1. Principal Component Analysis (PCA)- Marketing Analysis

(1) What’s eigenvalues and eigenvectors? How can you interpret from these eigenvalues?

```
Standard deviations (1, .., p=6):
[1] 2.0390121 1.0348535 0.5832983 0.5330762 0.3014807 0.2370872

Rotation (n x k) = (6 x 6):
              PC1         PC2         PC3         PC4         PC5          PC6
Cloth 0.4714976 -0.13415181  0.16243669 -0.08442307 -0.36562355 -0.769740096
Home  0.4600906 -0.24606042  0.09368100 -0.02432775 -0.57674845  0.621099157
House 0.4072388 -0.43111070 -0.05711805 -0.51840960  0.60811412  0.080537846
Sport 0.3381880  0.55672451 -0.73009336 -0.19485229 -0.06775409  0.009610765
Toys  0.4344633 -0.08672486 -0.11562190  0.82259982  0.33710429  0.006498018
Elect 0.3115024  0.64663572  0.64433763 -0.09431065  0.21365378  0.122943426
```

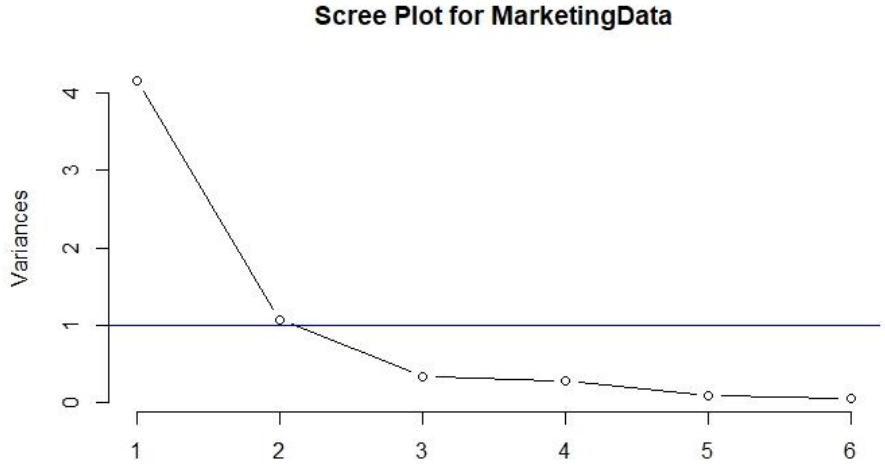
根據上圖，第一主成分特徵向量為(0.47, 0.46, 0.4, 0.34, 0.43, 0.31)，其餘主成分的特徵向量可由上圖類推。而 Standard deviation 則是 eigenvalue 的開根號，所以第一主成分的特徵值是 2.0390121 的平方，可得約為 4.158，其他主成分的特徵值如下圖。

```
> vars <- (pca$sdev)^2
> vars
[1] 4.15757024 1.07092166 0.34023690 0.28417026 0.09089061 0.05621032
```

特徵值代表著這份資料的變異（資訊量），由下圖每個主成分的解釋比例計算可以知道第一主成分占了所有變異的 0.69，因此若只取第一主成分就可以包含七成的資訊量。

```
> props <- vars / sum(vars)
> props
[1] 0.692928374 0.178486944 0.056706150 0.047361710 0.015148435 0.009368386
```

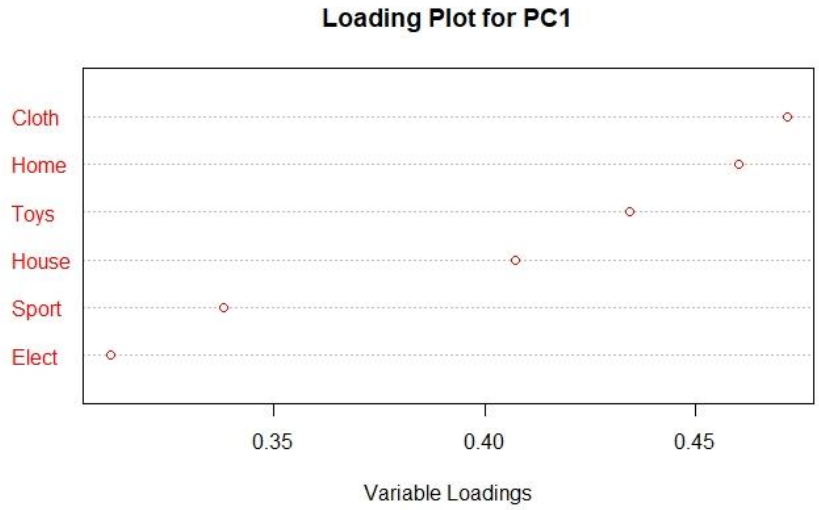
(2) Plot a scree plot and decide the most appropriate number of principal components to use.



