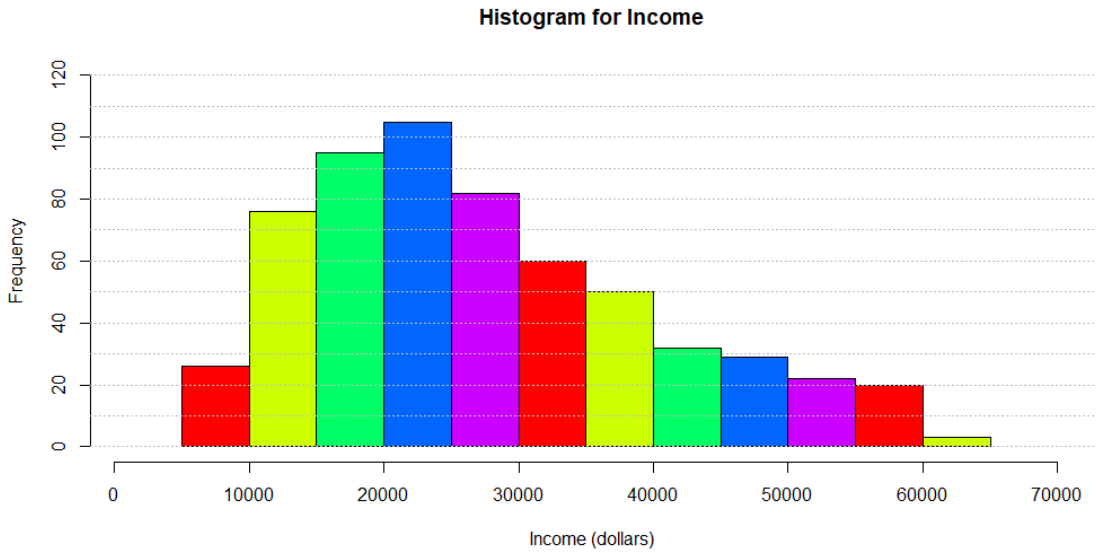
To observe the frequency of the numeric attributes, I created histograms for each numeric attribute. See below:

```
#histogram to show frequency of categorical and numeric attributes
hist(bank_data$age, main= "Histogram for Age", xlab ='Age (years)', border = "black",
     col = rainbow(5),ylim=c(0,80), xlim = c(10,70))
abline(h=seq(0,80,10), col= "grey", lty = "dotted")
```



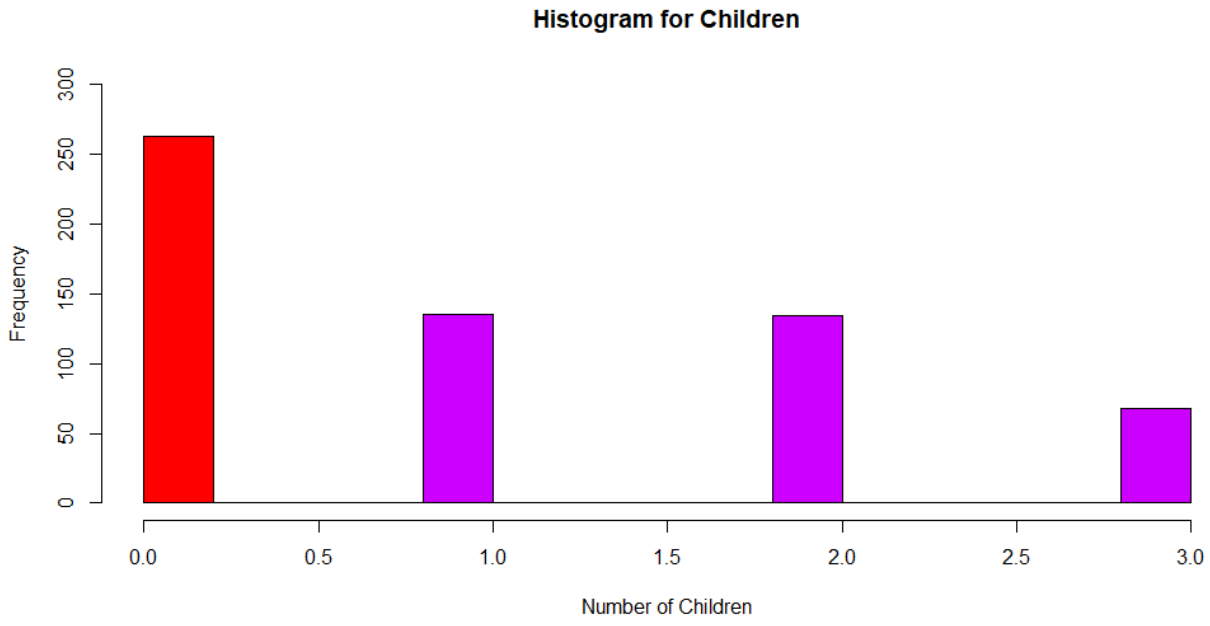
The data for the age variable is not symmetric and appears to be spread across a wide range of values.

```
#histogram for income
hist(bank_data$income, main= "Histogram for Income", xlab ='Income (dollars)', border = "black",
     col = rainbow(5),ylim=c(0,120), xlim = c(1000,70000))
abline(h=seq(0,120,10), col= "grey", lty = "dotted")
```



The data for the income variable appears to be positively skewed with the most observed values between 15,000 and 30,000.

```
#histogram for children
hist(bank_data$children, main= "Histogram for Children", xlab ='Number of Children', border = "black",
     col = rainbow(5),ylim=c(0,300), xlim = c(0,3))
```



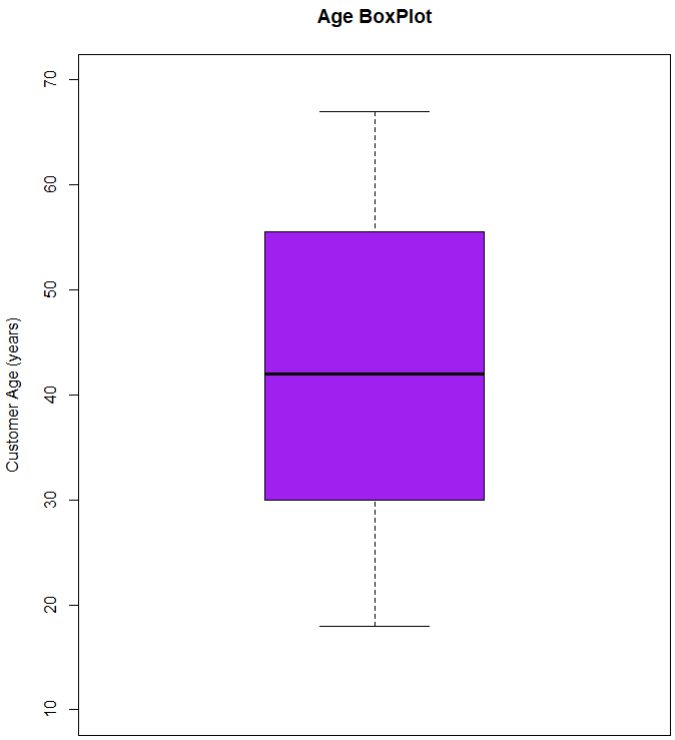
The value of zero appears to be the most frequently observed number of children.

