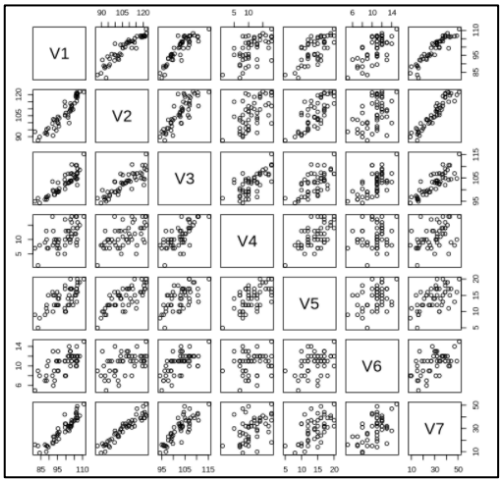
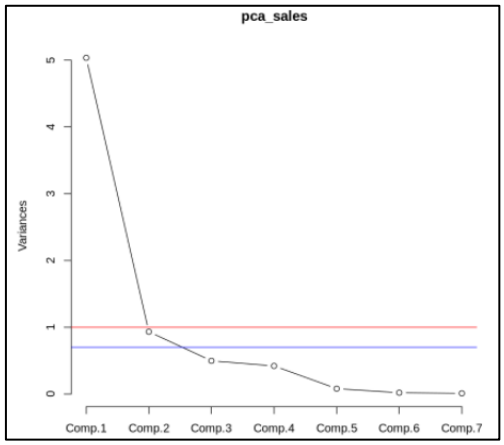
Q1: Perform a complete Principal Components Analysis for this data and interpret the result.





首先讀進資料，用散佈圖矩陣來觀看七個變量之間的關係，觀察到 V1~V3 兩兩成對的線性關係強烈，而 V4~V7 的部分由於能力測試的不同，這四種 test 沒有很明顯的線性關係，但比較特別的是 V7 Mathematics test 分值高的，在 V1~V3 上的表現也越好，有明顯線性關係。

原資料先標準化，經過分析後可以組成七個 Principal Components（PC），各自對變異的貢獻分別為 71%、13%、7%、6%……，繪製出 scree plot，以$\lambda = 1$的 criterion 來說應該只選一個 PC，再用 permutation test 去判斷每個 PC 的顯著性，可看出只有最大的那一個λ值顯著，所以應當只取第一個 PC，但是如此一來只保留了大約 71%的變異，而觀察第二個 PC 可以發現其λ值接近 1，或是從較寬鬆的 criterion（Jolliffe, 1971）來說，$\lambda = 0.7$以上也是可以考慮選取的範圍，但由於題目要求依據檢定結果，故而「不」取此 PC 以換取多 13%的變異被保留在模型裡，所以決定選擇只使用第一個 PC 來做後續分析，可以保留約 71.9%的變異。

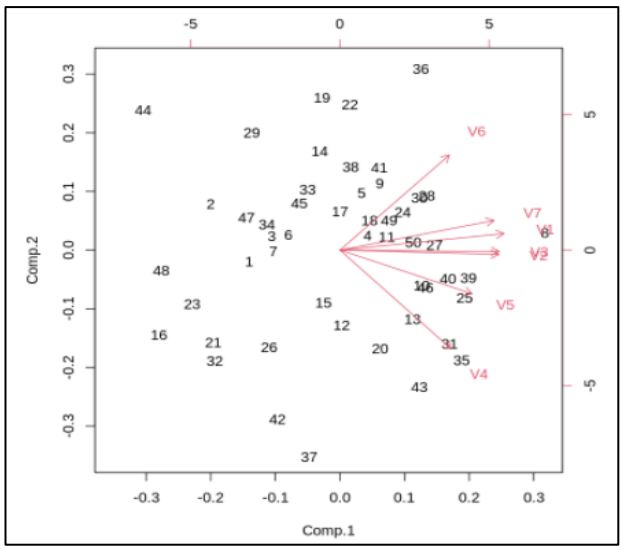
每一個主成份都是取原變量的 Linear Combination 後所得到的，觀察在原變數前面線性組合的係數，可以知道原變數對於主成份的正/負影響，影響程度有多大，我們前面所選的第一個 PCs 如下：

$$PC1 = 0.434V1 + 0.420V2 + 0.421V3 + 0.294V4 + 0.349V5 + 0.289V6 + 0.407V7$$

分析 PC1，在 V1~V3 的係數很接近，可見其應該是類似將 Sales growth、Sales profitability、New account sales 三個績效指標中取平均當作一個新的綜合指標的概念，而在 V4~V7，可以明顯感覺出 Mechanical reasoning test、Mathematics test 的突出相對另外兩項 test 來說，對於 PC1 來說是相對重要的，不過從此四個係數都是正的來判斷，在這些測試中越好，對 PC1 的影響都是正面的。

