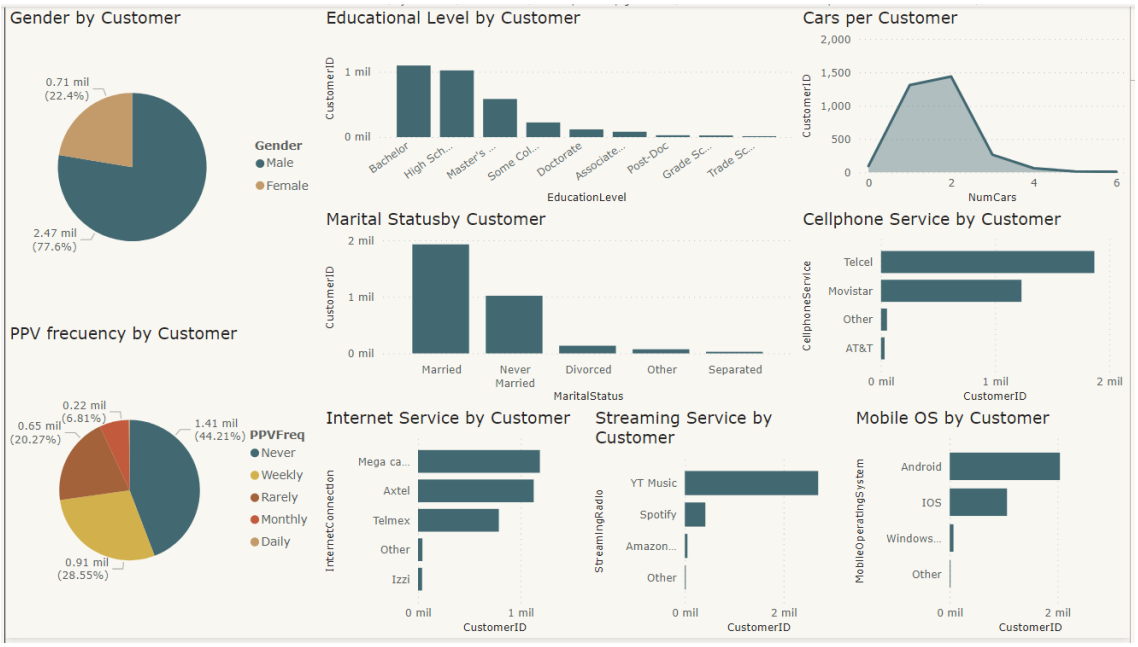
Segundo examen de BDTD

Nombre	Lilia Arceli Lobato Martínez
Carrera	Ing. Sistemas Computacionales
Expediente	IE706937
Fecha	29/11/2021

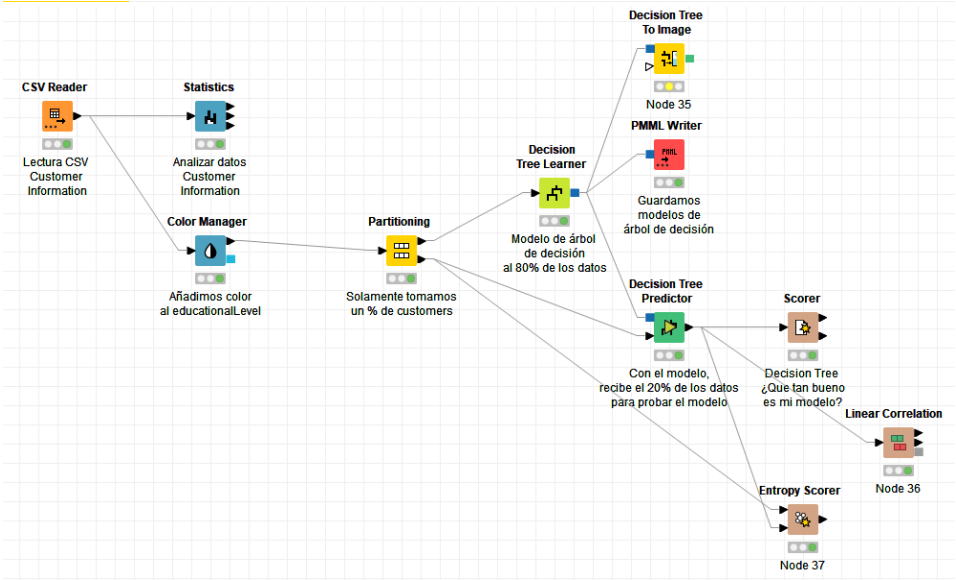
Lee todas las preguntas antes de responder el examen. En cada pregunta se indica lo que se espera como evidencia a tu respuesta. Responde en este documento, al terminar súbelo a Canvas junto con todos los archivos que utilizaste para responder. ¡Mucha suerte!

