Tarea 1

Garduño Martínez Eduardo

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: $M, \overline{x}, F_U, F_L, b_U, b_L, x^*, x_*$. Haga un análisis de estos datos usando los diagramas de cajas, desarrolle sus conclusiones.

	Variable 1	Variable 2
M	1	20
$\overline{x}$	1.4459	21.3243
F_U	2	24.5
F_L	1	18
b_U	3.5	34.25
b_L	-0.5	8.25
x^*	3	34
x_*	1	12

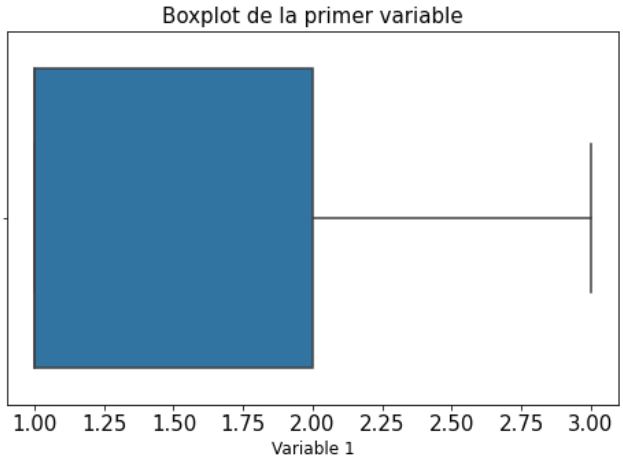

Figura 1:

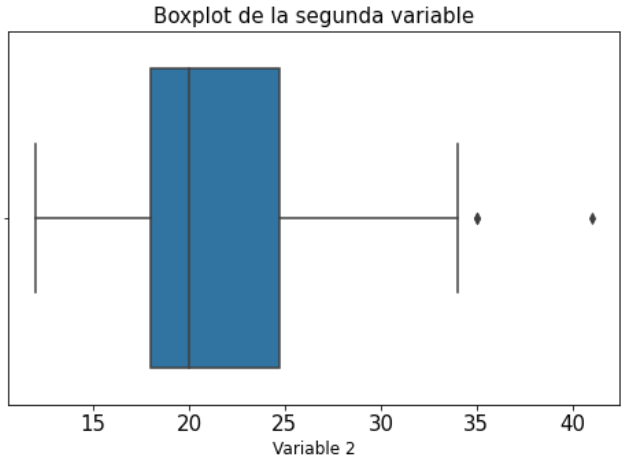

Figura 2:

Como podemos ver en el primer gráfico no se encuentra ningún tipo de dato atípico, además de que el mínimo se encuentra entre Q_3 y Q_1, esto quiere decir que hay una gran cantidad de unos en nuestros datos, por otro lado, segunda variable hay una mayor dispersión de los datos y podemos ver que se encuentran dos valores atípicos.

2. Producir un diagrama de cajas para los dos grupos: billetes genuinos y billetes falsos usando la componente X_1 de X. Calcular las estadísticas de $M, \overline{x}, F_U, F_L, b_U, b_L, x^*, x_*$ para los dos grupos usando la componente X_6. Comentar y comparar los dos análisis: las cajas para X_1 y las cajas para X_6.

	Variable 1	Variable 2
M	214.9	140.6
$\overline{x}$	214.896	140.4835
F_U	215.1	141.5
F_L	214.6	139.5
b_U	215.85	144.5
b_L	213.85	136.5
x^*	215.7	142.4
x_*	213.9	137.8

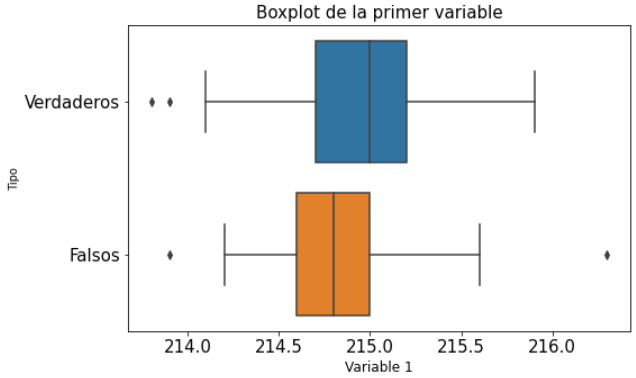

Figura 3:

