Plan social utilizando los indicadores de carencias económicas creada por el CONEVAL en el 2018.

Se requiere diseñar un plan social utilizando los indicadores presentados en la tabla. Dicho plan va a consistir en proveer de becas a estudiantes que más lo necesiten para su implementación en el 2022.

1. Construir un análisis descriptivo, con gráficas y estadísticas relevantes del set de datos, escribir sus conclusiones en el reporte.

1.- Análisis exploratorio de datos.

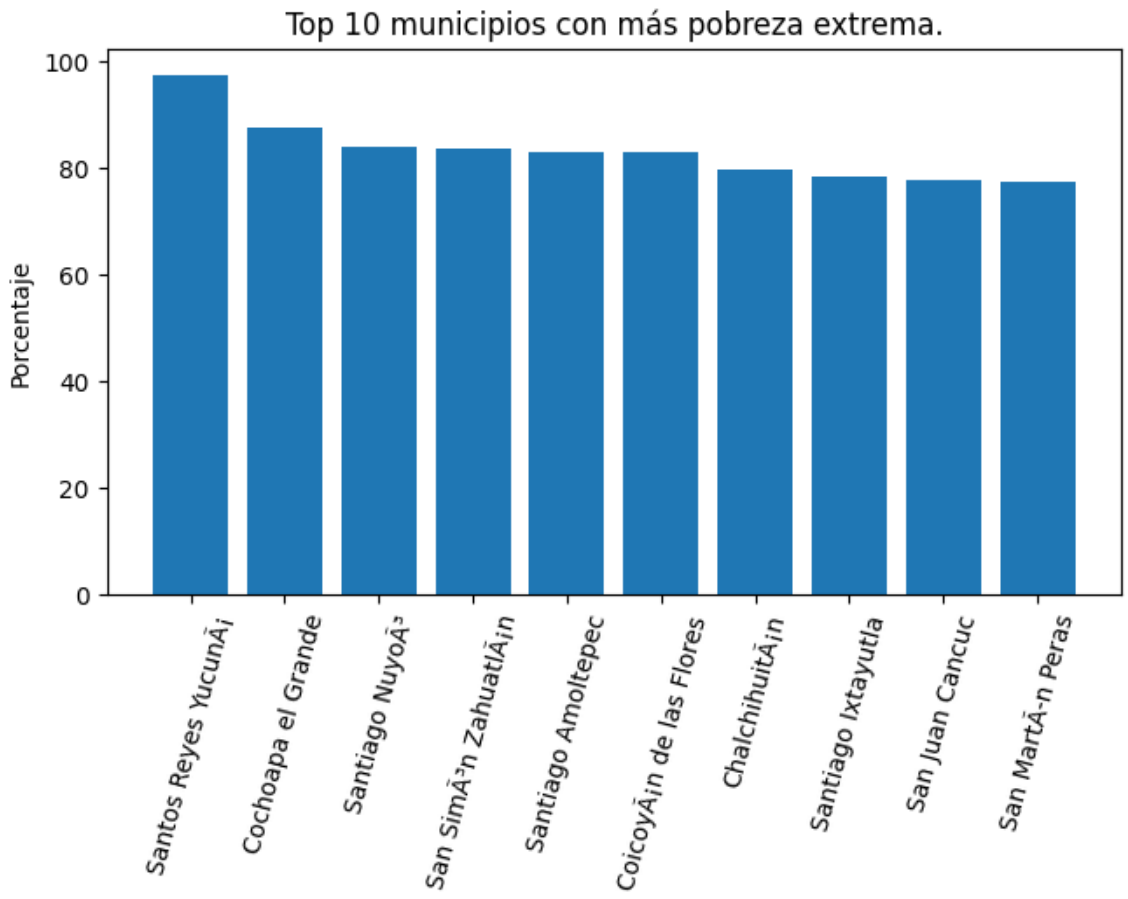
Variable	Tipo		Variable	Tipo
entidad_federativa	object		ic_asalud	float64
mun_name	object		ic_asalud_pob	int64
poblacion	int64		ic_segsoc	float64
pobreza	float64		ic_segsoc_pob	int64
pobreza_pob	int64		ic_cv	float64
pobreza_e	float64		ic_cv_pob	int64
pobreza_e_pob	int64		ic_sbv	float64
pobreza_m	float64		ic_sbv_pob	int64
pobreza_m_pob	int64		ic_ali	float64
vul_car	float64		ic_ali_pob	int64
vul_car_pob	int64		carencias	float64
vul_ing	float64		carencias_pob	int64
vul_ing_pob	int64		carencias3	float64
npnv	float64		carencias3_pob	int64
npnv_pob	int64		plb	float64
ic_rezedu	float64		plb_pob	int64
ic_rezedu_pob	int64		plbm	float64
			plbm_pob	int64

Previo al chequeo de variables, retiré las variables "clave_municipio","clave_entidad","long","lat" y "direccion".

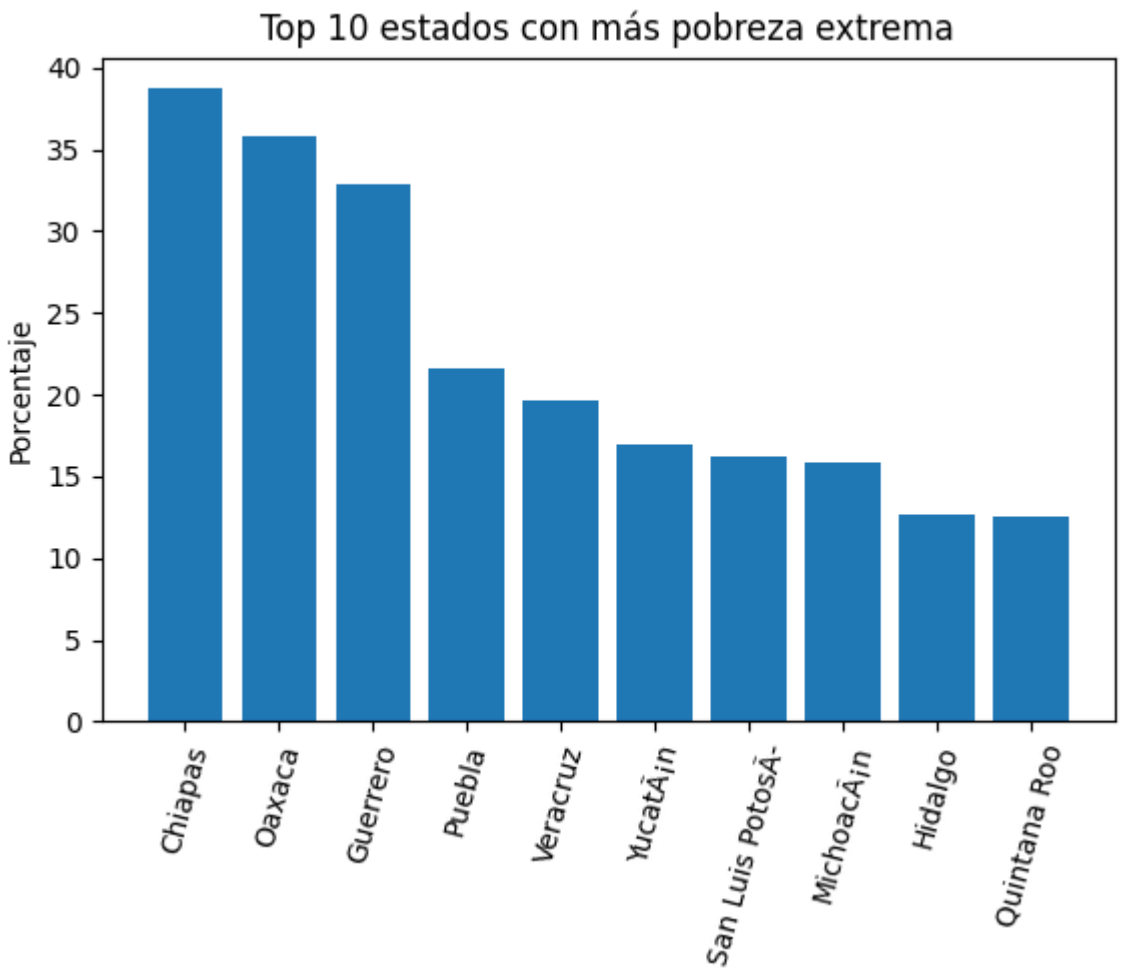
Se mantiene la entidad federativa y nombre del municipio para poder realizar un análisis descriptivo de los datos, aparte de éstas 2 variables que son objeto, las demás son variables numéricas.

Al no haber variables categóricas, por el momento no parece haber necesidad de convertir ninguna variable.

Se verificó la posible existencia de datos nulos, los cuales no hubo. También con la ayuda de la función describe() se verificó que no hubiera datos con valores negativos, porcentajes mayores a 100, ni datos que no concordaran.