1. Utiliza el archivo **CustomerInformation.csv**, para crear un tablero que permita revisar los siguientes elementos en una sola pantalla. Evidencia esperada, una captura de pantalla con los 9 elementos **[20 puntos]**
   - *1.1. Gender by Customer*
   - *1.2. MaritalStatus by Customer*
   - *1.3. EducationalLevel by Customer*
   - *1.4. CellphoneService by Customer*
   - *1.5. InternetConnection by Customer*
   - *1.6. StreamRadio by Customer*
   - *1.7. NumCars by Customer*
   - *1.8. PPVFreq by Customer*
   - *1.9. MobileOperatingSystem by Customer*



2. Utiliza el archivo **CustomerInformation.csv**, para crear un modelo un modelo para predecir *EducationalLevel*. Se espera, **[20 puntos]**

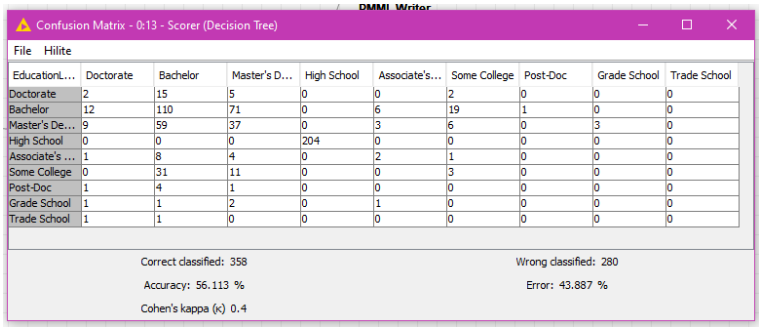
   2.1. Que describas la forma en la que funciona y cómo aseguras que tu modelo no está sobreentrenado.
   Evidencia esperada, captura del flujo de datos del modelo y tu respuesta concisa
   Tomamos el CSV con la información de los clientes, añadimos un color solamente al nivel de educación. Particionamos el dataset, usamos la partición del 80% para entrenar al árbol y la partición del 20% para revisar que tan bueno es nuestro modelo.



   2.2. Que interpretes el árbol resultante. Evidencia esperada, captura pantalla del árbol y tu respuesta concisa.
   El árbol toma como variable principal el estado marital:

2.3. Que indiques cuál es la precisión de tu modelo. Evidencia esperada, captura de pantalla de la matriz de confusión y tu respuesta concisa.

Tenemos una precisión de 56.113%



EducationL...	Doctorate	Bachelor	Master's D...	High School	Associate's...	Some College	Post-Doc	Grade School	Trade School
Doctorate	2	15	5	0	0	2	0	0	0
Bachelor	12	110	71	0	6	19	1	0	0
Master's De...	9	59	37	0	3	6	0	3	0
High School	0	0	0	204	0	0	0	0	0
Associate's ...	1	8	4	0	2	1	0	0	0
Some College	0	31	11	0	0	3	0	0	0
Post-Doc	1	4	1	0	0	0	0	0	0
Grade School	1	1	2	0	1	0	0	0	0
Trade School	1	1	0	0	0	0	0	0	0

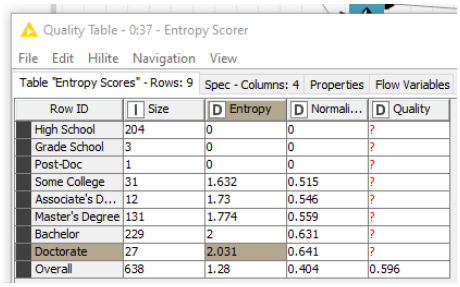
Correct classified: 358 Wrong classified: 280

