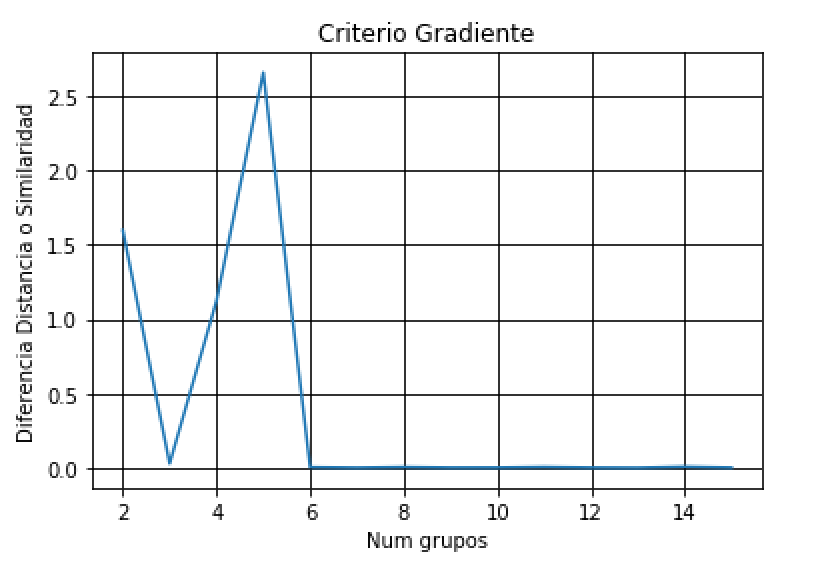
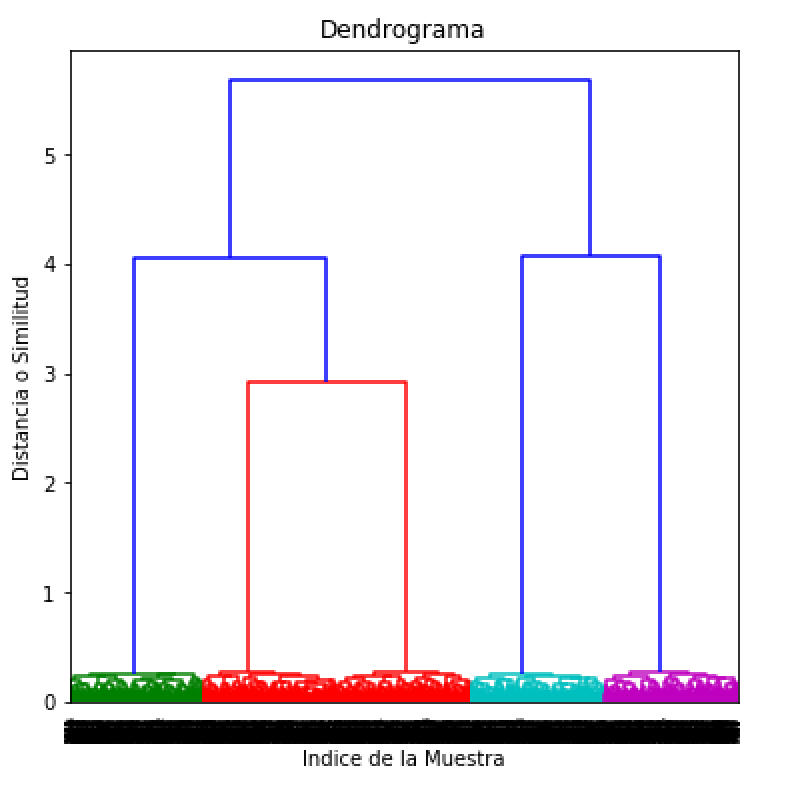


Nombre: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Israel Castillo Herrera If708348***

Lea detenidamente los reactivos y responda con claridad. Si se requiere hacer uso de más hojas para la realización de cálculos, es necesario que se adjunten a este cuando se haga entrega del examen.

