MOVIEALPES

- Samuel Freire 202111460 (K-medoids)
- Juan Felipe Garcia 202014961 (K-means)
- Lucciano Franco Márquez 202111458 (DbScan)

```
df_movies.describe()
```
✓ 0.0s

	#	index	isAdult	startYear	averageRating	numVotes	ordering	isOriginalTitle
count	7471.000000	7471.000000	7470.0	7470.000000	7470.000000	7.470000e+03	7470.000000	7470.0
mean	4043.482666	5440.458439	0.0	2013.157296	7.475676	7.282830e+04	16.567604	0.0
std	2192.753689	2752.820924	0.0	6.979151	2.771444	1.640233e+05	12.761147	0.0
min	1.000000	1.000000	0.0	1990.000000	6.500000	1.000000e+01	1.000000	0.0
25%	2152.500000	3073.500000	0.0	2008.000000	6.900000	6.265250e+03	6.000000	0.0
50%	4033.000000	5421.000000	0.0	2015.000000	7.300000	1.527900e+04	14.000000	0.0
75%	5914.500000	7641.500000	0.0	2019.000000	7.800000	5.641975e+04	24.000000	0.0
max	7849.000000	10274.000000	0.0	2023.000000	92.000000	2.197234e+06	119.000000	0.0

ENTENDIMIENTO DEL NEGOCIO Y LOS DATOS

Datos importantes

1. Existen múltiples valores que no tienen significado en el diccionario o que no se consideran relevantes como # y el index que seran tratados en un apartado anterior.

2. Antes del año 2000 en start year se ve que no existe ninguna valoración por debajo de 7,7 por lo que se puede decir que estos valores faltan o que antes de ese año esa era la minima calificacion.

3. Todos los lenguajes seleccionados son en ingles por lo que se podria obviar esta columna al no ser un factor diferenciador en el analisis, se cree que todos son en ingles debido a la region a las que estan asociadas.

4. Lo mismo que sucede con lenguajes sucede con isAdult en esta columna todos los valores son 0 por lo que se podria obviar.

5. En cuanto al tipo de datos de las columnas, existen dos columna que no tienen el tipo mencionada en el diccionario por lo que esto se arreglara mas adelante.

Estos análisis nos ayudan a entender el conjunto de datos y a identificar posibles tareas de limpieza que se deben realizar en la etapa de preparación, antes de generar un modelo de agrupación.

Calidad de datos

```
numVotes          0.000134
ordering          0.000134
main_genre        0.000134
isOriginalTitle   0.000134
attributes        0.000134
types             0.000134
language          0.000134
region            0.000134
secondary_genre   0.000134
averageRating     0.000134
runtimeMinutes    0.000134
startYear         0.000134
isAdult           0.000134
index             0.000000
originalTitle     0.000000
titleType         0.000000
tconst            0.000000
#                 0.000000
dtype: float64
```

Análisis de calidad

- **Completitud:** Hay varias columnas que tienen un 0.000134 porciento de datos nulos por lo que al ser un mismo porcentaje se sospecha que pueda ser una misma fila, y del mismo modo al ser un número tan pequeño se considerara en rellenar la información con algun valor constante como puede ser la media, por ejemplo.
- **Unicidad:** En este caso existen valores duplicados, en tconst existen valores duplicados lo cual no tendria que suceder ya que este es el ID de cada uno de los datos por lo que es un claro candidato a ser eliminado. Aparte de esto existen 6 filas duplicadas por lo que estas seran candidatas para se eliminadas al ser un numero muy pequeño.

Análisis de calidad

- **Consistencia:** En este apartado se pueden diferenciar varios problemas de con la consistencia, en el apartado de titleType existen distintas escrituras para el valor Movie por lo que es importante revisarlo mas tarde y validar los que estan mal escritos. Del mismo modo el tipo de unas columnas esta mal por lo que se revisara el problema mas tarde, un ejemplo de esto es la columna runtimeMinutes.
- **Validez:** Existen valores en averageRating que no funcionan del modo que deberian ya que el rango deberia ser de 0 a 10 con decimales y hay valores que se pasan de este valor, por lo que esto tendra que cambiarse mas adelante. Por ejemplo como se ve arriba el maximo valor es 92. Aparte de esto la validez de los datos de isAdult es bastante poca ya que todos los valores son el mismo. Algo similar sucede en el caso de la columa language y originaltype

Manipulación de datos

```
numVotes            0.000134
ordering            0.000134
main_genre          0.000134
isOriginalTitle     0.000134
attributes          0.000134
types               0.000134
language            0.000134
region              0.000134
secondary_genre     0.000134
averageRating       0.000134
runtimeMinutes      0.000134
startYear           0.000134
isAdult             0.000134
index               0.000000
originalTitle       0.000000
titleType           0.000000
tconst              0.000000
#                   0.000000
dtype: float64
```

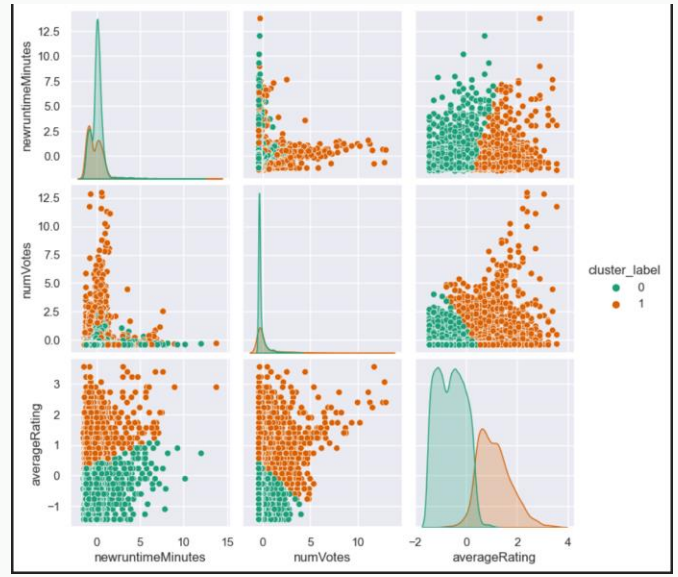
Manipulacion

- Eliminar columnas que no tienen utilidad para el análisis de los datos
- Hay filas con atributos que no coinciden con el tipo de diccionario debido a esto deben ser cambiados
- Aparte de los mencionados anteriormente la columna avarageRating tiene numero que se estan evaluando del 1 al 100 no del 1 al 10 con decimales por lo que esto es necesario ajustarlo
- En titleType hay valores que son de tipo Movie pero escrito de distintos modos, debido a esto es necesario cambiarlo para que no existan distintos tipos de Movie.

Manipulacion

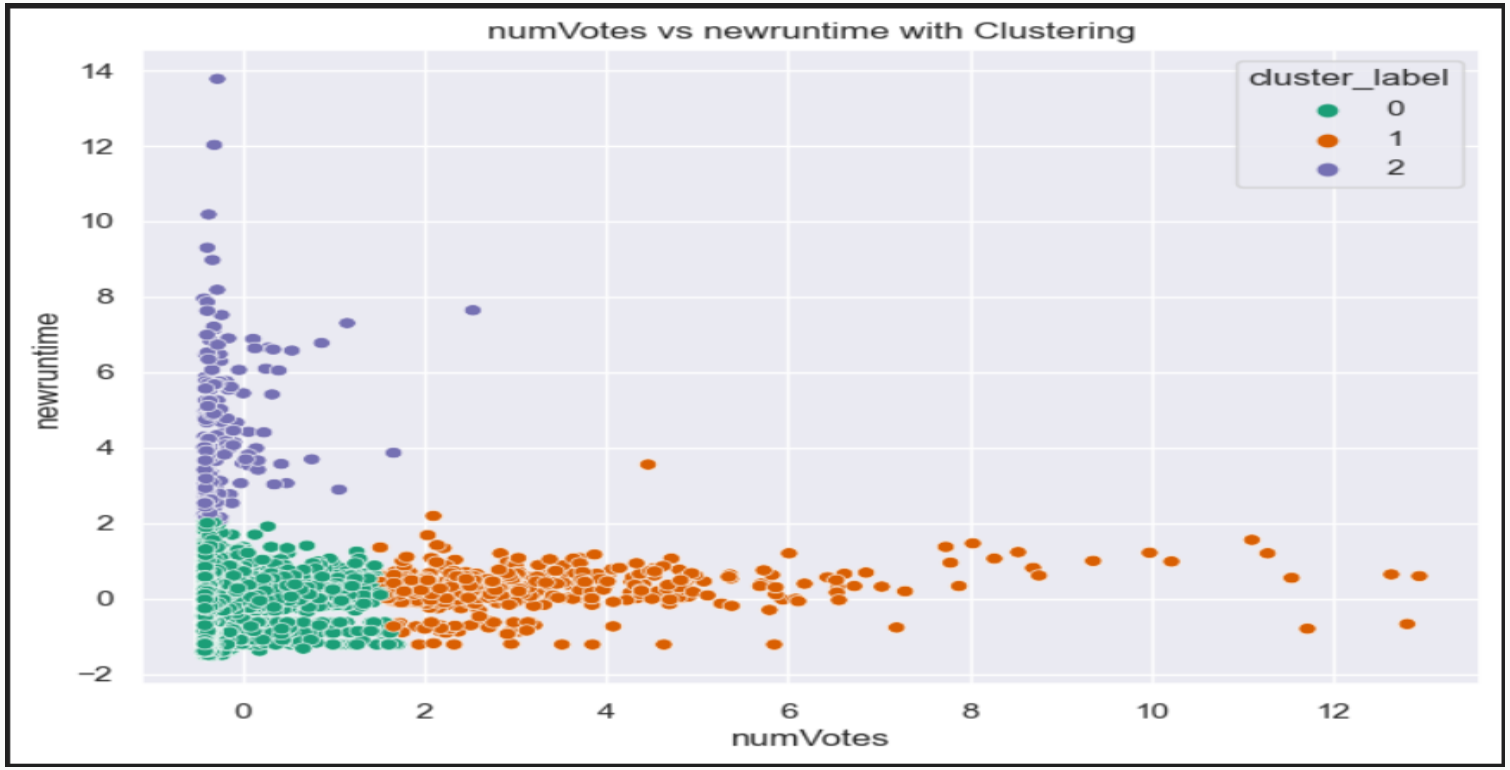
- "#": no se aclara que es lo que hace la columnas y se puede pensar que son numeros sueltos sin un sentido claro.
- "index": mismo razonamiento que "#"
- "isAdult": isAdult puede que fuese relevante si no todos los valores fueran 0, se cree que hay valores que son distintos a 0 realmente por lo se cree que la columna esta mal y de igual modo si esta correcta es irrelevante para el analisis de datos
- "isOriginalTitle": mismo razonamiento que isAdult
- "language": del mismo modo que isAdult todos los valores son iguales por lo que no se considera relevante para el analisis de datos.
- "Attribute": Se hizo el mismo razonamiento que para isAdult, todos los datos son ¥N por lo que se considera que no es relevante para el analisis de datos