I also wanted to visualize the data distribution by creating boxplots for the attributes. After creating the box plots, I calculated the IQR and determined if there were any outliers in the data using the following formulas:

IQR= Q3-Q1
Outlier lower value = Q1- 1.5 * IQR

Outlier higher value = Q3 + 1.5 * IQR

```
#boxplot to show the five number summary for numeric attributes
boxplot(bank_data$age,main= "Age BoxPlot",ylab="Customer Age (years)",
        ylim=c(10,70),col= "purple", border= "Black", horizontal = FALSE)
```



Data distribution for the boxplot of the age attribute (as shown in figure 1) is as follows:

Q1= 30, Q3= 55.25, Median = 42, min= 18, max= 67

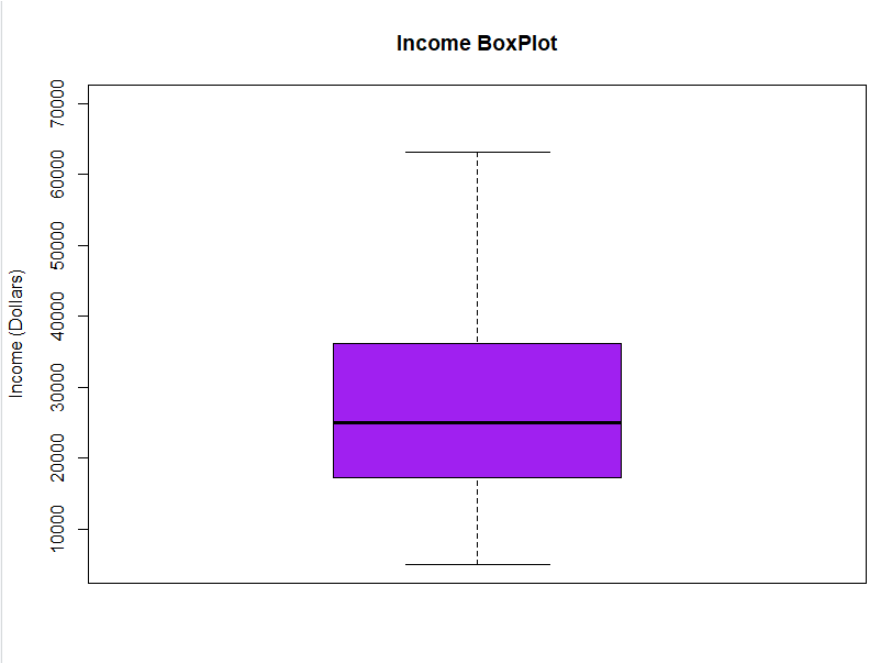
IQR= 55.25 – 30 = 25.25

High outlier = 55.25 + (1.5*25.25) = 93.125

Low outlier = 30 - (1.5*22.25) = -3.38

No outliers appear to be present, which means no values were above or below 1.5*IQR. the values also appear to be fairly well distributed between the low and high values.

```
boxplot(bank_data$income,main= "Income BoxPlot",ylab="Income (Dollars)", ylim=c(5000, 70000),
        col= "purple", border= "Black", horizontal = FALSE)
```



The data distribution for the boxplot of the income attribute (as shown in figure 1) is as follows:

```
    income
Min.   : 5014
1st Qu.:17265
Median :24925
Mean   :27524
3rd Qu.:36173
Max.   :63130
```

IQR= 36173 – 17265 = 18908

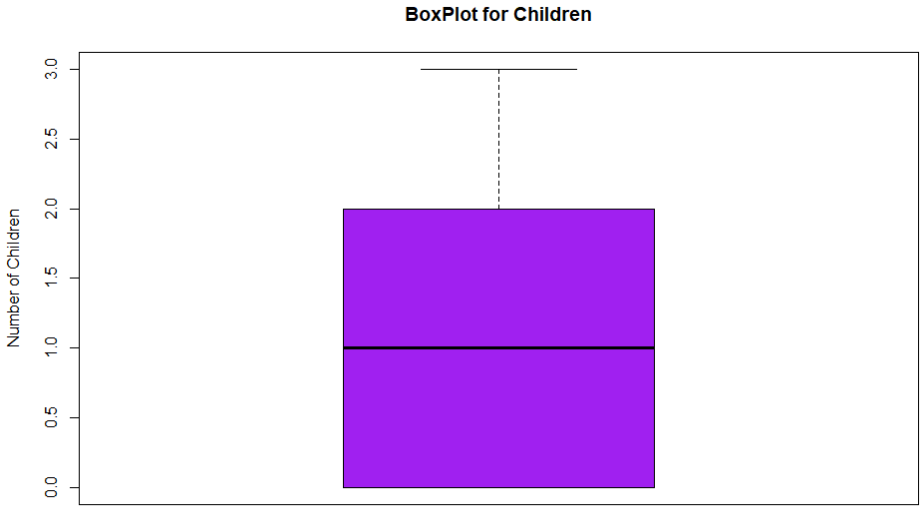
High value = 36173 + (1.5*18908) = 64535

Low value = 17265 - (1.5*18908) = -11097

No outliers appear to be present as there are no values more than or less than 1.5 times the IQR. Many of the values appear to be above the 3rd quartile.

```
boxplot(bank_data$children,main= "BoxPlot for Children",ylab="Number of Children", ylim=c(0,3),
        col= "purple", border= "Black", horizontal = FALSE)
```


```
boxplot(bank_data$children,main= "BoxPlot for Children",ylab="Number of Children", ylim=c(0,3),
        col= "purple", border= "Black", horizontal = FALSE)
```



Data distribution for the boxplot of the children attribute (as shown in figure 1) is as follows:

```
    children
Min.   :0.000
1st Qu.:0.000
Median :1.000
Mean   :1.012
3rd Qu.:2.000
Max.   :3.000
```

IQR= 2-0=2

No outliers appear to be present. In this case the number of children cannot be less than zero, so many values are above the 3rd quartile, and it appears that many people have more than 2 children in this data set.

Variance and standard deviation are shown below for the numerical attributes. The Variance and standard deviation will show the measures of the data dispersion and show how spread out the data distribution is. A low standard deviation will show that the data tends to be close to the mean value. A high standard deviation will show that the data is spread out over a large range of values.

```
> #variance of numeric variables
> var(bank_data$age)
[1] 208.0791
> var(bank_data$income)
[1] 166396281
> var(bank_data$children)
[1] 1.116725
>
```

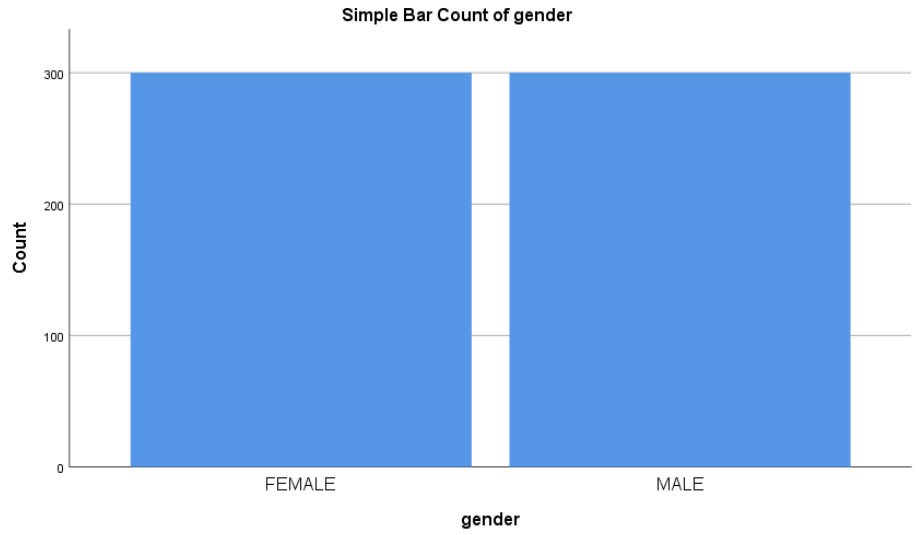
```
> #standard deviation of numeric variables
> sd(bank_data$age)
[1] 14.42495
> sd(bank_data$income)
[1] 12899.47
> sd(bank_data$children)
[1] 1.056752
```