Accuracy: 56.113 % Error: 43.887 %

Cohen's kappa (κ) 0.4

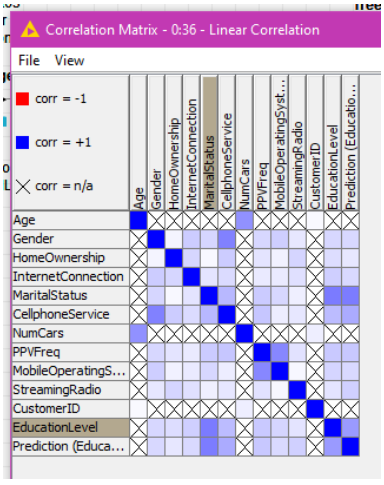
2.4. Que indiques cuál es la variable con mayor entropía. Evidencia esperada, nombre de la variable y captura de pantalla que apoye tu decisión.

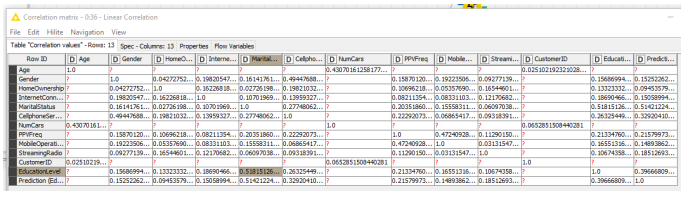
La entropía es la medida de la incertidumbre existente. La que mayor entropía tiene es *"Doctorate"*, esto lo podemos validar con un Entropy Scorer:



Row ID	Size	Entropy	Normali...	Quality
High School	204	0	0	?
Grade School	3	0	0	?
Post-Doc	1	0	0	?
Some College	31	1.632	0.515	?
Associate's D...	12	1.73	0.546	?
Master's Degree	131	1.774	0.559	?
Bachelor	229	2	0.631	?
Doctorate	27	2.031	0.641	?
Overall	638	1.28	0.404	0.596

2.5. Que utilices el nodo *Linear Correlation* y revises la opción *Correlation Measure* en busca de las dos variables con mayor *Correlation Value*. Evidencia esperada, nombre de las dos variables y captura de pantalla.

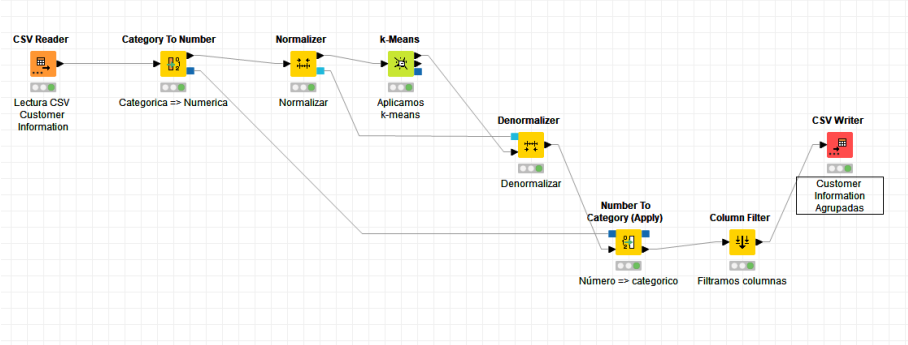


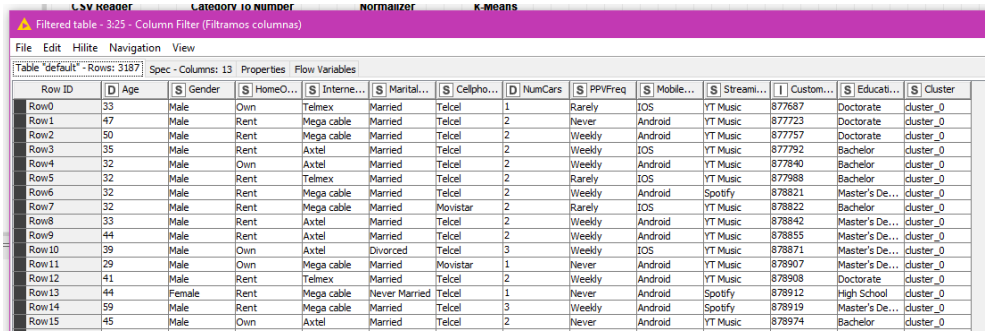


MaritalStatus con EducationLevel, tienen una índice de relación del 0.518, las demás relaciones en color morado/lila fuerte están cerca con un índice de aproximadamente 0.45

3. Utiliza el archivo **CustomerInformation.csv**, para crear cinco agrupamientos y realiza al menos lo siguiente **[20 puntos]**

   3.1. Exporta el set de datos con el agrupamiento correspondiente como un archivo CSV. Evidencia esperada, captura del flujo de datos del modelo y captura de pantalla de las primeras 5 filas con el cluster asignado a cada una.





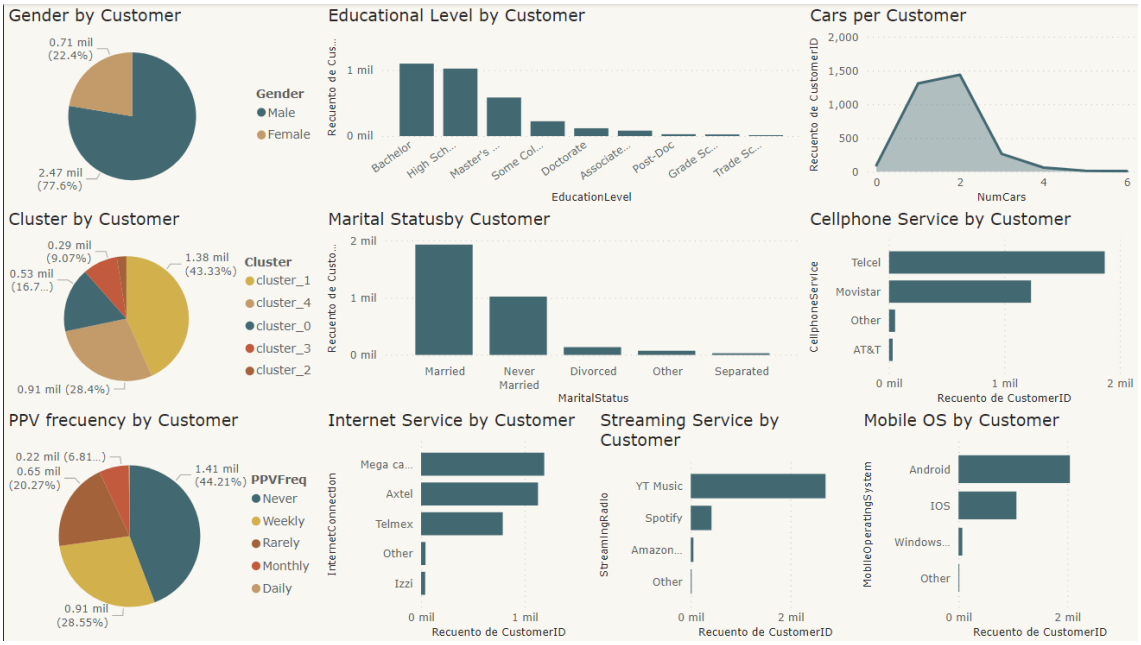
Filtered table - 3:25 - Column Filter (Filtramos columnas)

File Edit Hilite Navigation View

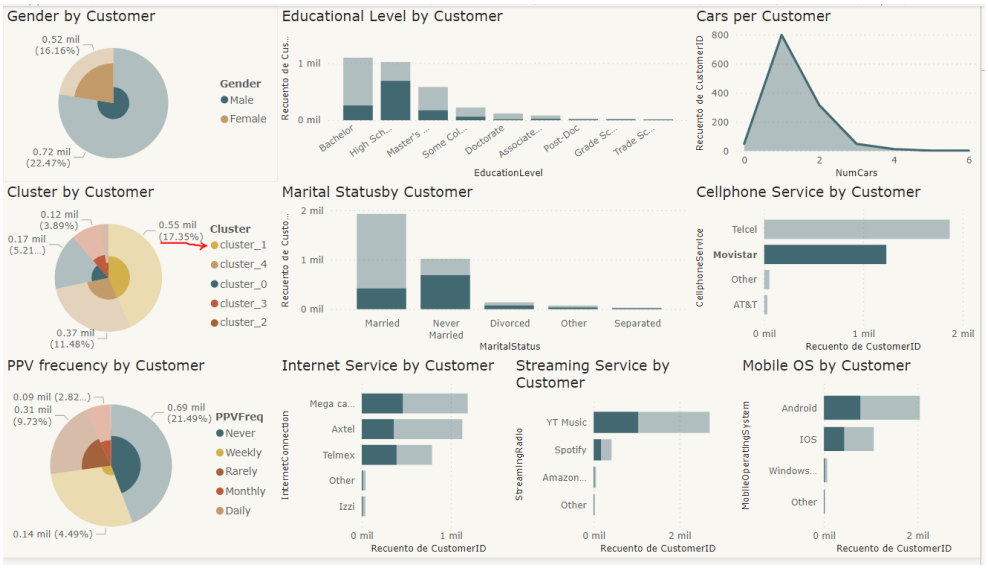
Table "default" - Rows: 3187 | Spec - Columns: 13 | Properties | Flow Variables

Row ID	Age	Gender	HomeO...	Interne...	Marital...	Cellpho...	NumCars	PPVFreq	Mobile...	Streami...	Custom...	Educati...	Cluster
Row0	33	Male	Own	Telmex	Married	Telcel	1	Rarely	IOS	YT Music	877687	Doctorate	cluster_0
Row1	47	Male	Rent	Mega cable	Married	Telcel	2	Never	Android	YT Music	877723	Doctorate	cluster_0
Row2	50	Male	Rent	Mega cable	Married	Telcel	2	Weekly	Android	YT Music	877757	Doctorate	cluster_0
Row3	35	Male	Rent	Axtel	Married	Telcel	2	Weekly	IOS	YT Music	877792	Bachelor	cluster_0
Row4	32	Male	Own	Axtel	Married	Telcel	2	Weekly	Android	YT Music	877840	Bachelor	cluster_0
Row5	32	Male	Rent	Telmex	Married	Telcel	2	Rarely	IOS	YT Music	877988	Bachelor	cluster_0
Row6	32	Male	Rent	Mega cable	Married	Telcel	2	Weekly	Android	Spotify	878821	Master's De...	cluster_0
Row7	32	Male	Rent	Mega cable	Married	Movistar	2	Rarely	IOS	YT Music	878822	Bachelor	cluster_0
Row8	33	Male	Rent	Axtel	Married	Telcel	2	Weekly	Android	YT Music	878842	Master's De...	cluster_0
Row9	44	Male	Rent	Axtel	Married	Telcel	2	Weekly	Android	YT Music	878855	Master's De...	cluster_0
Row10	39	Male	Own	Axtel	Divorced	Telcel	3	Weekly	IOS	YT Music	878871	Master's De...	cluster_0
Row11	29	Male	Own	Mega cable	Married	Movistar	1	Never	Android	YT Music	878907	Master's De...	cluster_0
Row12	41	Male	Rent	Telmex	Married	Telcel	2	Weekly	Android	YT Music	878908	Doctorate	cluster_0
Row13	44	Female	Rent	Mega cable	Never Married	Telcel	1	Never	Android	Spotify	878912	High School	cluster_0
Row14	59	Male	Rent	Mega cable	Married	Telcel	3	Weekly	Android	Spotify	878919	Master's De...	cluster_0
Row15	45	Male	Own	Axtel	Married	Telcel	2	Never	Android	YT Music	878974	Bachelor	cluster_0

   3.2. Integra la información de *Cluster by Customer* al tablero que creaste en el punto 1. Evidencia esperada, captura del tablero con la información del cluster.



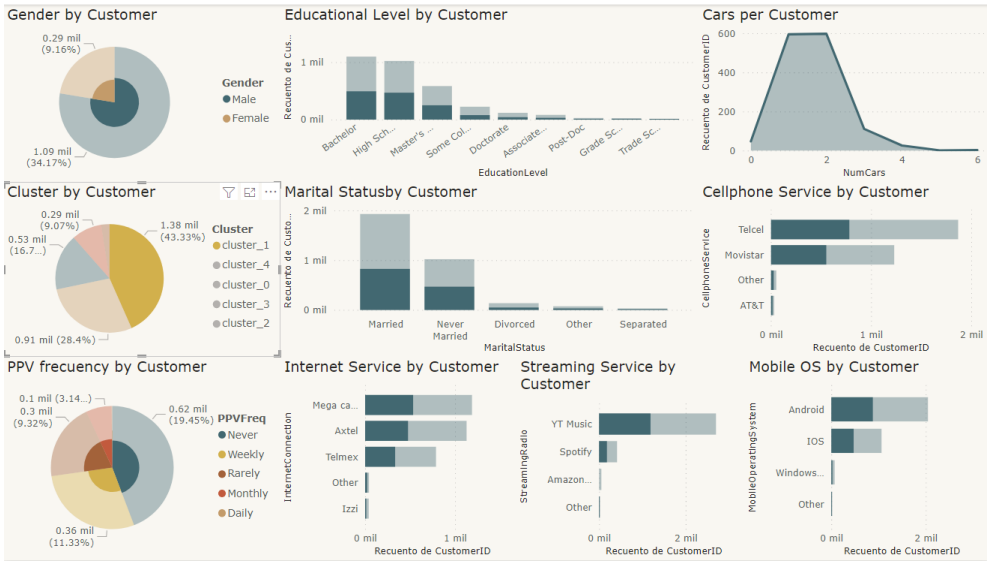
3.3. Describe la información únicamente del *cluster* que tenga más usuarios de Movistar. Evidencia esperada, el tablero con la información solo del cluster que responde a la pregunta.



El cluster con más usuarios en Movistar es el cluster. La mayoría de los usuarios son mujeres, con escolaridad media en preparatoria. La mayoría tienen 1 carro y se encuentran solteras. Su servicio de internet lo llevan con Mega Cable, la mayoría usa un celular con OS android y usan YT Music como principal reproductor musical.

Estos cluster me sirven para hacer una segmentación de clientes.

3.4. ¿Cuántos clientes tiene el cluster?, ¿cuál es el mayor EducationalLevel?, ¿Cuál es el mayor MaritalStatus?, ¿Cuál es el mayor PPVFreq? Evidencia esperada, respuesta a las preguntas y captura de pantalla de apoyo.

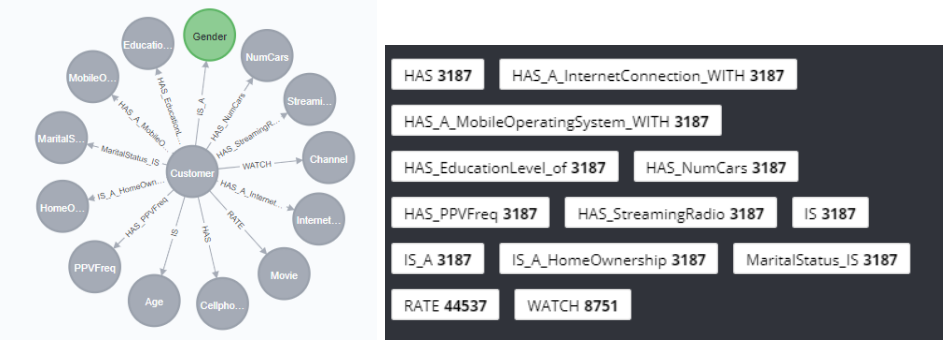


El cluster 1 tiene un total de 1382 usuarios, tienen un promedio de escolaridad entre preparatoria y universitario, la mayoría tienen entre 1 a 2 carros y se encuentran predominantemente casados.

su proveedor de celular principal es Telcel, su proveedor de internet principal es Mega Cable, usan YT Music como principal método de streaming y la mayoría cuentan con un celular android.

4. Utilizando la base de datos basada en grafos de Customer, que creaste en la última clase, realiza lo siguiente **[40 puntos]**

   4.1. Describe el grafo. Posibles evidencias, capturas de pantalla del esquema general del grafo, el grafo completo, los nodos, las relaciones, etc.



El grafo parte de un nodo Customer, de donde se puede llegar a los demás nodos.

   4.2. Crea una consulta con cypher que muestre el CustomerID, el número de películas que ha calificado y una colección con los nombres de las películas sin repetirse. Únicamente muestra los tres clientes que más películas han calificado. Evidencia esperada, código cypher legible y captura de la pantalla con el resultado.

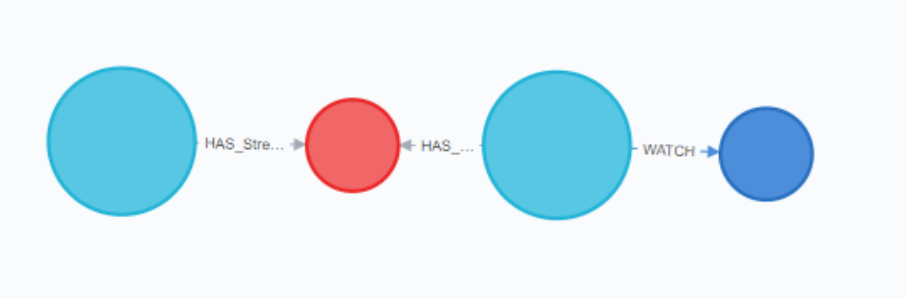
```
MATCH (n:Customer),(a:Movie)
WHERE (n)-[:RATE]->(a)
RETURN DISTINCT n.CustomerID,count(a.Movie), collect(distinct a.Movie)
ORDER BY count(a.Movie) DESC
LIMIT 3
```

"n.CustomerID"	"count(a.Movie)"	"collect(distinct a.Movie)"
"917231"	106	["9 to 5","A Fish Called Wanda","Ace Ventura Pet Detective", Flash","Ladyhawke","League of Their Own, A","Legally Blonde" ","Twister","Undercover Blues","V.I. Warshawski","Victor Vic
"885227"	104	["13th Warrior, The","Ace Ventura Pet Detective","Alien","Al ,"From Dusk Till Dawn","Gandhi","Gladiator","Glengarry Glen icide Kings","Taxi Driver","Ten Commandments, The","There's
"891480"	104	["12 Monkeys","8mm","After Hours","After Life","Almost Famou ch","Lord of the Rings: The Fellowship of the Ring, The","Ma Short Cuts","Silent Running","silkwood","Sliding Doors","Sli

MAX COLUMN WIDTH:

   4.3. Una consulta en cypher –o con bloom- que muestre la ruta más corta entre el cliente 885341 y el canal ESPN. Evidencia esperada, código cypher legible y captura de la pantalla con el resultado. En caso de usar Bloom solo la captura de pantalla.

```
MATCH (n:Customer{CustomerID:"885341"}), (m:Channel{Channel:"ESPN"}),
    path = shortestpath((n)-[*]-(m))
RETURN path
ORDER BY LENGTH(path) DESC
LIMIT 1
```



Customer[885341] -Has_Streaming-> StreamingRadio[YT Music] <-Has_Streaming- Customer[912653] -Watch-> Channel[ESPN]

4.4. Una consulta que muestre la ruta más corta entre el cliente que ha calificado más películas y el que ha calificado menos. Evidencia esperada, código cypher legible y captura de la pantalla con el resultado. En caso de usar Bloom solo la captura de pantalla.

Obtenemos el Customer que más películas ha calificado:

```
MATCH (n:Customer),(a:Movie)
WHERE (n)-[:RATE]→(a)
RETURN DISTINCT n.CustomerID,count(a.Movie)
ORDER BY count(a.Movie) DESC
LIMIT 1
```

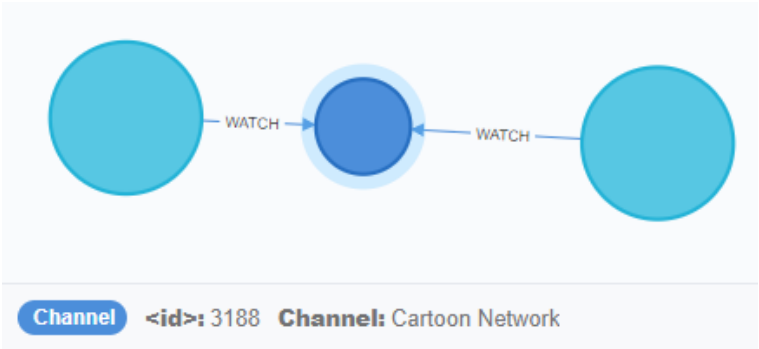
	n.CustomerID	count(a.Movie)
1	"917231"	106

Obtenemos el Customer que menos películas ha calificado:

```
MATCH (n:Customer),(a:Movie)
WHERE (n)-[:RATE]→(a)
RETURN DISTINCT n.CustomerID,count(a.Movie)
ORDER BY count(a.Movie) ASC
LIMIT 1
```

	n.CustomerID	count(a.Movie)
1	"888978"	1

Ruta más corta entre ellos se da por el canal “Cartoon Network”



4.5. Recomienda una película a los clientes que hayan visto "101 Dalmatians" y que sean casados (MaritalStatus:"Married"). Evidencia esperada, código cypher legible y captura de la pantalla con el resultado.

```
MATCH (M1:Movie{Movie:"101 Dalmatians"})<-[:RATE]-(oS1:Customer)-[:MaritalStatus_IS]->(s:MaritalStatus{MaritalStatus:"Married"})<-[:MaritalStatus_IS]-(oS2:Customer)-[:RATE]->(M2:Movie)
WHERE oS1.CustomerID <> oS2.CustomerID
RETURN M2.Movie
LIMIT 5
```

	M2.Movie
1	"Sleepless in Seattle"
2	"Star Wars Episode I: The Phantom Menace"
3	"Star Wars Episode V: Empire Strikes Back"
4	"Star Wars Episode VI: Return of the Jedi"
5	"Ben-Hur"

4.6. Recomienda una película a los clientes que ven el canal "Cartoon Network" y que EducationLevel sea "Bachelor". Evidencia esperada, código cypher legible y captura de la pantalla con el resultado.

```
MATCH (M1:Channel{Channel:"Cartoon Network"})<-[:WATCH]-(oS1:Customer)-[:HAS_EducationLevel_of]->(s:EducationLevel{EducationLevel:"Bachelor"})<-[:HAS_EducationLevel_of]-(oS2:Customer)-[:WATCH]->(M2:Channel)
WHERE oS1.CustomerID <> oS2.CustomerID
RETURN M2.Channel
LIMIT 5
```

M2.Channel
"WB"
"Other"
"CNN"
"Cartoon Network"
"Studio Universal"

4.7. Una recomendación a tu elección. Evidencia esperada, redacción de la pregunta, código cypher legible de la respuesta y captura de la pantalla con el resultado.

Justamente ayer vi “12 Angry Men”, quiero saber que peliculas me recomiendan para este fin

```
MATCH (M1:Movie{Movie:"12 Angry Men"})<-[:RATE]-(oS1:Customer)-[:MaritalStatus_IS]->(s:MaritalStatus{MaritalStatus:"Never Married"})<-[:MaritalStatus_IS]-(oS2:Customer)-[:RATE]->(M2:Movie)
WHERE oS1.CustomerID <> oS2.CustomerID
RETURN M2.Movie
LIMIT 5
```

M2.Movie
"Meet the Parents"
"Scent of a Woman"
"Fantasia"
"Whole Nine Yards, The"
"Army of Darkness"

Entregables

1. Este documento grabado con tu número de expediente e iníciales como nombre. **[100 PUNTOS MENOS SI NO SE INCLUYE O NO SE PUEDE LEER EL ARCHIVO]**
2. Archivo de KNIME, si es más de uno comprímelos en un archivo ZIP o 7Z. **[50 PUNTOS MENOS SI NO SE INCLUYE O NO SE PUEDE LEER EL ARCHIVO]**
3. Archivo de POWER BI, si es más de uno comprímelos en un archivo ZIP o 7Z. **[50 PUNTOS MENOS SI NO SE INCLUYE O NO SE PUEDE LEER EL ARCHIVO]**
4. Archivo con tus consultas de *cypher*. **[50 PUNTOS MENOS SI NO SE INCLUYE O NO SE PUEDE LEER EL ARCHIVO]**