Modelado

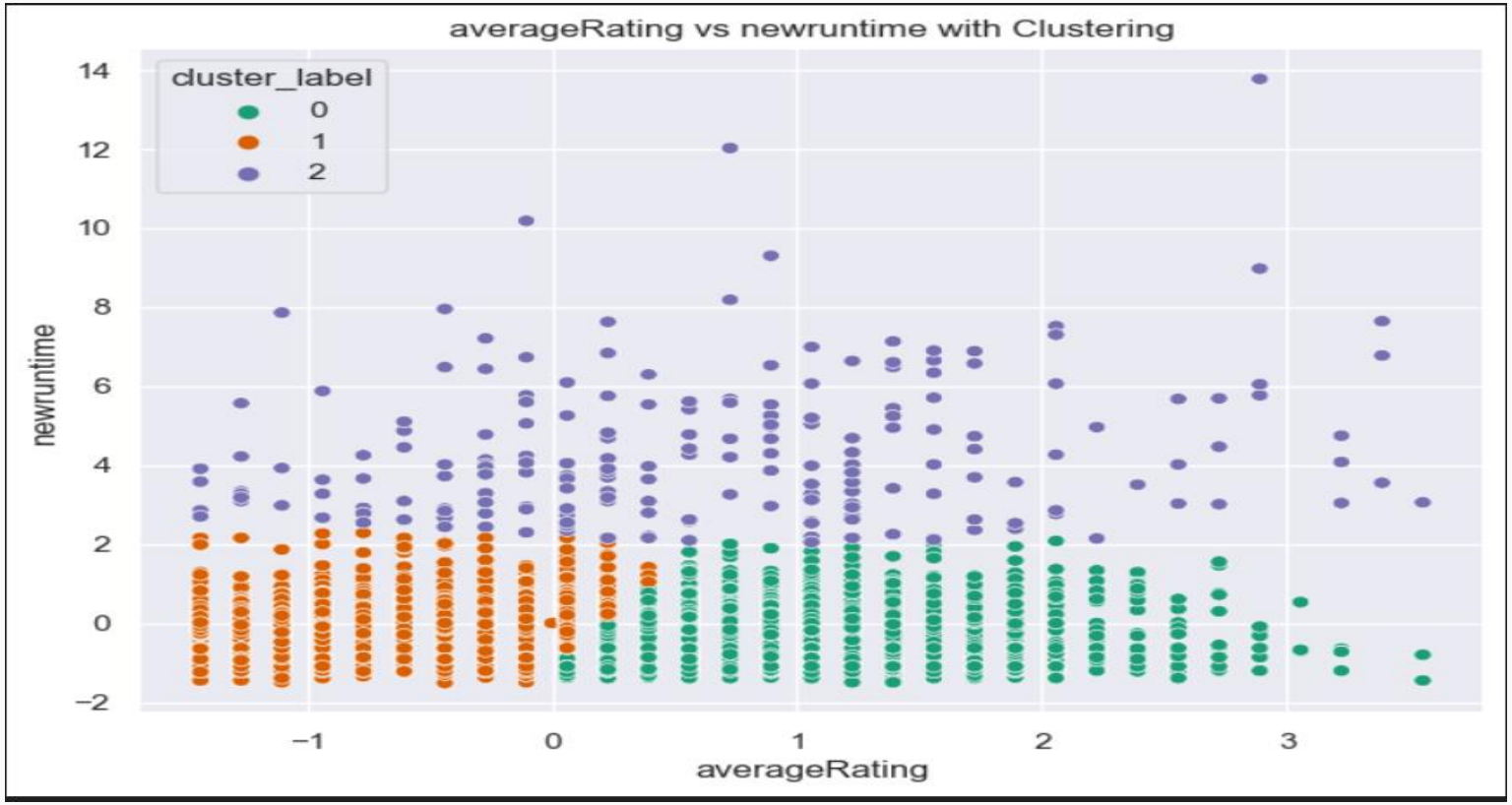

newruntimeMinutes
numVotes
averageRating
cluster_label
0
1

K-medians

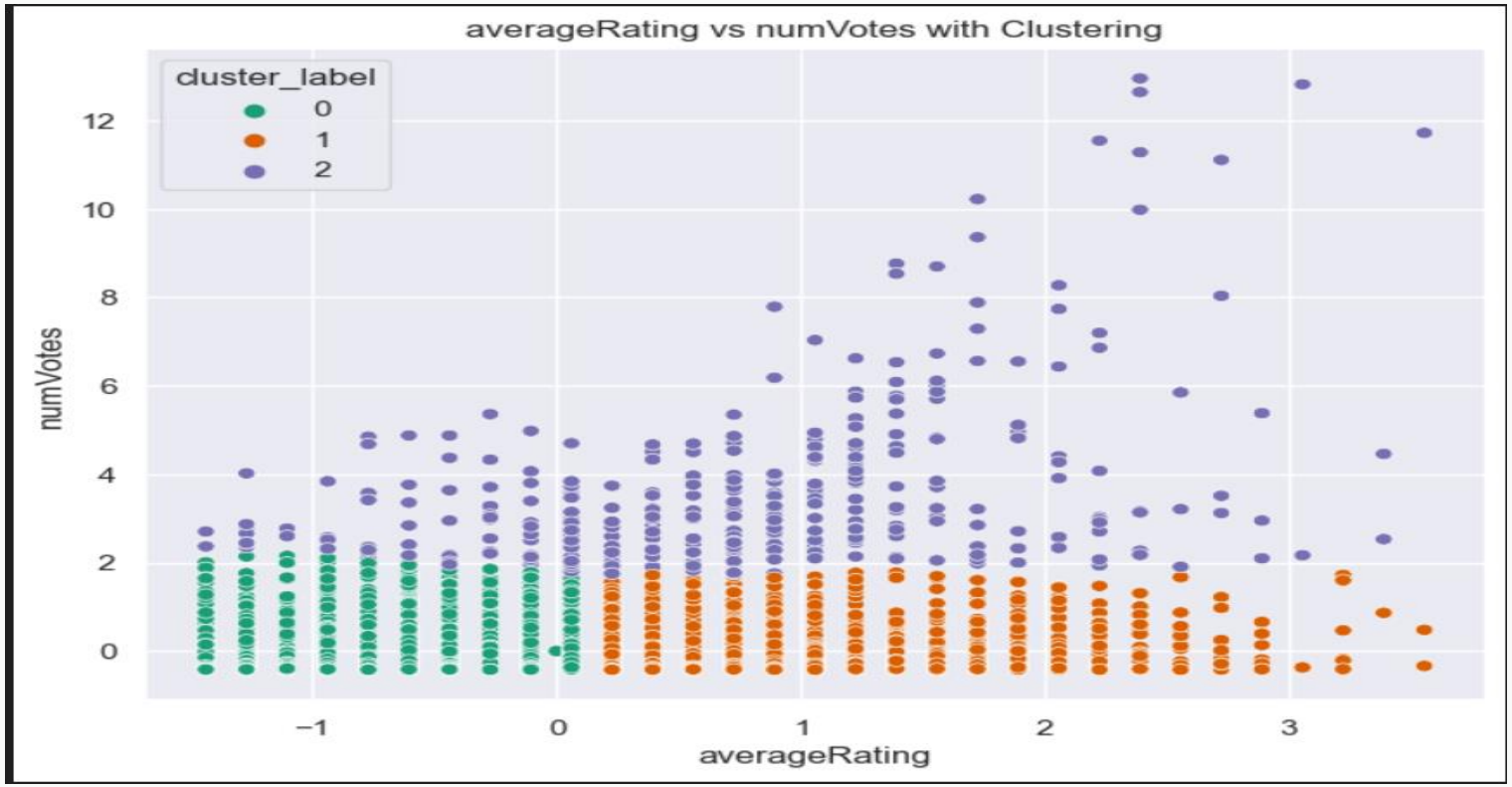
NumVotes vs runTime



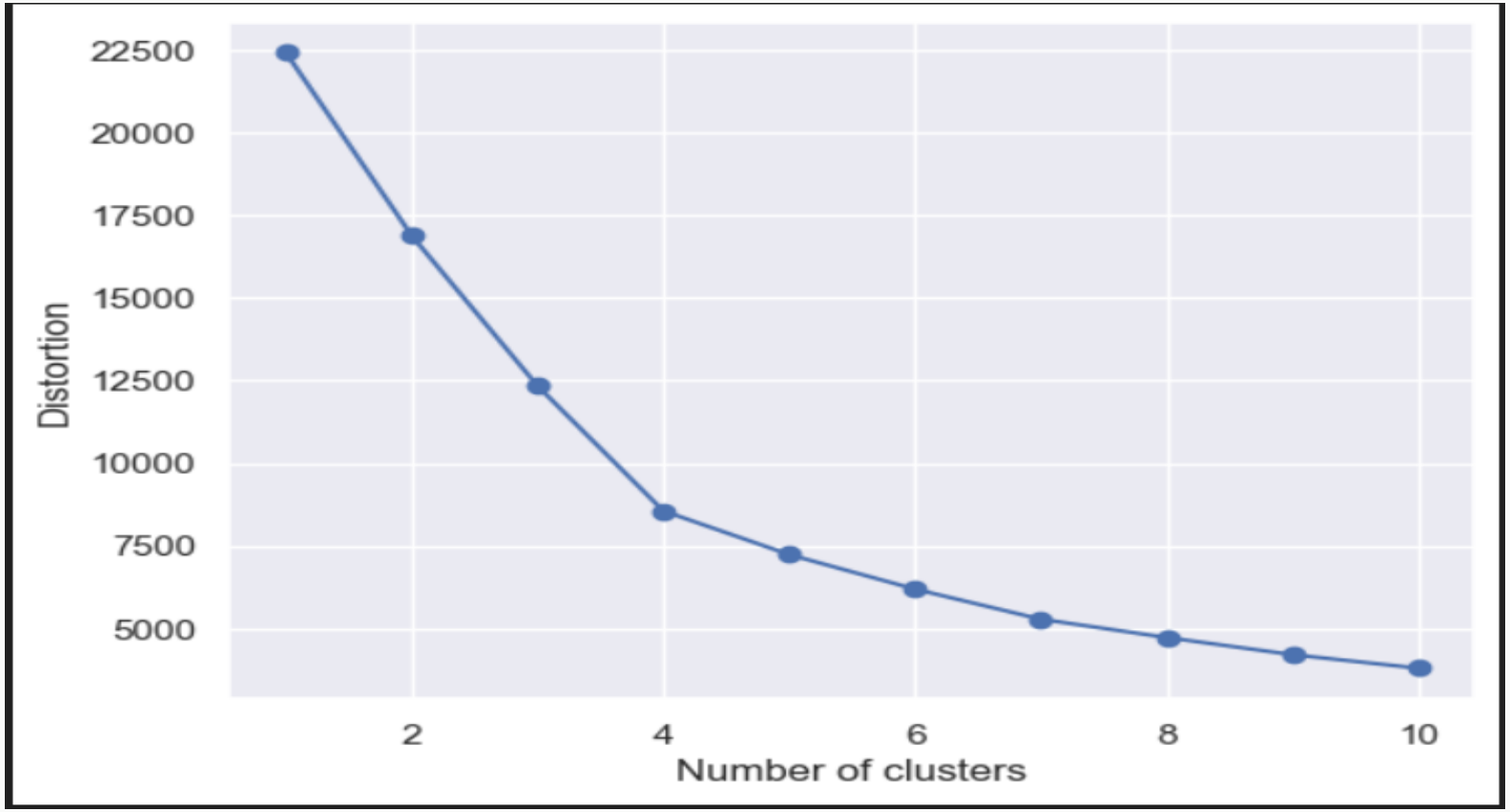
averageRating vs runTime



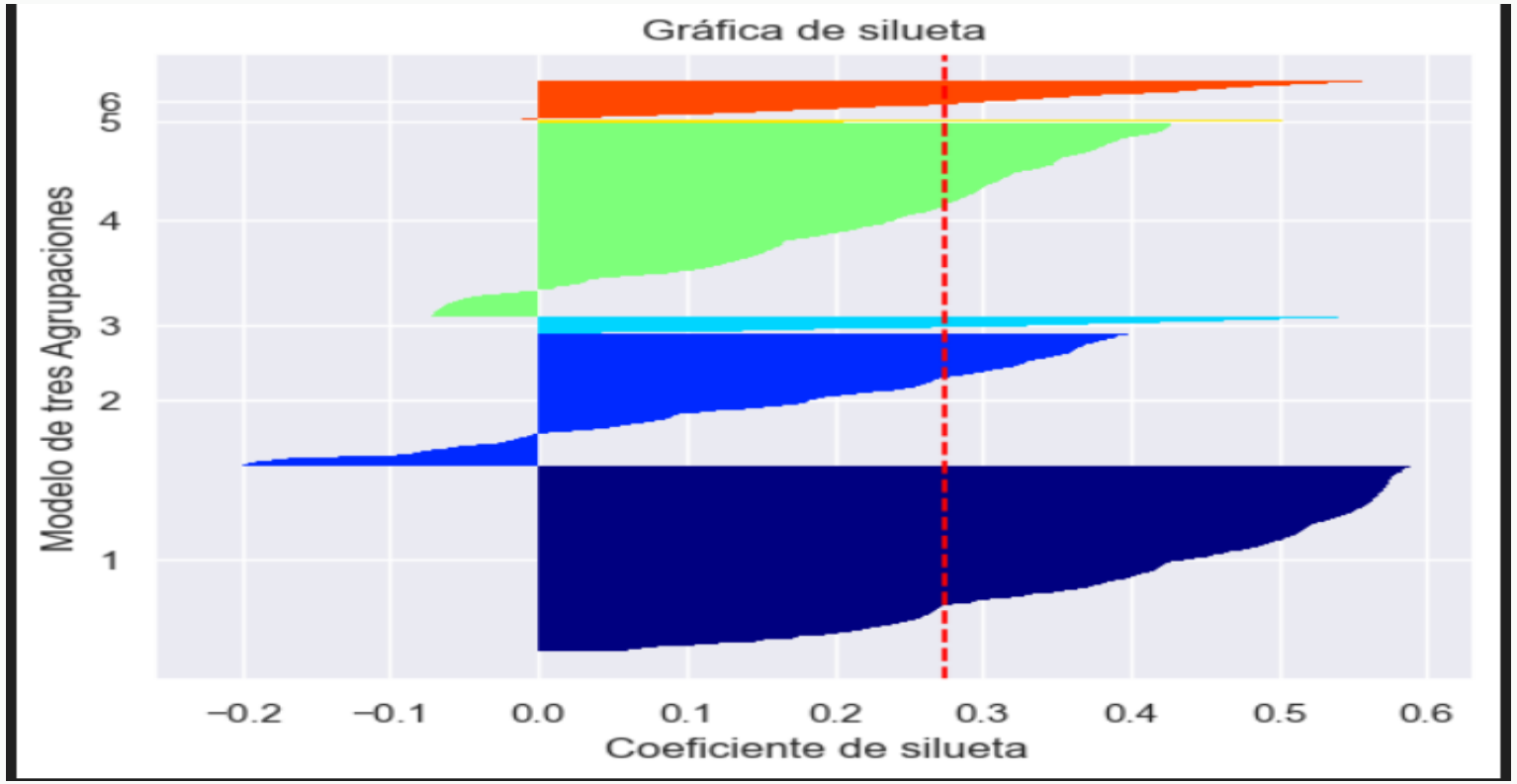
averageRating vs numVotes



Validacion Cuantitativa

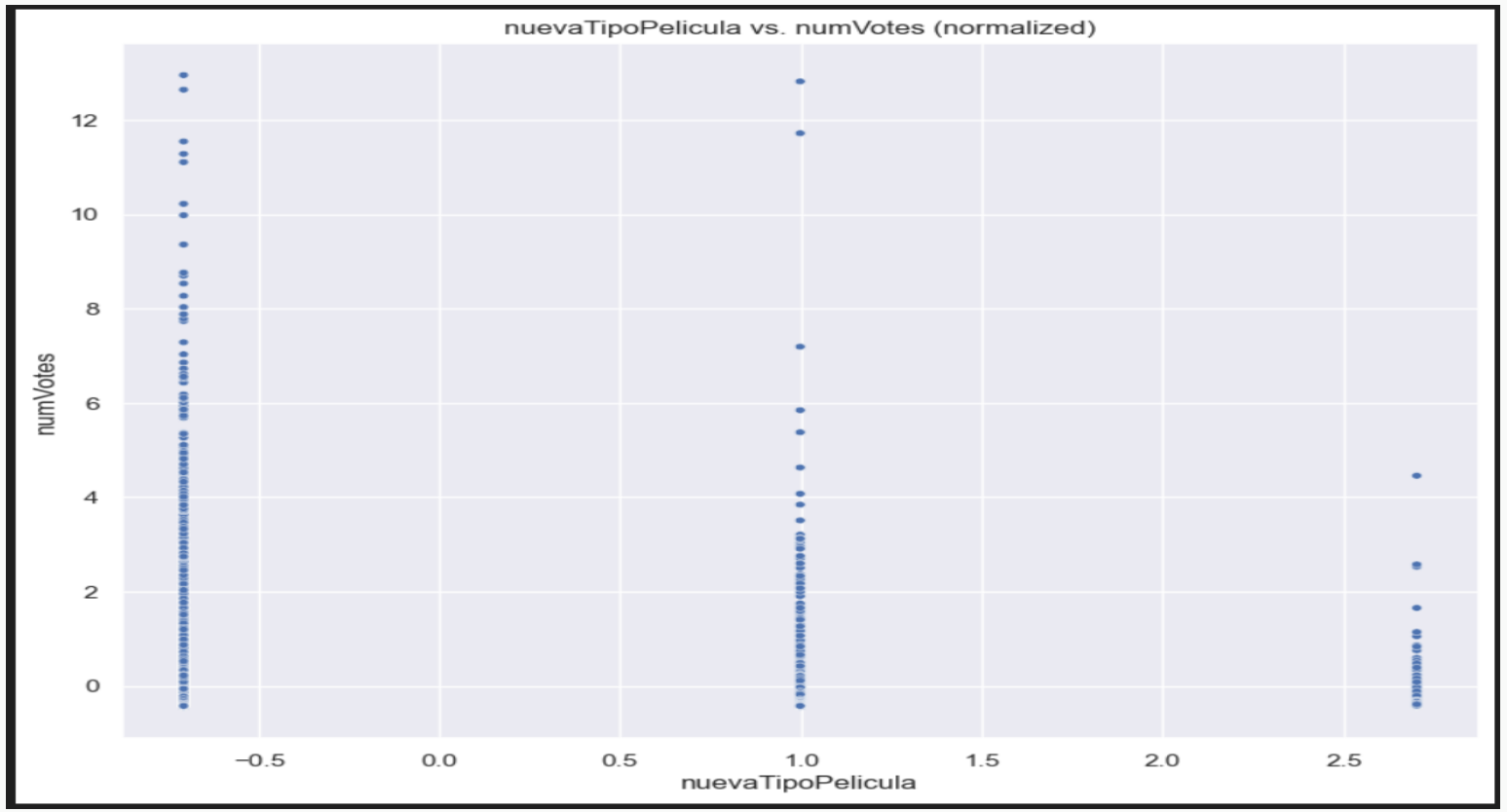