As a reminder see the mean values of the numeric variables below:

age	income	children
Mean :42.40	Mean :27524	Mean :1.012

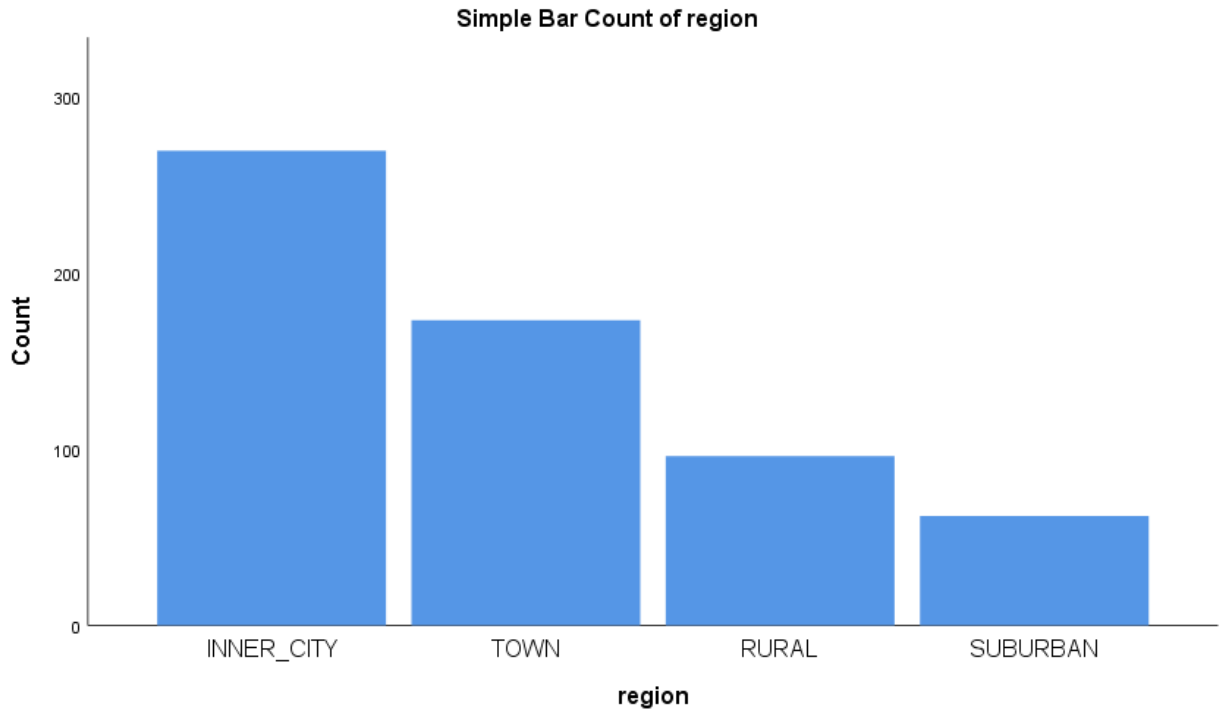
It appears that the age variable has a higher standard deviation which indicates that the data is spread out over a large range. The histogram shown previously for the age attribute also shows this. The standard deviation for the income variable is high and indicates that the data is spread out over a large range, the histogram for the attribute showed that the data is positively skewed. The standard deviation for the children attribute is low which indicates the data is very close to the mean.

For the categorical variables I used SPSS to visualize the distribution of the attributes. See the graphs below:



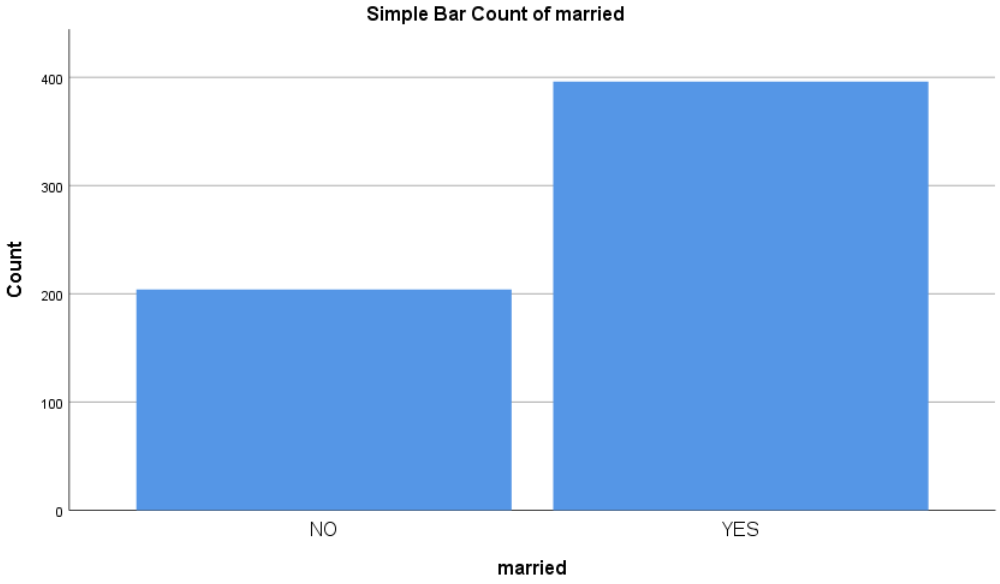
In the above graph, the data attribute for gender is separated into male and female. The graph above shows that the gender was evenly distributed amongst male and female.

The next graph shows the distribution of the different regions that people are from



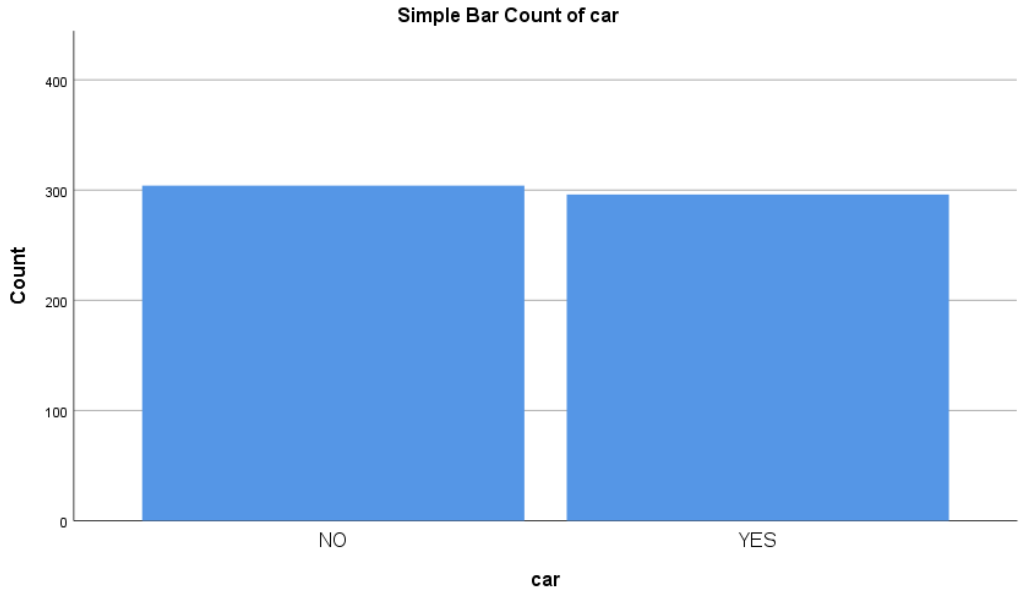
Most people appear to be from an inner city, with town being the next most frequent.

