Universidad Nacional Autónoma de México

Tarea 1

Analisis multivariado

Acosta Imandt Daniel

15 de septiembre de 2022

Código

Para poder asegurar la reproducibilidad de los datos y las gráficas, el código y los datasets se encuentran aquí.

Ejercicio 1

Para los datos correspondientes al rendimiento, millas por galón de combustible, de los automóviles provenientes de Japón, Norte América y Europa, encontrar en cada caso las estadísticas:

$$x_*, x^*, M, F_L, F_U, b_L, b_U, \bar{x}$$

Haga un análisis de estos datos usando diagramas de cajas, desarrolle sus conclusiones.

Solución

Algunos cálculos se hicieron con ayuda de python con, se obtuvieron los siguientes:

count	74.000000
mean	16.979730
std	6.178346
min	6.000000
25%	12.125000
50%	18.000000
75%	20.875000
max	34.000000

Figura 1: Cálculos estadísticos de las millas por litro

Primero vamos a calcular la media, para eso tenemos que:

$$\underline{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 16.98}$$

Vamos a calcular M, tenemos 74 muestras en el data set, por lo que para sacar la mediana hacemos lo siguiente:

$$\underline{M = \frac{1}{2}\{x_{(n/2)} + x_{(n/2+1)}\} = 18}$$

Ahora calculamos sus quartiles, para eso tenemos que calcular la mediana de:

$$c_1 = \{x_1, x_2, ..., x_{n/2}\}$$

$$c_2 = \{x_{n/2+1}, x_{n/2+2}, ..., x_n\}$$

Por lo tanto llegamos a que: $\underline{F_L = 12.125 \text{y} F_U = 20.875}$.
Ahora queremos ver las barras externas, estas se calculan de la siguiente forma:

$$\underline{b_U = F_U + 1.5(F_U - F_L) = 20.875 + 1.5(20.875 - 12.125) = 34}$$

$$\underline{b_L = F_L - 1.5(F_U - F_L) = 12.125 - 1.5(20.875 - 12.125) = -1}$$

Ahora por los bigotes, los cuales sabemos que están dados por las siguientes formulas:

$$\underline{x^* = max\{x \in \{x_{(1)}, ..., x_{(n)}\} : x \leq b_U\} = 34}$$

$$\underline{x_* = min\{x \in \{x_{(1)}, ..., x_{(n)}\} : x \geq b_L\} = 6}$$

Por ultimo vamos a hacer un análisis de los datos utilizando el box plot.
Primero logramos notar que la media es menos a la mediana, esto nos puede decir que por lo generar hay mayor cantidad de coches que riden menos de 18 millas por litro, pero que al comprar un coche esperamos que su rendimiento se encuentre rondando estos valores es decir alrededor de las 17millas por litro,pero por otro lado notamos que la barra externa derecha es mucho más alargada que la izquierda, esto nos dice que hay unos datos anómalos mas dispersos por el lado superior, es decir que existen pocos coches que tienen una gran eficiencia comparado por el otro lado que no hay tantos coches con muy mala eficiencia.

Ejercicio 2

1. Producir un diagrama de cajas para los dos grupos;billetes genuinos y billetes falsos usnado la componente X_1 y $\mathbb{X}$