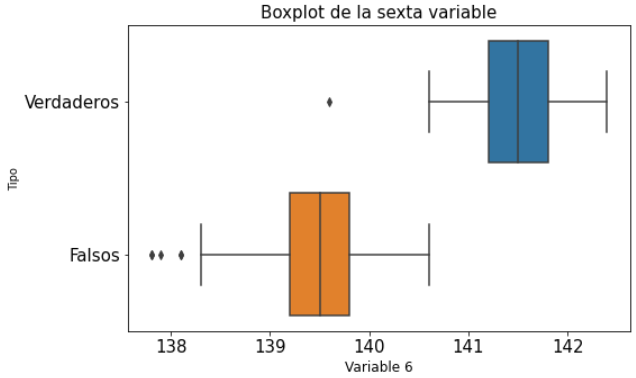

Figura 4:

Como podemos ver con el uso de la cuarta variable es posible separar la mayoría de los billetes verdaderos con respecto de los falsos con ayuda de esta variable ya que solo un billete verdadero cae en el rango de los billetes falsos, mientras que para la primer variable podemos ver que las cajas se solapan y no podemos sacar conclusiones con esta variable.

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.

   Un histograma es una gráfica de barras la cual en el eje x se encuentran los valores de una variable, mientras que el tamaño de las barras son el número de apariciones que aparecieron en un rango determinado, esto nos ayuda cuando trabajamos con variables aleatorias a aproximar la función de densidad de estas, siempre y cuando se escalen al número total de observaciones.

   Como podemos ver en el gráfico siguiente se simuló una normal con media igual a 30 y desviación estándar igual a 3 y como podemos ver el histograma se ajusta muy bien a la función teórica por lo que podemos usar este recurso para conocer la distribución de nuestros datos.

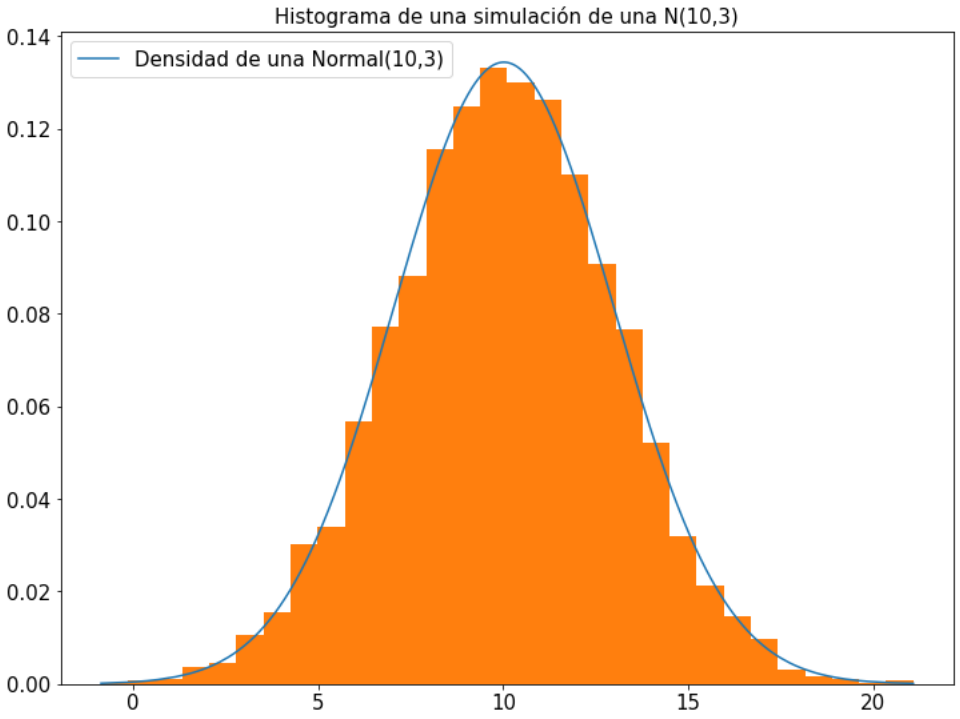

Figura 5:

4. Calcula de nuevo las componentes pero usando un reescalamiento de los datos $\widetilde{X}$. Por ejemplo si se asume que las variables X_1, X_2, X_3, X_6 fueron medidas en cms y que X_4, X_5 se quedan como estaban originalmente, osea en escala de mm, esto sería equivalente a re-escalar $\widetilde{X_1} = \frac{X_1}{10}, \widetilde{X_2} = \frac{X_2}{10}, \widetilde{X_3} = \frac{X_3}{10}, \widetilde{X_6} = \frac{X_6}{10}$. Compare sus resultados con la Figura C (obtenida al calcular las componentes principales usando los datos originales sin ningún re-escalamiento), ¿Qué se observa?

   Primero mostraremos la proyección de las componente principales de los datos son reescalar.

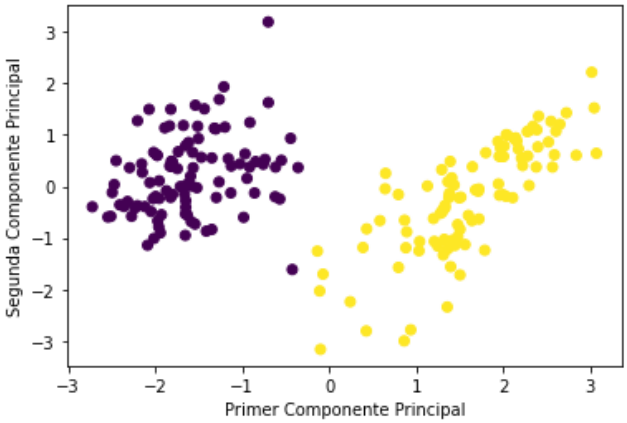

Figura 6:

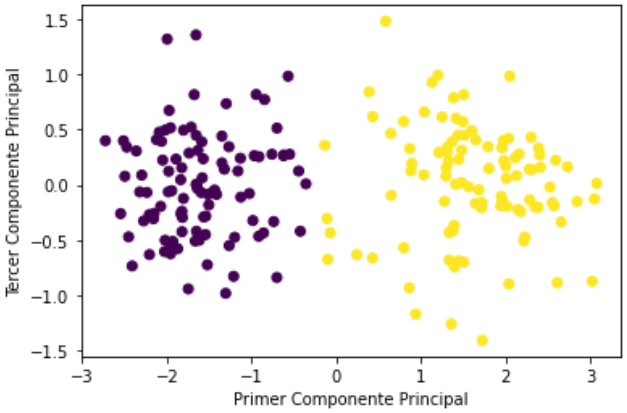

Figura 7:

   Posteriormente mostremos las nuevas componentes principales

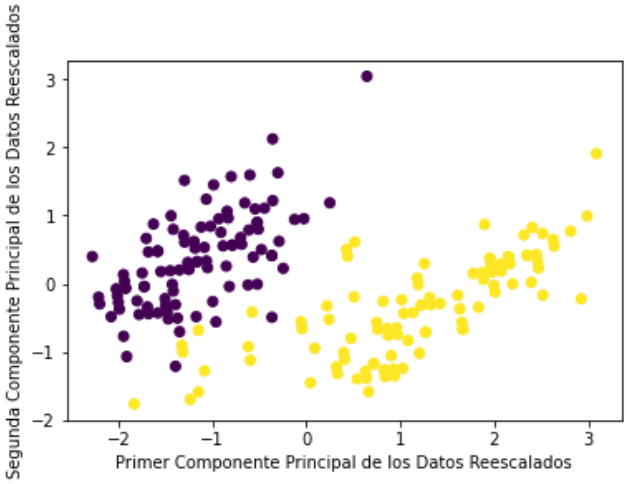

Figura 8:

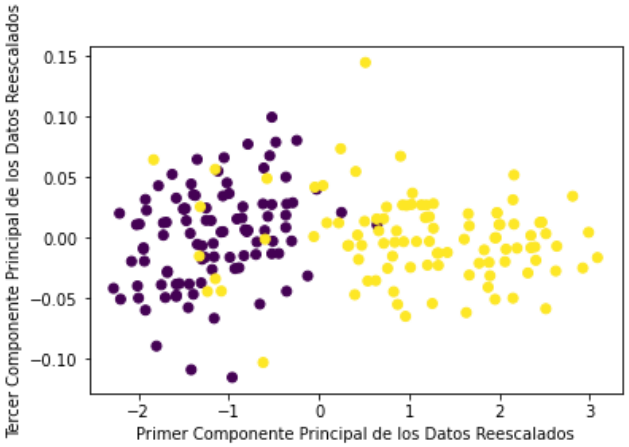

Figura 9:

Como era de esperarse las componentes principales cambian, ya que como algunos de los datos están divididos entre 10, esto quiere decir que su varianza en 100 veces menor, por lo que ahora de aplicar las componentes principales las direcciones de máxima varianza cambian, por lo que las proyecciones cambian como se puede ver en las gráficas.

5. Teorema: Sea $\mathbb{X}$ un vector aleatorio de dimensión px1 tal que $\mathbb{E}(\mathbb{X}) = \mathbb{M}$ y $Var(\mathbb{X}) = \Sigma$, sea $\mathbb{Y}$ el vector de componentes principales de $\mathbb{X}$, entonces:

   (i) $Cov(\mathbb{X}, \mathbb{Y}) = \Gamma\Lambda$, donde Γ es la matriz de dimensiones pxp cuyas columnas son los vectores propios de Σ (en la descomposición de Jordán de Σ y Λ es una matriz diagonal de dimensiones pxp que tiene los valores propios de Σ como elementos de la diagonal.

   (ii) La correlación de $\rho_{x_i y_j}$ entre la variable x_i y la componente principal y_j está dada por

$$\rho_{x_i y_j} = \gamma_{ij} \left(\frac{\lambda_j}{\sigma^2_{x_i x_i}} \right)^{\frac{1}{2}}$$

   donde $\sigma^2_{x_i x_i} = Var(X_i) = \Sigma_{ii}$, es decir, la entrada ii de la matriz Σ y γ_{ij} es la entrada ij de la matriz Γ

   DEM:
   (i) Consideremos $Cov(\mathbb{X}, \mathbb{Y})$, entonces

$$\begin{aligned} Cov(\mathbb{X}, \mathbb{Y}) &= Cov(\mathbb{X}, \Gamma^t(\mathbb{X} - \mathbb{M})) \\ &= Cov(\mathbb{X}, \Gamma^t(\mathbb{X} - \mathbb{M})) \\ &= Cov(\mathbb{X}, (\mathbb{X} - \mathbb{M}))\Gamma \\ &= Cov(\mathbb{X}, \mathbb{X})\Gamma \\ &= Var(X)\Gamma \\ &= \Sigma\Gamma \\ &= \Gamma\Lambda\Gamma^t\Gamma \end{aligned}$$

$$= \Gamma\Lambda$$

(ii) Primero sabemos que $Corr(X_i, Y_j) = \frac{Cov(X_i,Y_j)}{\sqrt{Var(X_i)}\sqrt{Var(Y_j)}}$, además sabemos que $Var(X_i) = \Sigma_{ii}$ y que $Var(Y_j) = \lambda_j$ y esto último viene dado por el teorema de la pagina 24, entonces

$$Corr(X_i, Y_j) = \frac{Cov(X_i, Y_j)}{\sqrt{Var(X_i)}\sqrt{Var(Y_j)}}$$

$$= \frac{Cov(X_i, Y_j)}{\sqrt{\Sigma_{ii}}\sqrt{\lambda_j}}$$

, por otro lado sabemos que $Cov(X, Y) = \Gamma\Lambda$, entonces la correlación de X_i con Y_j está dada por la multiplicación de la fila i de Γ con la columna j de Λ, sin embargo como Λ es diagonal con los valores propios en esta entonces la multiplicación de la fila i con la columna j queda de la siguiente forma $\gamma_{i1}0_{1,j} + \cdots + \gamma_{ij}\lambda_j + \cdots + \gamma_{in}0_{n,j} = \gamma_{ij}\lambda_j$, por lo que tenemos que