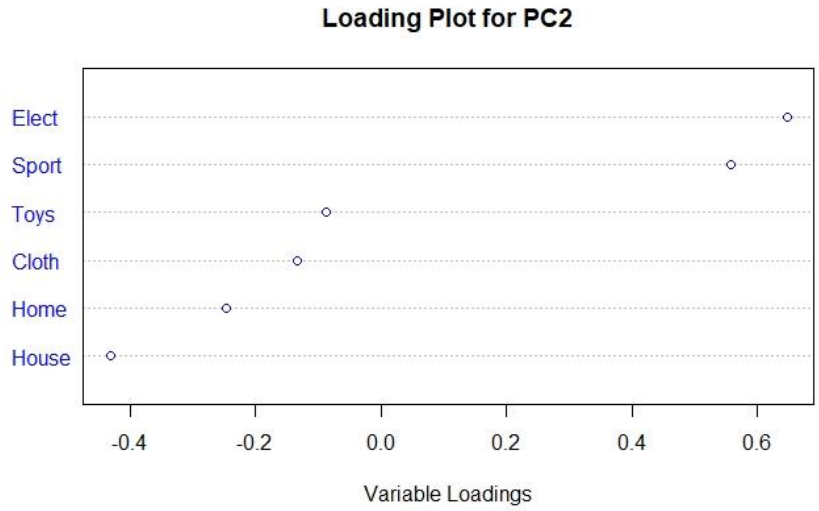
由圖可知，第三個主成分開始變異開始趨於平緩，而且只有第一主成分跟第二主成分的特徵值大於 1，因此取前兩個主成分即可。

(3) How to NAME the principal component? What conclusions can you draw from this analysis?

針對較為重要的第一主成分以及第二主成分繪製 Loading Plot。



從上圖可得知，每一個項目都跟 PC1 呈現正相關，因此每個項目的增加都會使得總體增加，但若細看的話可以發現 Cloth、Home、Toys 以及 House 四個項目有較高的 Variable Loading，而 Sport 及 Elect 則有較低的 Variable Loading，由於前四項比較偏向是家庭會購買的產品，後兩項雖然家庭也會買但應該需求相對較少，推測 PC1 比較傾向是跟「家庭消費」有關。



由圖可知，Elect 和 Sport 跟 PC2 呈現正相關，一般情況下男性相較於女性可能比較會去購買運動用具或是電子產品，因此推測 PC2 應該與男性顧客的消費比較有關係。

2. Stepwise Regression for Variable Selection

(1) Identify the important variable by linear regression (i.e. ordinary least squares, OLS)

將資料進行線性迴歸分析，根據最小平方法結果顯示，共三個變數呈現顯著，分別是 Cloth、House 及 Sport 三個變數，其中又以 Sport 的顯著性更高。

```
Call:
lm(formula = Elect ~ Cloth + Home + House + Sport + Toys, data = marketing_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46284 -0.26575  0.00986  0.30433  1.73774

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.09269    0.59358   3.526  0.00148 **
Cloth        0.13302    0.05535   2.403  0.02312 *
Home        -0.01201    0.15604  -0.077  0.93920
House       -0.40012    0.17395  -2.300  0.02910 *
Sport        0.30723    0.10398   2.955  0.00628 **
Toys        -0.04745    0.09116  -0.520  0.60683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6482 on 28 degrees of freedom
Multiple R-squared:  0.5896,    Adjusted R-squared:  0.5163
F-statistic: 8.044 on 5 and 28 DF,  p-value: 8.359e-05
```

(2) Identify the important variable by stepwise regression

若採用 Forward stepwise 的方式，則過程中依序會將 Sport、Cloth 以及 House 三個變數加入模型當中，可知這三個變數為篩選出來的重要變數，最終結果如下圖。

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.12540    0.56066   3.791 0.000676 ***
Sport        0.29758    0.09934   2.996 0.005449 **
Cloth        0.11606    0.03663   3.169 0.003511 **
House       -0.40234    0.15411  -2.611 0.013969 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.63 on 30 degrees of freedom
Multiple R-squared:  0.5845,    Adjusted R-squared:  0.543
F-statistic: 14.07 on 3 and 30 DF,  p-value: 6.633e-06
```

若採用 Backward stepwise 的方式，則過程會依序將 Home、Toys 從模型當中移除，剩下重要變數 Sport、Cloth 及 House，最終結果如下圖。

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.12540    0.56066   3.791 0.000676 ***
Cloth        0.11606    0.03663   3.169 0.003511 **
House       -0.40234    0.15411  -2.611 0.013969 *
Sport        0.29758    0.09934   2.996 0.005449 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.63 on 30 degrees of freedom
Multiple R-squared:  0.5845,    Adjusted R-squared:  0.543
F-statistic: 14.07 on 3 and 30 DF,  p-value: 6.633e-06
```

若採 Both stepwise，以 null 開始的加入移除變數順序恰跟 forward 一樣；而以 full 開始則加入移除變數順序恰跟 backward 相同，結果截圖同上兩圖。

(3) Give a comparison between (1) and (2). The results are consistent?

不論是使用簡單線性迴歸的方式或是使用 stepwise 方式，最終的重要變數都是 Sport、Cloth 以及 House，結果相同。