2. Calcular las estadísticas $M, F_U, F_L, b_U, x^*, x_*$ y para los dos grupos usando la componente X_6

3. Comentar y comparar los dos análisis: las cajas para X_1 y las cajas para X_6

Soluciones

1. A continuación viene el diagrama de cajas que se pide

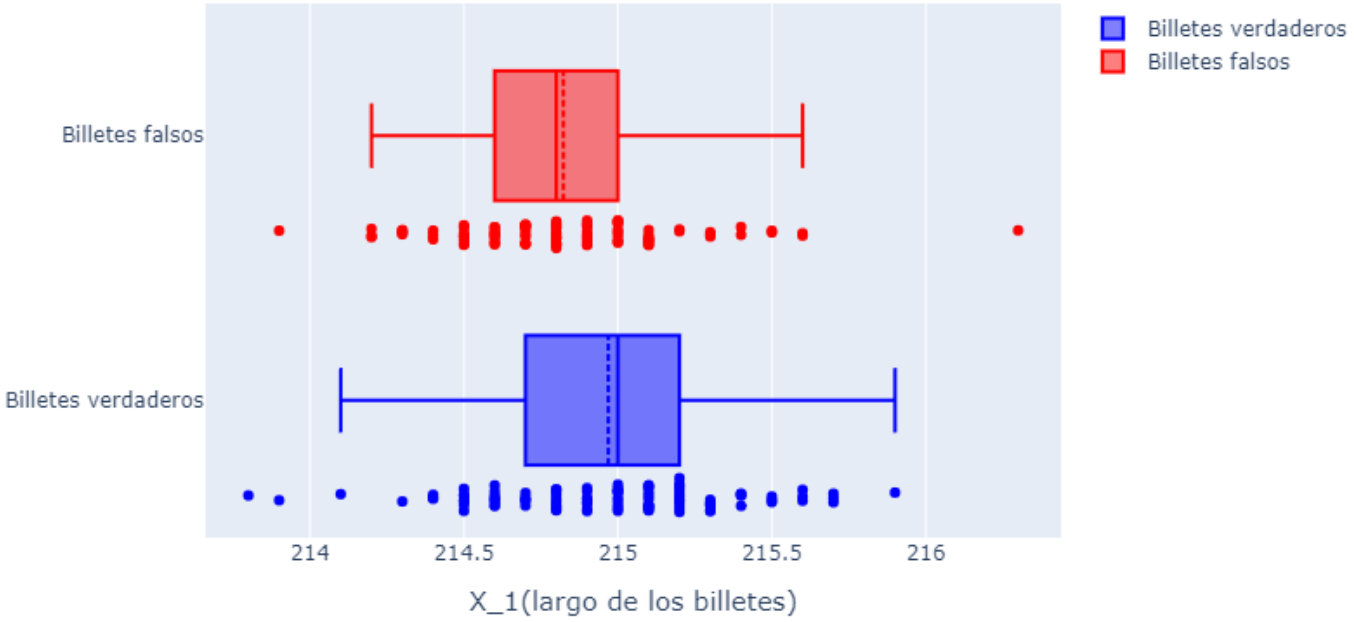

Figura 2: Diagrama de caja para la anchura de los billetes falsos y verdaderos

2. Ahora estamos interesados en calcular las estadísticas para los billetes falsos y verdaderos en la componente X_6.
   Primero vamos a hacer las box plots para esta componente.

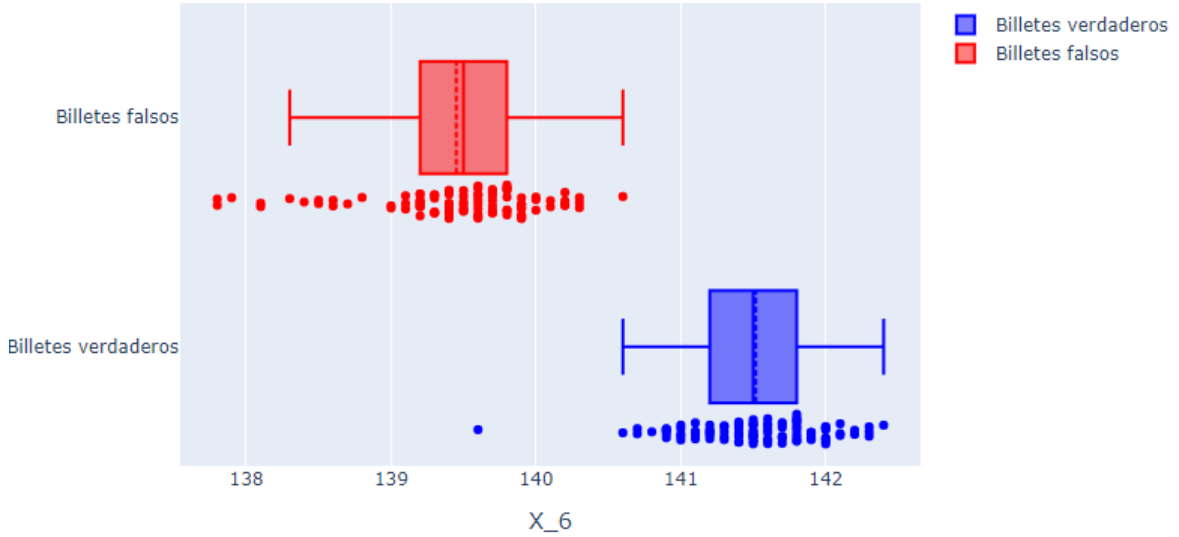

Figura 3: Diagrama de caja para la variable X_6 de los billetes falsos y verdaderos