1. **(3 puntos) En un experimento se logró identificar que 6 variables podían ser consideradas como importantes, las cuales determinaban el comportamiento de este. Después de hacer muchos experimentos se lograron recoger las muestras de diferentes condiciones de trabajo y se encuentran en el archivo “*ex2c\_1\_1.csv”*. Determine cuántos grupos o patrones se encuentran en los datos recopilados y justifique su respuesta con código, figuras o mediciones.**

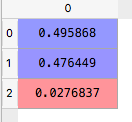
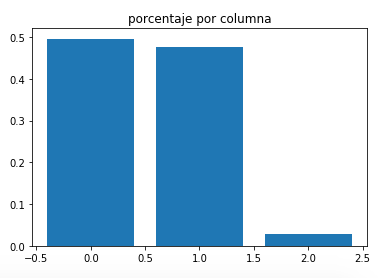


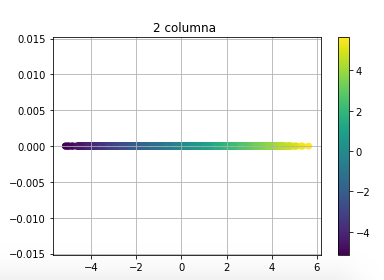
Mediante el uso de los algoritmos de clustering, y utilizando las mediciones creadas por los métodos de codo y del gradiente, podemos observar que al momento de generar los cluster con hierarchy, se generan el número óptimo de grupos (una vez que se haya normalizado). En los métodos de codo y de gradiente podemos ver que la primera inflexión se genera en 3 grupos, sin embargo, a mayor inspección bajo el dendograma, se muestra que el número óptimo de grupos es 5. Y en el codo y gradiente también se hace una inflexión en 5 grupos. Por lo que podemos concluir que el número óptimo de grupos para esa base de datos es 5.

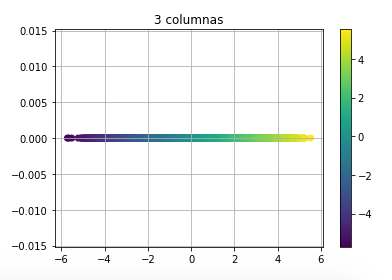


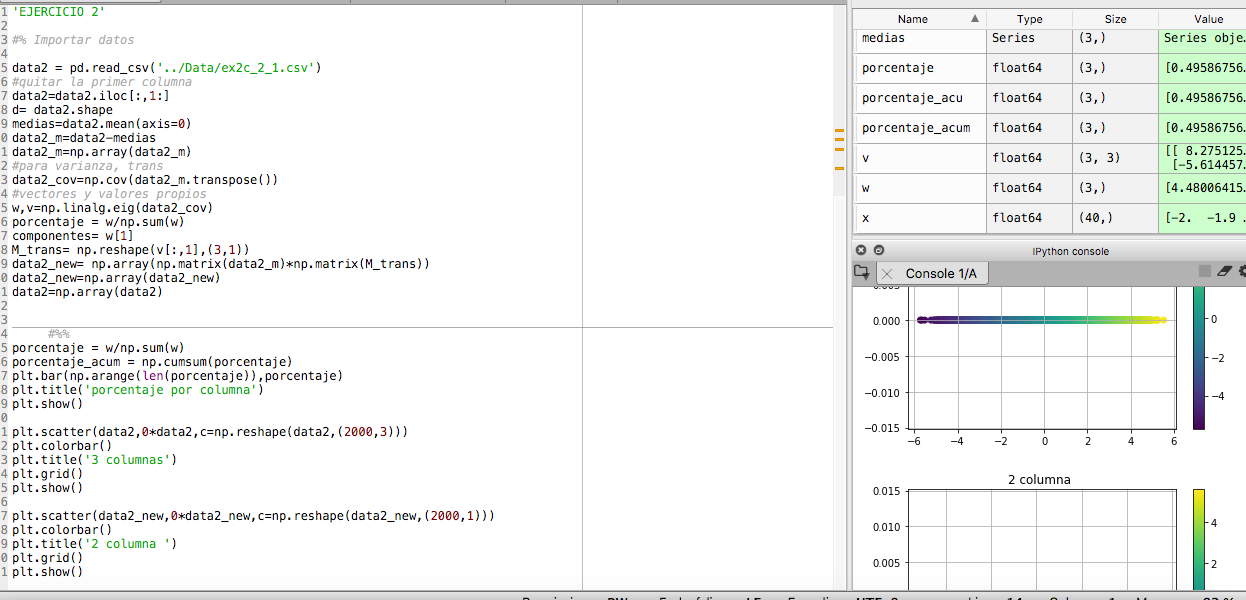
1. **(2 puntos) En la base de datos “ex2c\_2\_1.csv” se cuenta una base de datos en 3 dimensiones. Explique y justifique si ¿es posible hacer una reducción de las variables por medio del “PCA”?. Y muestre en un gráfico, como serían los datos después de la reducción.**