The next bar chart shows the distribution amongst married and non-married



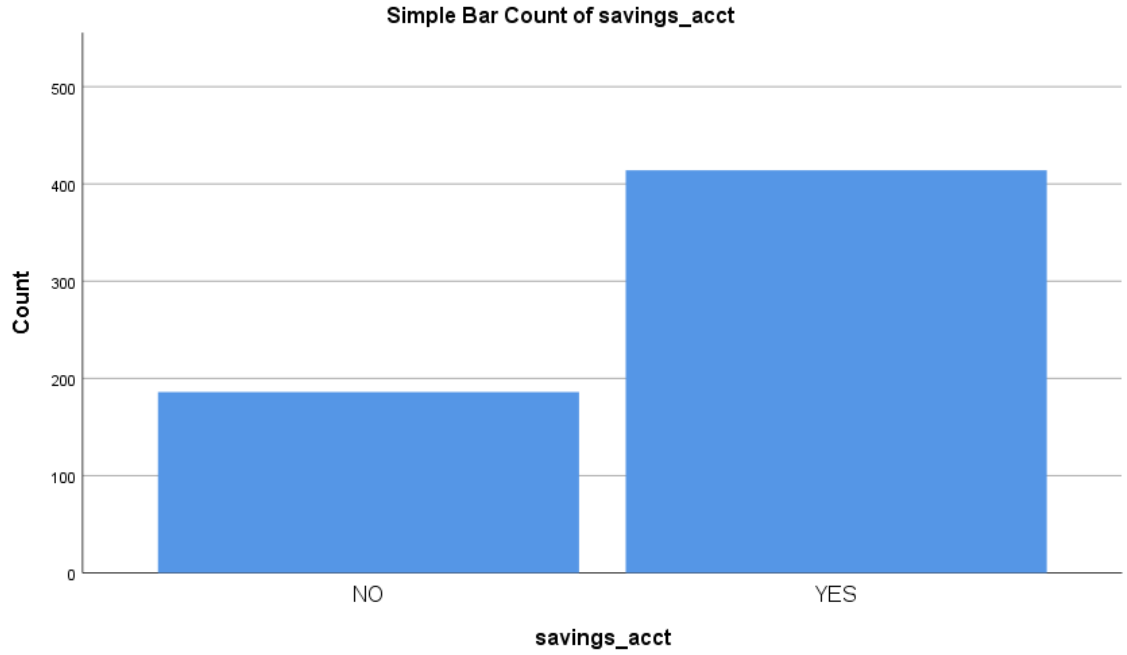
It appears that most of the people are married in this data set.

Next up is the distribution of attribute Car.



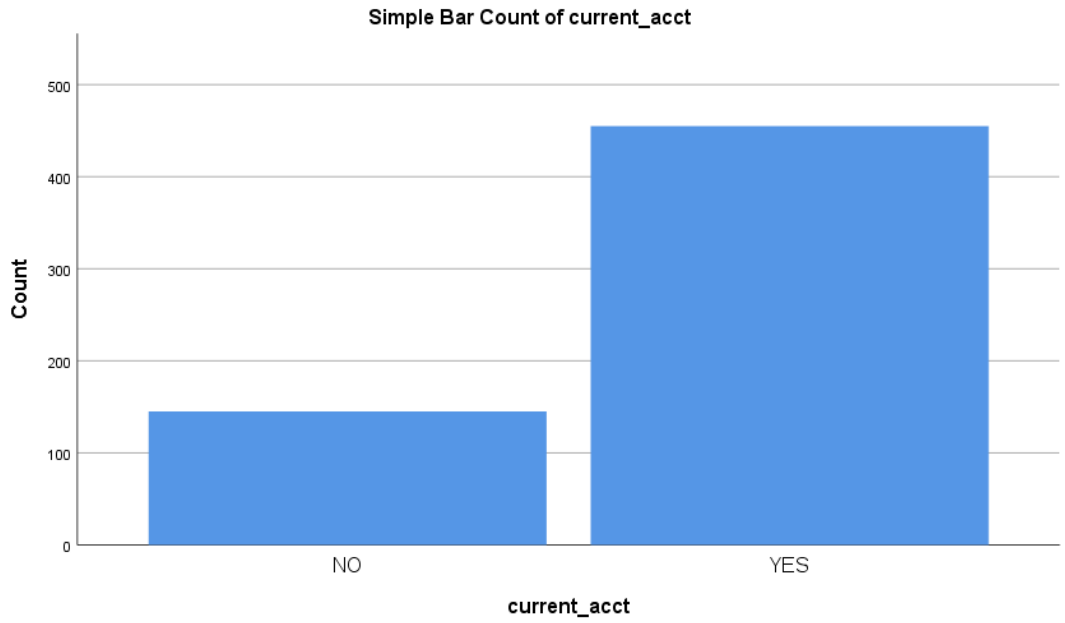
It appears that the number of people who own a car and the number of people who don't is even.

Next is the savings account attribute.



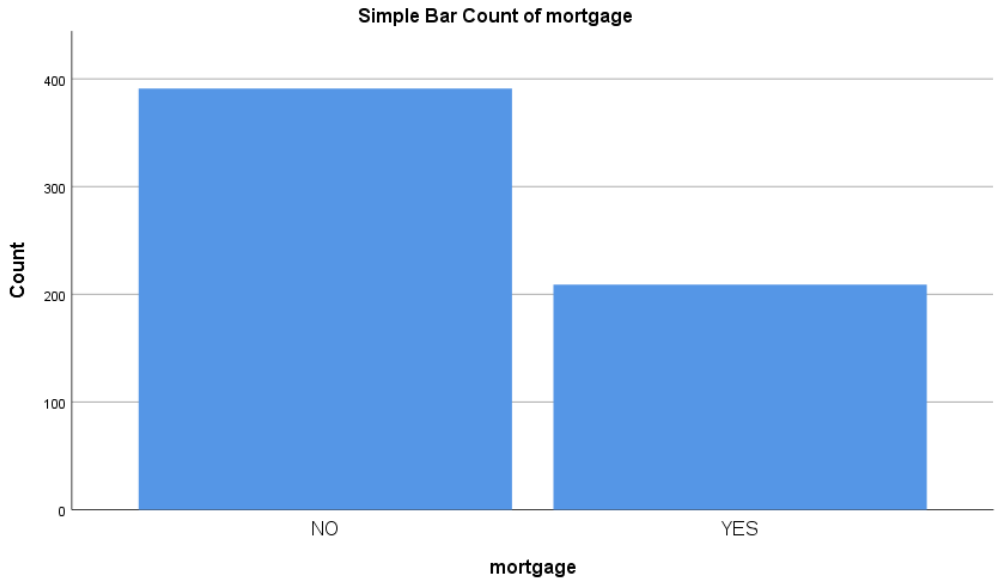
It appears that more people have a savings account, verses those that do not in the data set.

