

Análisis de datos NoSQL



INTRODUCCIÓN

Este trabajo consiste en elegir un Dataset, incorporarlo a MongoDB y hacer un análisis del mismo, realizando las consultas necesarias para ello. El análisis ha sido realizado en la aplicación MongoDB Compass.

Presentación y estructura del Dataset

Para la realización del análisis he escogido un Dataset que contiene las **aplicaciones disponibles en Google Play Store y las características de cada una**, obtenidas a través de la realización de *web scraping* en la Google Play Store por el usuario que publicó este Dataset.

El archivo descargado fue publicado en 2019 por el usuario *LAVANYA* en los Datasets disponibles en Kaggle¹

El Dataset utilizado contiene 10.841 observaciones de 13 variables. Considerar la limitación de que todos los datos son los que había en el momento en el que se realizó el *web scraping* en 2019, ya que no se ha vuelto a actualizar.

Dimensiones o variables:

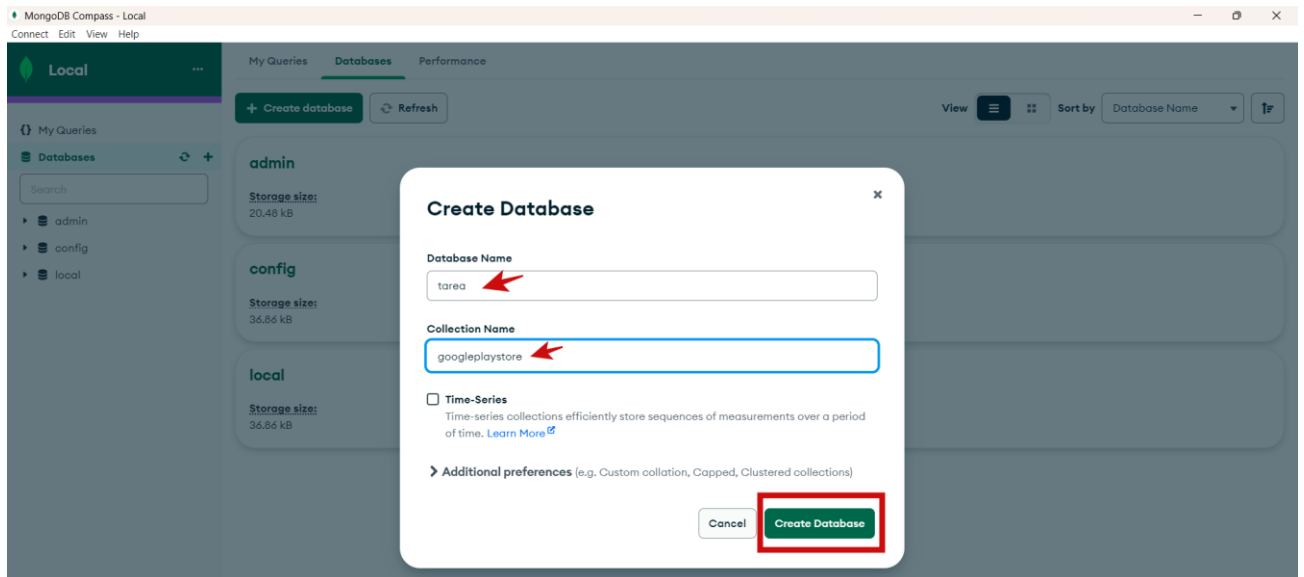
- App: Nombre de la aplicación.
- Category: Categoría a la que pertenece la aplicación.
- Rating: Puntuación de los clientes a la aplicación.
- Reviews: Número de valoraciones de la aplicación por parte de los usuarios.
- Size: Tamaño de la aplicación.
- Installs: Número de descargas de la aplicación por parte de los usuarios.
- Type: Si la aplicación es de pago o gratuita.
- Price: Precio de la aplicación (en \$)

¹ <https://www.kaggle.com/datasets/lava18/google-play-store-apps?select=googleplaystore.csv>

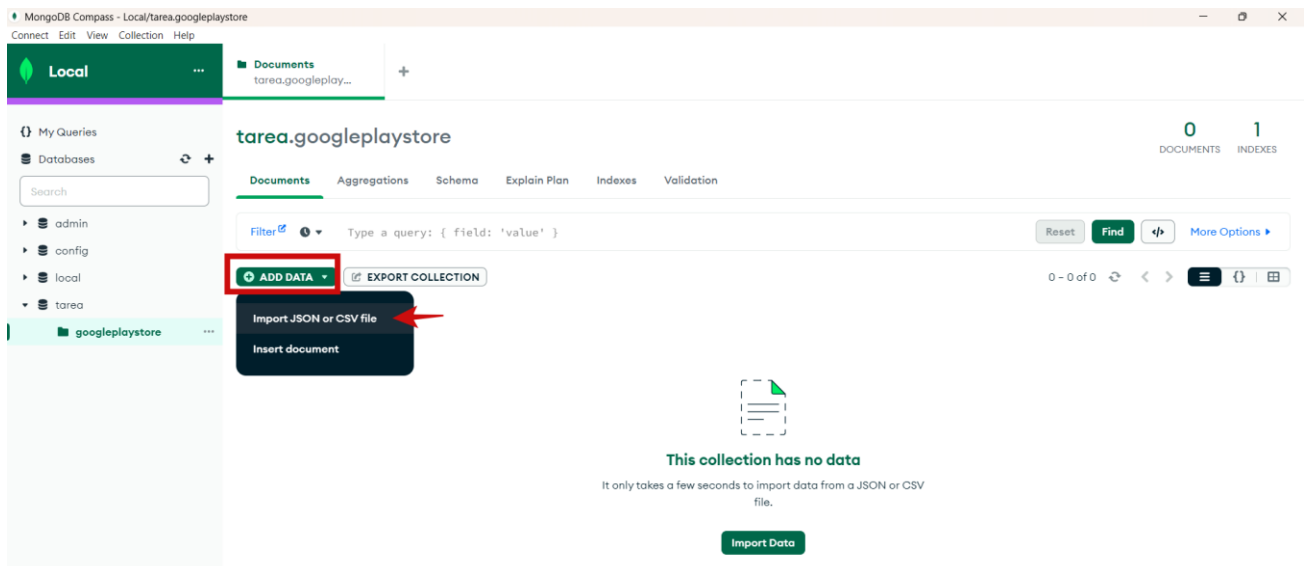
- Content Rating: Grupo de edad objetivo (*target*) de la aplicación.
- Genres: Una aplicación puede pertenecer a varios géneros (aparte de su categoría principal). Por ejemplo, un juego musical familiar pertenecerá a los géneros Music, Game, Family.
- Last Updated: Fecha de la última actualización de la aplicación en Play Store.
- Current Ver: Versión actual de la aplicación en Play Store.
- Android Ver: Versión de Android mínima requerida para la aplicación.

1. CARGAR/IMPORTAR DATASET

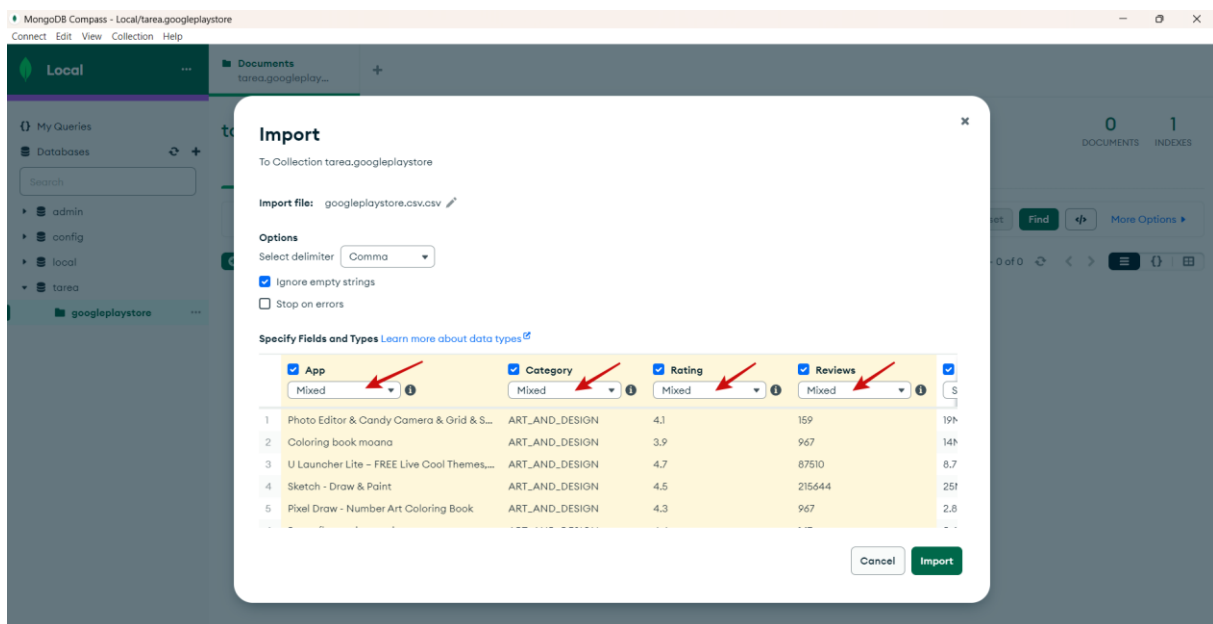
En primer lugar, creamos la *Database* y la *Collection*:



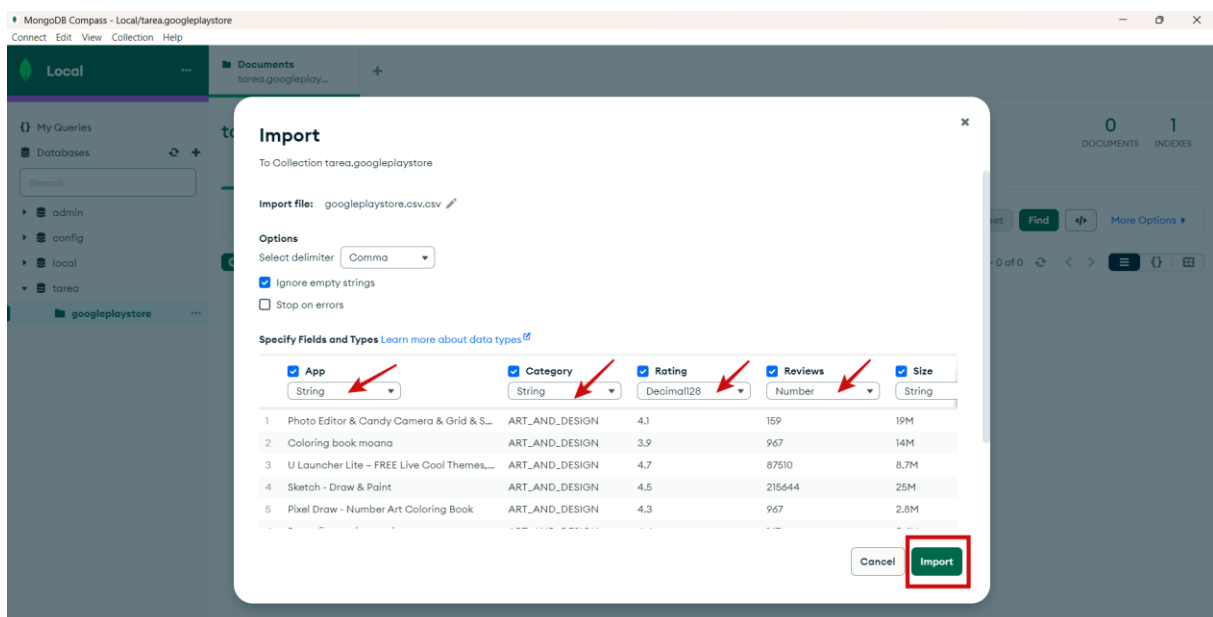
Dentro de esta ruta, importamos el csv o json del *dataset* que vamos a utilizar (en nuestro caso un csv):



Nos aparece una pantalla donde debemos establecer el delimitador (coma) y los tipos de campos que tenemos:

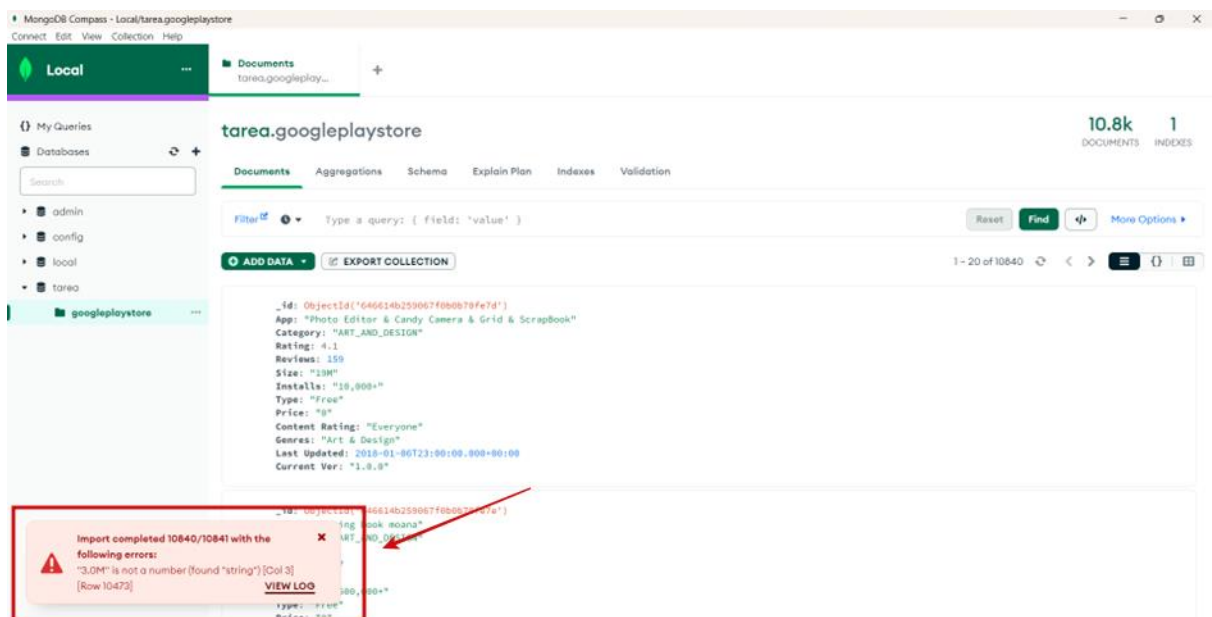


Los tipos de campos están incorrectos, por lo que los cambiamos por los correctos para poder realizar luego el análisis y finalmente importamos el *dataset*:

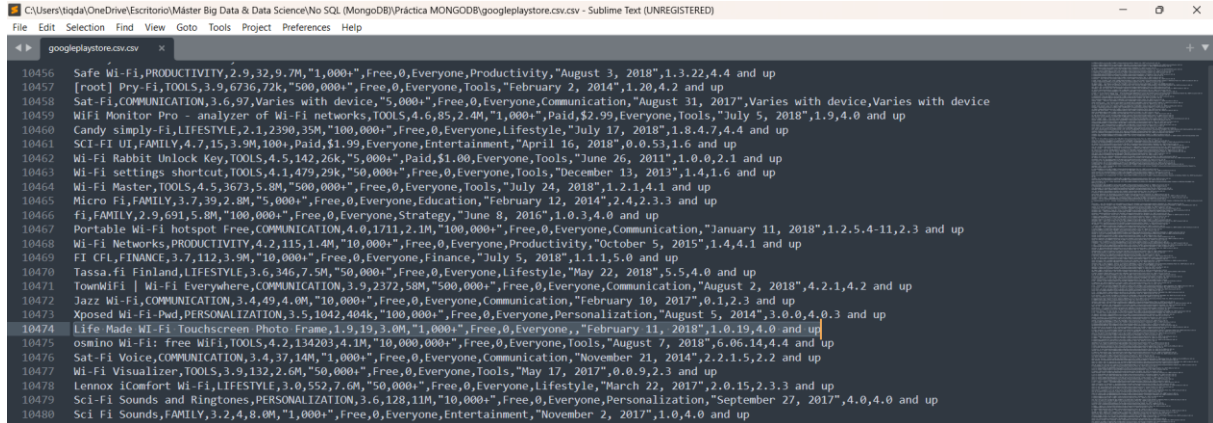


NOTA: hay una limitación porque en varios campos del *dataset* no se ha podido poner el tipo de dato correcto debido a que en la escritura se han mezclado números y letras, como se puede ver por ejemplo, en el campo “Size” o “Installs” que ponen 3.0M, o el campo “Price” que ponen \$4.99, por lo que MongoDB no los reconoce como números.

(1) 646614b459067f0b0b7128d4 { Category : "LIFESTYLE", Type : "Free" } (13 fields)			Document
_id	646614b459067f0b0b7128d4		ObjectId
App	iHoroscope - 2018 Daily Horoscope & Astrology		String
Category	LIFESTYLE		String
Rating	4.5		Decimal
Reviews	398.307 (0.40M)		Int32
Size	19M		String
Installs	10,000,000+		String
Type	Free		String
Price	0		String
Content Rating	Everyone		String
Genres	Lifestyle		String
Last Updated	25/7/2018 0:00:00 - 5 years ago		Date
Current Ver	Varies with device		String



Podemos observar que **se han importado exitosamente 10840 de 10841 documentos**, debido a que un registro del csv está mal registrado y **no encajan los tipos de campo con los datos introducidos en ese registro**:



```

10456 Safe Wi-Fi,PRODUCTIVITY,2.9,32,9.7M,"1,000+",Free,0,Everyone,Productivity,"August 3, 2018",1.3.22,4.4 and up
10457 [root] Pry-Fi,TOOLS,3.9,6736,72k,"500,000+",Free,0,Everyone,Tools,"February 2, 2014",1.20,4.2 and up
10458 Sat-Fi,COMMUNICATION,3.6,97,Varies with device,"5,000+",Free,0,Everyone,Communication,"August 31, 2017",Varies with device,Varies with device
10459 WiFi Monitor Pro - analyzer of Wi-Fi networks,TOOLS,4.6,85,2.4M,"1,000+",Paid,$2.99,Everyone,Tools,"July 5, 2018",1.9,4.0 and up
10460 Candy simply-Fi,LIFESTYLE,2.1,2390,35M,"100,000+",Free,0,Everyone,Lifestyle,"July 17, 2018",1.8.4.7,4.4 and up
10461 Sci-Fi UI,FAMILY,4.2,15,3.9M,100k,"Paid,$1.99,Everyone,Entertainment,"April 16, 2018",0.0.53,1.6 and up
10462 Wi-Fi Rabbit Unlock Key,TOOLS,4.5,142,26k,"5,000+",Paid,$1.00,Everyone,Tools,"June 26, 2011",1.0.0,2.1 and up
10463 Wi-Fi settings shortcut,TOOLS,4.1,479,29k,"50,000+",Free,0,Everyone,Tools,"December 13, 2013",1.4,1.6 and up
10464 Wi-Fi Master,TOOLS,4.5,3673,5.8M,"500,000+",Free,0,Everyone,Tools,"July 24, 2018",1.2.1,4.1 and up
10465 Micro Fi,FAMILY,3.7,39,2.8M,"5,000+",Free,0,Everyone,Education,"February 12, 2014",2.4,2.3.3 and up
10466 fi,FAMILY,2.9,691,5.8M,"100,000+",Free,0,Everyone,Strategy,"June 8, 2016",1.0.3,4.0 and up
10467 Portable Wi-Fi hotspot Free,COMMUNICATION,4.0,1711,2.1M,"100,000+",Free,0,Everyone,Communication,"January 11, 2018",1.2.5.4-11,2.3 and up
10468 Wi-Fi Networks,PRODUCTIVITY,4.2,115,1.4M,"10,000+",Free,0,Everyone,Productivity,"October 5, 2015",1.4,4.1 and up
10469 Fi CPL,FINANCE,3.7,112,3.9M,"10,000+",Free,0,Everyone,Finance,"July 5, 2018",1.1.1,5.0 and up
10470 Tassa.fi Finland,LIFESTYLE,3.6,346,7.5M,"50,000+",Free,0,Everyone,Lifestyle,"May 22, 2018",5.5,4.0 and up
10471 TownWiFi | Wi-Fi Everywhere,COMMUNICATION,3.9,2372,58M,"500,000+",Free,0,Everyone,Communication,"August 2, 2018",4.2.1,4.2 and up
10472 Jazz Wi-Fi,COMMUNICATION,3.4,49,4.0M,"10,000+",Free,0,Everyone,Communication,"February 10, 2017",0.1,2.3 and up
10473 Xposed Wi-Fi-Pad,PERSONALIZATION,3.5,1042,404k,"100,000+",Free,0,Everyone,Personalization,"August 5, 2014",3.0.0,4.0.3 and up
10474 Life Made Wi-Fi Touchscreen Photo Frame,1.9,19,3.0M,"1,000+",Free,0,Everyone,"February 11, 2018",1.0.19,4.0 and up
10475 osmino Wi-Fi: free WiFi,TOOLS,4.2,134203,4.1M,"10,000,000+",Free,0,Everyone,Tools,"August 7, 2018",6.06.14,4.4 and up
10476 Sat-Fi Voice,COMMUNICATION,3.4,37,14M,"1,000+",Free,0,Everyone,Communication,"November 21, 2014",2.2.1.5,2.2 and up
10477 Wi-Fi Visualizer,TOOLS,3.9,132,2.6M,"50,000+",Free,0,Everyone,Tools,"May 17, 2017",0.0.9,2.3 and up
10478 Lennox iComfort Wi-Fi,LIFESTYLE,3.0,552,7.6M,"50,000+",Free,0,Everyone,Lifestyle,"March 22, 2017",2.0.15,2.3.3 and up
10479 Sci-Fi Sounds and Ringtones,PERSONALIZATION,3.6,128,11M,"10,000+",Free,0,Everyone,Personalization,"September 27, 2017",4.0,4.0 and up
10480 Sci-Fi Sounds,FAMILY,3.2,4,8.0M,"1,000+",Free,0,Everyone,Entertainment,"November 2, 2017",1.0,4.0 and up

```

Vemos que en ese registro no se introdujo nada en el campo categoría, tampoco un “null”, por lo que como categoría reconoce el dato que tiene el campo de la derecha (ratings) y así sucesivamente.

Posteriormente en la parte de *queries* añadiremos este registro a mano para que no se pierda esa información y esté correcto.

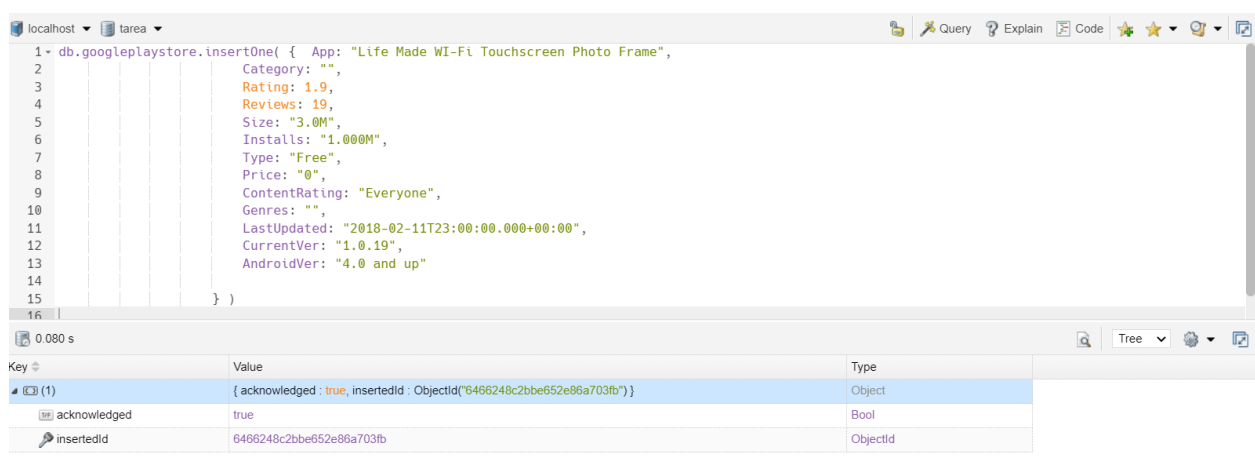
2. QUERIES

NOTA: Consultas realizadas en NoSQLBooster for MongoDB.

FASE 1: EXPLORACIÓN Y LIMPIEZA DEL DATASET

2.1. En primer lugar, añadimos manualmente el registro que no se ha podido importar porque los datos no estaban correctamente introducidos:

```
db.googleplaystore.insertOne(  
  { App: "Life Made WI-Fi Touchscreen Photo Frame",  
    Category: "",  
    Rating: 1.9,  
    Reviews: 19,  
    Size: "3.0M",  
    Installs: "1.000M",  
    Type: "Free",  
    Price: "0",  
    ContentRating: "Everyone",  
    Genres: "",  
    LastUpdated: "2018-02-11T23:00:00.000+00:00",  
    CurrentVer: "1.0.19",  
    AndroidVer: "4.0 and up"  
  } )
```

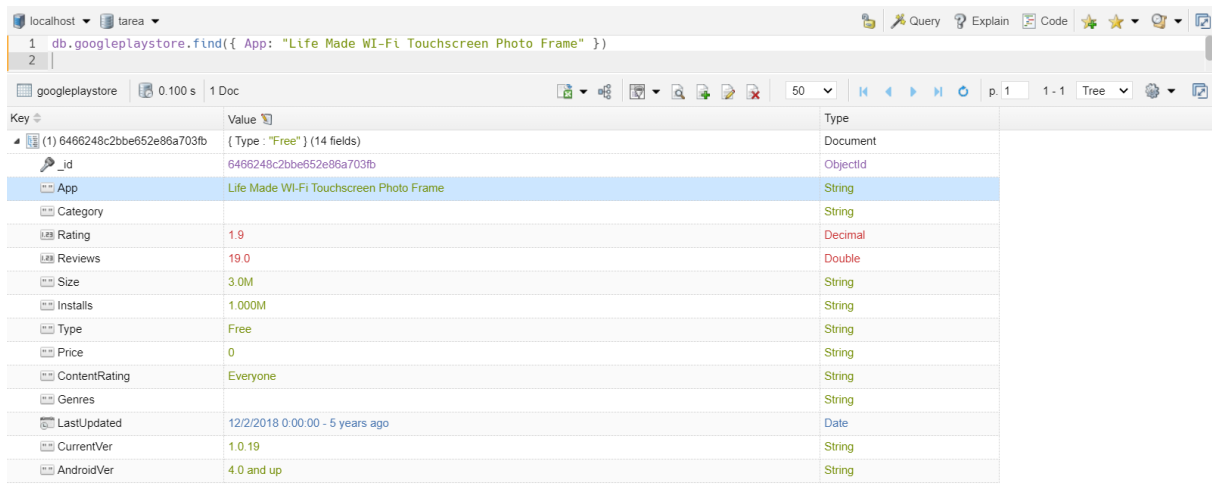


The screenshot shows the NoSQLBooster for MongoDB interface. The top panel displays the MongoDB command used to insert a document into the 'googleplaystore' collection. The bottom panel shows the result of the operation, indicating that the document was successfully inserted with an 'insertedId' of '6486248c2bbe652e86a703fb'.

Key	Value	Type
(1)	{ acknowledged: true, insertedId: ObjectId("6486248c2bbe652e86a703fb") }	Object
acknowledged	true	Bool
insertedId	6486248c2bbe652e86a703fb	ObjectId

Comprobamos que se ha registrado bien:

```
db.googleplaystore.find({ App: "Life Made WI-Fi Touchscreen Photo  
Frame" })
```



Key	Value	Type
(1) 6466248c2bbe652e86a703fb	{ Type: "Free" } (14 fields)	Document
_id	6466248c2bbe652e86a703fb	ObjectId
App	Life Made WI-Fi Touchscreen Photo Frame	String
Category		String
Rating	1.9	Decimal
Reviews	19.0	Double
Size	3.0M	String
Installs	1.000M	String
Type	Free	String
Price	0	String
ContentRating	Everyone	String
Genres		String
LastUpdated	12/2/2018 0:00:00 - 5 years ago	Date
CurrentVer	1.0.19	String
AndroidVer	4.0 and up	String

2.2. Ahora contamos cuántos **registros tenemos en total en la base de datos de Google Play Store.**

```
db.googleplaystore.find().count()
```



Key	Value	Type
1	10841	Number

Obtenemos que hay un total de 10841 registros en la base de datos de Google Play Store.

2.3. En segundo lugar, contamos el número de registros del campo “App” no repetidos para saber el **total de aplicaciones** que hay en la Google Play Store.

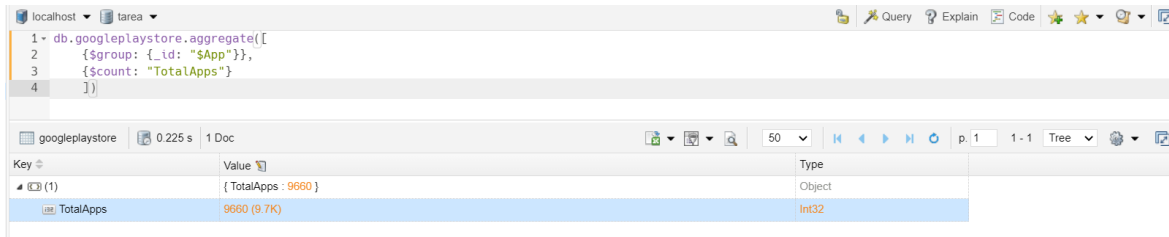
```
db.googleplaystore.distinct("App").length
```



Key	Value	Type
1	9660	Number

También lo podemos obtener de la siguiente manera:

```
db.googleplaystore.aggregate([
  {$group: {_id: "$App"}},
  {$count: "TotalApps"}
])
```

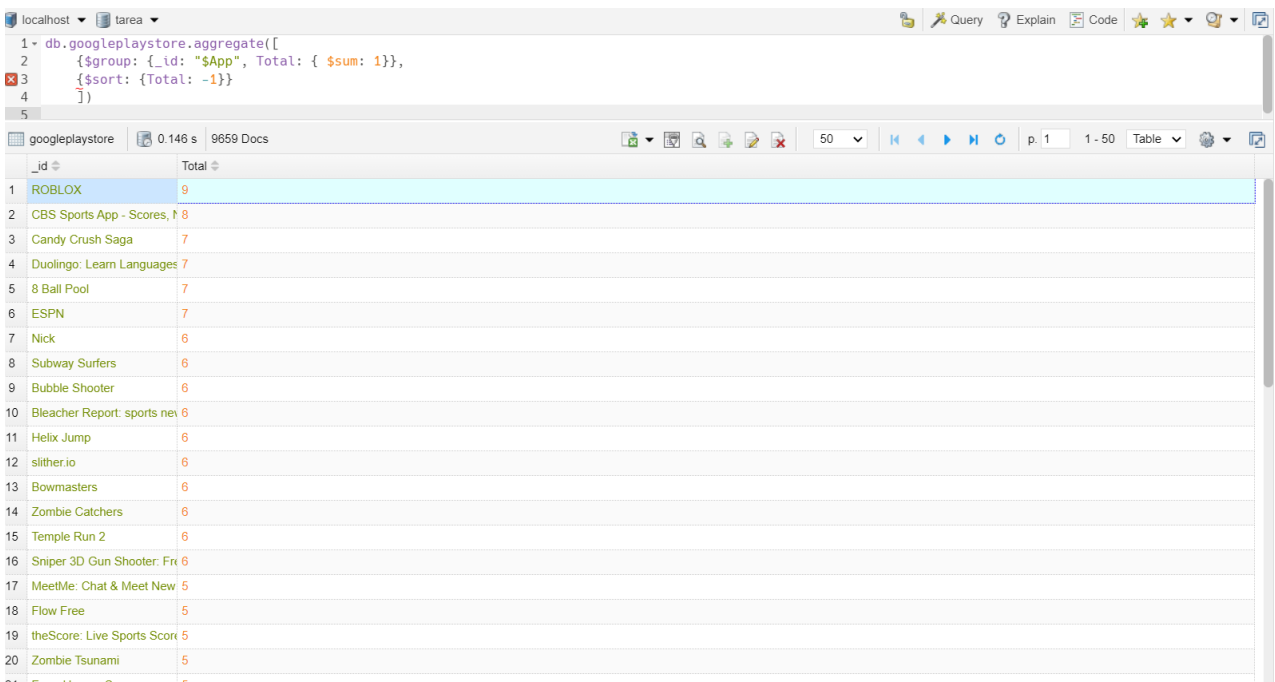
```
1 db.googleplaystore.aggregate([
2   {$group: {_id: "$App"}},
3   {$count: "TotalApps"}
4 ])
```

Key	Value	Type
TotalApps	9660 (9.7K)	Int32

Obtenemos que había un total de 9660 aplicaciones diferentes en la Google Play Store, es decir, **hay varios registros sobre la misma aplicación** en la base de datos.

Podemos ver **cuántas veces se repiten las aplicaciones** a continuación:

```
db.googleplaystore.aggregate([
  {$group: {_id: "$App", Total: { $sum: 1 }}},
  {$sort: {Total: -1}}
])
```

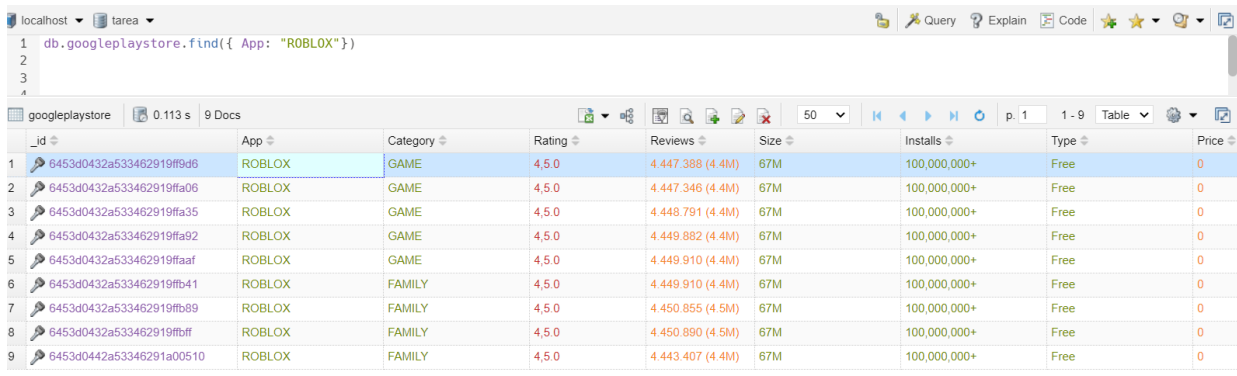


```
1 db.googleplaystore.aggregate([
2   {$group: {_id: "$App", Total: { $sum: 1 }}},
3   {$sort: {Total: -1}}
4 ])
```

_id	Total
ROBLOX	9
CBS Sports App - Scores, News	8
Candy Crush Saga	7
Duolingo: Learn Languages	7
8 Ball Pool	7
ESPN	7
Nick	6
Subway Surfers	6
Bubble Shooter	6
Bleacher Report: sports news	6
Helix Jump	6
slither.io	6
Bowmasters	6
Zombie Catchers	6
Temple Run 2	6
Sniper 3D Gun Shooter: Free	6
MeetMe: Chat & Meet New	5
Flow Free	5
theScore: Live Sports Scores	5
Zombie Tsunami	5
Farm Heroes Saga	5

Comprobamos si todos los registros repetidos de cada aplicación son registros con la misma información, por ejemplo, en la aplicación que más se repite (9 veces) “ROBLOX”:

```
db.googleplaystore.find({ App: "ROBLOX" })
```



_id	App	Category	Rating	Reviews	Size	Installs	Type	Price
6453d0432a533462919ff9d6	ROBLOX	GAME	4,5,0	4.447.388 (4.4M)	67M	100,000,000+	Free	0
6453d0432a533462919ffa06	ROBLOX	GAME	4,5,0	4.447.346 (4.4M)	67M	100,000,000+	Free	0
6453d0432a533462919ffa35	ROBLOX	GAME	4,5,0	4.448.791 (4.4M)	67M	100,000,000+	Free	0
6453d0432a533462919ffa92	ROBLOX	GAME	4,5,0	4.449.882 (4.4M)	67M	100,000,000+	Free	0
6453d0432a533462919ffaaf	ROBLOX	GAME	4,5,0	4.449.910 (4.4M)	67M	100,000,000+	Free	0
6453d0432a533462919ffb41	ROBLOX	FAMILY	4,5,0	4.449.910 (4.4M)	67M	100,000,000+	Free	0
6453d0432a533462919ffb89	ROBLOX	FAMILY	4,5,0	4.450.855 (4.5M)	67M	100,000,000+	Free	0
6453d0432a533462919ffbff	ROBLOX	FAMILY	4,5,0	4.450.890 (4.5M)	67M	100,000,000+	Free	0
6453d0442a53346291a00510	ROBLOX	FAMILY	4,5,0	4.443.407 (4.4M)	67M	100,000,000+	Free	0

Apreciamos que **todos los registros tienen prácticamente la misma información**, no está más actualizada en unos o en otros, tampoco tenemos un campo “fecha” que nos muestre cuál de todos es el más actualizado, además en el contexto y estructura del *dataset* se deja claro que el *web scraping* se hizo en 2019 y no se volvió a realizar, asique simplemente **nos quedaremos con un registro por aplicación** para poder realizar el análisis de los datos sin interferencias.

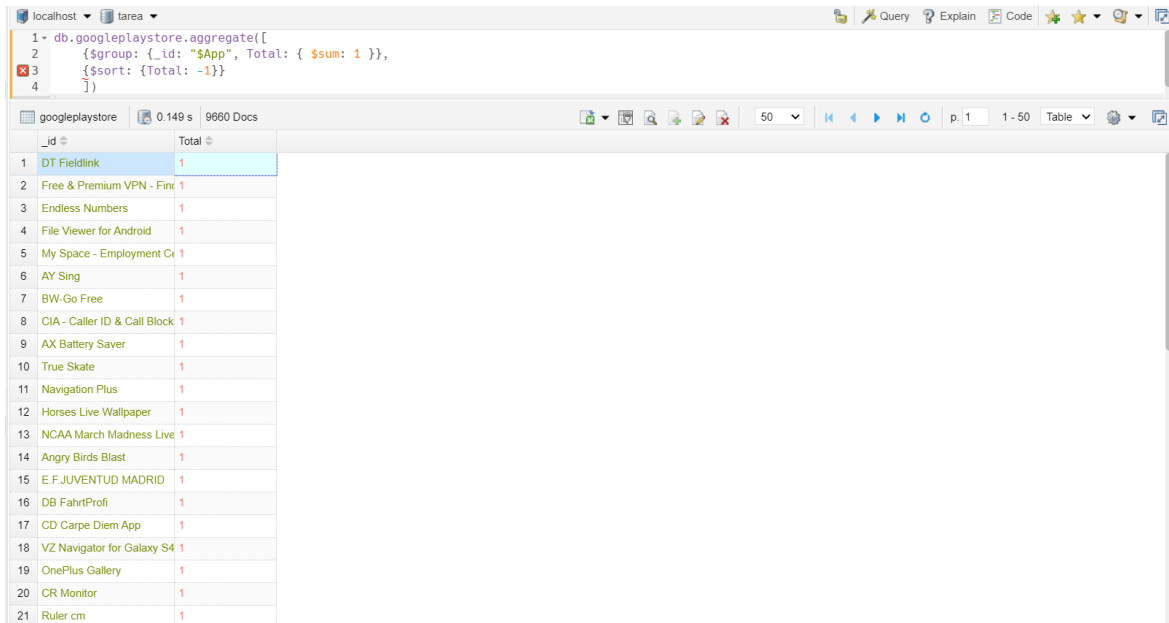
2.4. Con todo lo dicho anteriormente, procedemos a realizar la eliminación de los registros duplicados del *dataset*:

```
db.googleplaystore.aggregate([
  { "$group": {
    "_id": { "App": "$App" },
    "dups": { "$push": "$_id" },
    "count": { "$sum": 1 }
  } },
  { "$match": { "count": { "$gt": 1 } } }
]).forEach(function(doc) {
  doc.dups.shift();
  db.googleplaystore.remove({ "_id": {"$in": doc.dups } });
});
```



Comprobamos que se han eliminado correctamente los duplicados:

```
db.googleplaystore.aggregate([
  {$group: {_id: "$App", Total: { $sum: 1 } }},
  {$sort: {Total: -1}}
])
```



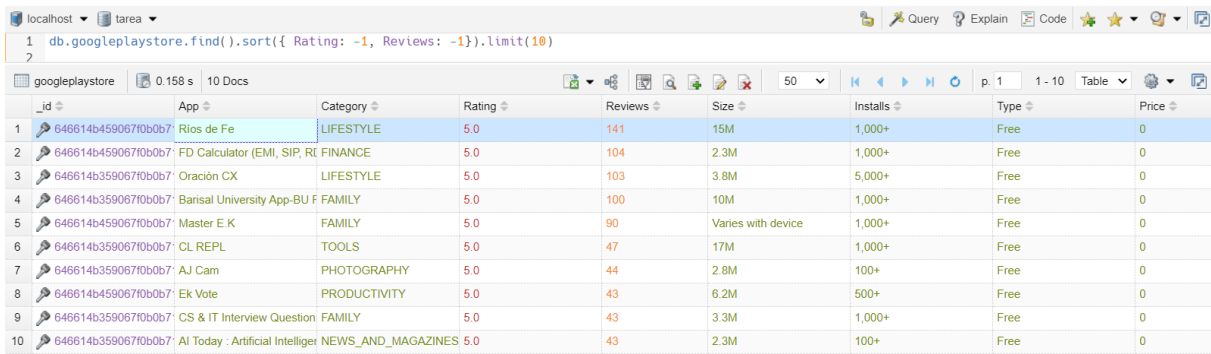
_id	Total
DT Fieldlink	1
Free & Premium VPN - Fin...	1
Endless Numbers	1
File Viewer for Android	1
My Space - Employment Co...	1
AY Sing	1
BW-Go Free	1
CIA - Caller ID & Call Block	1
AX Battery Saver	1
True Skate	1
Navigation Plus	1
Horses Live Wallpaper	1
NCAA March Madness Live	1
Angry Birds Blast	1
E.F.JUVENTUD MADRID	1
DB FahrtProfi	1
CD Carpe Diem App	1
VZ Navigator for Galaxy S4	1
OnePlus Gallery	1
CR Monitor	1
Ruler cm	1

Podemos ver que ahora efectivamente hay **solo 9660 registros totales, 1 por aplicación**. Ya podemos proceder a realizar el análisis del *dataset*.

FASE 2: ANÁLISIS DEL DATASET

2.5. A continuación, obtenemos el **Top 10 de aplicaciones con mayor puntuación y número de reseñas**, porque no significa prácticamente nada que una aplicación tenga una puntuación muy alta si solo procede de una valoración.

```
db.googleplaystore.find().sort({ Rating: -1, Reviews: -1}).limit(10)
```

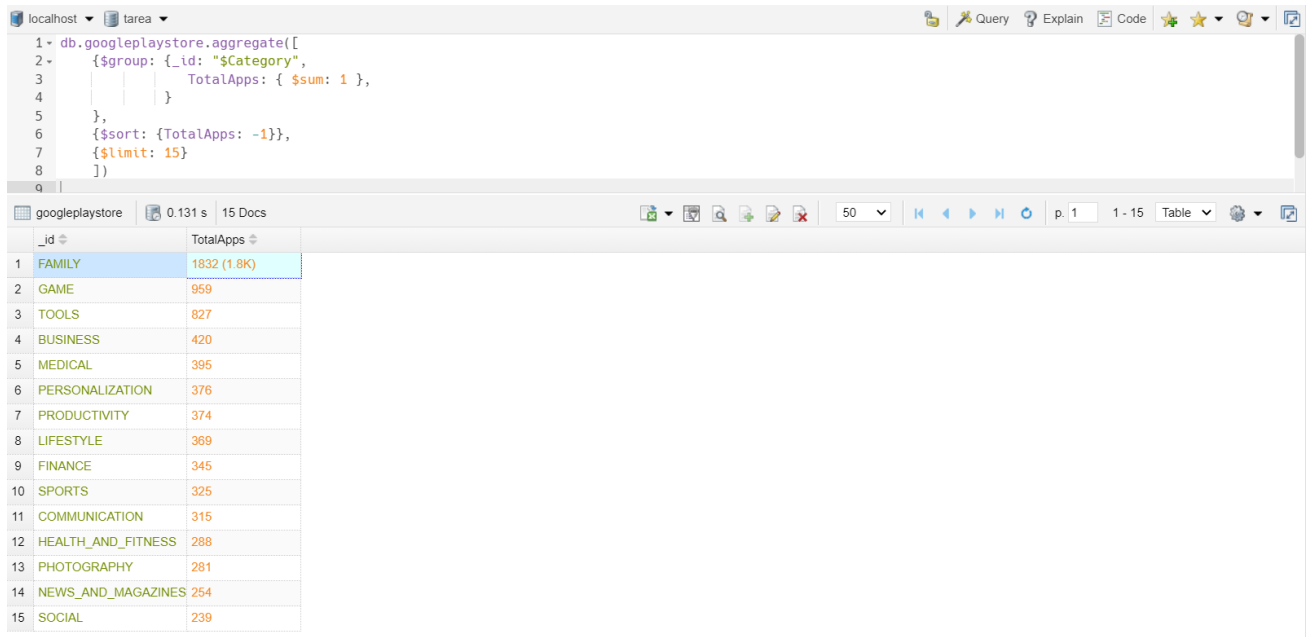


	_id	App	Category	Rating	Reviews	Size	Installs	Type	Price
1	646614b459067f0b0b7	Ríos de Fe	LIFESTYLE	5.0	141	15M	1,000+	Free	0
2	646614b459067f0b0b7	FD Calculator (EMI, SIP, R	FINANCE	5.0	104	2.3M	1,000+	Free	0
3	646614b359067f0b0b7	Oración CX	LIFESTYLE	5.0	103	3.8M	5,000+	Free	0
4	646614b359067f0b0b7	Barisal University App-BU F	FAMILY	5.0	100	10M	1,000+	Free	0
5	646614b459067f0b0b7	Master E.K	FAMILY	5.0	90	Varies with device	1,000+	Free	0
6	646614b359067f0b0b7	CL REPL	TOOLS	5.0	47	17M	1,000+	Free	0
7	646614b359067f0b0b7	AJ Cam	PHOTOGRAPHY	5.0	44	2.8M	100+	Free	0
8	646614b459067f0b0b7	Ek Vote	PRODUCTIVITY	5.0	43	6.2M	500+	Free	0
9	646614b359067f0b0b7	CS & IT Interview Question	FAMILY	5.0	43	3.3M	1,000+	Free	0
10	646614b359067f0b0b7	AI Today - Artificial Intellig	NEWS_AND_MAGAZINES	5.0	43	2.3M	100+	Free	0

Obtenemos que “Ríos de Fe” fue la aplicación con un mayor número de valoraciones y a la vez con el máximo *rating*, seguida de “FD Calculator” y “Oración CX”.

2.6. Top 15 Categorías más frecuentes

```
db.googleplaystore.aggregate([
  {$group: {_id: "$Category",
    TotalApps: { $sum: 1 },
  }},
  {$sort: {TotalApps: -1}},
  {$limit: 15}
])
```



```

1 db.googleplaystore.aggregate([
2   {$group: {_id: "$Category",
3             TotalApps: { $sum: 1 },
4             }},
5 ],
6 {$sort: {TotalApps: -1}},
7 {$limit: 15}
8 ])
  
```

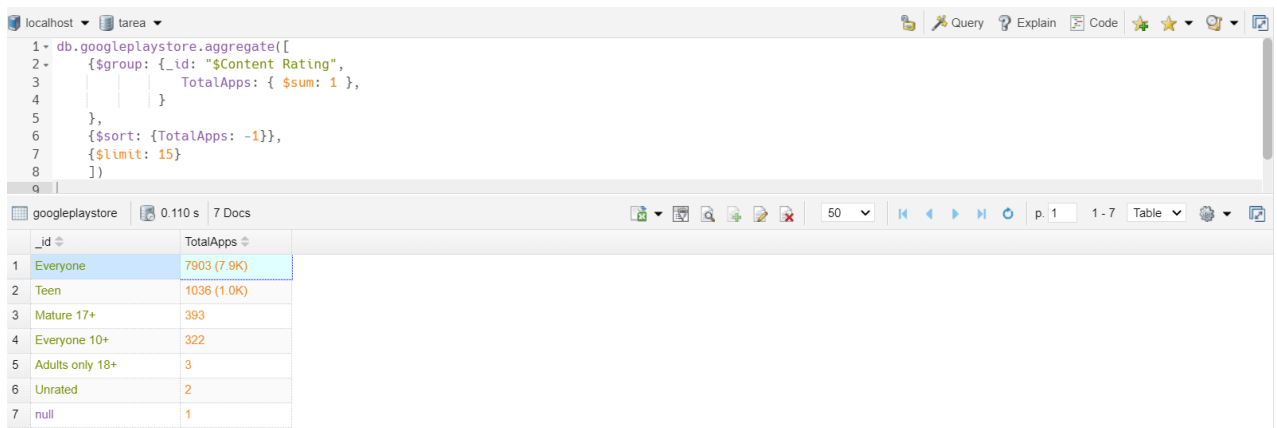
_id	TotalApps
1 FAMILY	1832 (1.8K)
2 GAME	959
3 TOOLS	827
4 BUSINESS	420
5 MEDICAL	395
6 PERSONALIZATION	376
7 PRODUCTIVITY	374
8 LIFESTYLE	369
9 FINANCE	345
10 SPORTS	325
11 COMMUNICATION	315
12 HEALTH_AND_FITNESS	288
13 PHOTOGRAPHY	281
14 NEWS_AND_MAGAZINES	254
15 SOCIAL	239

De las 15 categorías disponibles, la categoría más frecuente es la de “Familia”, seguida de la de “Juegos”, “Herramientas” y “Negocios”.

2.7. Grupo de edad objetivo más frecuente.

```

db.googleplaystore.aggregate([
  {$group: {_id: "$Content Rating",
            TotalApps: { $sum: 1 },
            }},
  ],
  {$sort: {TotalApps: -1}},
  {$limit: 15}
])
  
```



```

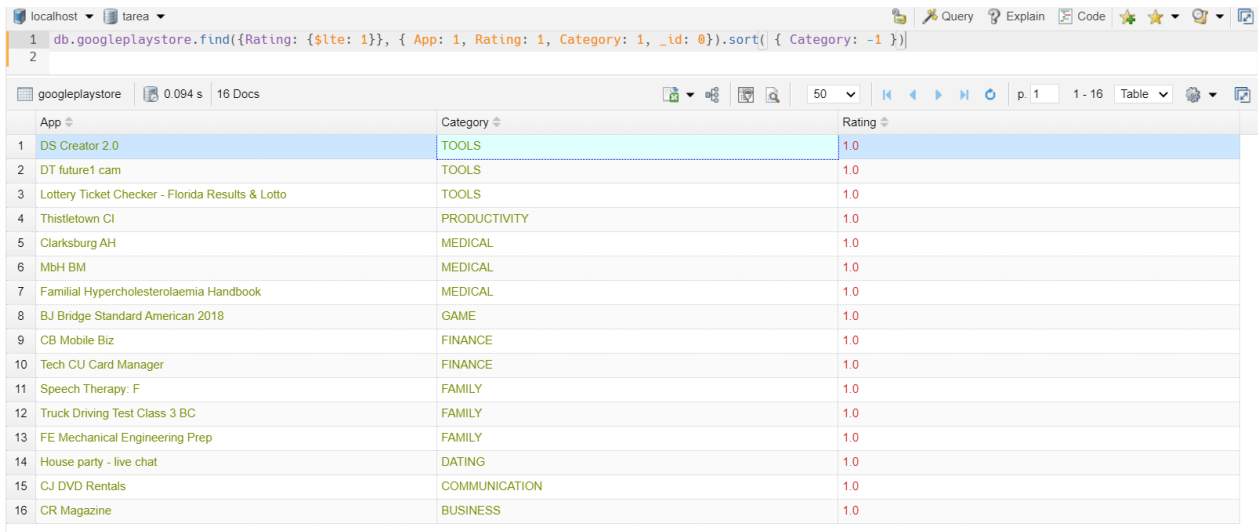
1 db.googleplaystore.aggregate([
2   {$group: {_id: "$Content Rating",
3             TotalApps: { $sum: 1 },
4             }},
5 ],
6 {$sort: {TotalApps: -1}},
7 {$limit: 15}
8 ])
  
```

_id	TotalApps
1 Everyone	7903 (7.9K)
2 Teen	1036 (1.0K)
3 Mature 17+	393
4 Everyone 10+	322
5 Adults only 18+	3
6 Unrated	2
7 null	1

El grupo de edad objetivo o *target* más frecuente es el **general “para todos”, seguido por los adolescentes.**

2.8. Aplicaciones con una puntuación menor o igual a 1 sobre 5, ordenado por categoría y mostrando solo los campos “App”, “Categoría” y “Rating”, sin el campo `_id`.

```
db.googleplaystore.find({Rating: {$lte: 1}}, { App: 1, Rating: 1,
Category: 1, _id: 0}).sort( { Category: -1 })
```



	App	Category	Rating
1	DS Creator 2.0	TOOLS	1.0
2	DT future1 cam	TOOLS	1.0
3	Lottery Ticket Checker - Florida Results & Lotto	TOOLS	1.0
4	Thistle town CI	PRODUCTIVITY	1.0
5	Clarksburg AH	MEDICAL	1.0
6	MbH BM	MEDICAL	1.0
7	Familial Hypercholesterolaemia Handbook	MEDICAL	1.0
8	BJ Bridge Standard American 2018	GAME	1.0
9	CB Mobile Biz	FINANCE	1.0
10	Tech CU Card Manager	FINANCE	1.0
11	Speech Therapy: F	FAMILY	1.0
12	Truck Driving Test Class 3 BC	FAMILY	1.0
13	FE Mechanical Engineering Prep	FAMILY	1.0
14	House party - live chat	DATING	1.0
15	CJ DVD Rentals	COMMUNICATION	1.0
16	CR Magazine	BUSINESS	1.0

Obtenemos 16 aplicaciones con una estrella de puntuación.

3. CONCLUSIONES

Tras depurar y analizar este *dataset*, podemos concluir que:

- “Ríos de Fe” fue la aplicación con un mayor número de valoraciones y a la vez con el máximo *rating*, seguida de “FD Calculator” y “Oración CX”, donde dos de estas tres pertenecen a la **categoría “Lifestyle”**.
- La categoría de aplicaciones más frecuente es la de **“Familia”**, seguida de la de “Juegos”, “Herramientas” y “Negocios”.
- El grupo de edad objetivo o *target* más frecuente de las aplicaciones es el **público general**, seguido por el público adolescente.
- Obtenemos que, del total de 9660 aplicaciones, tan solo 16 tienen una puntuación de una estrella o menos.