從 biplot 來分析，圖 1-3. biplot 上面有七個紅色箭頭，箭頭的方向都指向了 PC1 正向的地方，也就是右半部，代表這些變量都有正的係數做線性組合產生了 PC1，結合上面的 PC1 式子的係數皆為正可以相互映證，而任兩紅色箭頭之

間的夾角小於 90 度代表兩原變量之間正相關，夾角的餘弦值的絕對值越大代表關係越強，可以從圖 1-1 看出一樣的結果。



Q2: Perform a complete exploratory Factor Analysis based on MLE and interpret the result.

```
          Chi-squared type tests for Multivariate Normality

data : sales

McCulloch (S2)             : 2.96392
p-value.S2                 : 0.08514131

Nikulin-Rao-Robson (Y2)    : 7.556563
p-value.Y2                 : 0.1092408

Dzhaparidze-Nikulin (U2)   : 4.592643
p-value.U2                 : 0.2041741

Result  : Data are multivariate normal (sig.level = 0.05)
```

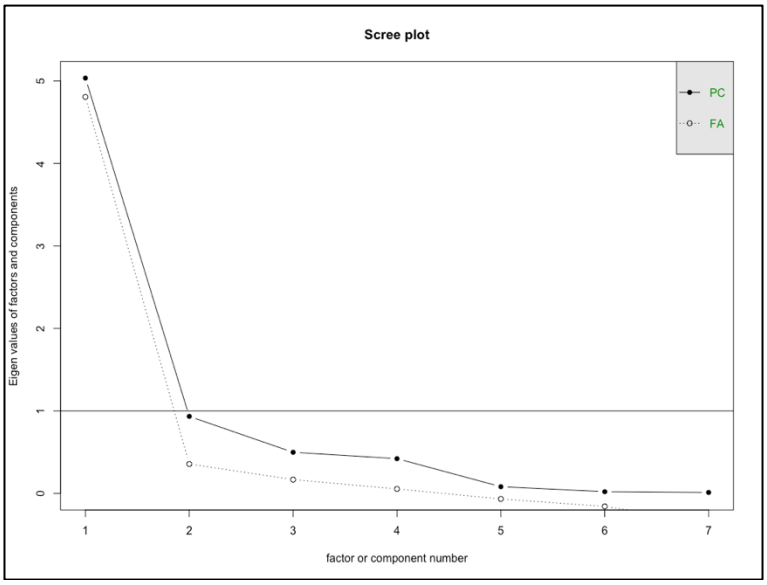
由於題目要求使用 MLE 法來做，MLE 有分配假設，所以先對資料做檢定，觀察其是否為 multivariate normal distribution，在 MVN package 中的檢定均顯示並無服從常態分配，改使用 mvnTest package 中的檢定，最後由 S2.test（Chi-square type tests of Multivariate Normality）中獲得資料服從 multivariate normal 的結果。

FA with 1 factor	FA with 2 factors	FA with 3 factors
Loadings: Factor1 V1 0.975 V2 0.959 V3 0.902 V4 0.567 V5 0.712 V6 0.615 V7 0.953 Factor1 SS loadings 4.799 Proportion Var 0.686	Loadings: Factor1 Factor2 V1 0.852 0.452 V2 0.868 0.419 V3 0.717 0.602 V4 0.148 0.987 V5 0.501 0.525 V6 0.619 V7 0.946 0.277 Factor1 Factor2 SS loadings 3.545 2.071 Proportion Var 0.506 0.296 Cumulative Var 0.506 0.802	Loadings: Factor1 Factor2 Factor3 V1 0.793 0.374 0.438 V2 0.911 0.317 0.185 V3 0.651 0.544 0.438 V4 0.255 0.964 V5 0.542 0.465 0.207 V6 0.299 0.950 V7 0.917 0.180 0.298 Factor1 Factor2 Factor3 SS loadings 3.175 1.718 1.453 Proportion Var 0.454 0.245 0.208 Cumulative Var 0.454 0.699 0.906