Next is the current account attribute:



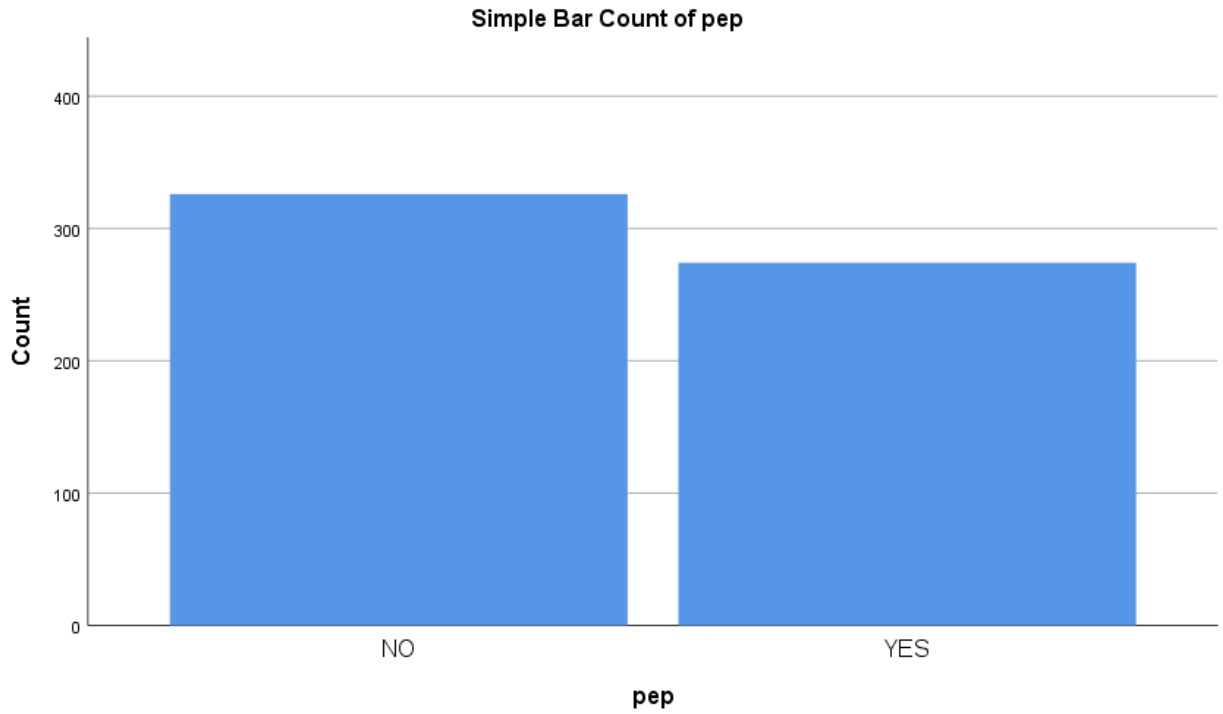
It appears that more people have a current account, verses those that do not in the data set.

Next is the mortgage attribute.



More people appear to not have a mortgage versus those that have a mortgage in the dataset.

And lastly is the PEP attribute

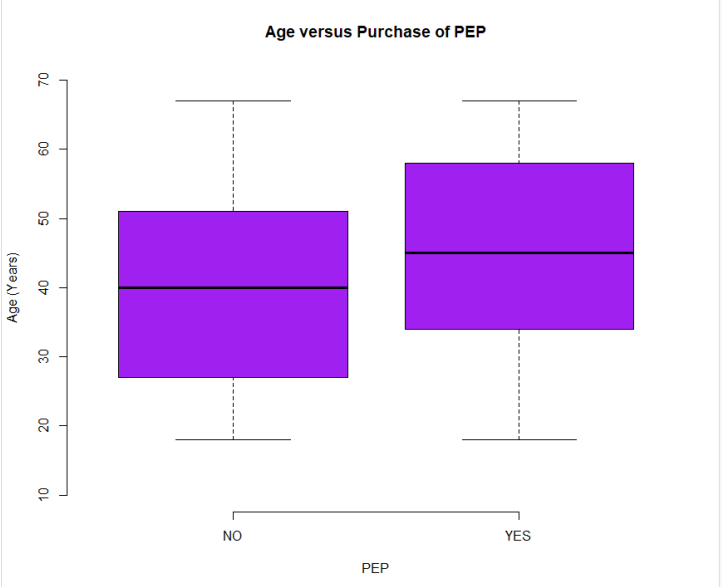


The count of people who do not have a personal equity plan is greater than those that do, although the number does not seem significantly higher.

2. **Suppose that the hypothetical bank is particularly interested in customers who buy the PEP (Personal Equity Plan) product. Compare and contrast the subsets of customers who buy and don't buy the PEP. Compute summaries (as in part 1) of the selected data with respect to all other attributes. Can you observe any significant differences between these segments of customers? Discuss your observations.**

   The subsets of customers who buy the PEP were observed using box plots, chi-square tests(for categorical attributes), and correlation and variance(for numeric attributes). Boxplots were observed for the age and income variables to visual the data distribution. A box plot was not examined for the children attribute because the values are too small.

```
> summary(bank_data$age,bank_data$pep)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   30.00   42.00   42.40   55.25   67.00
> summary(bank_data$income,bank_data$pep)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5014   17265   24925   27524   36173   63130
> summary(bank_data$children,bank_data$pep)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   1.012   2.000   3.000
> summary(bank_data$gender,bank_data$pep)
```



For purchase of a PEP compared to the age of a customer it is observed that the 1st quartile, median, and the 3rd quartile values of the age attribute of the customers appears to be higher when a PEP is purchased. The ages of the customers that do not have a PEP appear to be lower.

A chi square test was also performed using SPSS to determine if there is a correlation relationship amongst age and purchase of a PEP. See the output below:

pep * age

Crosstab

			age															
			18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
pep	NO	Count	9	6	11	5	10	11	5	7	7	12	4	7	11	3	6	4
		Expected Count	6.0	5.4	8.7	4.3	8.2	9.2	5.4	5.4	5.4	10.3	4.9	4.9	6.5	5.4	4.9	4.3
	YES	Count	2	4	5	3	5	6	5	3	3	7	5	2	1	7	3	4
		Expected Count	5.0	4.6	7.3	3.7	6.9	7.8	4.6	4.6	4.6	8.7	4.1	4.1	5.5	4.6	4.1	3.7
Total		Count	11	10	16	8	15	17	10	10	10	19	9	9	12	10	9	8
		Expected Count	11.0	10.0	16.0	8.0	15.0	17.0	10.0	10.0	10.0	19.0	9.0	9.0	12.0	10.0	9.0	8.0

33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
4	6	7	9	9	7	6	12	7	4	11	3	6	8	11	8	7	4
4.3	6.5	7.6	8.7	6.5	6.5	6.5	10.9	6.5	5.4	10.9	6.0	4.9	7.6	9.2	8.7	6.5	3.8
4	6	7	7	3	5	6	8	5	6	9	8	3	6	6	8	5	3
3.7	5.5	6.4	7.3	5.5	5.5	5.5	9.1	5.5	4.6	9.1	5.0	4.1	6.4	7.8	7.3	5.5	3.2
8	12	14	16	12	12	12	20	12	10	20	11	9	14	17	16	12	7
8.0	12.0	14.0	16.0	12.0	12.0	12.0	20.0	12.0	10.0	20.0	11.0	9.0	14.0	17.0	16.0	12.0	7.0

50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	Total
4	5	8	0	5	1	9	6	8	2	3	6	2	6	7	7	4	4	326
3.8	3.8	8.2	4.3	4.3	1.6	7.1	5.4	10.3	5.4	4.3	7.6	4.3	7.1	10.9	6.0	5.4	7.6	326.0
3	2	7	8	3	2	4	4	11	8	5	8	6	7	13	4	6	10	274
3.2	3.2	6.9	3.7	3.7	1.4	5.9	4.6	8.7	4.6	3.7	6.4	3.7	5.9	9.1	5.0	4.6	6.4	274.0
7	7	15	8	8	3	13	10	19	10	8	14	8	13	20	11	10	14	600
7.0	7.0	15.0	8.0	8.0	3.0	13.0	10.0	19.0	10.0	8.0	14.0	8.0	13.0	20.0	11.0	10.0	14.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	60.887[a]	49	.119
Likelihood Ratio	66.623	49	.048
N of Valid Cases	600		

a. 35 cells (35.0%) have expected count less than 5. The minimum expected count is 1.37.

