Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
CC3094 - Security Data Science

Laboratorio 04

Evelyn Andrea Amaya Malin - 19357
Brandon Josue Hernández Marroquín - 19376

Guatemala, Ciudad de Guatemala 25 de marzo de 2023

Análisis Exploratorio

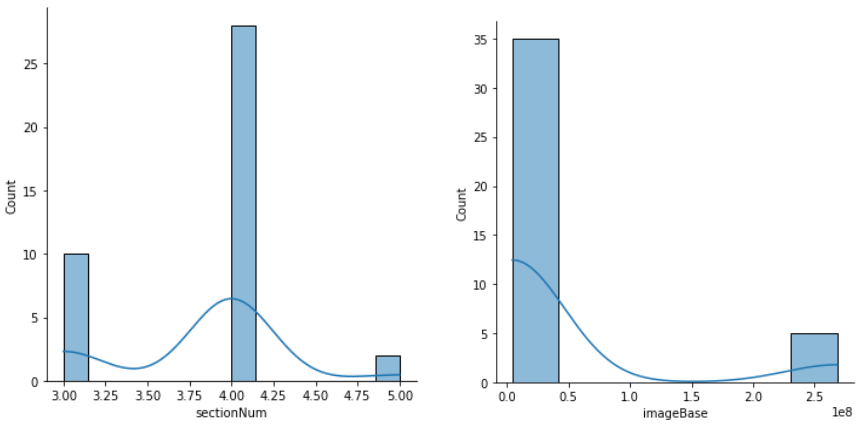
Para este laboratorio fue necesario crear un conjunto de datos propios con base al malware proporcionado. Luego de observar los nombres de las secciones se descubrió que estos estaban empaquetados por lo que se descomprimieron antes de identificar qué información se pasaría al conjunto de datos. Las variables que se incluyeron dentro de la información fueron las siguientes: índice, nombre del archivo, nombres de las secciones, direcciones virtuales, tamaños virtuales, tamaño de la información, número de secciones, base del código, base de la imagen, alineación de la sección, tamaño inicial de la información, tamaño de código, características DLL, dirección del entry point, fecha de creación, DLLs y llamadas a función.

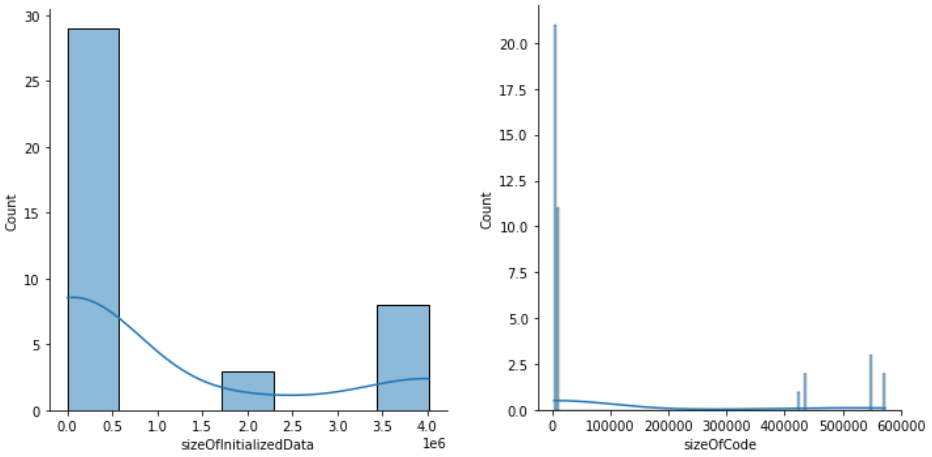
Luego se separaron las variables cuantitativas y las cualitativas, primero se describieron las cuantitativas las cuales cuentan con la siguiente descripción:

	sectionNum	baseOfCode	imageBase	sectionAlignment	sizeOfInitializedData	sizeOfCode	dllCharacteristics	addressOfEntryPoint
count	40.000000	40.0	4.000000e+01	40.0	4.000000e+01	40.000000	40.0	4.000000e+01
mean	3.800000	4096.0	3.722445e+07	4096.0	9.958016e+05	106969.600000	0.0	3.422618e+09
std	0.516398	0.0	8.850283e+07	0.0	1.613229e+06	206344.889444	0.0	0.000000e+00
min	3.000000	4096.0	4.194304e+06	4096.0	2.560000e+03	3584.000000	0.0	3.422618e+09
25%	3.750000	4096.0	4.194304e+06	4096.0	3.072000e+03	4096.000000	0.0	3.422618e+09
50%	4.000000	4096.0	4.194304e+06	4096.0	3.072000e+03	4096.000000	0.0	3.422618e+09
75%	4.000000	4096.0	4.194304e+06	4096.0	2.011136e+06	10752.000000	0.0	3.422618e+09
max	5.000000	4096.0	2.684355e+08	4096.0	4.012032e+06	572928.000000	0.0	3.422618e+09

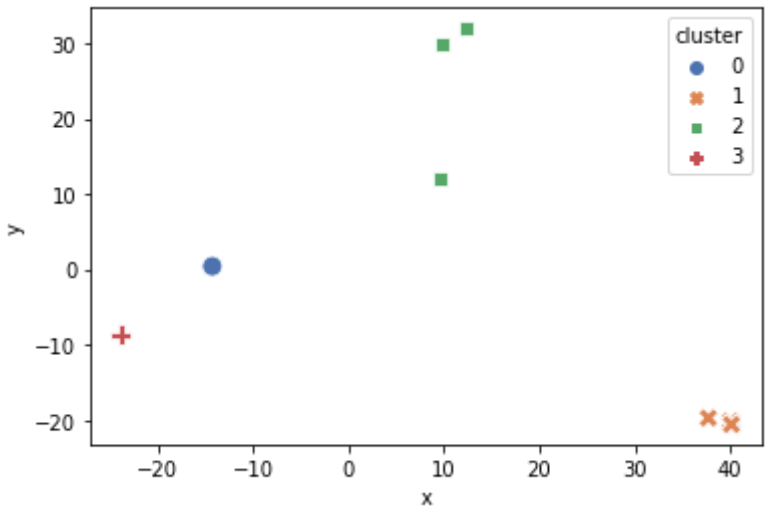
Con base a esto se puede determinar que las variables que cuentan con una desviación estándar de 0 no son significativas para la clasificación, por lo que se procede a eliminarlas.

De estas se obtuvieron los siguientes gráficos para observar más su comportamiento:





Luego de observar todas las gráficas se tiene una mayor idea de cómo es que están distribuidos los datos cuantitativos. Luego de esto se utilizó TfidfVectorizer para obtener la relevancia de funciones y DLL de los malwares. Luego a las variables numéricas y la vectorización de las funciones y DLL se les aplicó la reducción de dimensiones. Con base a estas variables reducidas se realizó el entrenamiento y clasificación con el Kmeans, del cual se obtuvieron cinco familias. Al observar la siguiente imagen:

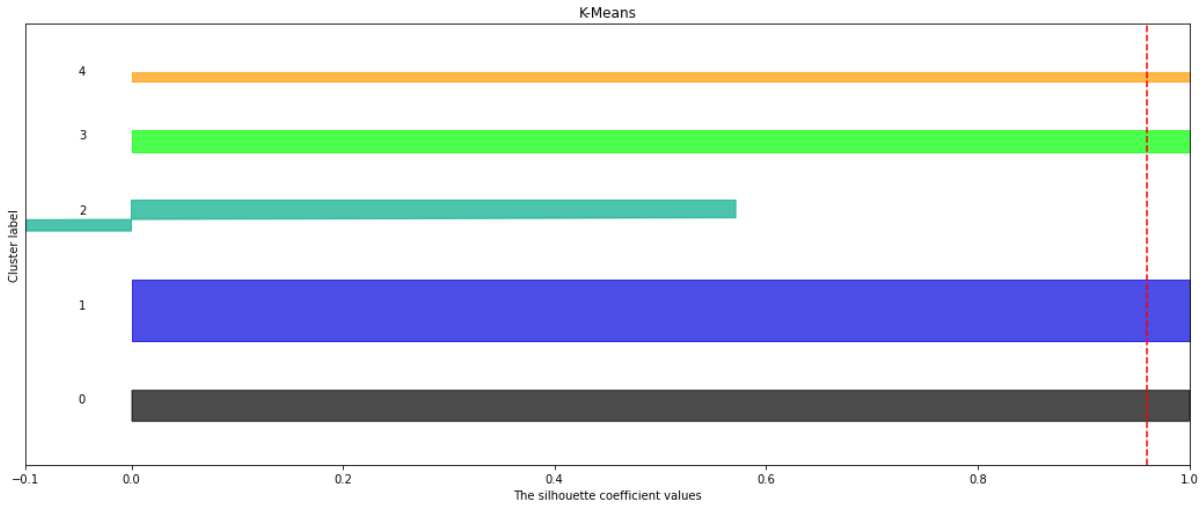


Se explica que se observan pocos puntos en la imagen, por la proximidad de los malwares entre sí.