Al principio importamos el código, y vemos que tenemos una columna que no aporta ninguna información a nuestra base de datos. Una vez eliminada esta columna hacemos calculamos los valores y los vectores propios y nos damos cuenta que nuestras dos primeras columnas son las que nos aportan la mayor parte de la información, porcentualmente nos aporta un 96% de la información esto nos lleva a concluir que si es posible hacer una reducción mediante PCA de la última columna.



al momento de hacer la reducción de datos podemos observar que no tenemos una pérdida grande de información

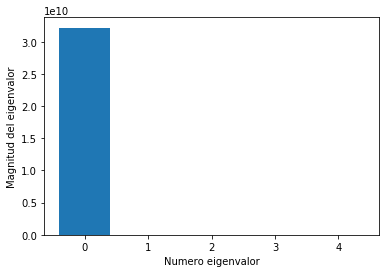




1. **Considere la siguiente base de datos pequeña (“*enfermedades.csv”*):**

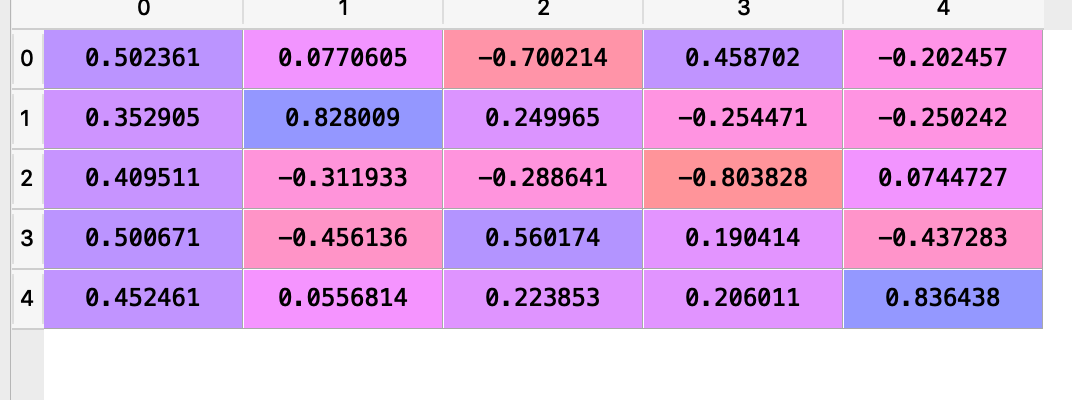
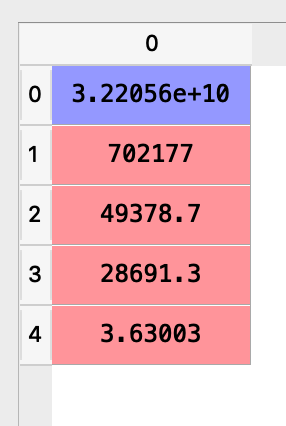
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nombre** | **2014** | **2013** | **2012** | **2011** | **2010** |
| **Casos de Dengue** | **1446** | **2584** | **560** | **175** | **1171** |
| **Casos de Influenza A H1N1** | **608** | **14** | **592** | **6** | **108** |
| **Casos de VIH/SIDA** | **506** | **630** | **978** | **665** | **578** |
| **Egresos hospitalarios** | **221364** | **155789** | **180462** | **220280** | **199288** |
| **Esperanza de vida al nacer** | **75.36** | **75.36** | **75.89** | **77.28** | **77.07** |
| **Muertes maternas** | **52** | **33** | **35** | **36** | **48** |