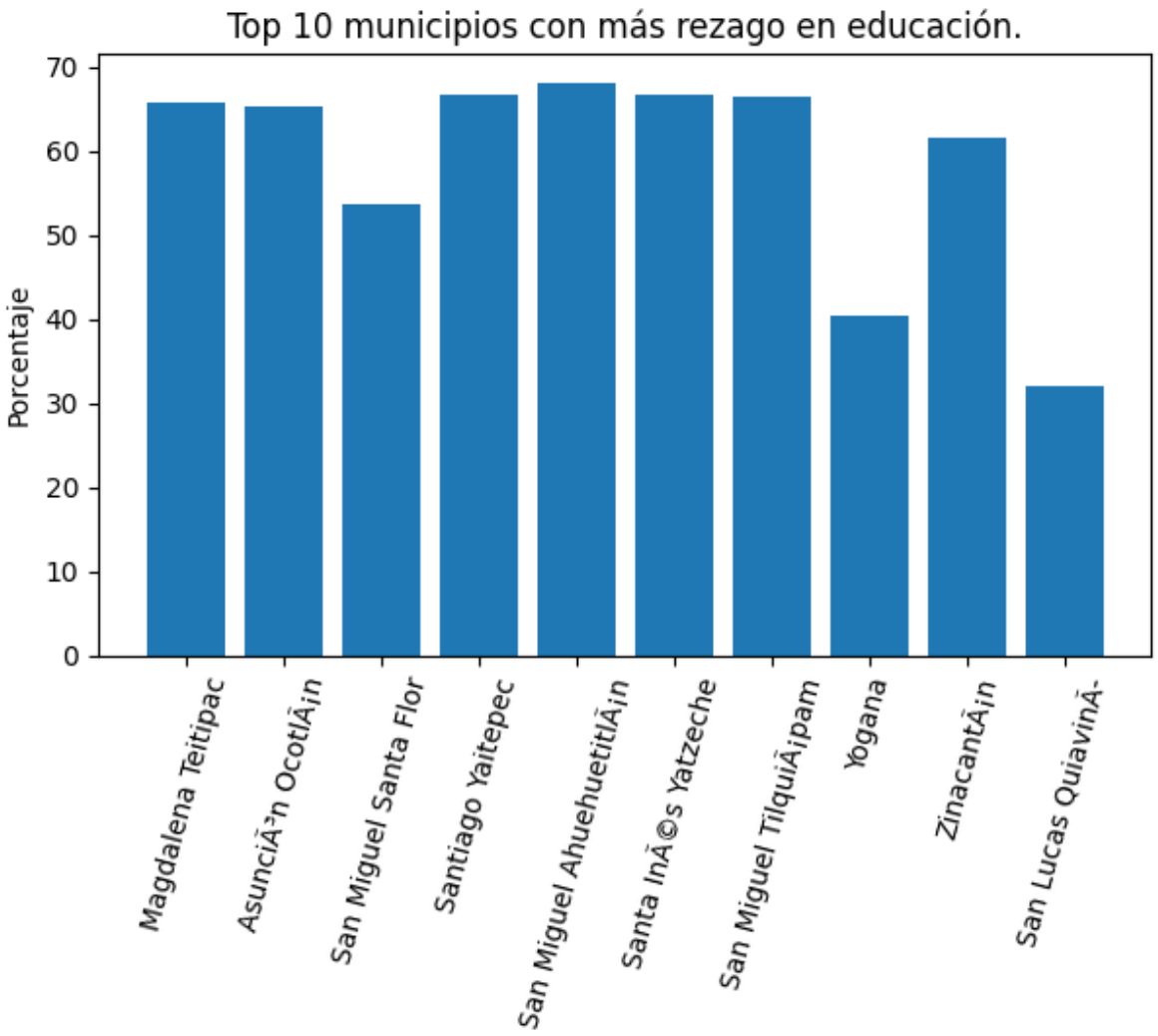
Encabezado por el municipio de Santos Reyes Yucuná perteneciente al edo. De Oaxaca, se muestran los 10 municipios con más porcentaje de pobreza extrema en México. Cabe mencionar que de los 10, solamente 1 pertenece al Edo. De Guerrero, 2 a Chiapas y el resto a Oaxaca.



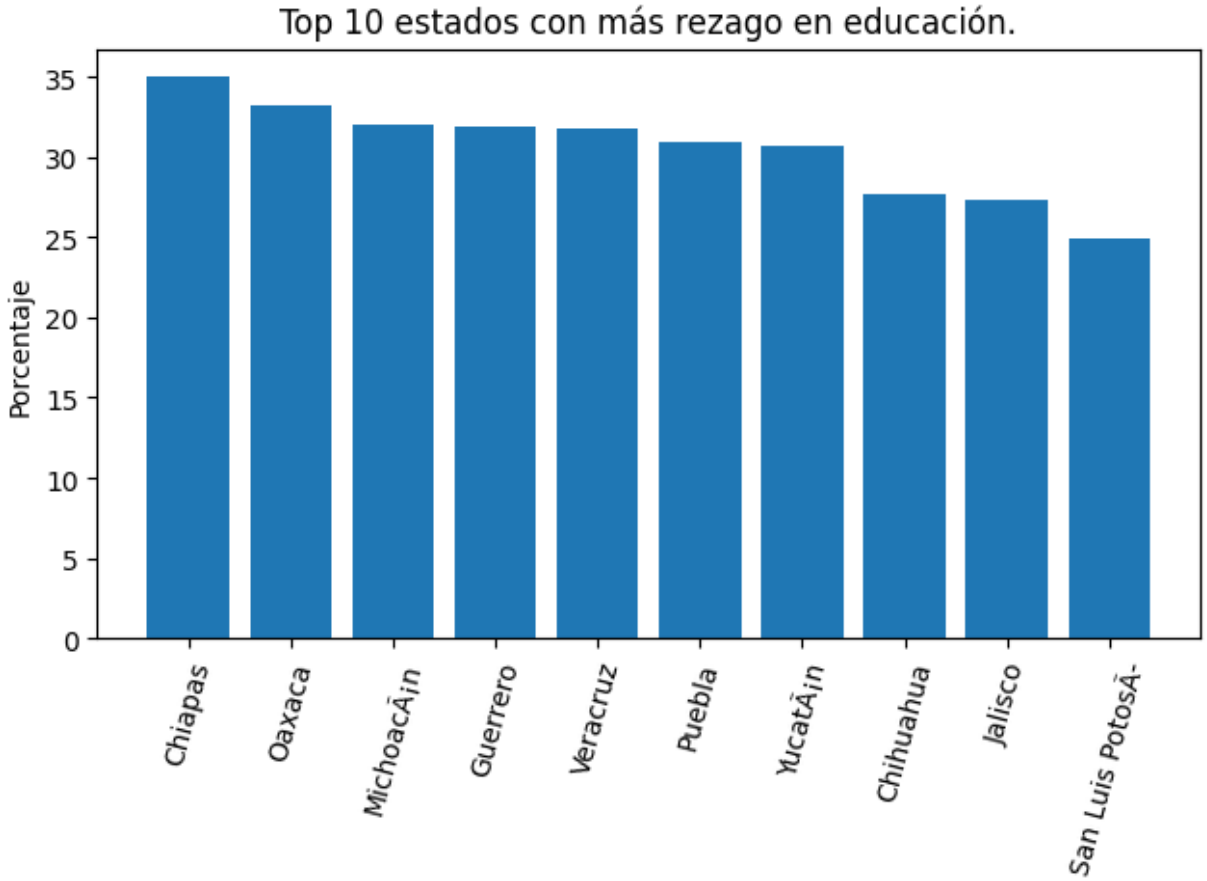
entidad_federativa	pobreza_e
Chiapas	38.705085
Oaxaca	35.822261
Guerrero	32.901235
Puebla	21.556481
Veracruz	19.669811
YucatÃ¡n	16.984906
San Luis PotosÃ	16.248276
MichoacÃ¡n	15.882301
Hidalgo	12.704762
Quintana Roo	12.560000

Chiapas, Oaxaca y guerrero son los que mayor pobreza extrema tienen, a partir del 4to puesto baja un 50% el promedio de pobreza a comparación del top 3.

La diferencia del 10mo puesto vs. el 1ro es de un 26% lo cual muestra una grande acumulación de pobreza en esos estados.



En esta gráfica solamente el puesto número nueve pertenece al edo. De Chiapas, todos los municipios de la lista son pertenecientes a Oaxaca una vez más. Cabe mencionar que en su mayoría el porcentaje de rezago en la educación oscila entre el 60% y 70%.