Validacion Cualitativa

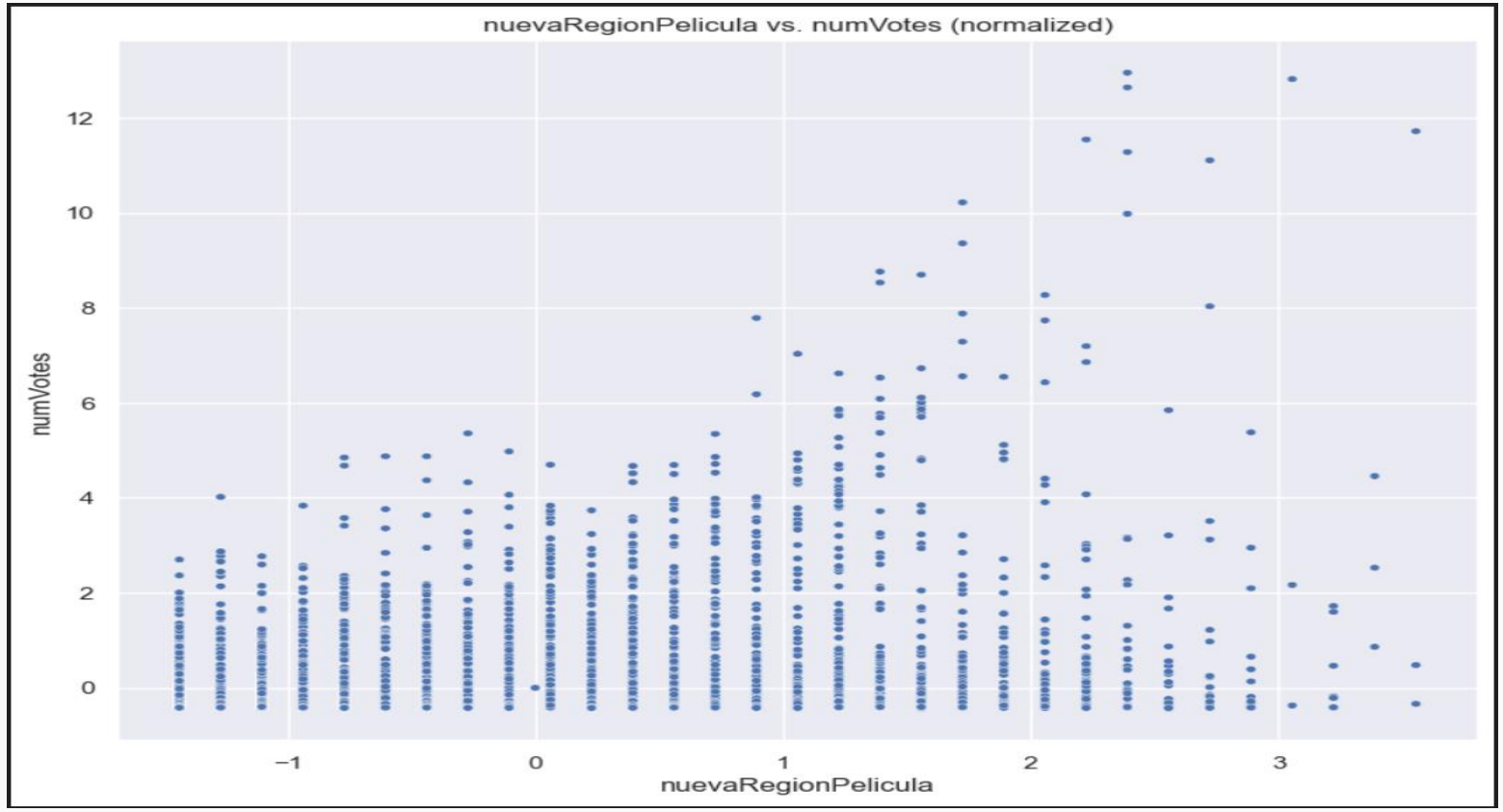


K-medoids

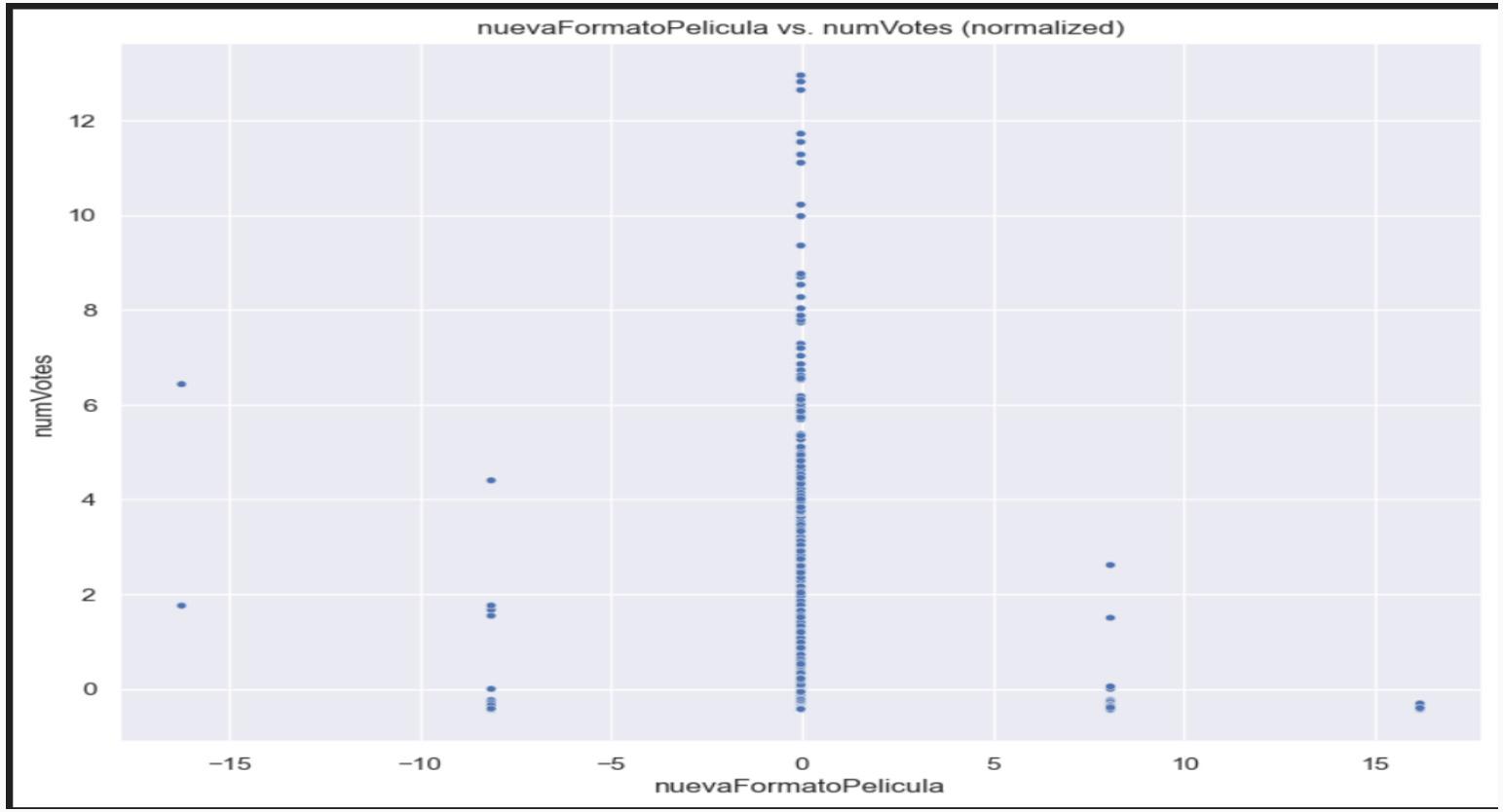
Numero de votos vs Tipo de pelicula



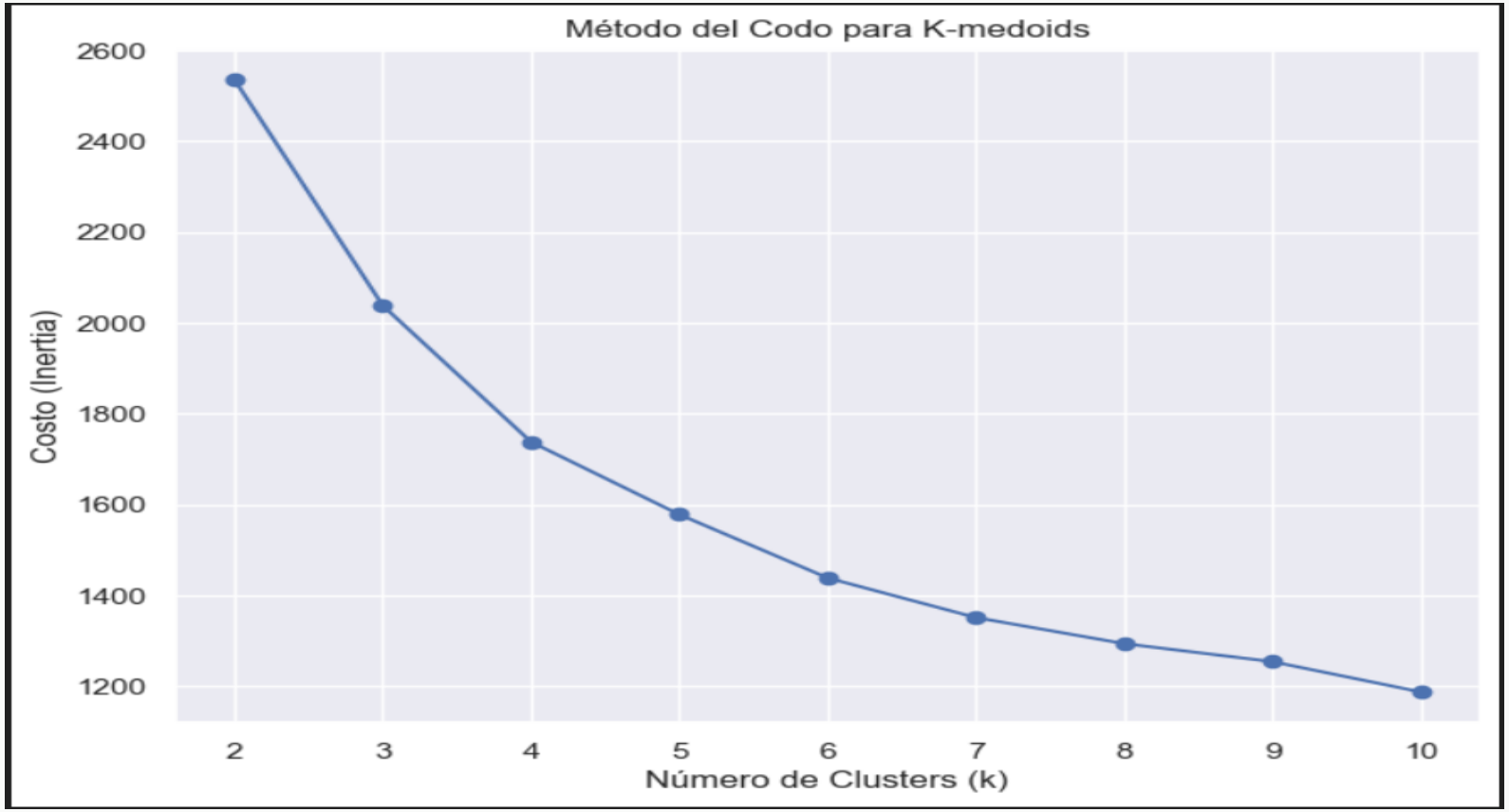
Numero de votos vs AverageRating



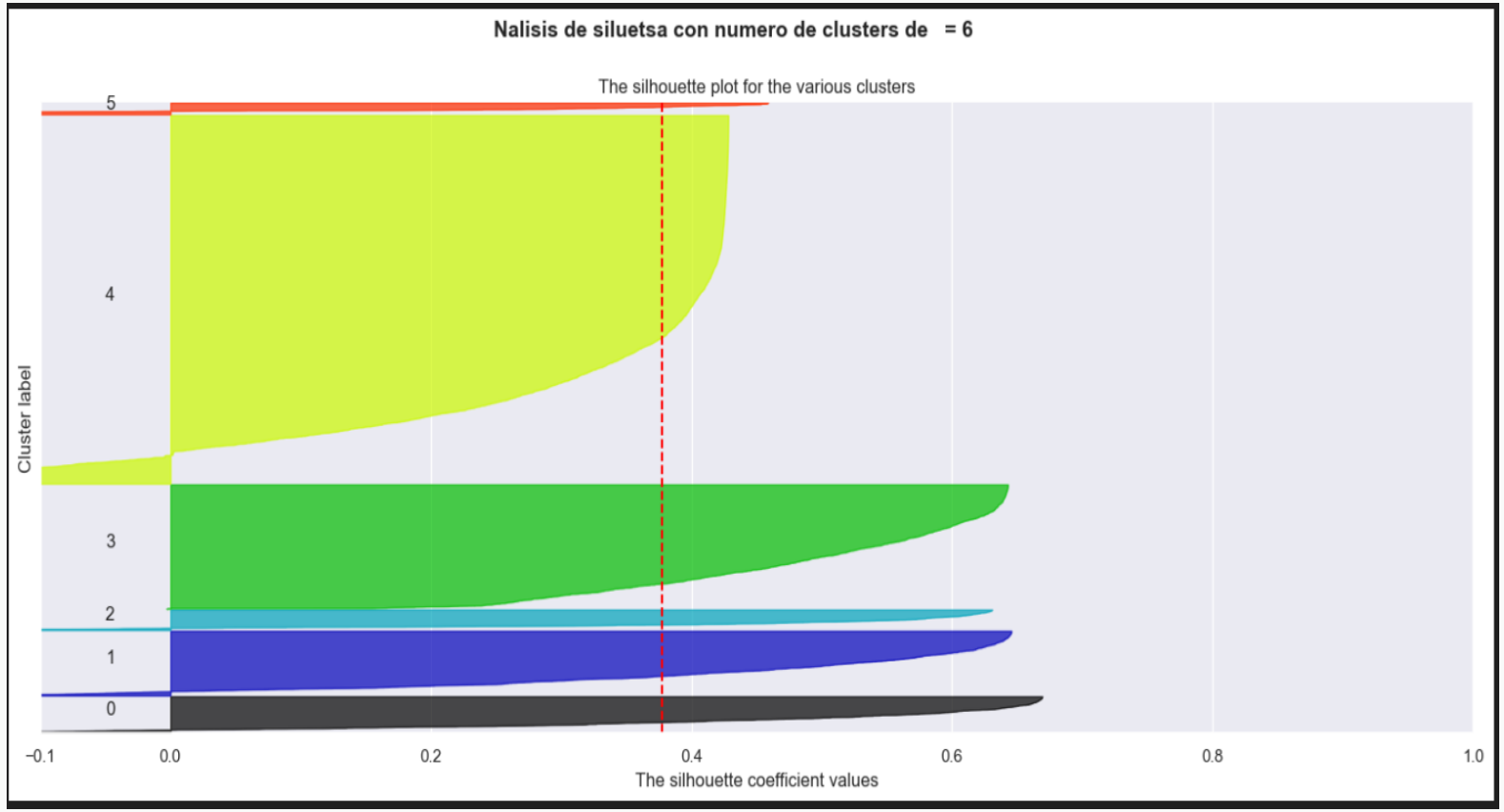
Numero de votos vs Formato de pelicula



Validacion Cuantitativa

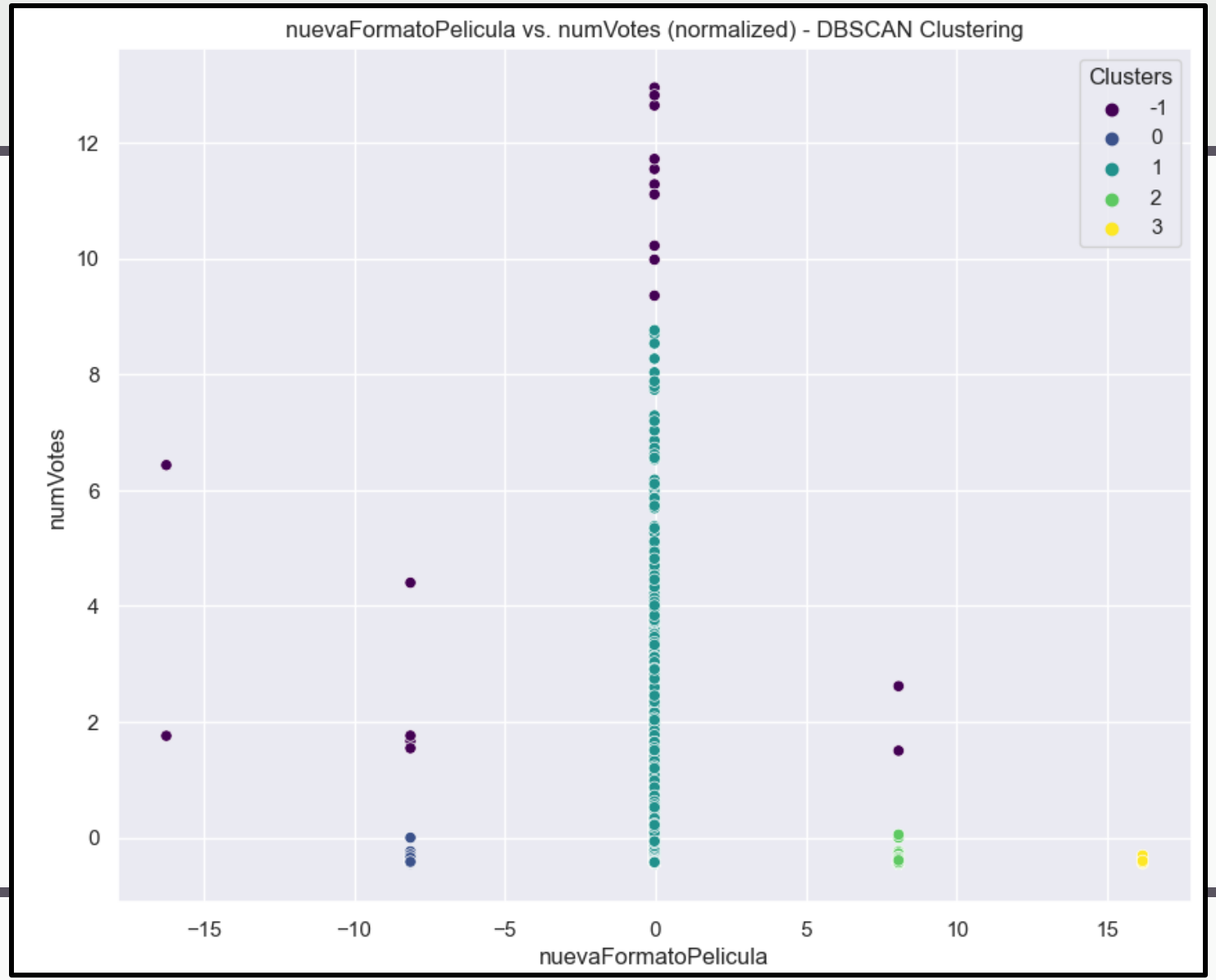