Como se hizo en el ejercicio anterior y utilizando *python* llegamos a que, para los billetes verdaderos y falsos se tienen las siguientes estadísticas:

count	100.000
mean	141.517
std	0.447
min	139.600
25%	141.200
50%	141.500
75%	141.800
max	142.400

Figura 4: billetes verdaderos

count	100.000000
mean	139.450000
std	0.557864
min	137.800000
25%	139.200000
50%	139.500000
75%	139.800000
max	140.600000

Figura 5: billetes falsos

Por lo tanto llegamos a que para los billetes verdaderos tenemos:

$$M = 141.5$$
$$F_U = 141.8$$
$$F_L = 141.2$$
$$b_U = 142.7$$
$$b_L = 140.3$$
$$x^* = 142.4$$
$$x_* = 140.6$$

Y para los billetes falsos tenemos que:

$$M = 139.5$$
$$F_U = 139.9$$
$$F_L = 139.2$$
$$b_U = 140.7$$
$$b_L = 138.3$$
$$x^* = 140.6$$
$$x_* = 138.3$$

3. Por ultimo vamos a comparar y comentar los dos análisis:
   Logramos notar que para el largo del billete, los verdaderos y los falsos se comportan de una forma muy similar, lo que hace bastante difícil lograr diferenciar una muestra no etiquetada, ya que justo los bigotes de los verdaderos abarcan el área de los falsos, por otro lado si nos fijamos en la variable X_6 se nota que hay una clara tendencia entre los billetes falsos y los verdaderos, donde los falsos son menores que los verdaderos, de hecho algo bastante bueno es que el F_U de los falsos es casi el mismo que el F_L de los verdaderos por lo que esta variable nos sirve mucho más para saber si un billete es falso o verdadero que la de el largo de los billetes.

Ejercicio 3

Haga un resumen respecto a cómo funcionan los histogramas y las estimaciones de la densidad para unos datos. Explique con cuidado qué es lo que los paquetes dibujan y describa un ejemplo con datos

Resumen

Un estimador de densidad como lo es el histograma, nos da una buena impresión de la distribución de los datos.La idea es poder representar la densidad de los datos contando el numero de observaciones en una secuencia de intervalos consecutivos(bins) con origen en x_0 donde $B_j(x_0, h)$ denota la altura del bin del elemento que empieza en x_0

$$B_j(x_0, h) = [x_0 + (j-1)h, x_0 + jh) \quad j \in \mathbb{Z}$$

Ahoera si $\{_i\}_{i=1}^n$ es independiente e idénticamente distribuida con densidad f, el histograma se define de esta forma:

$$\hat{f}_h(x) = n^{-1}h^{-1} \sum_{j \in \mathbb{Z}} \sum_{i=1}^{n} \mathbf{I}\{x_i \in B_j(x_0, h)\}\mathbf{I}\{x \in B_j(x_0, h)\}$$

Donde $\mathbf{I}\{x_i \in B_j(x_0, h)\}$ cuenta el número de observaciones que caen en el bin $B_j(x_0, h)$ y el segundo indicador es responsable de contabilizar los puntos alrededor de x.
Por ultimo la detección de modas, requiere de métodos de alisado para encontrar el "óptimo" h para n observaciones:

$$h_{opt} = (\frac{24\sqrt{\pi}}{n})^{1/3}$$

Ahora utilizando los datos de de cars, queremos ver que tan económicos, buen diseño y seguros creen que son los coches vistos por los usuarios,para eso hacemos los siguientes histograma.