Symmetric Measures

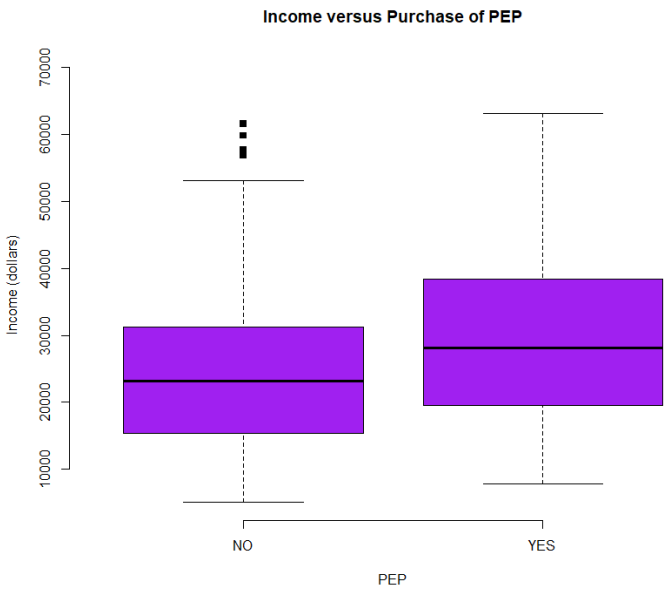
		Value	Approximate Significance
Nominal by Nominal	Phi	.319	.119
	Cramer's V	.319	.119
N of Valid Cases		600	

The chi square value is: 60.887 and the p^2 value is 0.119. There are some differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.119 is less than 0.5 so the data suggests some correlation amongst the age and purchasing of a PEP. The phi and Cramer's V values also indicate that the age and purchase of a PEP are correlated.

Next is the income attribute verses PEP. The boxplot below shows that customers with a higher income have purchased a PEP

```
> summary(bank_data$income,bank_data$pep)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5014   17265   24925   27524   36173   63130
```



The IQR and outlier calculations are listed below(values are approximate):

The IQR for when a PEP is not purchased
IQR= ~30000-15000= ~15000
Low Outlier = 15000- (1.5*15000) = ~ -1.5
High Outlier = ~30000 +(1.5*~15000) = ~52500
There appears to be outliers that have values more than 1.5 times the IQR of approximately 15000.

The IQR when a PEP is purchased is as follows:
IQR= 36173-20000= ~16173
Low Outlier = 20000- (1.5*16173) = -16389.25
High Outlier = ~36173 +(1.5*16173) = ~60432.5
There do not appear to be any values that fall significantly more or less than the approximated values of more than or less than 1.5 * IQR.

Chi square was also calculated to see if there was any correlation amongst the income and PEP attributes. See the contingency table below:

pep * income

Crosstab

Count

		income																
		5014.21	6294.21	7304.20	7549.38	7606.25	7723.93	7756.36	7948.62	8020.19	8062.73	8143.75	8162.42	8562.86	8639.24	8877.07	9316.98	9362.58
pep	NO	1	1	1	1	1	1	0	1	1	0	0	1	1	1	0	1	0
	YES	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	1
Total		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