計算 Communality 的部分，如下表：

# of factor	V1	V2	V3	V4	V5	V6	V7
1	0.9512075	0.9191309	0.8128338	0.3214652	0.5074988	0.3777667	0.9088059
2	0.9308084	0.9296196	0.8766912	0.995	0.5264151	0.3863614	0.9711830
3	0.9614284	0.9655193	0.9118782	0.995	0.5533795	0.995	0.9624902

在選擇 common factor 數量上，由於對此組資料沒有其他 theory、研究可以去輔助選擇 common factor，故決定用類似 PCA 挑選 PC 的方式去決定 common factor 的個數，由下圖可知，我會選 Eigenvalues of factor 大於一的數目當作我的 common factor 選擇標準，故決定選擇 common factor 的數目為 1，此外，根據檢定結果（"Test of the hypothesis that 1 factor is sufficient."），其 p-value 為 2.02e-27，一個 common factor 足矣，在此想做保守的選擇，故而不選擇其他 common factor 的數目。



在 communality 的部分，一個 common factor 可以在 V1、V2、V3、V7 解釋的很好，對 V5（Mechanical reasoning test）的解釋表現平平，而對 V4（Creativity test）、V6（Abstract reasoning test）的解釋表現略差。

還是可以多分析一下選定兩個 common factor，依舊可以看出 V1、V2、V3、V6、V7 大致上屬於第一個 Factor 所影響，而 V4 則由第二個 Factor 所影響，而 V5 同時被兩個 Factors 影響，已經初步的發現，除了 V5 外，
其餘的原變量都可以被不同的 Factor 解釋的很好，而兩個 Factors 可以解釋的變異約有 80%。

選定三個 common factor 的情況下，可以解釋的變異來到了 90%，可以發現 V1、V2、V7 會受到 Factor 1 所影響，V4 會被 Factor 2 影響，V6 會被 Factor 3 影響，V3、V5 同時受到 Factor 1、2 影響。

綜合三個情況的討論統整，比較明顯的是 V1、V2、V3、V7 受到同一個

Factor 的影響明顯，而在 V4、V6 各受到一種 Factor 所影響，V5 則受到 V4、V6 背後的 Factor 同時影響，那在沒有其他額外資訊幫助選擇 factor 個數的情況下，會選擇一個 common factor，粗略的分出了 V1、V2、V3、V7 與其他，模型較為簡單，可以解釋的變異也不差；而當如果有其他額外資訊提供做分析的話，我會選擇 3 個 Factors，可以幾乎看出不同 Factor 各自掌控的不同的原變量，且解釋變異高。

Q3: How do the factors obtained in Q2 compare to the principal components obtained in Q1?

從 Q1、Q2 可知，恰好都選擇了相同數目的 PC 跟 common factor，在此比較同樣數目下的 PC 跟 common factor 的表現。

首先從可解釋的變異上的差異來論述，PCA 及 FA 可解釋的變異比例在同一組資料下分別為 71.9%與 68.6%，PCA 專注在找一個投影方向可以使保留的變異最多，在解釋變異的表現上自然相較 FA 好，而 PCA 的理論方法就是為了要在降維下保存資料的最大變異，故 PCA 在這點上表現比 FA 好。

再來從結構的概念來分析，在 PCA 中，每一個 PC 都是由原資料（V1~V7）的線性組合所構成，而在 FA 中，是由各種不同的 common factors 與 factor loading 去組成原資料（V1~V7）的變量，簡而言之就是 PCA 是利用原資料的變量去尋找線性組合形成 Principal Components，而 FA 就是藉由原資料想要找到背後的 Factors，而這些 Factors 的某種組合組成原資料的變量。

PCA 可以找到唯一的一組且能保留最大變異的軸，而 FA 中在$LL' + \Psi$的結構下 L 不唯一，所以可以藉由 factor rotation 找到方便我們解釋的 L，而在 R 裡面已經預設使用 VARIMAX 的 rotation。