**Si aplicamos el análisis de componentes principales (PCA) a los datos de los años del 2010 al 2014 y obtenemos los siguientes eigenvalores, , los cuales pueden ser graficados como:**

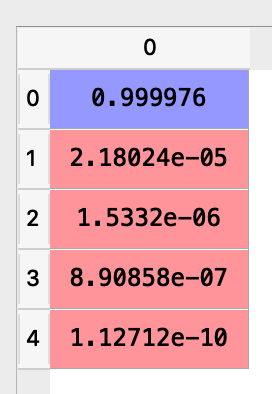
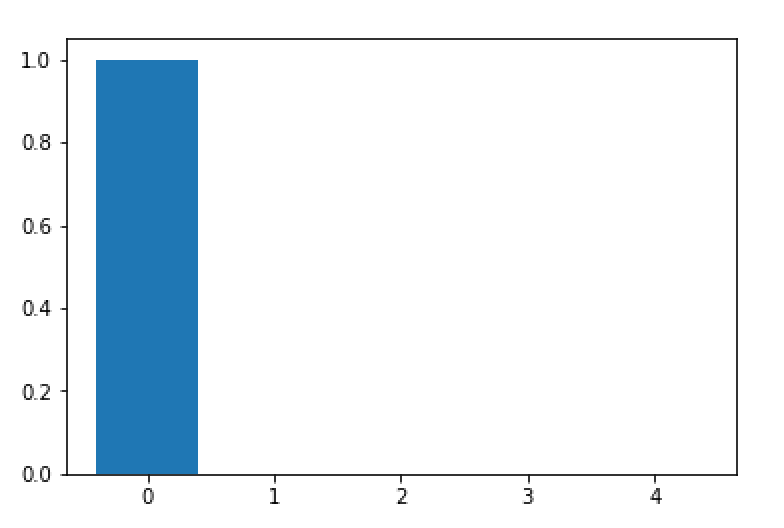


* 1. **(1 punto) ¿Es correcto pensar que con solo los datos del año 2014 de todas categorías es suficiente para poder distinguirlas o se conserva más del 90% de información?**

Obtuvimos los vectores y valores propios de los datos de las enfermedades.



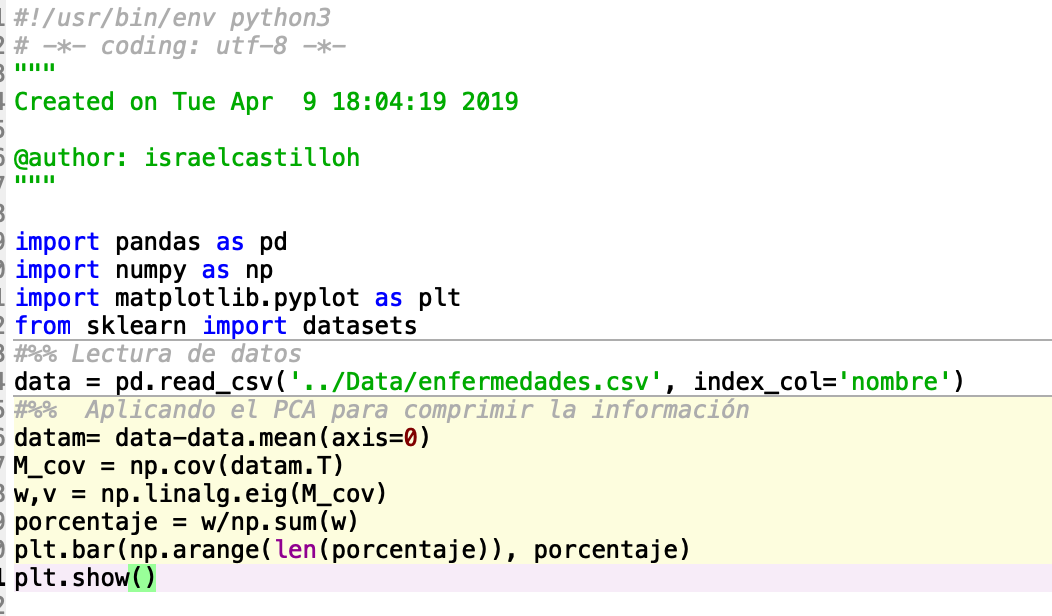
Con estos valores obtenemos la importancia de cada uno. Por lo que vemos que:

****

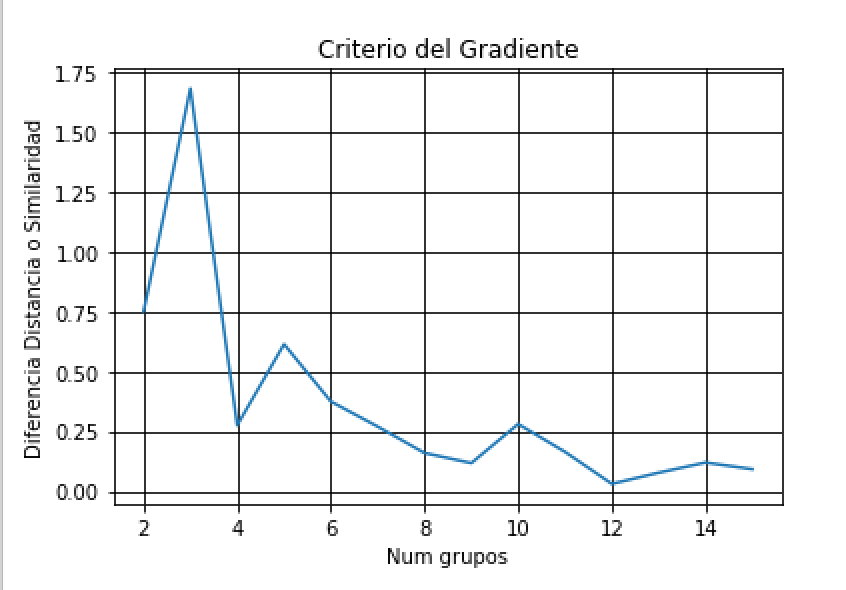
El valor con mayor peso es el 1, que tiene el 99.99% de la importancia de los datos, por lo que podemos descartar los demás, y decir que si se conserva el 90% de la información con sólo el año 2014.

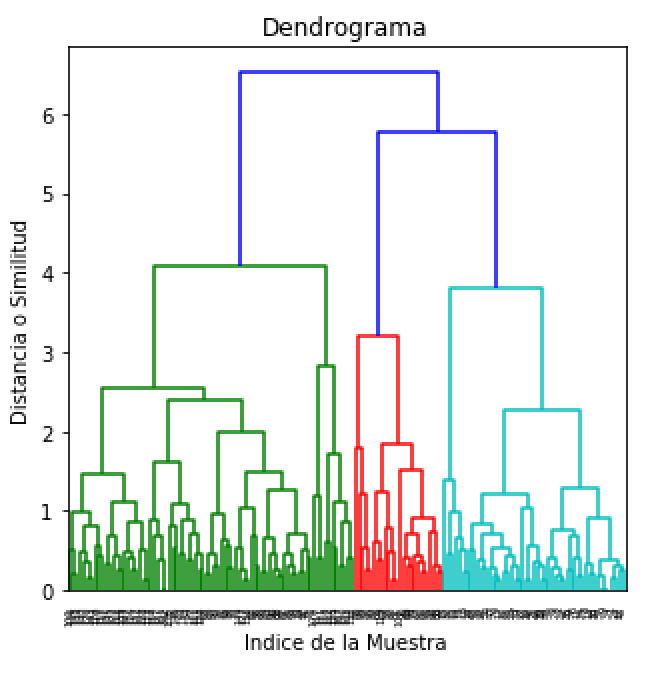
* 1. **(1 punto) Normalmente el PCA se aplica a las columnas de una base de datos. ¿Qué interpretación se le puede dar a los resultados de aplicar el PCA a las filas de la base de datos?**