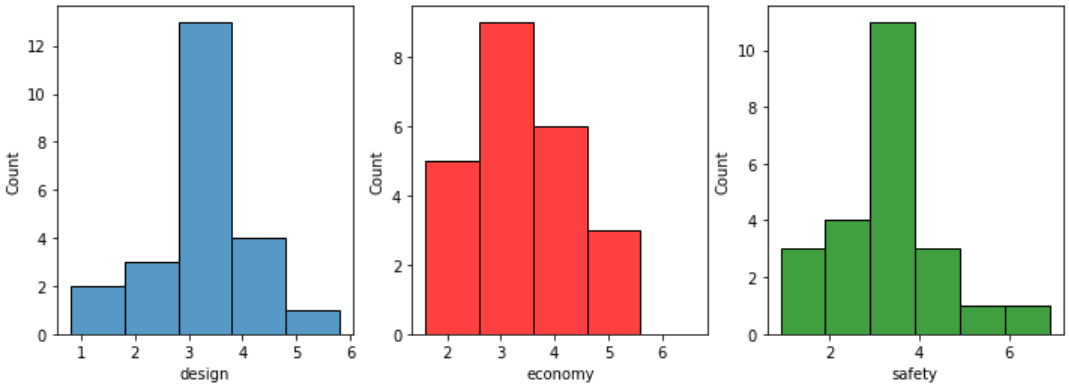

Figura 6: Histogramas

Ejercicio 4

Calcular de nuevo las componentes principales pero usando un re-escalameinto de los datos $\bar{X}$. Por ejemplo si se asume que las variables x_1, x_2, x_3 y x_6 fueron medidas en cm y que x_4, x_5 se quedan como estaban originalmente, o sea en escala de mm, esto sería equivalente a re-escalar $\bar{x_1} = x_1/10, \bar{x_2} = x_2/10, \bar{x_3} = x_3/10$ y $\bar{x_6} = x_6/10$ y usar la matriz de datos $\bar{X}$ Compare sus resultados con la Figura C(obtenido al calcular las componentes principales usando los datos originales sin ningún re-escalamiento,¿Qué se observa?)

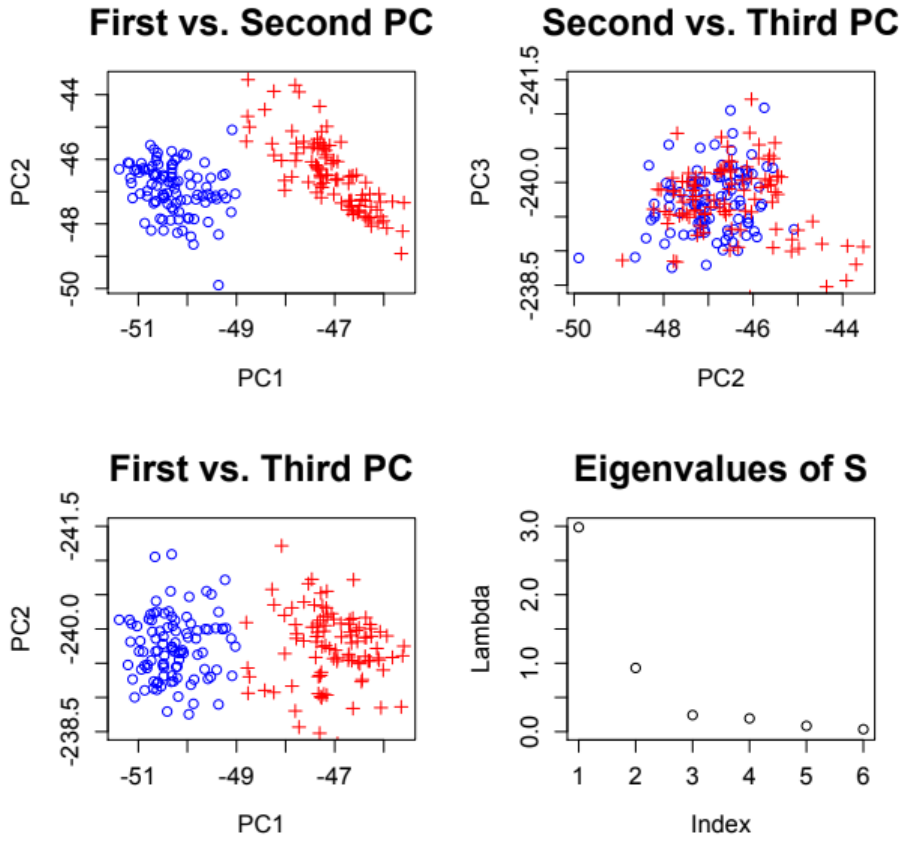

Figura 7: Figura C

Una vez hecho el re-escalamiento y aplicando PCA para tener sus componentes principales llegamos a las siguientes gráficas.