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

$$= \frac{\gamma_{ij}\lambda_j}{\sqrt{\sigma^2_{x_i x_i}}\sqrt{\lambda_j}}$$

$$= \gamma_{ij}\left(\frac{\lambda_j}{\sigma^2_{x_i x_i}}\right)^{\frac{1}{2}}$$

6. Verifíque que:

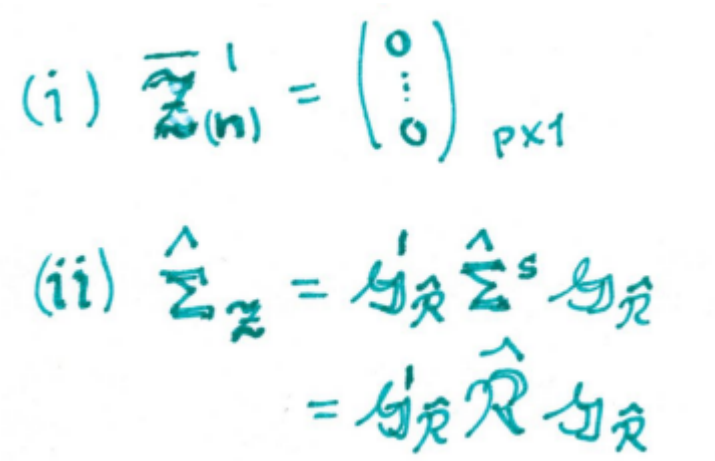

Figura 10:

(i) Consideremos $\widetilde{Z_{(n)}}$, entonces $\widetilde{Z_{(n)}} = \frac{1}{n}\mathbb{Z}^t\mathbb{1}_{(1)}$, pero $\mathbb{Z} = X_s M_{\hat{R}}$, entonces tenemos que

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

$$= \frac{1}{n}M^t_{\hat{R}}X^t_s\mathbb{1}_{(1)}$$

$$= M^t_{\hat{R}}(\frac{1}{n}X^t_s\mathbb{1}_{(1)})$$

$$= M_{\hat{R}}^t(0_{px1}) = 0_{px1}$$

ya que lo que está adentro de los paréntesis es el vector de medias pero como en este caso son 0 al multiplicarse dan 0's.

(ii) Consideremos $\hat{\Sigma_Z}$, entonces $\hat{\Sigma_Z} = \frac{1}{n}Z^tZ - \overline{X}_{(n)}\overline{X}_{(n)}^t$ pero como la medias son 0, entonces podemos prescindir de estos por lo que tenemos

$$\hat{\Sigma_Z} = \frac{1}{n}Z^tZ$$

$$= \frac{1}{n}(X_sM_{\hat{R}})^t(X_sM_{\hat{R}})$$

$$= \frac{1}{n}M_{\hat{R}}^tX_s^tX_sM_{\hat{R}}$$

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

$$= M_{\hat{R}}^t(\hat{\Sigma}^s)M_{\hat{R}}$$

$$= M_{\hat{R}}^t(\hat{R})M_{\hat{R}}$$

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 aperiencia)	Design
X_6	6	aspecto deportivo	Sporty
X_7	7	Grado de Seguridad del vehículo	Safety
X_8	8	Facilidad de Manejo	Easy

Figura 11:

Las 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 sí? o ¿Porqué no?