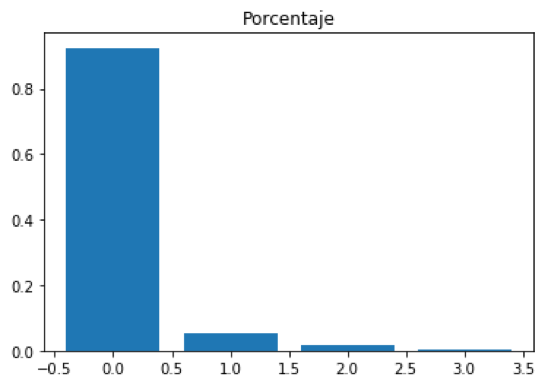
En este caso se puede interpretar que estamos tomando las enfermedades de un año con respecto al otro, es decir, comparando años tomando como conjunto las covarianzas y relaciones existentes entre enfermedades.



1. **En el archivo llamado “ex2c\_4.csv” se encuentra un set de datos con información sobre características de flores que fueron capturadas por un botánico. Basado en estos realicé lo siguiente:**
   1. **(1.5 puntos) Por medio de un algoritmo de clustering determine cuántos grupos se patrones se pueden determinar en este set de datos.**
   2. **(1.5 puntos) Aplique el análisis de PCA en la base de datos original y con la base de datos reducida (como mínimo 90% de la información mantenida), vuelva a determinar el número de patrones por medio de un algoritmo de clustering. ¿Se mantuvo el número de grupos encontrados? Valide su respuesta con gráficas y mediciones.**

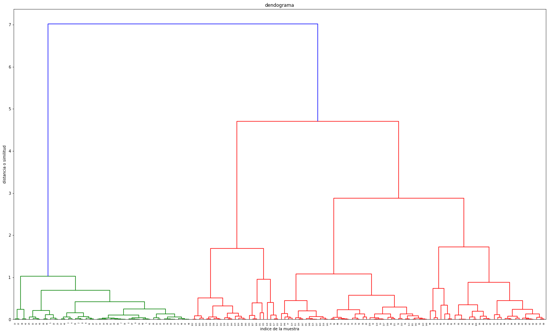




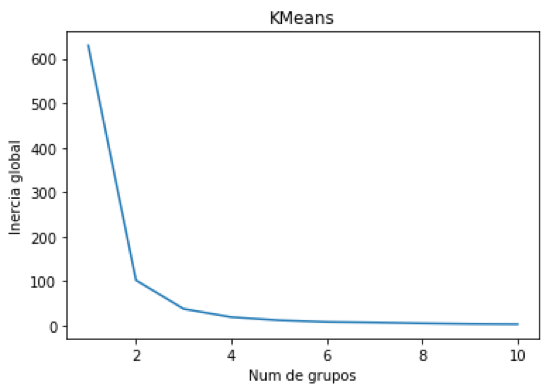
La interpretación directa de las gráficas es que existen 3 grupos base.

b.

Para confirmar esto, usaremos el algoritmo de clustering y confirmaremos que sí son dos grupos.



De acuerdo con dendograma, no podemos concluir nada ya que es muy grande, por lo cual, utilizaremos KMeans.



De acuerdo a KMeans, confirmamos que el número óptimo de grupos es dos.