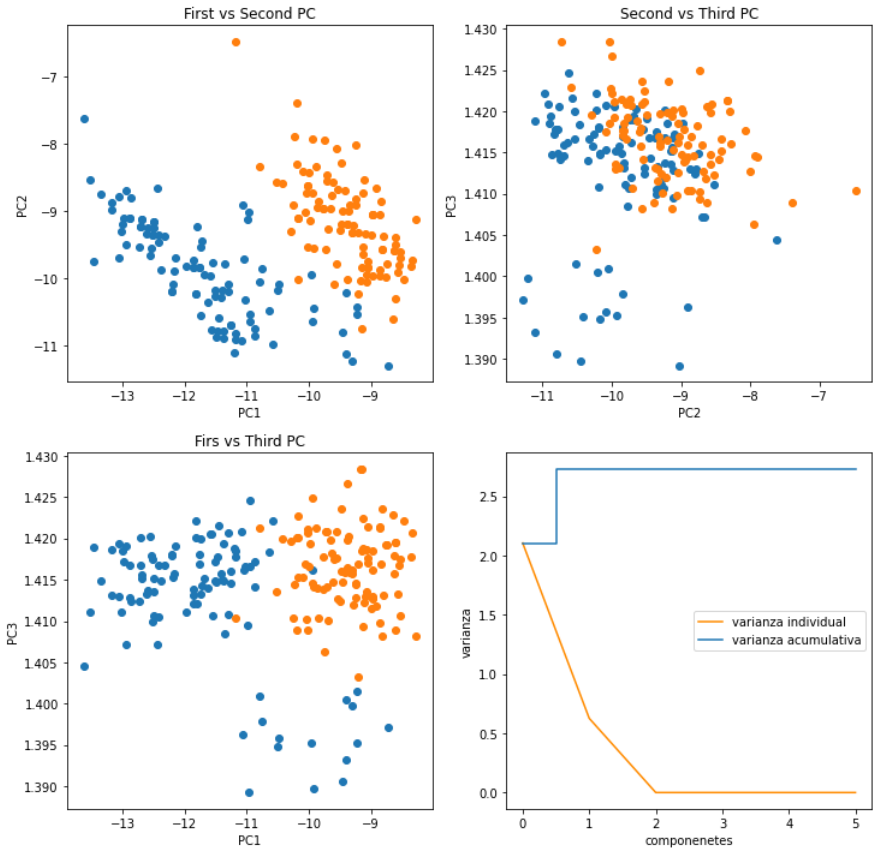

Figura 8: Gráficas de comparación de las componentes principales con los datos re-estandarizados

A partir de lo obtenido se pueden llegar a resultados interesantes,se logra apreciar de una manera muy contundente que las gráficas son muy diferentes a cuando no había un re-escalamiento, esto se debe a que las componentes principales son muy sensibles a cambios de escala de sus variables, por lo que es de gran importancia que al aplicar PCA se debe de hacer con componentes que tengan una escala parecida.
De hecho al observar los eigenvectores logramos observar que las componentes X_4, X_5 tiene mucho más peso y de hecho dominan, a las demas, con los datos re-escalados,cosa que no pasa con los datos sin re-escalar.

Ejercicio 5

Teorema: Sea X un vector aleatorio de dimensión $pX1$ tal que $E(X) = \mu$ y $Var(X) = \Sigma$. Sea Y el vector de componentes principales de X, entonces:

1. $COV(X, Y) = T\Delta$, donde T es la matriz de dimensiones pXp cuyas columnas son los vectores propios de Σ (en la descomposición de Jordan de Σ) y Δ es una matriz diagonal(de dimensiones pXp) que tiene los valores propios de Σ, $\lambda_1, ..\lambda_p$ como elementos de la diagonal.

2. La correlacción px_iy_i entre varaibles X_i y la componente principal Y_j esta dada por:

$$P_{X_iY_j} = \gamma_{ij}(\frac{\lambda_j}{\delta^2_{X_iX_i}})^{1/2}$$

Donde $\delta X_i^2 x_i = Var(X_i) = \Sigma_{ii}$ la entrada i, i de la matriz Σ y γ_{ij} es la entrada i,j de la matriz T.

Demostración

1. Por el **teorema** tenemos que Y es el vector de componentes principales en X, por lo tanto llegamos a que $Y = T^t(X - M)$, de aquí llegamos a que:

$$Cov(X, Y) = Cov(X, T^t(X - M))$$

Ahora por propiedades de la covarianza tenemos que:

$$Cov(X, T^t(X - M))) = Cov(X, (X - M))T = Cov(X, (X))T = Var(X)T$$

Recordando el **Teorema** tenemos que:

$$Var(X)T = \Sigma T = T\Delta T^t T$$

Ahora utilizando la propiedad de las matrices que $AA^t = I$ llegamos a que

$$Cov(X, Y) = T\Delta$$

■

2. Sabemos que la correlación esta dada por $Corr(X_i, Y_j) = \frac{Cov(X_i, Y_j)}{\sqrt{Var(X_i)}\sqrt{Var(Y_j)}}$, ahora por las hipótesis podemos decir que $\delta X_i^2 x_i = Var(X_i) = \Sigma_{ii}$ y además por las notas de la pagina 24 del tema de PCA vistas en clase, podemos decir que $Var(Y_j) = \lambda_j$, por ultimo sabemos que $Cov(X, Y) = T\Delta$ por la parte 1 de este ejercicio,entonces llegamos a que

$$Corr(X_i, Y_j) = \frac{T\Delta}{\sqrt{\Sigma_{ii}}\sqrt{\lambda_j}}$$

Por lo que tenemos que la correlación entre X_i y Y_j viene dada por la multiplicación de la fila i de T por la columna j de Δ, pero por las hipótesis sabemos que Δ es diagonal con los valores propios de Σ como elementos, por lo que llegamos a que:

$$Corr(X_i, Y_j) = \frac{\gamma_{ij}\lambda_j}{\sqrt{\Sigma_{ii}}\sqrt{\lambda_j}} = \frac{\gamma_{ij}}{\sqrt{\delta^2_{x_i x_i}}\sqrt{\lambda_j}} = \gamma_{ij}(\frac{\lambda_j}{\delta^2_{X_i X_i}})^{1/2}$$

■

Ejercicio 6

Verifique que:

- $$\bar{Z_{(n)}}^T = \begin{pmatrix} 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}_{pX1}$$

- $$\hat{\Sigma_Z} = G_{\hat{R}}^T \hat{\Sigma}^S G_{\hat{R}} = G_{\hat{R}}^T \hat{R} G_{\hat{R}}$$

  ¿Cuanto vale esta ultima matriz?

Verificación

1. Sabemos que $Z = X_s M_{\hat{R}}$ y además $\bar{Z_{(n)}} = \frac{1}{n} Z^t \mathbb{1}_{(1)}$ por lo tanto llegamos a que:

$$\bar{Z_{(n)}} = \frac{1}{n}(X_s M_{\hat{R}})^t \mathbb{1}_{(1)} = M_{\hat{R}}^t \left(\frac{X_s^t \mathbb{1}_{(1)}}{n}\right)$$

Ahora sabemos que $\frac{X_s^t \mathbb{1}_{(1)}}{n} = 0_{pX1}$, ya que está dentro el os paréntesis del vector de medias, pero sabemos que son 0,por lo tanto llegamos a que

$$\bar{Z_{(n)}}^t = 0_{pX1} = \underline{\begin{pmatrix} 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}_{pX1}}$$

2. Sabemos que $\hat{\Sigma_Z} = \frac{Z^t Z}{n} - \bar{X}_{(n)} \bar{X}_{(n)}^t$, pero sabemos que las medias son cero llegando a que $\hat{\Sigma_Z} = \frac{Z^t Z}{n}$ ahora recordando que $Z = X_s M_{\hat{R}}$ tenemos que

$$\hat{\Sigma_Z} = \frac{1}{n}(X_s M_{\hat{R}})^t (X_s M_{\hat{R}})$$

$$= \frac{1}{n} M_{\hat{R}}^t X_s^t X_s M_{\hat{R}} = M_{\hat{R}}^t (\hat{\Sigma^s}) M_{\hat{R}} = M_{\hat{R}}^t (\hat{R}) M_{\hat{R}}$$

Ejercicio 7

Considere los datos en el archivo *cars.dat*. Estos datos corresponden a calificaciones promedio que cuarenta personas le asignaron a 23 modelos de automóviles y se tiene que las variables(cualidades evaluadas fueron)

	Columna	Cualidad evaluada	Abreviatura (nombre)
X_1	1	Economía	Economy
X_2	2	Servicio	Service
X_3	3	No depreciación de su valor	Value
X_4	4	Precio	Price
X_5	5	El diseño (la apariencia)	Design
X_6	6	aspecto deportivo	Sporty
X_7	7	Grado de Seguridad del vehículo	Safety
X_8	8	Facilidad de Manejo	Easy

Figura 9: La calificaciones van desde 1 (muy bueno) a 6 (muy malo)

Haga un análisis de componentes principales de estos datos. Interprete y obtenga conclusiones relevantes usando las primeras 2 componentes principales.¿Es necesario utilizar la tercera componente principal? ¿Porqué si? o ¿porqué no?