解釋意義上，PCA 每一個 loadings 數值相去不遠，就像是一個新的綜合指標，將所有原資料每個變量的做一個平均的概念，FA 的 loading 更像是一個想要找出不同變量背後那隻看不見的手的概念，最好的結果就是想要找到各個因素，可以讓相似性質的原變量被 Factor 所掌控，然後不同 Factor 掌控不同的原變量，最好能各司其職，盡量不要一個原變量同時被多個 Factor 掌握，以較精簡的 Factor 去代表原來較複雜的原變量的結構。

Q4: Suppose a researcher is interested in finding how the density of air pollution contents (PT, CO, SO2) is related to the set of variables (Temp, Man, Pop, Rain). Perform a complete Canonical Correlation Analysis for these two groups of variables and interpret the result.

Test	p-value	Result
Mardia	<NA>	No
Henze-Zirkler	2.159284e-08	No
Royston	2.208885e-20	No

Doornik-Hansen	7.067618e-62	No

由 MVN package 中的四種 Multivariate normal 檢定可知，在這組資料中，四種檢定的結果都顯示了此資料不滿足多元常態的假設，畫出原資料各個變量的分佈圖，可以看出本題所用到的變量「幾乎」都呈現右偏，故決定對每個變量都取對數，取完對數之後再畫一次圖，發現比較接近對稱跟常態的形狀，再一次使用 MVN package 檢定轉換後的資料是否服從多元常態分配，結果依舊是失敗，改嘗試 mvnTest package，在 S2.test（Chi-square type tests of Multivariate Normality teest）以及 CM.test（Cramer-von Mises test for Multivariate Normality test）下，p-value 分別為 0.1447711 以及 0.1465853 不顯著，可認為經過對數轉換的資料服從多元常態分配。

經過轉換再檢定後，檢定結果顯示資料已經服從多元常態分配假設，接著對這筆新資料做 Rao's Approximate F-test 去檢定 Canonical correlation 的顯著性，呈現結果如下：

Wilk's Lambda, using F-approximation (Rao's F)					
	stat	approx	df1	df2	p.value
1 to 3:	0.5287778	2.04793024	12	90.24705	0.02858857
2 to 3:	0.9252070	0.46239841	6	70.00000	0.83378923
3 to 3:	0.9967748	0.05824196	2	36.00000	0.94351036

進行 CCA，可以發現第一組的相關性約為 65.46%、第二組約為 26.80%，經過 Rao's Approximate F-test 可知，只需要取第一組即可（p-value = 0.0286），部分報表整理如下：

$cor	0.6545810	0.2679539	0.0567911	
$xcoef	0.01077289	-0.01747191	-0.14884873	
$ycoef	0.11268998	-0.12706441	0.08189457	-0.04695085

整理出關係式如下：

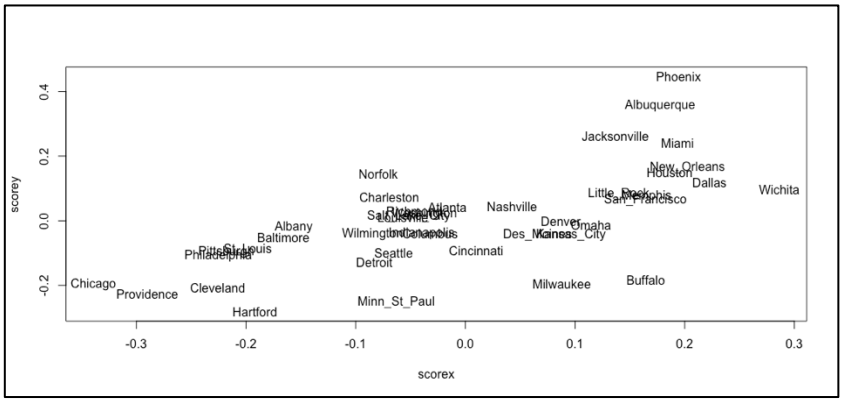
$$u1 = 0.01077289 * PT - 0.01747191 * CO - 0.14884873 * SO2$$

$$v1 = 0.11268998 * Temp - 0.12706441 * Man + 0.08189457 * Pop - 0.04695085 * Rain$$

首先觀察 u1，越往右邊代表空中的懸浮微粒（PT）越多，CO 和 SO2 越少，因為係數 PT 是正的其餘兩個是負的；觀察 v1，越上方的城市，均溫越高，人口越多，工廠可能較少也較少下雨。

此份資料由於不確定年代，所以只好依照現況來推測，先從橫向來看，Wichita 的主要產業有飛機製造、航空運輸等產業，New Orleans 為工業城市，有石油化工等重工業，所以在排煙等方面造成懸浮微粒（PT）較高，CO、SO2 等溫室氣體可能也較多，但是根據懸浮微粒與 CO、SO2 等溫室氣體的特性，溫室氣體容易被氣候帶到不同地方，而懸浮微粒的地區性較強，比較容易留在原地方，所以造成了在 u1 上的值比較大的原因，所以位在下圖的右側，而在左側，Chicago 為金融中心，其產業主要為金融、商業中心，在懸浮微粒、溫室氣體上相較於工業城市來說可能會較少，故羅列在下圖的左方，推測造成橫軸部

分不同的原因大致上有當地產業以及當地氣候造成的風向會影響懸浮微粒、溫室氣體的多寡；而從縱軸來看，亦跟當地產業影響人口聚集，以及城市所在位置的氣候影響溫度與雨量。



Q5: Perform a Simple Correspondence Analysis on this data set and interpret the result.

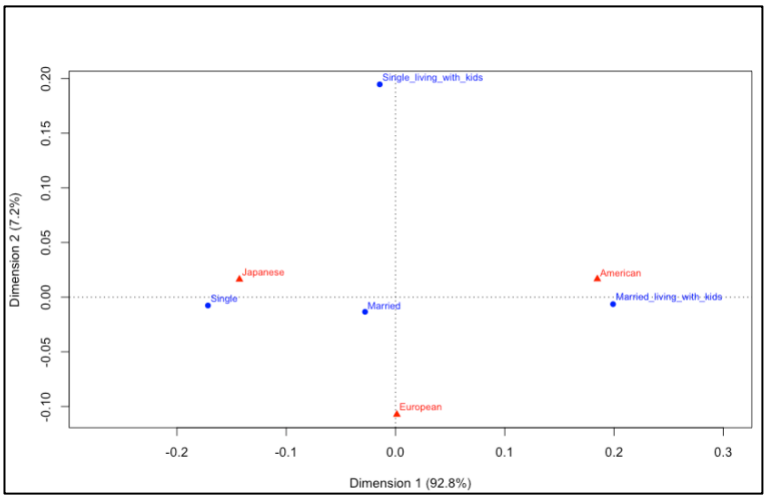
將資料讀進，是一個 4X3 的列連表，solution 的最大維度是 min (4-1, 3-1) = 2，決定選擇 nd=2 進行分析。

```
 Principal inertias (eigenvalues):
           1        2
Value      0.022866 0.001764
Percentage 92.84%   7.16%


 Rows:
          Married Married_living_with_kids    Single Single_living_with_kids
Mass     0.300885                 0.327434  0.327434                 0.044248
ChiDist  0.030886                 0.199222  0.171767                 0.195231
Inertia  0.000287                 0.012996  0.009661                 0.001687
Dim. 1  -0.184062                 1.316796 -1.134789                -0.095231
Dim. 2  -0.318833                -0.152043 -0.181484                 4.636157


 Columns:
        American  European  Japanese
Mass    0.377581  0.132743  0.489676
ChiDist 0.185461  0.107350  0.143708
Inertia 0.012987  0.001530  0.010113
Dim. 1  1.221544  0.008525 -0.944224
Dim. 2  0.395307 -2.556024  0.388083
```

在此的 value 代表從 SVD 得到的 Square of singular values，第一個維度解釋了 0.022866 / (0.022866+0.001764) = 92.84%，第二個維度解釋以此類推，解釋了 7.16%，接著做圖如下，嘗試解釋其關係：



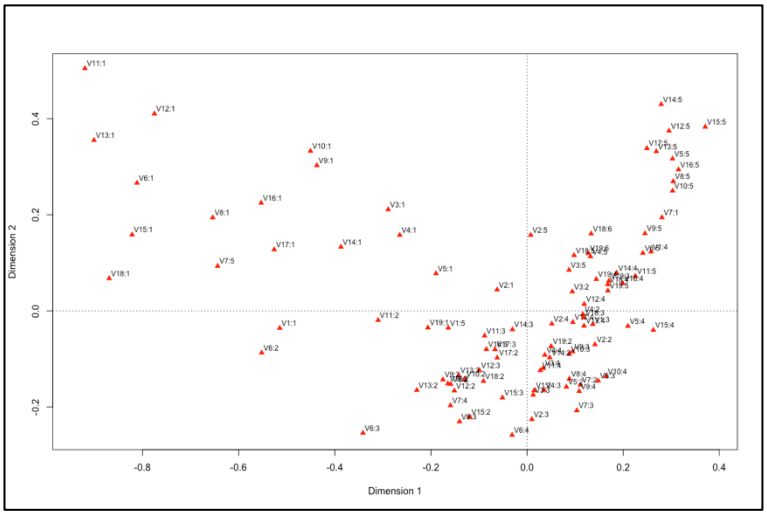
從圖來看，單身不跟小孩生活的車主偏好日本車，已婚且與小孩生活的車主偏好美國車，解釋是單身不跟小孩生活的人選日本車比較經濟實惠且空間不需要太大，正好是日本車的特點，而已婚與小孩生活的人選美國車可能是有出遊、採購等需要空間載人裝貨等要求所以需要空間較大的車款，比較符合美國車的特性，而已婚不小孩生活的車主距離三種車款的距離沒有差很多，略偏日本車一點，推測可能是只有夫妻兩人所以選日本車夠用就行，單親的部分跟不同 origin 的汽車距離都很遠，足見單親可能都不買車。

Q6: Perform a trustful Multiple Correspondence Analysis on this data set and interpret the result.

首先，最大的維度數目是 97（每一個變量內的類別數目加總）- 19（變量的數目）= 78，若要視覺化，兩的解釋比例相加必須超過一定比率才會 trustful，下表是不同方法下，前兩個維度的可解釋比例：

indicator	Burt	Adjusted	JCA
11.34%	33.70%	62.56%	64%

JCA 可以解釋的比例最多，達 64%，所以選擇 JCA。



從左半邊可以看出，11、13、12、6、15、18、16、17、10、9 等都是 1，而右半邊的上面幾乎都是 5，再觀察中間密集的部分可知，由左到右沿著 Dim. 1 的軸是從 1 慢慢過渡到 5，可以看出越往右越發認同、信任網購、電話傳真購物等方式，越往左則反之。

觀察第一象限，可以看出第一題回答 4 的人，也就是過去六個月花費超過五百元的人明顯地在使用 call、fax、web 等方式提供卡號以及購買資訊是足夠信任的，在 12、14、15、13、17、16、5 等題使用不同方式（call、fax、web）提供卡號等個人資訊的選項都選擇了 very likely，且右上角的點位都很密集，表示這些關聯強烈，此外，這些購買金額夠高的人，在購買的頻率上也很高。

觀察第二象限，這邊的人在藉由電話傳真、網購明顯的有不信任感，這邊的問題大多針對消費方式的安全性，而填答者都選一，此外這邊的點位排佈鬆散，可能是不同填答者只針對部分不同意，而非質疑所有消費方式的安全性。

觀察第三、四象限，這邊的填答者對購買安全性有疑慮，但是仍會使用這些方式購物，也有一定的購買頻率約一個月一次，但購買金額小，對於電話傳真、網購有一定的認知，也知道其便利性，但仍不太信任。

這邊我會主要將填答者分成三塊，第一象限的是對於電話傳真、網購等平台安全性足夠信任且購買頻率高、金額較多的消費者，基本上對於這些購買方式駕輕就熟，是企業商家需要爭取的對象，可以發展成穩定客源；第二象限的部分，這些人對平台安全性不信任，購買頻率跟金額均低，企業商家要爭取這些人可能要花很多成本但可能收效甚微，不建議以此象限的人當作客群；在三四象限的人，知道這些購買平台的好處，也願意小額購買，購買頻率普通，但部分對於平台安全性仍舊不夠信任，且這裡的點位更為密集，足見關聯很強，企業商家如果能拉動這群人，能在平台安全性的部分使人信任，那將會是很大的商機。