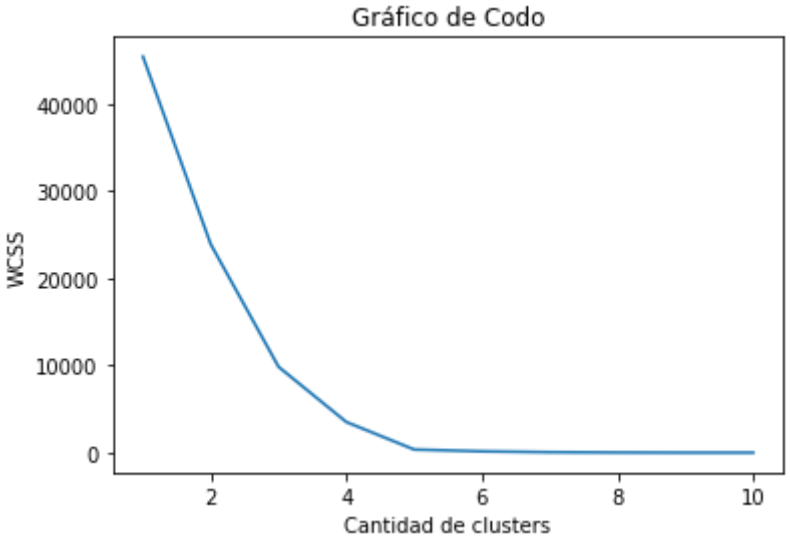
Discusión

¿Para qué número de clústeres se obtiene el coeficiente de Silhouette más alto?

El mejor resultado se obtuvo haciendo uso de 5 clusters dando un resultado de 0.9602. Al observar la gráfica se puede observar que el único cluster con datos mal clasificados es el número 2, debido a que tira resultados por debajo de 0, que es el cluster que baja el resultado del coeficiente de silhouette.



¿Coincide el coeficiente de Silhouette con el método del codo?
Sí, al hacer uso de 5 clusters (donde se nota el primer punto con pendiente distinta), se obtiene un k-means de 0.9602. Si se hace de 4 clusters (segundo punto con pendiente distinta) se obtiene un k-means de 0.8795. Por lo tanto, el método del codo coincide con el coeficiente de Silhouette.




```
For n_clusters = 3 The average silhouette_score of K-Means is: 0.7154488505104158
```

```
For n_clusters = 4 The average silhouette_score of K-Means is: 0.8795005811169923
```

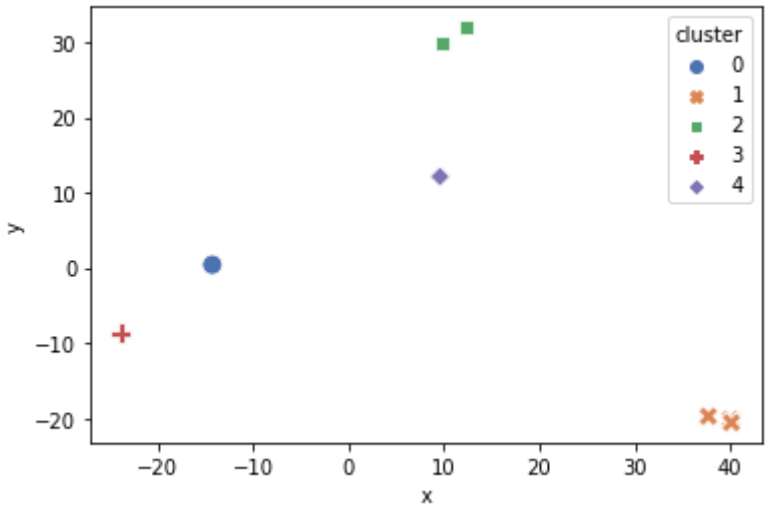
K-Means

```
[41]  ✓ 0.3s
For n_clusters = 5 The average silhouette_score of K-Means is: 0.960249680609244
```

K-Means

¿Cuántas familias existen entre los ejemplares de malware proporcionados?

Se identificaron 5 familias de malware entre los ejemplares.



cluster	algunas funciones
0	closehandle waitforsingleobject createeventa exitthread sleep getcomputernamea createpipe disconnectnamedpipe terminateprocess ...
1	globalmemorystatus getvolumeinformationa module first module next thread first getlocaltime getcomputernamea flushconsoleinputbuffer ...
2	createfilea localalloc sleep createthread createmutexa copyfilew getfilesize createprocessa getenvironmentvariablew getshortpathnamew ...
3	closehandle waitforsingleobject createeventa exitthread sleep getcomputernamea createpipe disconnectnamedpipe terminateprocess waitformultipleobjects terminatethread createthread ...
4	terminateprocess allocconsole multibytetowidechar getsystemtime writefile readfile peeknamedpipe sleep createprocessa getsystemdirectorya createpipe freeconsole getlasterror ...

¿Coincide el índice de Jaccard con las familias encontradas?

Se hizo uso de un índice de 0.75 y en el grafo resultante se puede observar que se tienen alrededor de 7 familias, de las cuales 2 familias son un solo malware individual. Esto coincide con las familias observadas si se toma en cuenta que el cluster 2 tiene mal clasificado algunos resultados.