Validacion Cualitativa

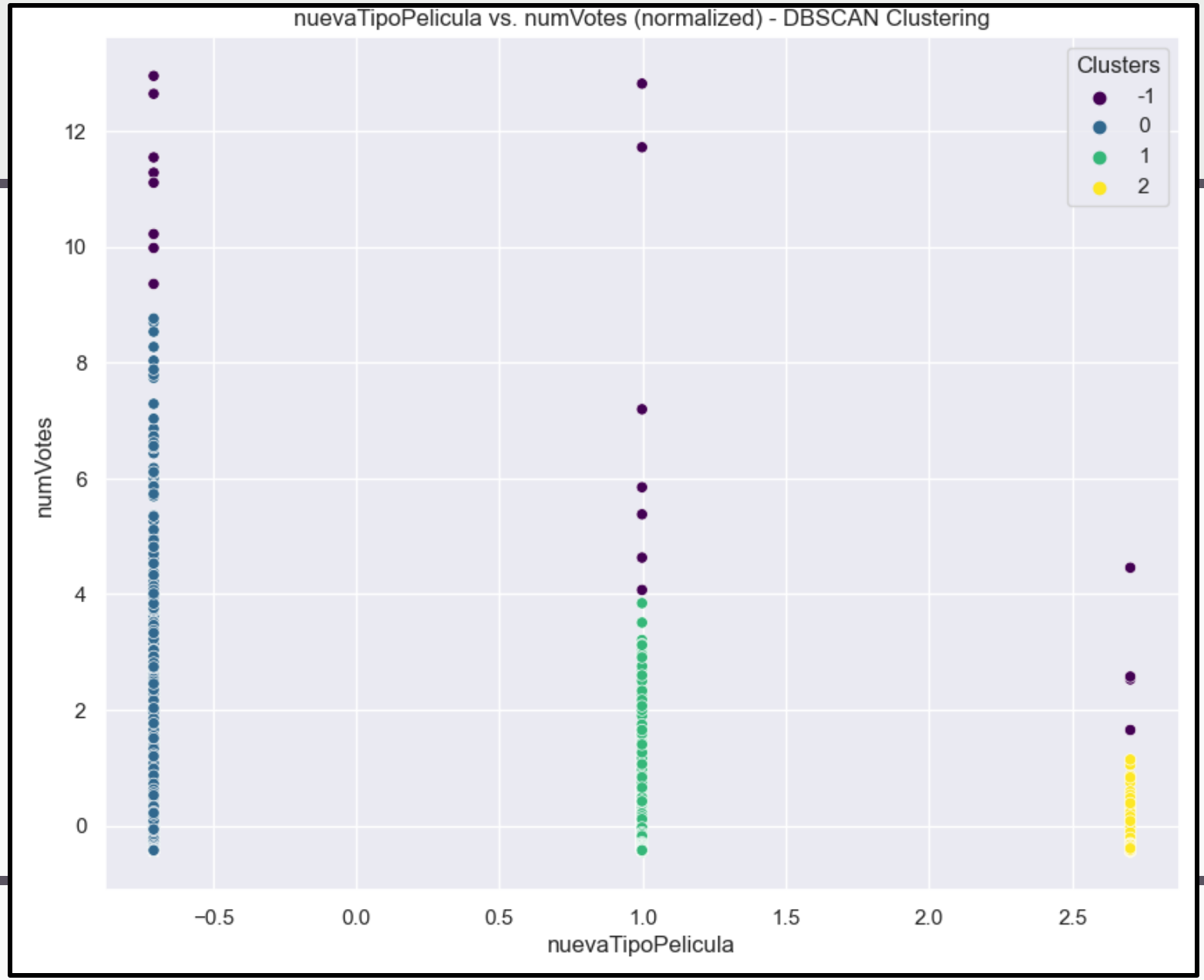


DbScan

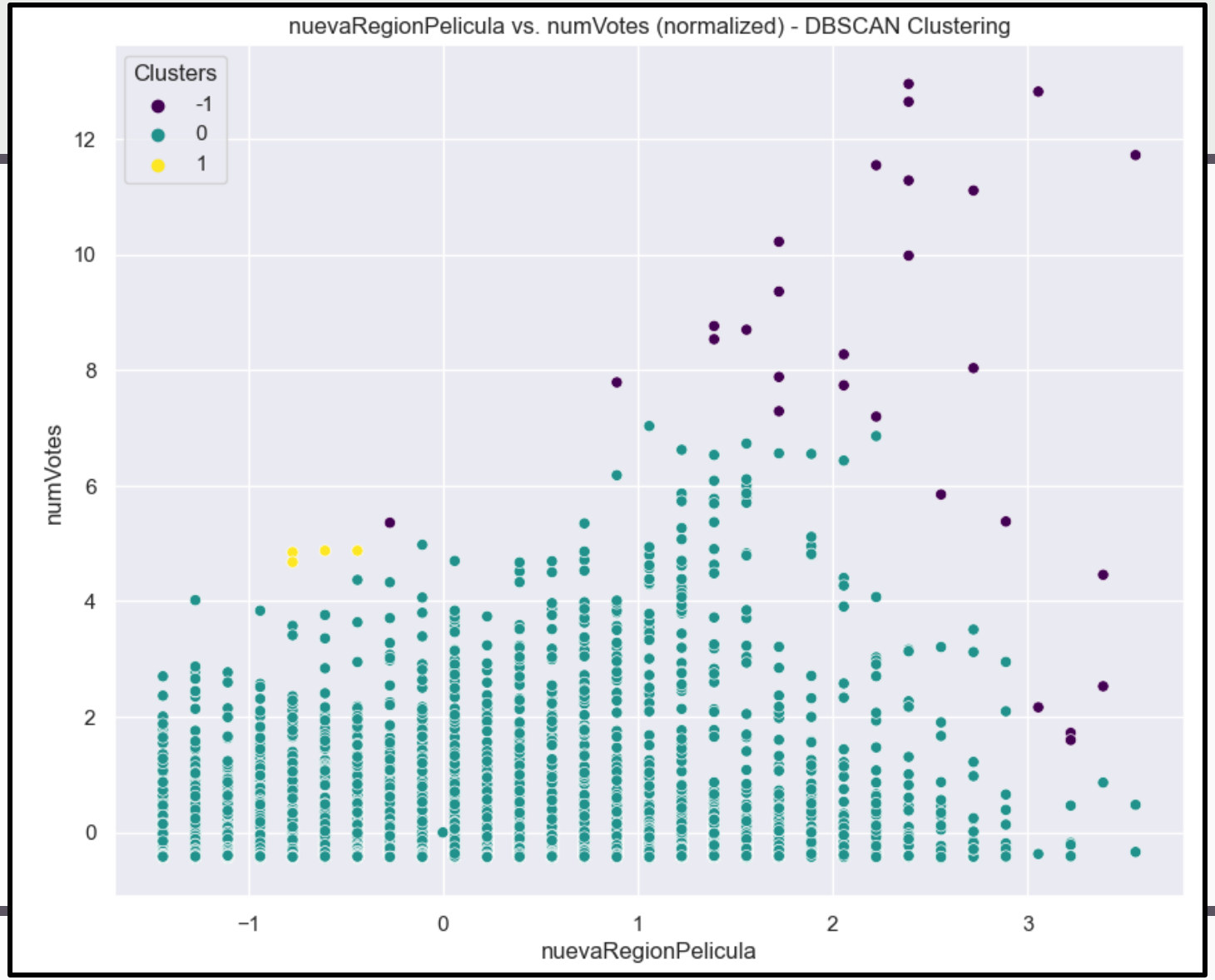
FormatoPeli VS NumVotes



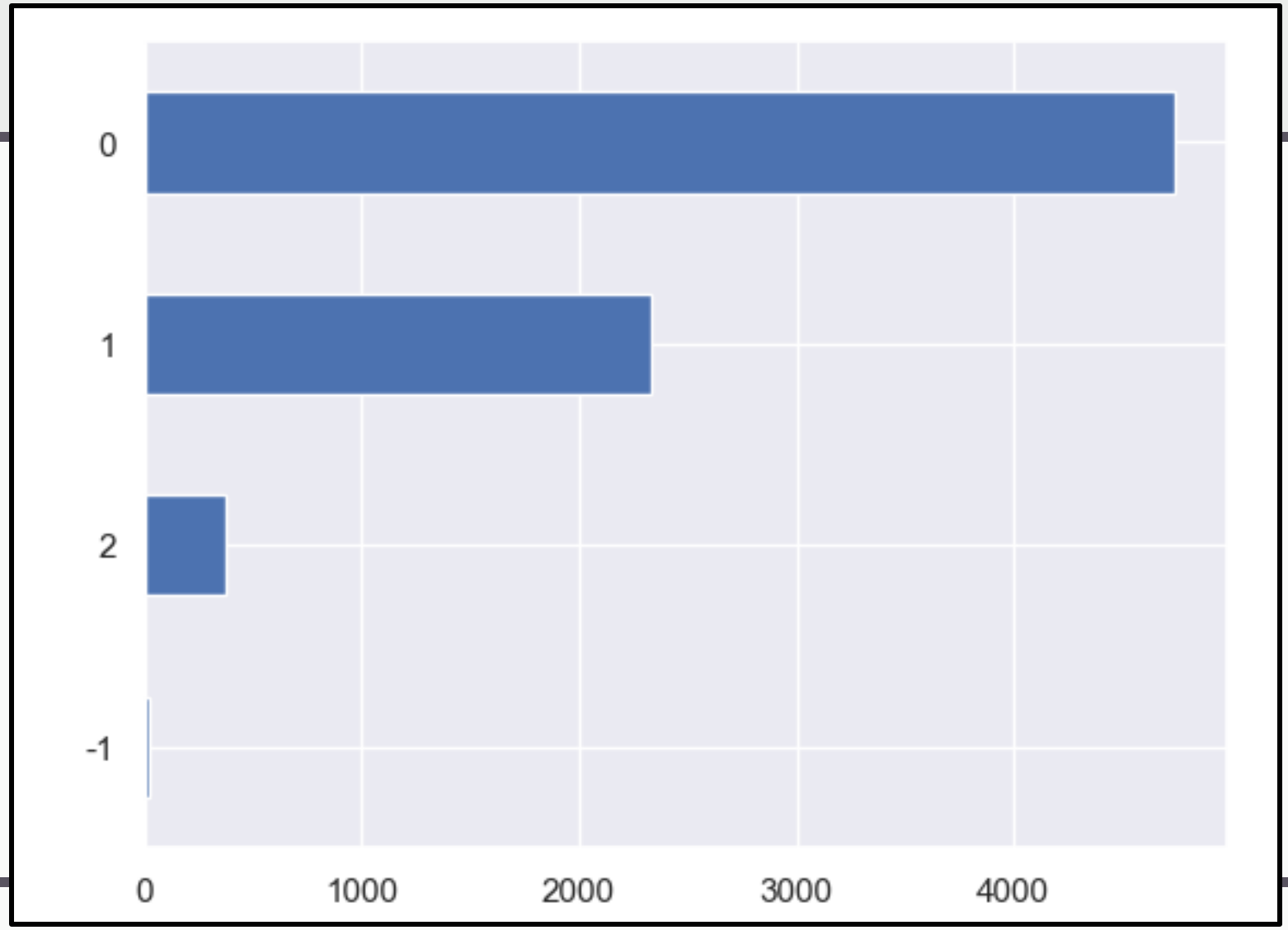
NumeroVotos VS RunTime



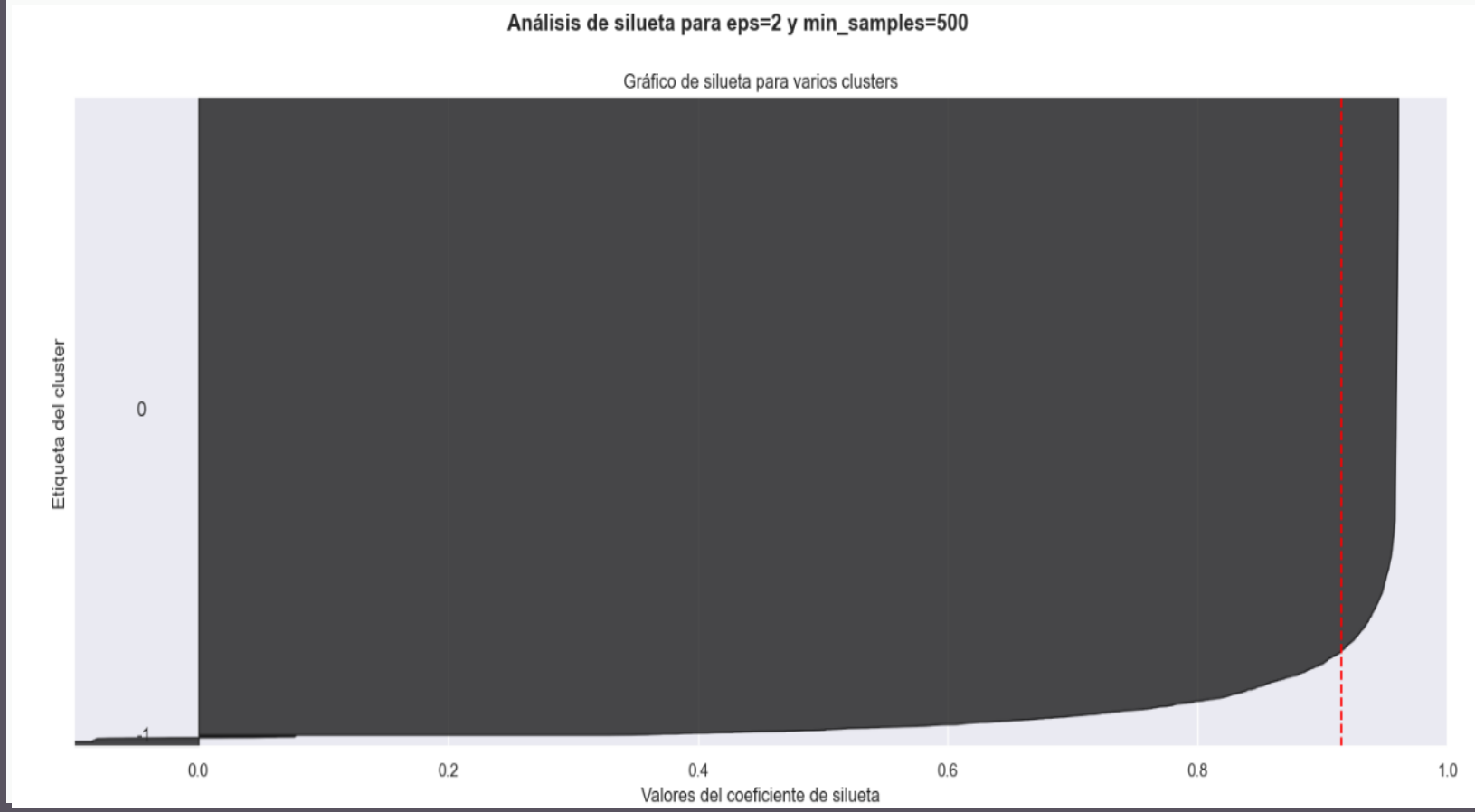
AverageRating VS NumVotes



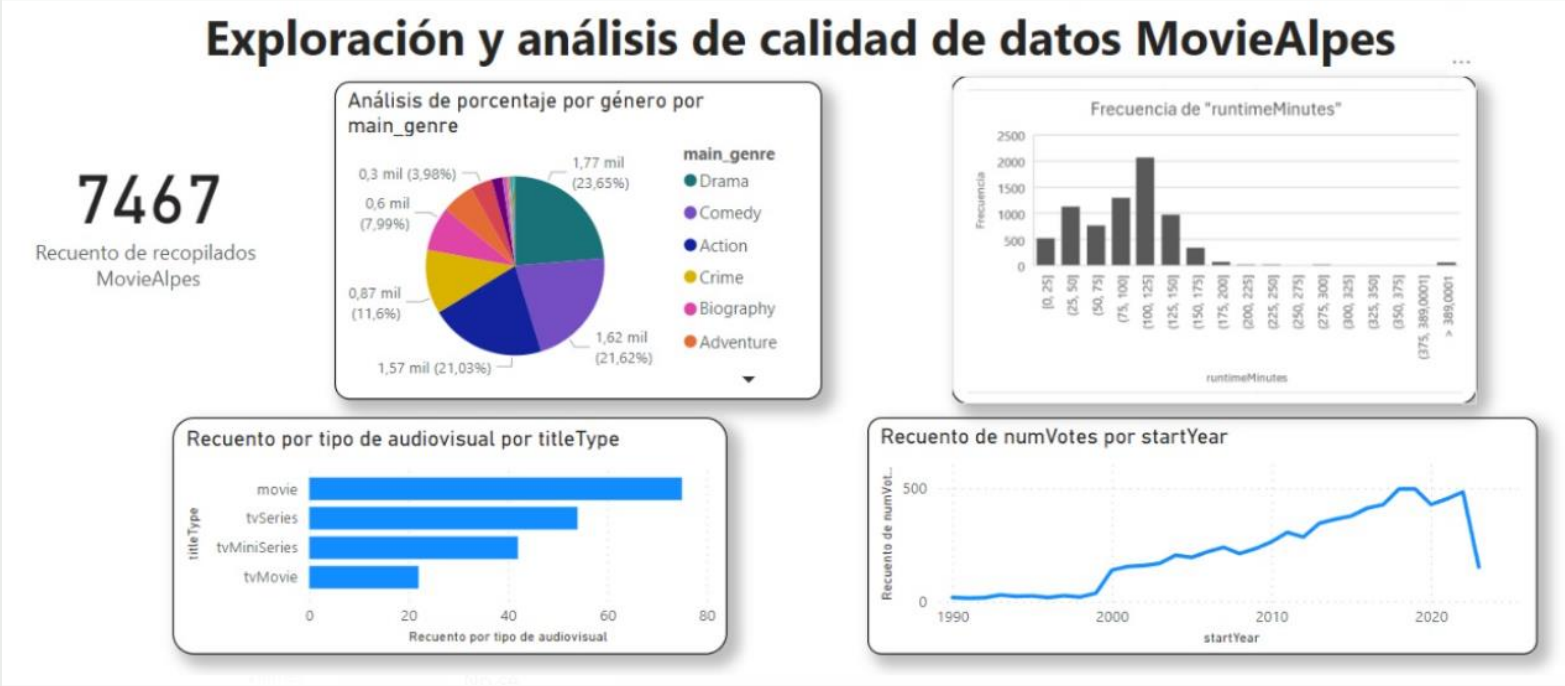
Validación cualitativa



Validación cuantitativa

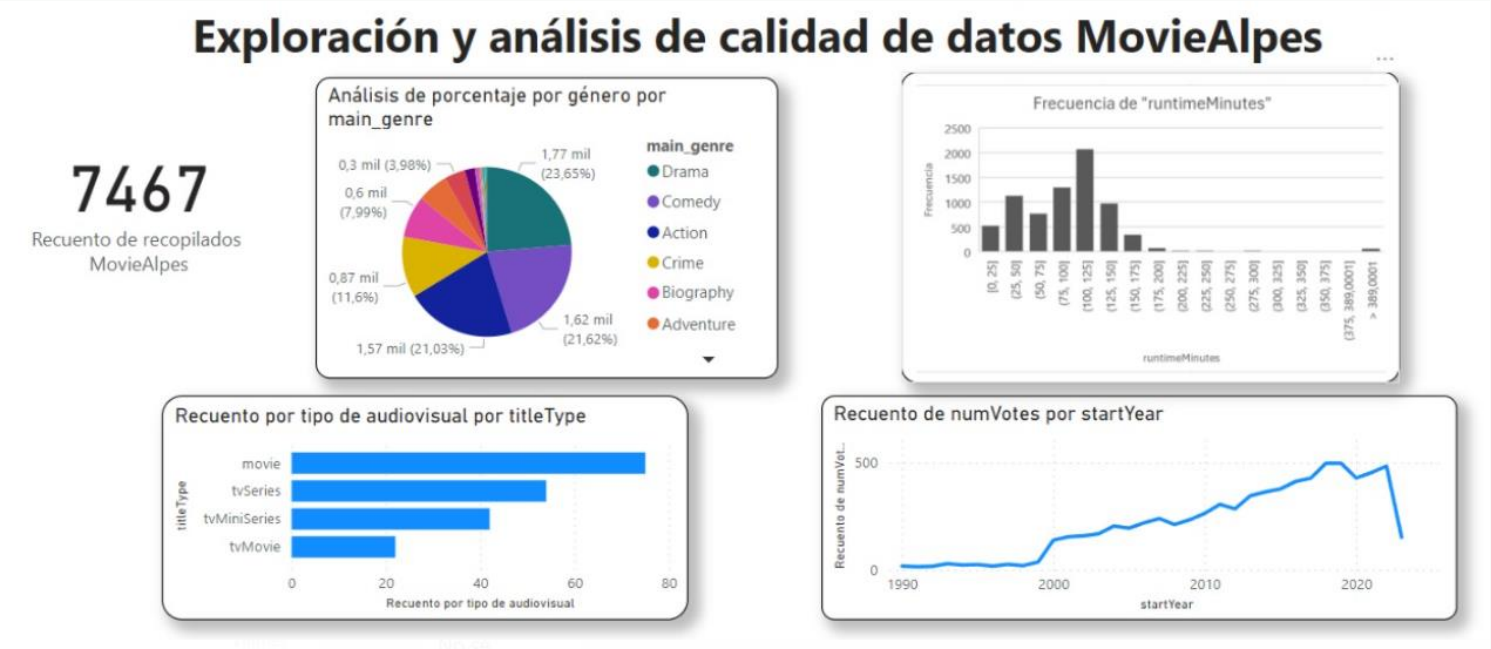


RESULTADOS / CONCLUSIONES



Resultados

- Falta variedad en cuanto a ciertos valores en los registros para lograr un análisis más variado
- Usar la variable de averageRating junto con la variable de número de votos, acompañada de otras variables, puede generar estrategias interesantes para la empresa.
- Como pilar en el número de clusteres pára los tres algoritmos se encontró que un valor entre 2 y 4 clusteres es ideal para el análisis.
- Las variables types main_genre y seondary_gnere son variables que tiene la mayor variedad en cuanto a la posibilidad de analisis para personalizacion de contenido



RECOMENDACIONES

Recomendaciones

- Se recomienda realizar publicidad o generar películas de tipo drama, con una duración entre 100 y 120 minutos.
- Realmente en este caso se recomienda que la duración, de sus películas principalmente vallan en un rango [75,150]
- Luego, enfocarse en las películas de tipo imbdDisplay que son las de mayor aceptación según los datos.
- Ahora, sería bueno hacer campañas para aumentar la visualización en los géneros de comedia y de aventura

Recomendaciones

- Se recomienda aumentar el contenido de los principales géneros, como drama, comedia y acción en formato de película.
- Incluir más contenido audiovisual entre los 100 y 125 minutos, puesto que es lo que más MovieAlpes proyecta.
- En general, MovieAlpes se desempeñade forma adecuada y mantiene calificaciones positivas, con el análisis mencionado previamente, se puede proporcionar recomendaciones más relevantes y personalizados