	5014.21	6294.21	7304.2	7549.38	7606.25	7723.93	7756.36	7948.62	8020.19	8062.73	8143.75	8162.42	8562.86	8639.24	8877.07	9316.98					
NO	1	1	1	1	1	1	0	1	1	0	0	1	1	1	0	1					
	9362.58	9465.21	9485.84	9516.91	9589.91	9592.73	9672.25	9824.37	9909.82	9990.11	10044.1	10072.6	10191.8	10441.9	10629.1	10672					
NO	0	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1					
	10861	10953	11043.7	11073	11215.3	11281.5	11299.3	11411	11520.8	11536.2	11595.4	11601.4	11604.4	11700.4	11736.9	11866.4	11920.7				
NO	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1				
	12117.3	12125.8	12163.9	12166.9	12178.5	12279.5	12533.2	12549	12591.4	12623.4	12640.3	12644.9	12681.9	12683.6	12764.8	12810.2					
NO	1	1	0	1	1	0	1	1	1	1	1	1	0	1	0	0					
	12823.7	12977.2	13039.9	13106.6	13175.5	13236.4	13267.6	13283.9	13327.8	13381	13519.2	13667.7	13700.2	13739	13740	13864.6					
NO	1	1	1	0	0	1	1	0	1	0	0	1	1	1	1	0					
	13950.4	14014.5	14048.9	14058.5	14064.9	14092.7	14136.5	14290.5	14309.7	14388.6	14433.4	14505.3	14511.8	14585.9	14606.6	14627.9					
NO	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1					
	14642.2	14711.8	14724.5	14960.2	14996.4	15109.4	15143.8	15156.2	15237.6	15254.8	15281.8	15308.2	15315.3	15348.9	15349.6	15417.1					
NO	0	1	1	1	1	0	1	1	1	0	0	1	0	1	1	1					
	15499.9	15525	15538.8	15610.2	15689.1	15735.8	15797.1	15848.7	15933.3	15976.3	16088.8	16109.9	16249.8	16259.7	16291	16325.8					
NO	0	1	1	1	1	0	0	1	1	0	1	1	0	0	1	1					
	16352.2	16394.4	16398.8	16403.8	16479.5	16497.3	16518.6	16575.4	16625.9	16662.5	16672.8	16711.3	16716.1	16849.3	16977.3	17139.5					
NO	1	0	0	1	0	1	0	1	0	0	1	0	1	1	1	1					
	17149.2	17180.2	17239.5	17240.6	17247.7	17270.1	17308.7	17364.8	17371.1	17390.1	17546	17610.3	17655	17729.8	17839.9	17861					
NO	0	1	0	1	1	0	0	0	1	1	0	1	0	1	0	1					
	17866.9	17867.3	17882.9	17921.8	17944.2	17986.8	18036.7	18050	18067.5	18158.5	18184.6	18275.5	18364.9	18504.3	18516	18555.9					
NO	1	1	0	0	0	0	1	0	1	0	0	0	0	1	1	1					
	18565.8	18707.3	18802.4	18860.3	18875.7	18912.2	18923	19012.8	19160.3	19166	19326.9	19403.1	19416.8	19474.6	19481.3	19563.8					
NO	0	0	0	1	1	0	0	1	0	0	0	1	1	1	0	1					
	19726.3	19868	19918.9	19968.1	20058.7	20114	20236.2	20262.6	20268	20347	20375.4	20409.3	20467.3	20555	20708.5	20736.2	20754.3				
NO	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1	0	0				
	20771.9	20799	20809.7	20819	20866.3	20950.7	21042	21096.2	21139.8	21184.7	21268.4	21332.3	21350.3	21384.4	21495.6	21506.2					
NO	0	0	1	0	1	1	0	0	1	0	0	1	0	1	1	0					
	21612.2	21612.6	21623.8	21730.3	21796.6	21821.4	21876.5	21951.3	21984	21984.4	22007.1	22052.1	22053.2	22110.1	22197.1	22234.7					
NO	1	1	1	1	1	0	1	1	0	1	1	0	1	1	0	1					
	22327.8	22342.1	22362.3	22366.1	22400.7	22446.5	22495.7	22522.8	22562.2	22678.1	22791.4	22792.3	22848.5	22882.9	22916.1	22942.9					
NO	0	1	0	0	1	0	1	1	1	0	1	0	0	0	1	1					
	23038.2	23092.1	23124.9	23171.8	23175	23197.5	23246.4	23287.9	23337.2	23356.1	23371	23443.2	23475.6	23485.9	23528.4	23638.1					
NO	1	0	0	1	0	1	1	0	1	1	1	0	0	1	0	1					
	23818.6	23894.8	24026.1	24027.6	24031.5	24042	24071.8	24212.1	24262.8	24270.1	24346.6	24424.3	24474.1	24477.5	24554.1	24583.4					
NO	1	1	0	0	1	0	0	1	0	0	1	1	0	1	1	1					
	24607.8	24675.7	24760.8	24763.3	24814.5	24823.5	24858.4	24867.6	24888.2	24904	24946.6	24977.5	25127.7	25132.9	25257.7	25304.3					
NO	1	0	0	0	0	1	0	1	0	1	1	0	1	1	1	1					

```
     25333.2 25334.3 25372.8 25391.5 25429.3 25468.5 25683.4 25699.4 25732.5 25768.6 25830.5 26077.8 26097.9 26106.7 26261.7 26281.4
NO         1       0       1       1       1       0       1       0       1       1       1       0       1       0       1       0

     26325.3 26462.5 26542.8 26658.8 26671.6 26688.1 26707.5 26707.9 26774.2 26900.6 26909.2 26920.8 26948 26952.6 26999.4 27022.6
NO         1       1       0       1       0       0       1       0       0       1       0       1     1       1       0       1

     27045.1 27056.5 27417.6 27434.8 27571.5 27642.9 27712.9 27756.3 27757.6 27765.8 27808.1 27825.5 27863.9 28021.6 28138.5 28193.6
NO         0       1       1       0       1       0       0       0       0       0       1       1       1       1       0       0       0

     28240.4 28253.6 28409.4 28413.8 28421.7 28469.9 28495.1 28598.7 28658.3 28702.7 28864.9 28882.3 28920.6 28938.6 28969.4 28981.1
NO         1       1       0       1       1       1       1       1       1       0       1       0       1       1       1       1

     29093.1 29231.4 29359.1 29414.6 29525.5 29541.7 29574 29622 29625.1 29714.4 29794.1 29866.3 29866.9 29921.3 30067.5 30085.1
NO         1       1       1       0       1       1     1       1       1       0       0       1       0       1       0       1

     30099.3 30157.7 30189.4 30198.5 30396.1 30404.3 30488 30488.7 30658.7 30760.4 30799.5 30870.8 30971.8 31095.6 31207.1 31273.8
NO         0       0       0       0       0       0       1     1       1       1       0       0       0       1       0       0       0

     31290.6 31334.8 31415.7 31473.9 31671.3 31683.1 31693.5 31774.1 31864.8 31982 32184.4 32245.4 32395.5 32548.9 32583.5 32669.9
NO         1       0       0       0       0       0       0       1       1     0       0       1       1       1       1       0

     32762.5 33007.3 33028.3 33088.5 33123.7 33204.3 33229 33302.8 33615.4 33630.6 33665.5 33886.4 34020.5 34061.4 34073.8 34182.2
NO         1       0       0       1       0       0     0       0       0       0       1       0       1       0       1       0

     34253.6 34410 34513.6 34524.9 34625.2 34836.8 34852.3 34866.5 34866.9 34892.9 35263.5 35610.5 36057.8 36086.1 36095.9 36166.2
NO         1     0       0       0       1       0       0       0       1       1       0       0       0       0       1       1

     36192.1 36256.9 36281 36432.8 36436.4 36589 36599 36646.4 36972.4 37094.2 37095.2 37162.1 37330.5 37389 37414.7 37521.9 37554.1
NO         0       0     1       0       1     0     0       0       0       1       1       0       0     0       1       0       0

     37558.5 37689.1 37706.5 37773.9 37850.6 37869.6 37930.9 38059.8 38080.9 38103.4 38248.3 38446.6 38453.7 38459.9 38540 38598.4
NO         1       0       1       1       0       0       1       1       1       0       2       1       0       0     1       1

     38784 39010.8 39175.8 39205.3 39253.6 39308.7 39358.3 39547.8 39666.6 39745.3 40949.9 40972.9 41016 41034 41107.2 41127.4 41438
NO       1       0       1       0       0       0       1       0       1       1       0       0     1     1       1       1     0

     41438.2 41462.3 41521.6 41558.1 41609.5 41627.1 42124.1 42173.9 42378.2 42579.1 42603.9 42628.3 43057 43228.2 43395.5 43499.5
NO         1       1       1       0       1       1       0       1       0       1       1       1     0       0       0       1

     43530 43719.5 43743.2 43799.6 43940.6 43943 44288.3 44658.6 44682.1 45031.9 45189.8 45342.5 45765 45856.1 46323.8 46358.4
NO       1       1       1       1       1     1       1       0       1       0       1       0     0       0       0       1

     46461.5 46587.9 46633 46870.4 46963.9 47025 47198.6 47750.2 47796.8 47835.8 48346.1 48720.3 48770.5 48950.9 48971.6 48974.8
NO         1       0     1       1       0     1       0       1       1       0       0       0       0       1       0       1

     49024.9 49175.7 49456.7 49673.6 49874.4 49917.3 50186.1 50409.9 50474.6 50576.3 50849.2 50897.6 51204.2 51284.3 51299.3 51417
NO         0       1       0       1       0       1       0       0       0       1       0       1       0       1       0     0

     51620.8 51879.3 52117.3 52255.9 52662.5 52670.6 52674 52769.9 53104.3 54314.5 54618.8 54863.8 55204.7 55263 55716.5 56031.1
NO         0       0       0       0       0       0     0       0       1       0       0       0       0     0       0       0

     56340.3 56394.3 56658.9 56842.5 57176.4 57398.1 57444.5 57671.7 57880.7 58092 58367.3 59175.1 59409.1 59503.8 59803.9 59805.6
NO         0       0       0       1       0       0       0       1       0     0       0       0       0       0       1       0
```

```
     56340.3 56394.3 56658.9 56842.5 57176.4 57398.1 57444.5 57671.7 57880.7 58092 58367.3 59175.1 59409.1 59503.8 59803.9 59805.6
NO         0       0       0       1       0       0       0       1       0     0       0       0       0       0       1       0

     60747.5 61554.6 63130.1
NO         0       1       0
 [ reached getOption("max.print") -- omitted 1 row ]
> print(chisq.test(bank.data))

        Pearson's Chi-squared test

data:  bank.data
X-squared = 600, df = 598, p-value = 0.4693
```

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	600.000[a]	598	.469
Likelihood Ratio	827.264	598	.000
N of Valid Cases	600		

a. 1198 cells (100.0%) have expected count less than 5. The minimum expected count is .46.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	1.000	.469
	Cramer's V	1.000	.469
N of Valid Cases		600	

The chi square value is: 600 and the p^2 value is 0.469. There are some slight differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.469 is less than 0.5 and suggests some correlation amongst the income and purchasing of a PEP but not by much. The phi and Cramer's V values also indicate there some significance in the relationship between income and purchasing a PEP.