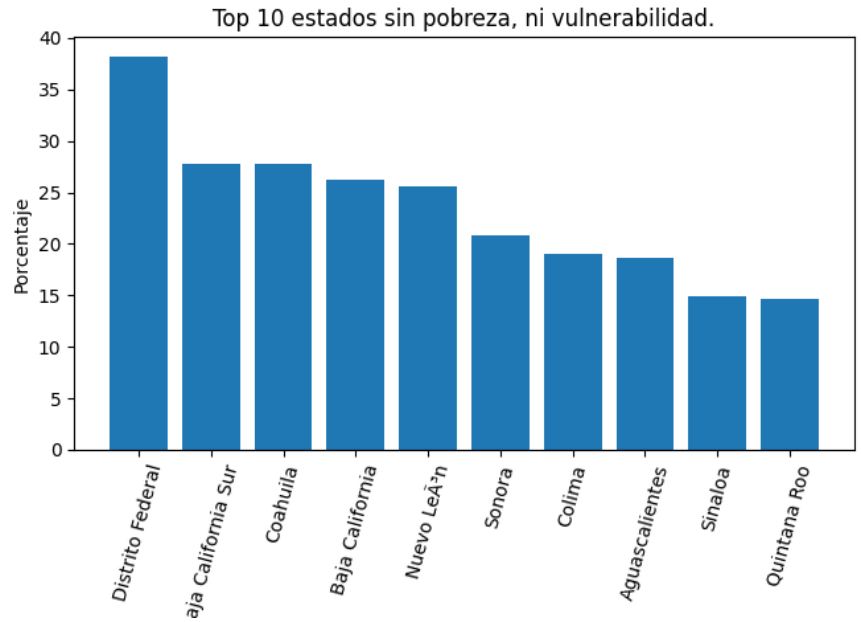


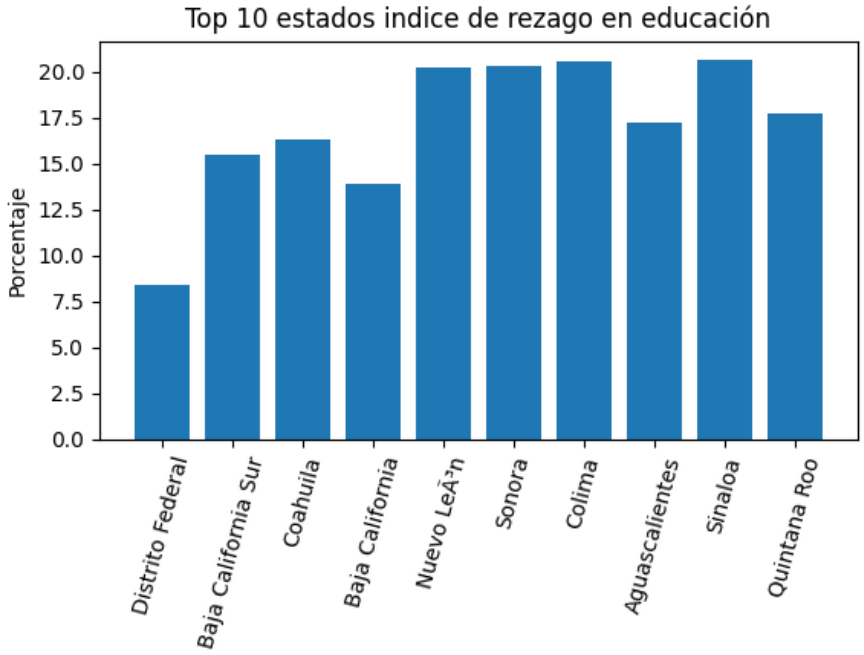
De nuevo se observan los 3 estados con más pobreza extrema: Oaxaca, Chiapas y Guerrero.

El porcentaje del Top 10 fluctúa entre el 25% y el 35%, hasta el momento se ha observado gráficamente relación entre pobreza y rezago en educación.

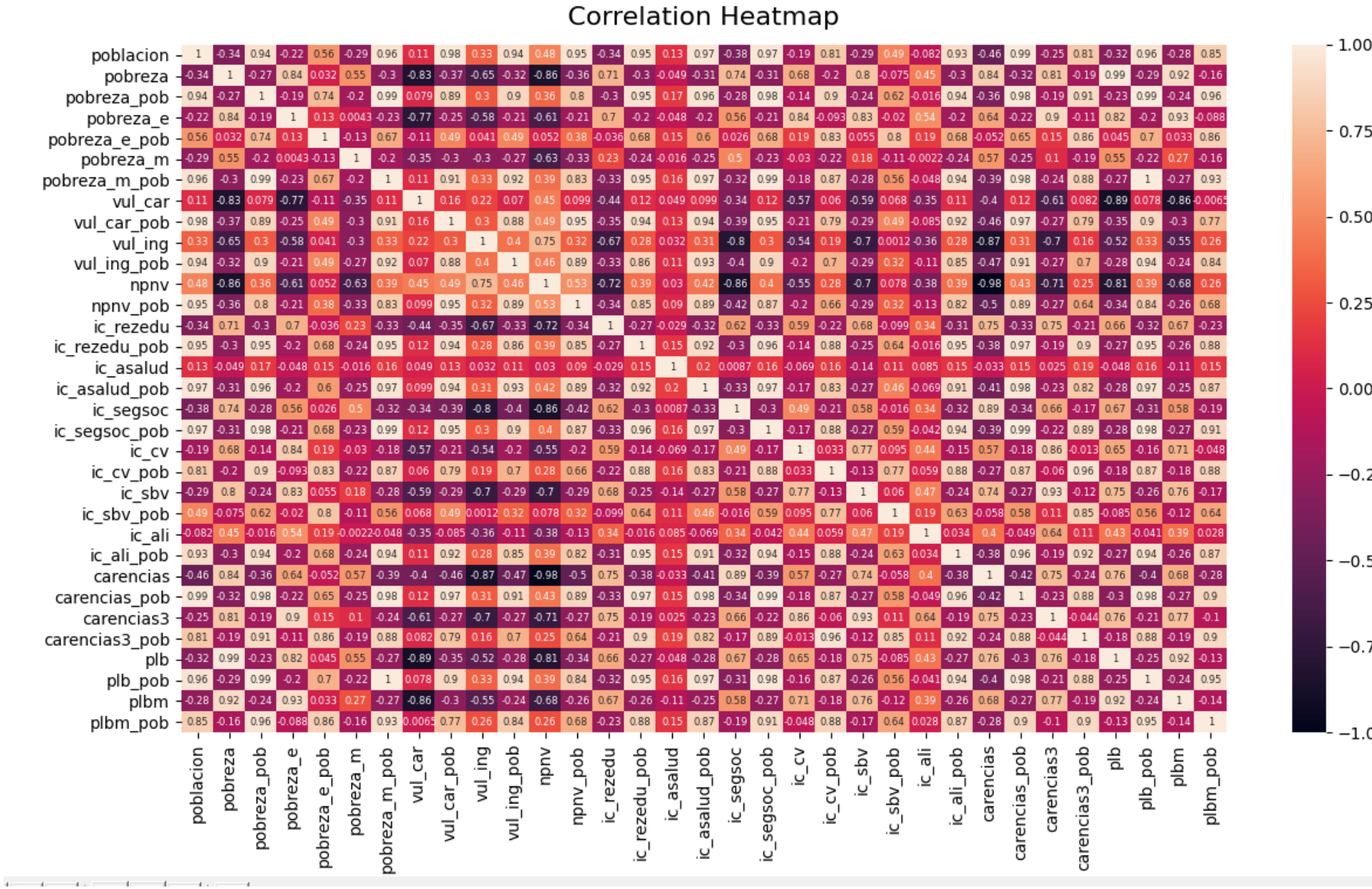
```
Población en situación de pobreza extrema:   7.96029119554835
Población en situación de pobreza moderada:  36.017832135390435
Población en situación de pobreza:           43.97812827455067

Porcentaje total de pobreza en México:       87.95625160548946
```

Con un total de pobreza del 87.95%, es más que notable la falta de oportunidades que se requiere para poder progresar como país.





De nuevo se observa gráficamente una correlación negativa entre bienestar y educación/pobreza.

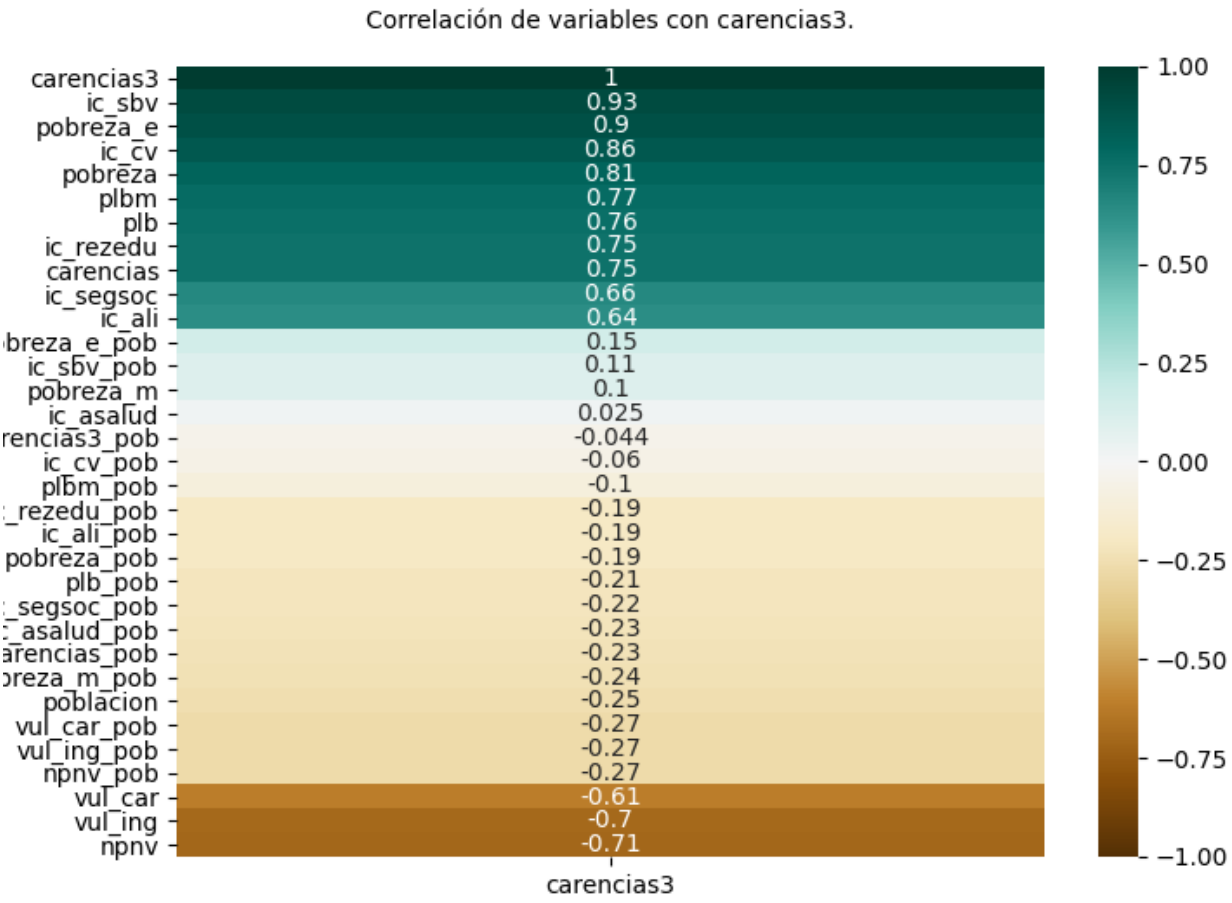

Correlation Heatmap
poblacion
pobreza
pobreza_pob
pobreza_e
pobreza_e_pob
pobreza_m
pobreza_m_pob
vul_car
vul_car_pob
vul_ing
vul_ing_pob
npnv
npnv_pob
ic_rezedu
ic_rezedu_pob
ic_asalud
ic_asalud_pob
ic_segsoc
ic_segsoc_pob
ic_cv
ic_cv_pob
ic_sbv
ic_sbv_pob
ic_ali
ic_ali_pob
carencias
carencias_pob
carencias3
carencias3_pob
plb
plb_pob
plbm
plbm_pob
1.00
0.75
0.50
0.25
0.00
−0.2
−0.5
−0.7
−1.0

En la gráfica de Heatmap se observa la correlación de todas las variables en nuestros datos.

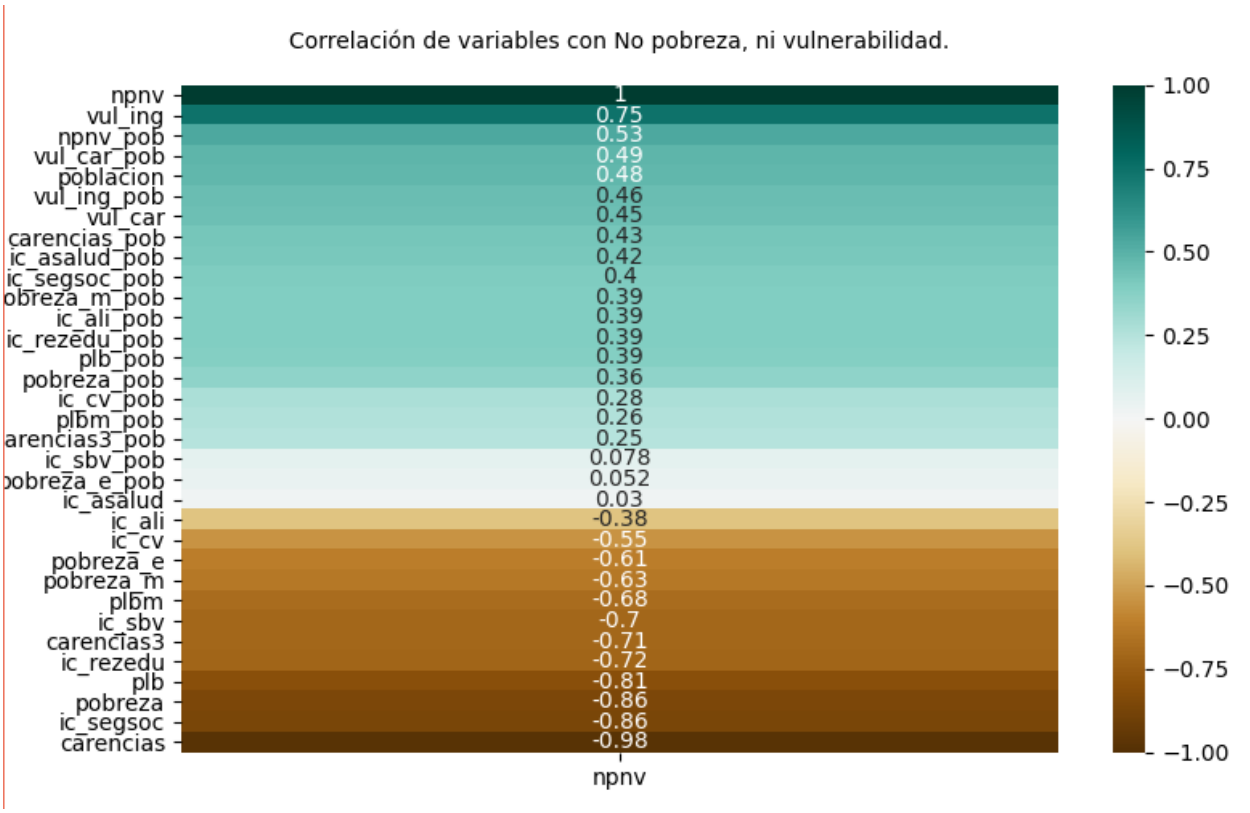
Aunado a nuestro análisis gráfico, ahora podemos comprobar analíticamente nuestras suposiciones.

La variable que indica el porcentaje de rezago en educación tiene una correlación de un .7 y .71 con la variable de pobreza y pobreza extrema respectivamente. Esto indica una correlación directa entre ellas.

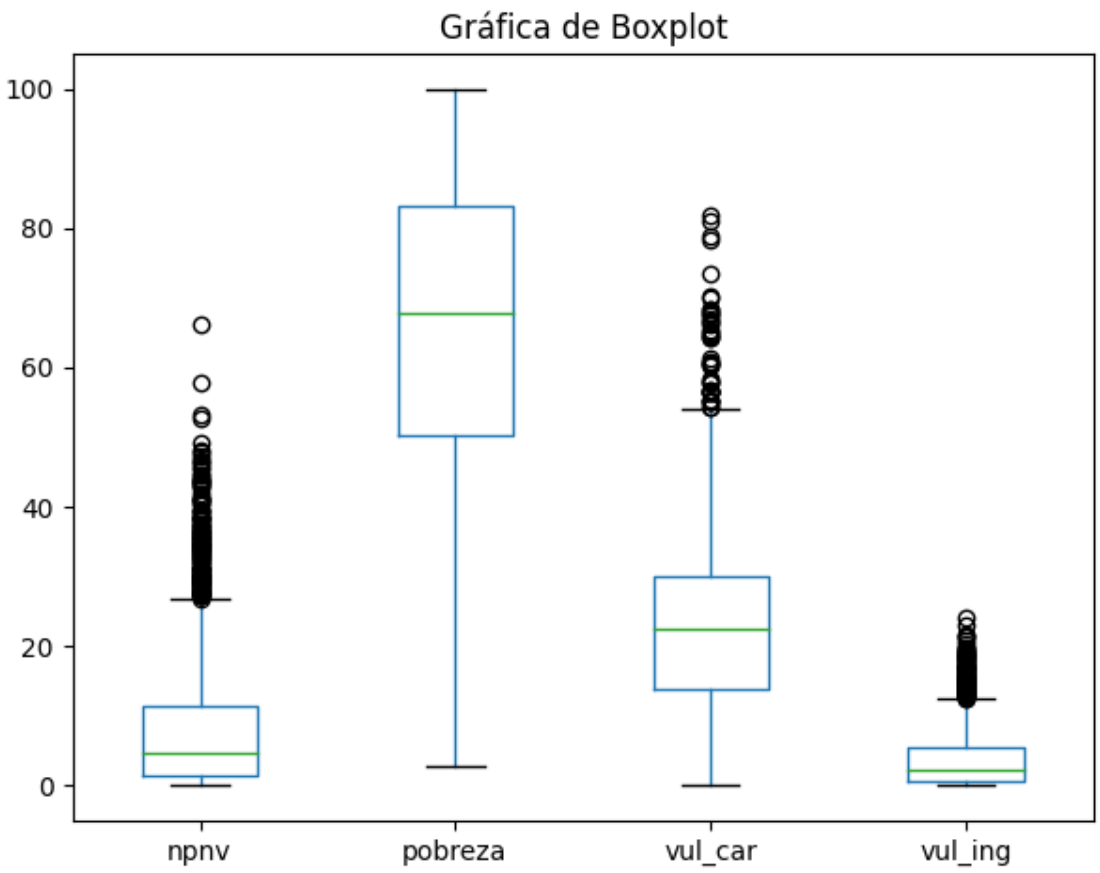
A continuación, se muestran más a detalle algunas de las variables antes mencionadas:



La variable que indica que al menos tiene 3 carencias sociales tiene una correlación de .75 con la variable de rezago en educación. Lo cual indica una correlación directa.

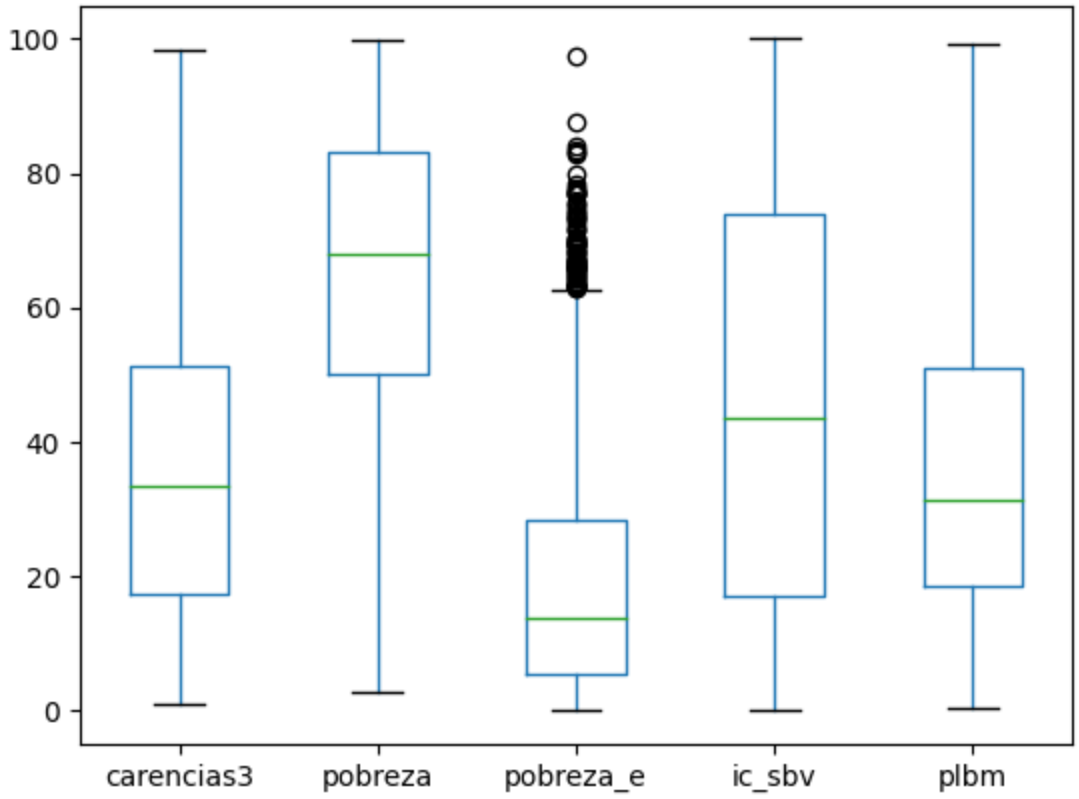


La variable npnv que indica no pobreza, ni vulnerabilidad muestra una correlación de -.86, -.72 y -.98 con las variables pobreza, rezago en educación y carencias respectivamente. Indica una correlación inversa.



Esta gráfica es muy interesante, ya que se puede observar que la brecha que existe entre las oportunidades que se obtienen al no tener pobreza, ni vulnerabilidad es aún mayor de lo que parece.

Esto debido a que, en la variable de no pobreza, ni vulnerabilidad se observa que se presentan ingresos superiores con valores atípicos al extremo superior, lo que contrasta con el comportamiento de las variables de vulnerabilidad por carencias e ingresos, donde los datos atípicos también son superiores al límite superior, lo cual indica que existen personas cuya vulnerabilidad es mucho mayor a la media.



De nuevo se observa el comportamiento de datos atípicos superior al límite superior de la variable que indica pobreza extrema.

Conclusión del análisis descriptivo.

Después de un análisis gráfico y técnico, se sabe que el rezago de educación, carencias y vulnerabilidad están fuertemente condicionadas por el nivel de ingresos y que se percibe.

También se observó que las condiciones de pobreza, rezago, carencias, etc. se concentran fuertemente en algunos estados de la república mexicana, de cierta manera condicionando a sus habitantes a no progresar por la falta de oportunidades.

Teniendo en cuenta que el bienestar se concentra en unos pocos y la pobreza está presente en la mayoría (El 87.95% de la población en el año 2018.), es necesario identificar de manera correcta a las personas que en realidad necesiten un apoyo para progresar, y de esa manera intentar reducir esa brecha existente de clases sociales apostando en la educación para un incremento en el bienestar.

Debido al enriquecimiento de la actividad informal y muchísimas variables más, quizás los datos no son los óptimos, lo cual podría causar sesgos en nuestros datos actuales.

Considero necesario investigar más a fondo este fenómeno, utilizando más datos, historia, decisiones políticas, inversión extranjera en el estado, etc.

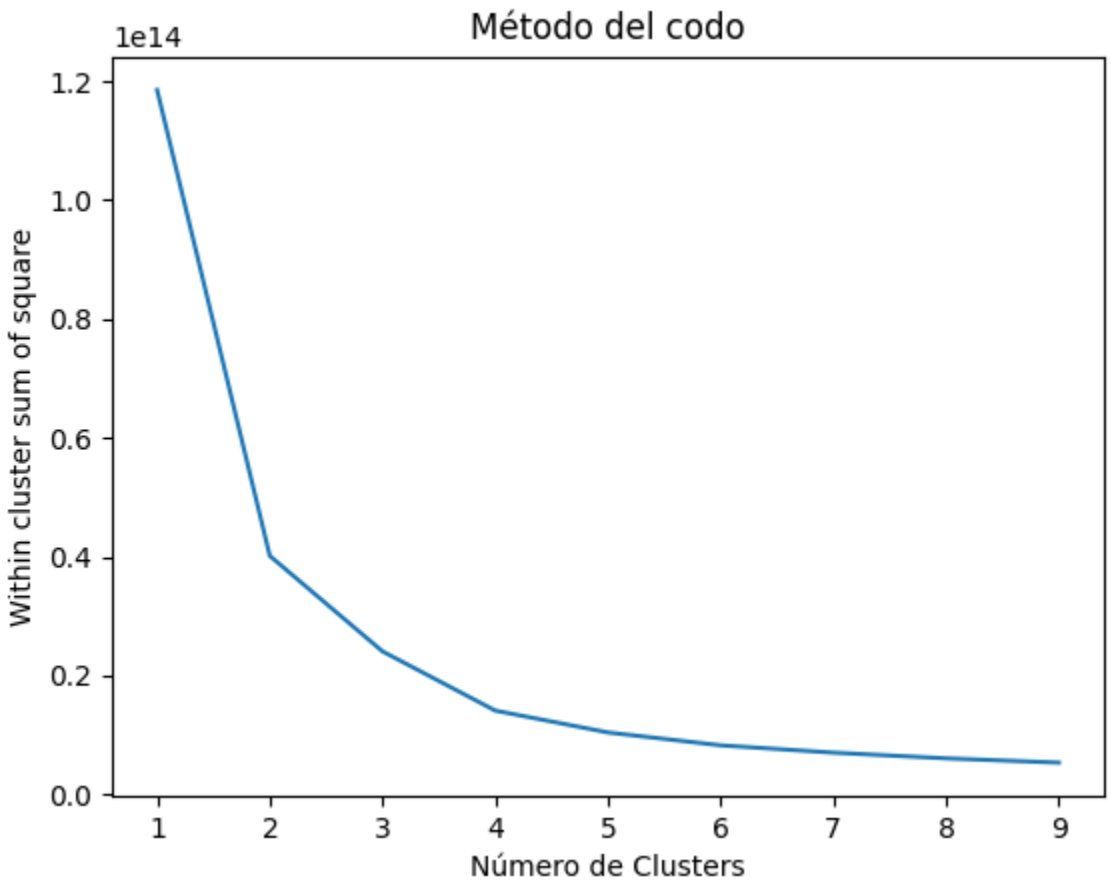
Ya que, si se destinan recursos para la educación en sectores donde por factores externos a los habitantes sean conflictos paramilitares, plantones, desplazamiento forzado por intereses personales, etc., podrían llegar a ser no tan eficientes debido a factores de los cuales no tenemos información al respecto en nuestros datos.

Definir un clasificador con las siguientes características:

Definir una variable objetivo para otorgar las becas

Al no contar con una variable objetivo, opté por un modelo de K-Means ya que es un algoritmo de clasificación no supervisada que agrupa objetos en k grupos basándose en sus características.

Para obtener un número de clusters sugerido, utilicé el método de codo (WCSS por sus siglas en inglés, within cluster sum of squares), el cual mide la distancia de los datos del cluster hacia su centroide.



Se debe iniciar el número de clusters donde la gráfica tenga forma de codo (En el punto en el que la pendiente deje de bajar de manera pronunciada).

En este caso al parecer es el número 4.

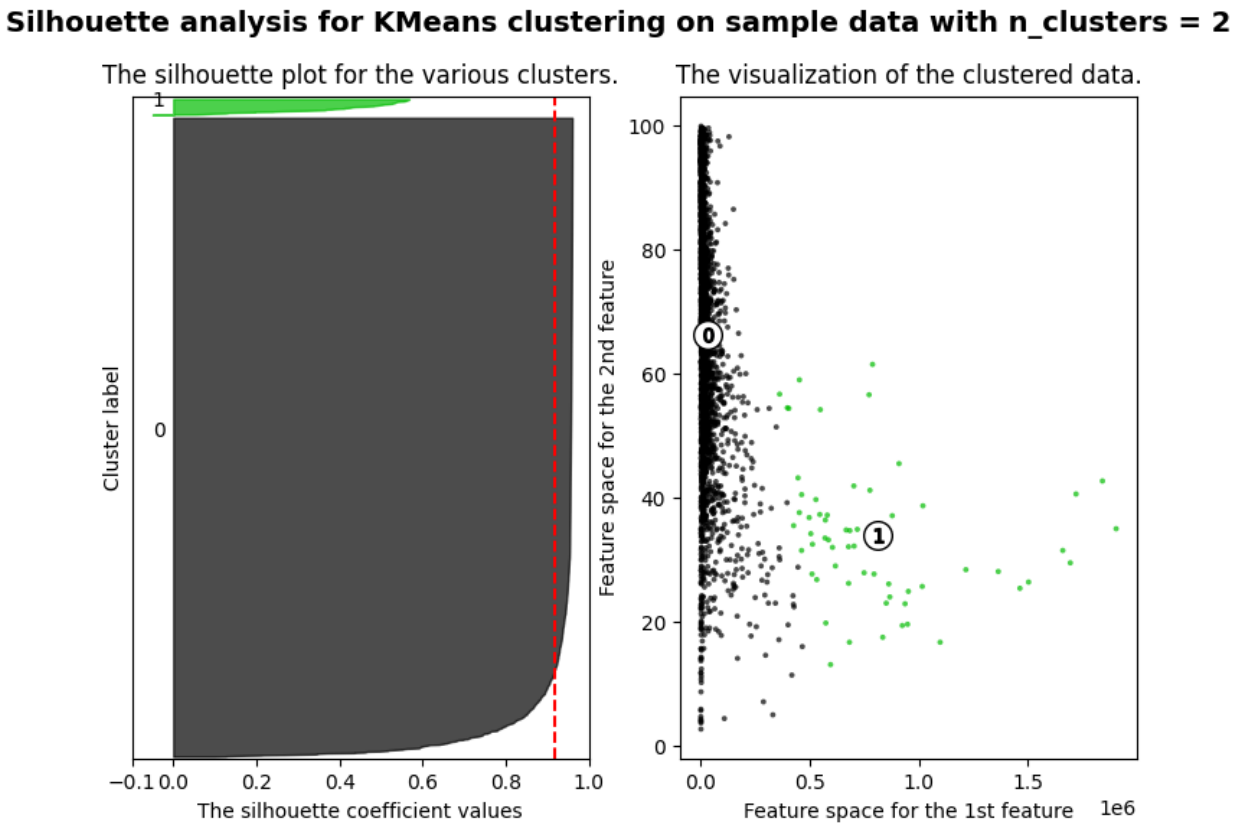
Debido a que no tenemos como comprobar la eficacia de la clasificación con datos previamente clasificados de manera correcta, se procede a determinar el número de clusters ideal en base a métricas medibles para una mejor toma de decisiones.

El método por utilizar es el de Silhouette score, la manera en la que mide la eficacia es la siguiente:

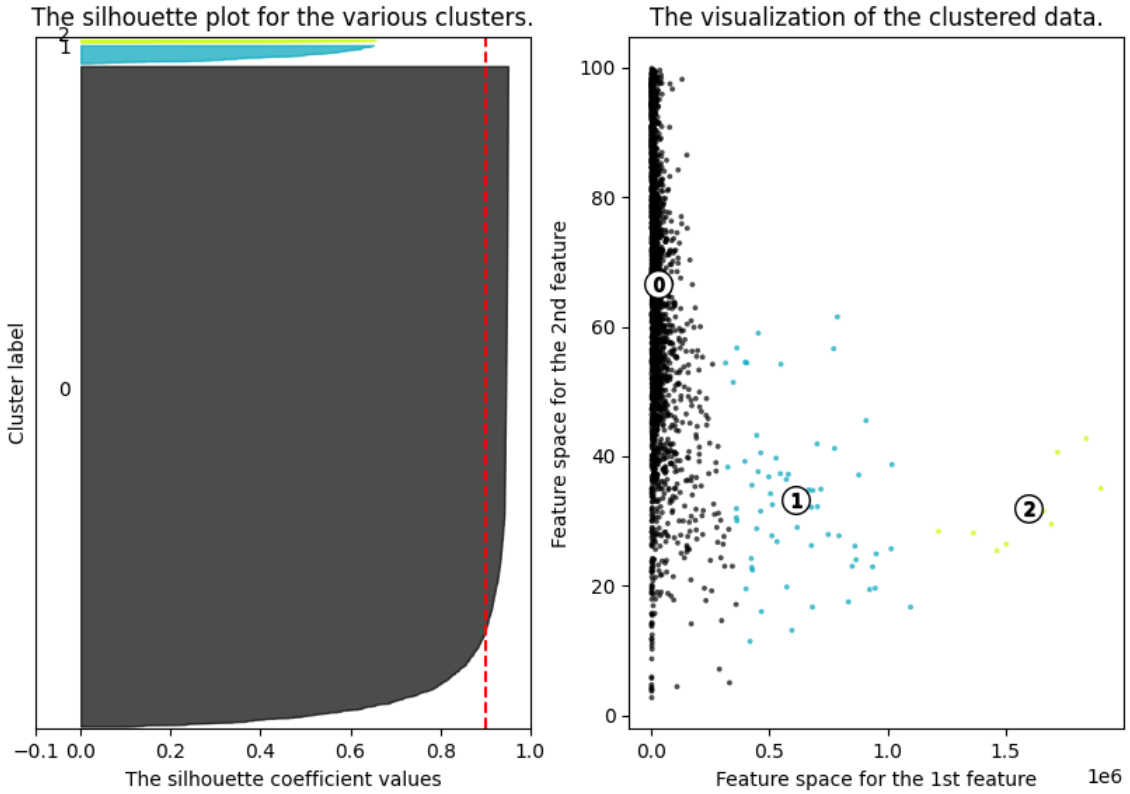
De cada cluster, se mide la distancia euclidiana de un dato vs. todos los demás dentro del cluster, una vez terminado, se pasa al siguiente punto y repite el proceso hasta terminar con todos los datos y se saca el promedio. A esto lo denotamos como **αi.**

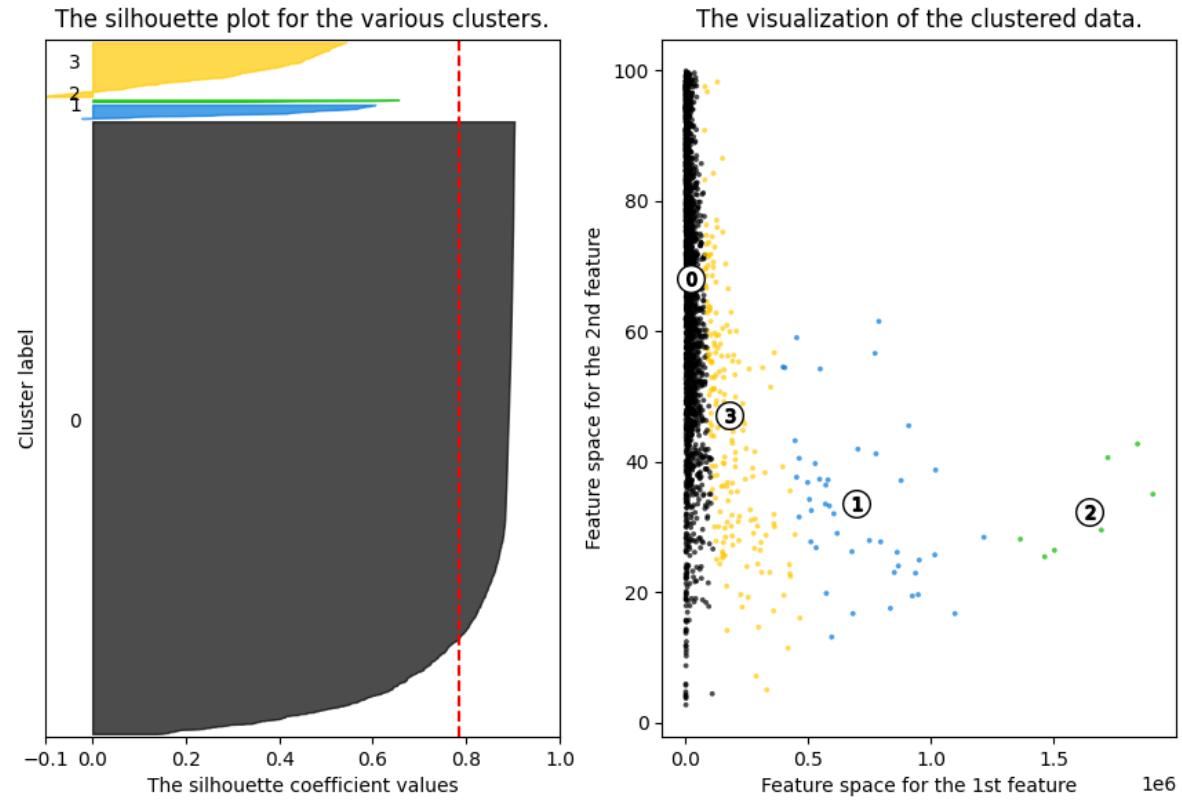
Después, hace el mismo proceso, pero ahora de un cluster a otro mide cada punto vs. los datos del siguiente cluster hasta terminar y de nuevo sacar un promedio. A esto lo denotamos con **βi**

Por lógica, **αi** siempre debe ser menor a **βi,** de lo contrario tendríamos un valor negativo indicando que el cluster está clasificando erróneamente algunos datos.

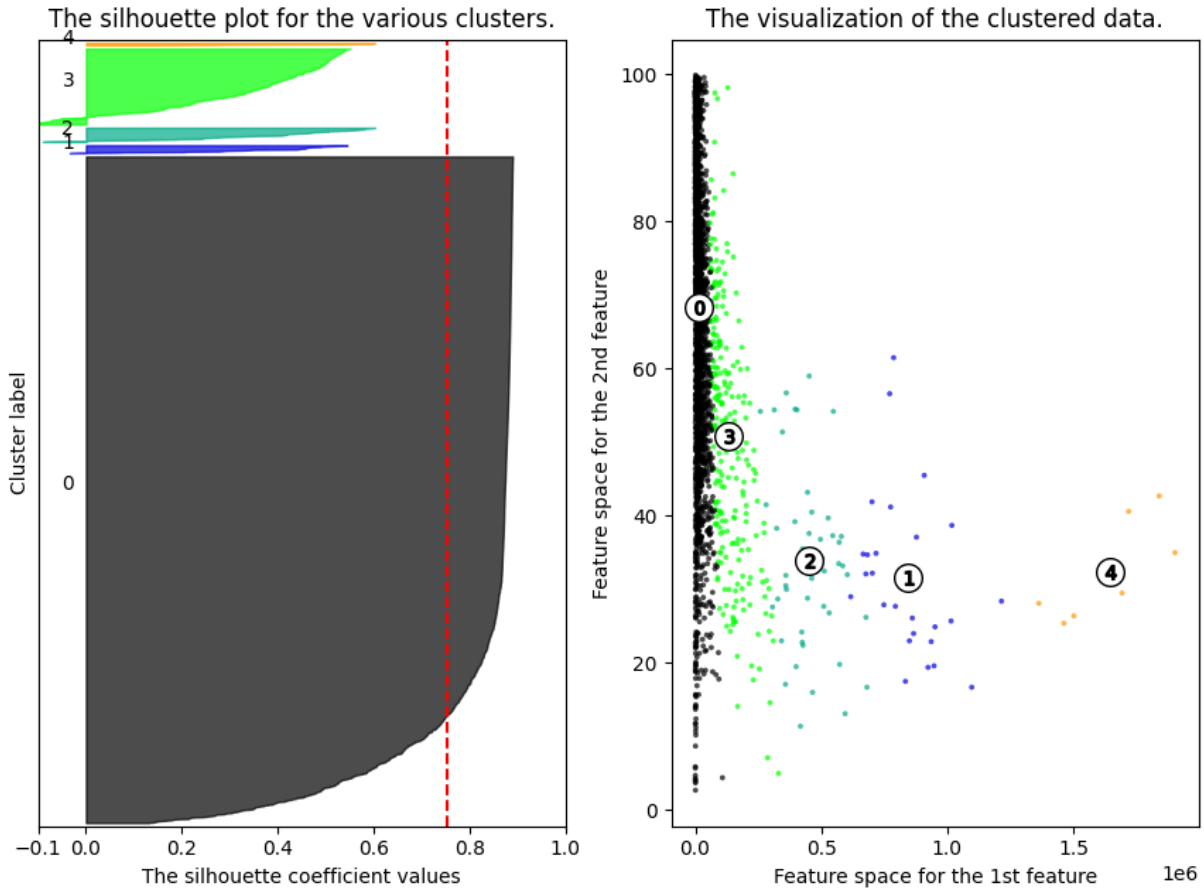


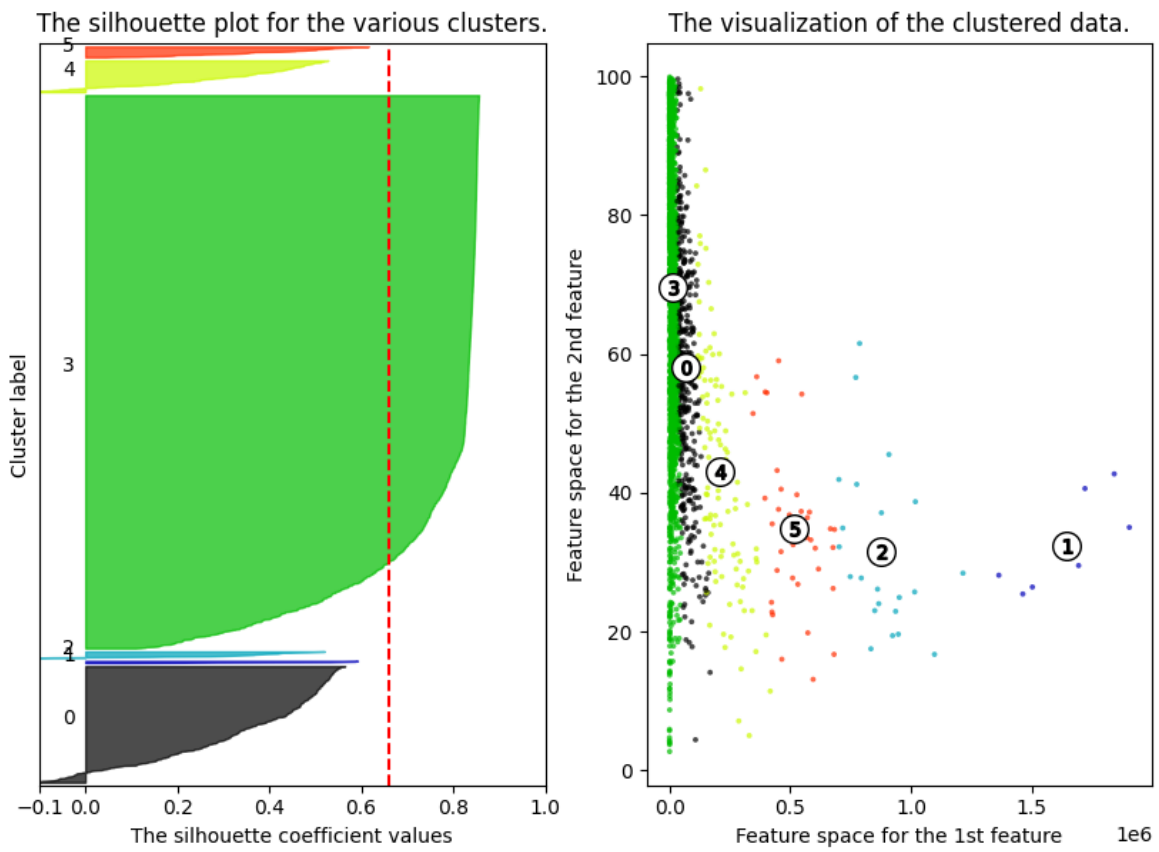
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3




Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
Feature space for the 2nd feature
100
80
60
40
20
0
0.0
0.5
1.0
1.5
1e6
Feature space for the 1st feature
0
1
2
3

Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



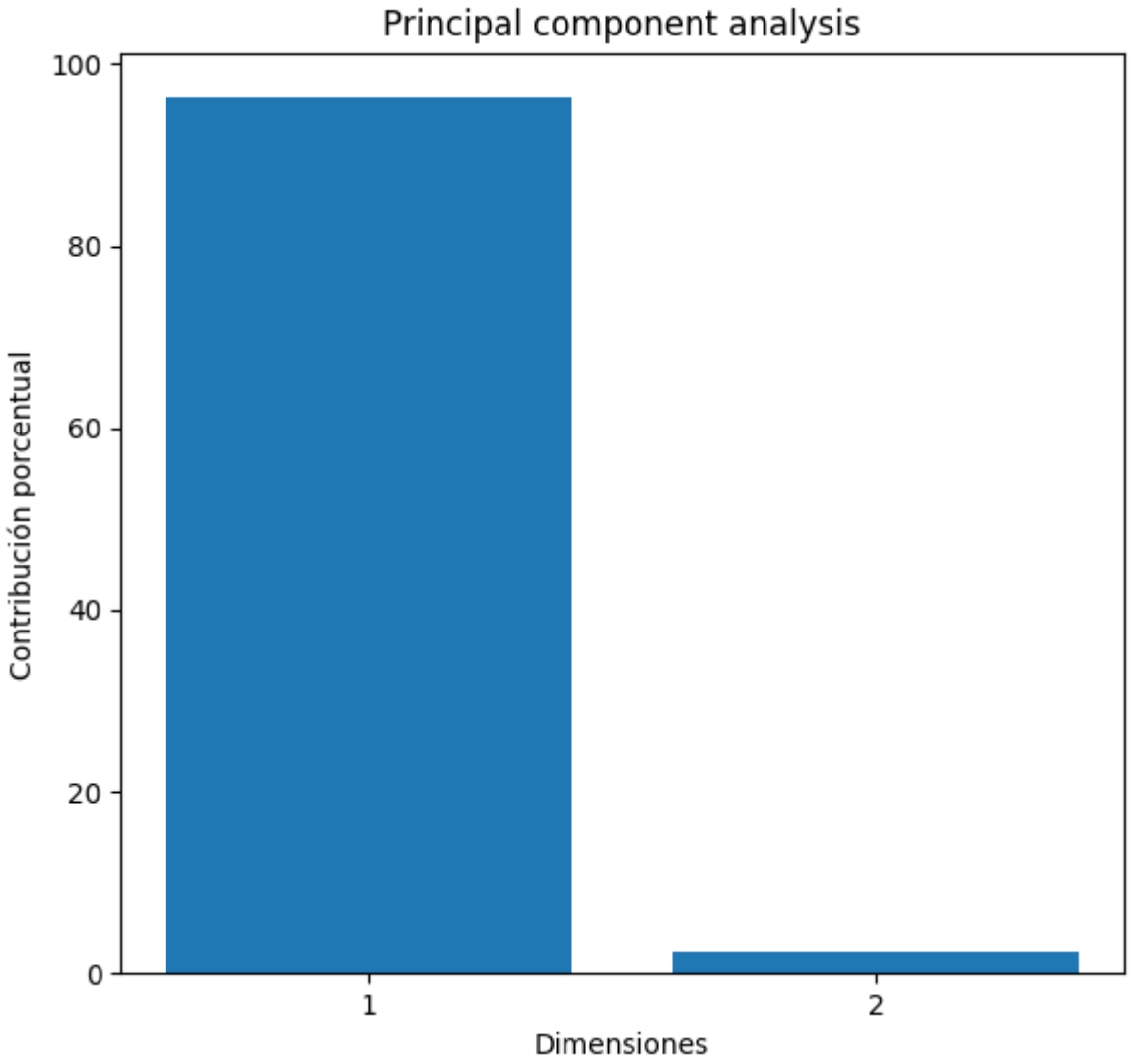


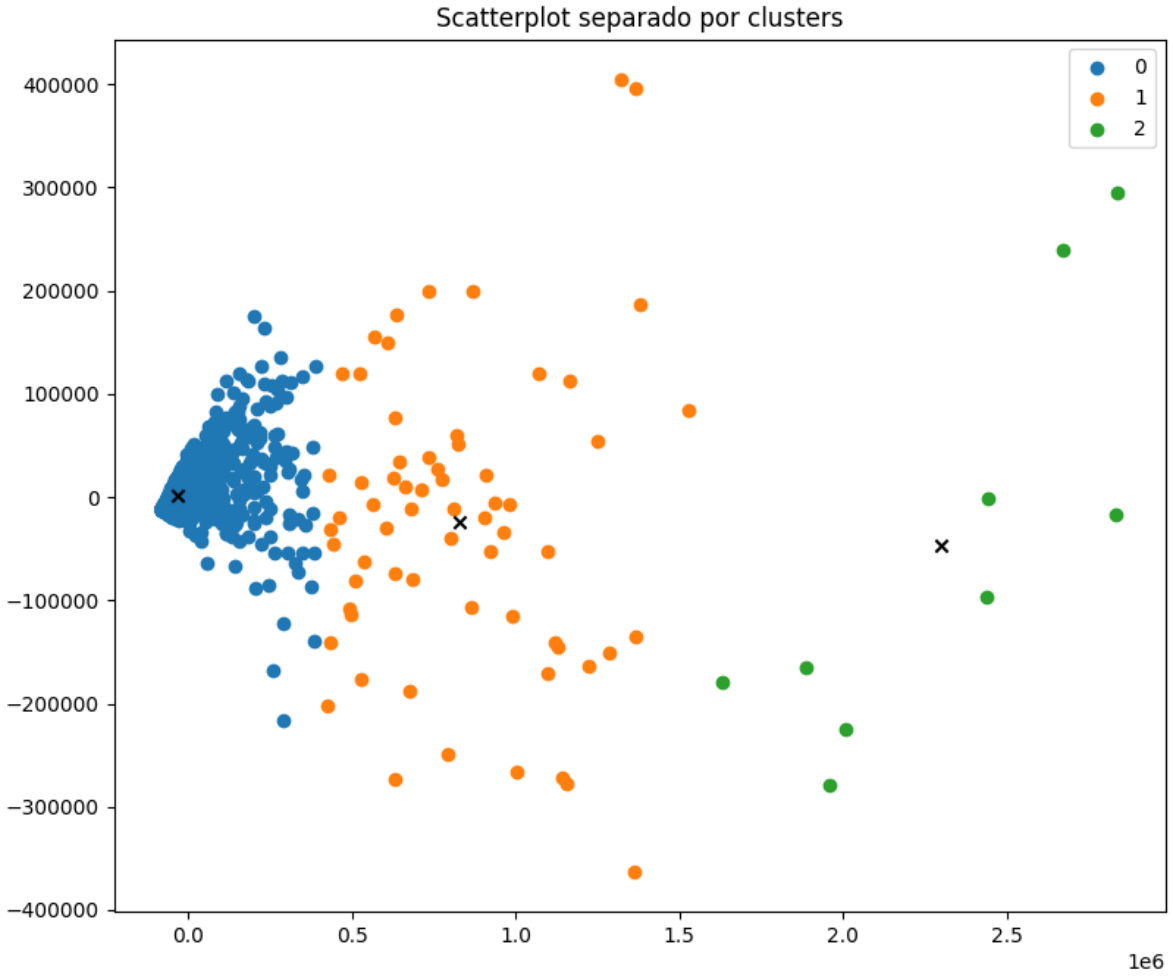

```
For n_clusters = 2 The average silhouette_score is : 0.9182619953358572
For n_clusters = 3 The average silhouette_score is : 0.9002392450617892
For n_clusters = 4 The average silhouette_score is : 0.7857412346885483
For n_clusters = 5 The average silhouette_score is : 0.7522856040891653
For n_clusters = 6 The average silhouette_score is : 0.6611584238631103
```

Como se puede observar en las gráficas, a pesar de tener un buen score, existen datos con valores negativos en los ejemplos con 2, 4, 5 y 6 clusters.

Por lo que utilizamos un modelo de K-Means con 3 clusters y un average silhouette_score de 0.9002392450617892

Para efectos de visualización de los clusters, se usó PCA.





Y ya que las variables que mayor correlación tienen con el rezago en educación son carencias3, pobreza extrema, porcentaje de la población con carencia por acceso a los servicios básicos en la vivienda y porcentaje de la población con ingreso inferior a la línea de bienestar mínimo. Se midió el promedio de estas variables dentro de cada cluster obteniendo los siguientes resultados:

Cluster	pobreza_e	carencias3	ic_sbv	plbm
0	20.227951	36.383137	47.211003	37.615725
1	3.370769	11.978462	6.760000	11.464615
2	2.211111	9.500000	3.655556	10.688889

Siendo el cluster 0, la variable objetivo para el programa de becas.

Y el de mayor ocurrencia en los datos:

Cluster	Cantidad
0	2372
1	65
2	9

3.- Definir una política de reentrenamiento y consideraciones para el clasificador.

Debido a que la fuente de datos en la que el modelo está basado se actualiza anualmente, el reentrenamiento tendría que ser anual a menos de que se agregue información de otra fuente.

Una vez que haya datos nuevos, el proceso sería aplicar la misma metodología y cambiar parámetros si así lo requiere el modelo.

Consideraciones:

Como se menciono con anterioridad, el análisis del lugar geográfico y su situación política, social, ocupación, etc., es fundamental para tomar una decisión que maximice la ayuda dada y no se mal gaste.

Por último, es recomendable analizar otros algoritmos de clasificación como Regresión logísitca o Naive Bayes para poder comparar distintas métricas y tomar el modelo que más convenga.