Solución

Primero vamos a ver como se ven los datos:

	marca	economy	service	value	price	design	sportly	safety	easy
0	BMW3	4.8	1.6	1.9	5.0	2.0	2.5	1.6	2.8
1	CiAX	3.0	3.8	3.8	2.7	4.0	4.4	4.0	2.6
2	Ferr	5.3	2.9	2.2	5.9	1.7	1.1	3.3	4.3
3	FiUn	2.1	3.9	4.0	2.6	4.5	4.4	4.4	2.2
4	FoFi	2.3	3.1	3.4	2.6	3.2	3.3	3.6	2.8

Figura 10: Una muestra de los datos

Ahora a estos datos los vamos a estandarizar y sacar sus eigenvalores y eigenvectores, los cuales vienen a continuación:

```
Eigenvals: [5.63376131 1.929085   0.44233152 0.0373529  0.02844299 0.06210169
 0.11076142 0.11979954]
eigenvecs: [[-0.26785698  0.46930681  0.68081587  0.45994141 -0.00859217  0.0644491
  -0.16944764  0.00271849]
 [ 0.38241266  0.28523414 -0.12174478  0.05966054 -0.21694056 -0.68094733
  -0.38477131  0.30857649]
 [ 0.41010002  0.18124889 -0.04561881  0.20736697  0.79018561 -0.13456854
   0.13716629 -0.30425006]
 [-0.40877357  0.17035644  0.09865125 -0.58160446  0.4959989  -0.1205613
  -0.1285386   0.42286515]
 [ 0.40251176 -0.11245185  0.2216015   0.14033168  0.03801809  0.23012079
   0.42077006  0.72463113]
 [ 0.38197195 -0.10932092  0.62842286 -0.54579695 -0.15355564 -0.13536773
   0.08807154 -0.31571571]
 [ 0.37134859  0.3245322  -0.12925231 -0.2065853  -0.01223097  0.65430719
  -0.51847878  0.01754092]
 [-0.0302867   0.71176014 -0.22164416 -0.2082425  -0.23933337  0.01074385
   0.57816736 -0.09119081]]
```

Figura 11: Una muestra de los datos Eigenvalores y eigenvectores de los datos estandarizados

Con esto ya podemos ver cual es la varianza total de los datos y la acumulativa de estos.

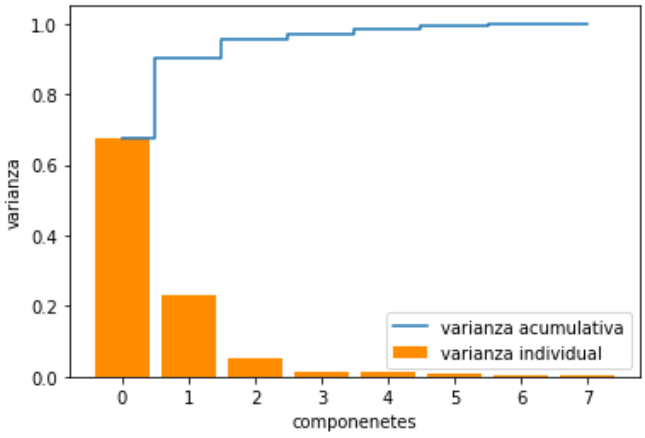

Figura 12: Varianza individual y acumulativa de los componentes principales

De aqui podemos ver que con solo las dos primeras componetnes principales ya se alcanza una confianza del 90 % la cual es muy buena y se logran eliminar otras 6 variables, haciendo que nuestros datos pesen mucho menos, pero ahora notamos que si también tomamos en cuenta la tercer componente principal se alcanza una confianza alrededor del 95 % de confianza, la cual es una muy aceptada y nos permite conseguir mejores resultados, por lo que si utilizaría las primeras tres componentes y ya no sería necesario utilizar las demás consiguiendo valores muy buenos.