In the next comparisons only the chi square test is observed:

pep * children

Crosstab

			children				
			0	1	2	3	Total
pep	NO	Count	167	25	79	55	326
		Expected Count	142.9	73.4	72.8	36.9	326.0
	YES	Count	96	110	55	13	274
		Expected Count	120.1	61.7	61.2	31.1	274.0
Total		Count	263	135	134	68	600
		Expected Count	263.0	135.0	134.0	68.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	99.164[a]	3	.000
Likelihood Ratio	104.902	3	.000
N of Valid Cases	600		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 31.05.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.407	.000
	Cramer's V	.407	.000
N of Valid Cases		600	

The chi square value is: 99.164 and the p^2 value is 0.000. There are some slight differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.000 is significantly less than 0.5 and suggests that the number of children and the purchasing of a PEP are correlated. The phi and Cramer's V values also indicates that children play a very significant role when a PEP is purchased.

pep * gender

Crosstab

			gender		
			FEMALE	MALE	Total
pep	NO	Count	170	156	326
		Expected Count	163.0	163.0	326.0
	YES	Count	130	144	274
		Expected Count	137.0	137.0	274.0
Total		Count	300	300	600
		Expected Count	300.0	300.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.317[a]	1	.251		
Continuity Correction[b]	1.135	1	.287		
Likelihood Ratio	1.317	1	.251		
Fisher's Exact Test				.287	.143
N of Valid Cases	600				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 137.00.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.047	.251
	Cramer's V	.047	.251
N of Valid Cases		600	

The chi square value is: 61.317 and the p^2 value is 0.251. There are some slight differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.251 is less than 0.5 and suggests some correlation amongst the gender and purchasing of a PEP. The phi and Cramer's V values indicate there is not much significance in the relationship between income and purchasing a PEP so there might be a chance that gender plays some role in when a PEP is purchased but it may not be very much.

pep * region

Crosstab

			region				
			INNER_CITY	RURAL	SUBURBAN	TOWN	Total
pep	NO	Count	146	50	28	102	326
		Expected Count	146.2	52.2	33.7	94.0	326.0
	YES	Count	123	46	34	71	274
		Expected Count	122.8	43.8	28.3	79.0	274.0
Total		Count	269	96	62	173	600
		Expected Count	269.0	96.0	62.0	173.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3.791[a]	3	.285
Likelihood Ratio	3.790	3	.285
N of Valid Cases	600		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 28.31.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.079	.285
	Cramer's V	.079	.285
N of Valid Cases		600	

The chi square value is: 3.791 and the p^2 value is 0.285. There are some slight differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.285 is less than 0.5 and suggests the region and purchasing of a PEP variable are correlated. The phi and Cramer's V values indicate there is slight significance in the relationship between region and purchasing a PEP. There might be a chance region isn't playing a big role in determining purchase of a PEP.

pep * married

Crosstab

			married		
			NO	YES	Total
pep	NO	Count	84	242	326
		Expected Count	110.8	215.2	326.0
	YES	Count	120	154	274
		Expected Count	93.2	180.8	274.0
Total		Count	204	396	600
		Expected Count	204.0	396.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	21.564[a]	1	.000		
Continuity Correction[b]	20.768	1	.000		
Likelihood Ratio	21.594	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	600				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 93.16.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-.190	.000
	Cramer's V	.190	.000
N of Valid Cases		600	

The chi square value is: 21.564 and the p^2 value is 0.000. There are some slight differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.000 is less than 0.5 and suggests that the marriage attribute and purchasing a PEP are correlated. The phi and Cramer's V values indicate there is not much significance in the relationship between marriage and purchasing a PEP so there might be a chance that married doesn't play a big role in purchase of a PEP.

pep car

Crosstab

			car		
			NO	YES	Total
pep	NO	Count	168	158	326
		Expected Count	165.2	160.8	326.0
	YES	Count	136	138	274
		Expected Count	138.8	135.2	274.0
Total		Count	304	296	600
		Expected Count	304.0	296.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.215[a]	1	.643		
Continuity Correction[b]	.145	1	.703		
Likelihood Ratio	.215	1	.643		
Fisher's Exact Test				.682	.351
N of Valid Cases	600				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 135.17.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.019	.643
	Cramer's V	.019	.643
N of Valid Cases		600	

The chi square value is: 0.215 and the p^2 value is 0.643. There are no differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.0643 is greater than 0.5 and suggests there is no correlation in owning a car and purchasing of a PEP. The phi and Cramer's V values indicate there is not much significance in the relationship between owning a car and purchasing a PEP, so the attributes are probably not correlated.

pep * savings_acct

Crosstab

			savings_acct		
			NO	YES	Total
pep	NO	Count	91	235	326
		Expected Count	101.1	224.9	326.0
	YES	Count	95	179	274
		Expected Count	84.9	189.1	274.0
Total		Count	186	414	600
		Expected Count	186.0	414.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.178[a]	1	.075		
Continuity Correction[b]	2.870	1	.090		
Likelihood Ratio	3.172	1	.075		
Fisher's Exact Test				.077	.045
N of Valid Cases	600				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 84.94.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-.073	.075
	Cramer's V	.073	.075
N of Valid Cases		600	

The chi square value is: 3.178 and the p^2 value is 0.75. There are some slight differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.75 is less than 0.5 and suggests some correlation amongst having a savings account and purchasing of a PEP. The phi and Cramer's V values indicate there is not much significance in the relationship between owning a savings account and purchasing a PEP so the savings account may not play a big role in determining if the customer will purchase a PEP.

pep * mortgage

Crosstab

			mortgage		
			NO	YES	Total
pep	NO	Count	209	117	326
		Expected Count	212.4	113.6	326.0
	YES	Count	182	92	274
		Expected Count	178.6	95.4	274.0
Total		Count	391	209	600
		Expected Count	391.0	209.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.351[a]	1	.554		
Continuity Correction[b]	.256	1	.613		
Likelihood Ratio	.351	1	.553		
Fisher's Exact Test				.606	.307
N of Valid Cases	600				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 95.44.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-.024	.554
	Cramer's V	.024	.554
N of Valid Cases		600	

The chi square value is: .351 and the p^2 value is 0.554. There are some slight differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.554 is greater than 0.5 and suggests there is no correlation amongst the mortgage variable and purchasing of a PEP. The phi and Cramer's V values indicate there is not much significance in the relationship between having a mortgage and purchasing a PEP, so they don't seem to be correlated.

pep * current_acct

Crosstab

			current_acct		
			NO	YES	Total
pep	NO	Count	82	244	326
		Expected Count	78.8	247.2	326.0
	YES	Count	63	211	274
		Expected Count	66.2	207.8	274.0
Total		Count	145	455	600
		Expected Count	145.0	455.0	600.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.379[a]	1	.538		
Continuity Correction[b]	.271	1	.603		
Likelihood Ratio	.380	1	.538		
Fisher's Exact Test				.566	.302
N of Valid Cases	600				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 66.22.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.025	.538
	Cramer's V	.025	.538
N of Valid Cases		600	

The chi square value is: .379 and the p^2 value is 0.538. There are some slight differences in the expected counts verses the actual counts in the data.

The result is significant if the p^2 is equal to or less than the alpha level (normally 0.5), the p^2 value of 0.538 is greater than 0.5 and suggests no correlation amongst the current account variable and purchasing of a PEP. The phi and Cramer's V values indicate there is not much significance in the relationship between current account and purchasing a PEP so there isn't much significance in the relationship of the current_acct variable and a PEP being purchased.

Summary

The attributes that appear to be correlated with purchasing a PEP include age, income, children, gender, region, married, and savings account. The variables that have the most significance appear to be age, married, and savings account. It can be concluded that the customers age and whether they are married will influence the purchase of a PEP. A customer who has a saving account will also be more likely to purchase a PEP. Lastly, the income variable also showed some outliers when compared to the purchase of a PEP