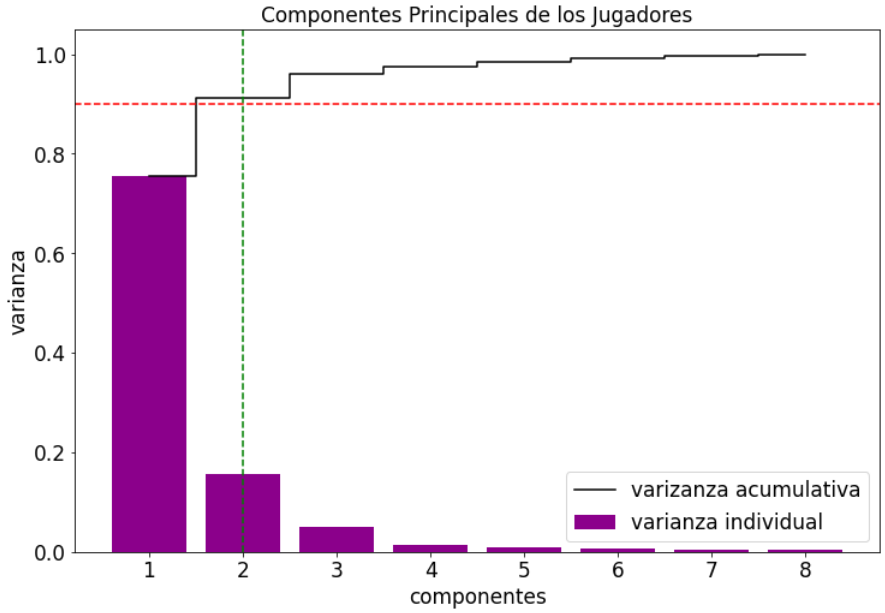

Figura 12:

Como podemos ver en la gráfica anterior podemos ver que el 90 % de la varianza se encuentra en las dos componentes principales por lo que tomar 2 componentes principales podría se un buen punto de partida, además de que si buscamos separar los datos y y con dos componentes es suficiente entonces basta con tomar estas os mientras que si queremos tener un poco mas de precisión podremos tomar 3 que en este caso representa un 95 % de la varianza.

Como podemos ver en la siguiente gráfica se dividieron los grupos de acuerdo con el promedio de las calificaciones obtenidas y como podemos ver que con el uso de dos componentes principales se ve que se pueden separar los conjuntos ya sea con el uso de elipse u otras curvas como funciones de decisión.

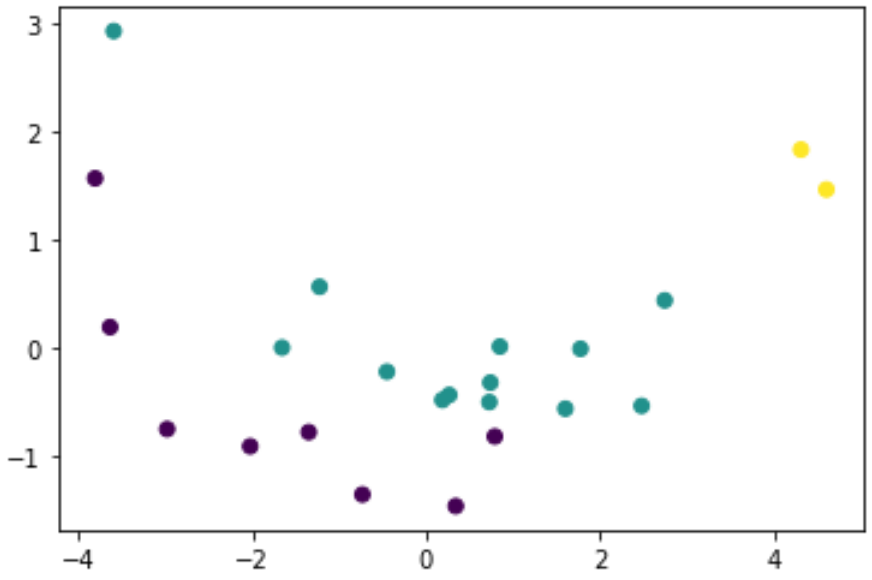

Figura 13: