Utilización de MongoDB

Módulo: Bases de datos NO SQL

AUTOR: MIGUEL MORENO MARDONES

FECHA 11/11/2021

1. Introducción

El objetivo de este trabajo es aprender los conceptos y fundamentos relacionados con las bases de datos NOSQL y la utilización de MongoDB para crear consultas que nos permitan familiarizarnos con el entorno y visualizar los resultados.

Para ello se nos pidió incorporar un dataset o colección de datos sobre el que realizar dichas consultas, siendo necesarios ejercicios sobre inserción, actualización, proyección y filtrado, así como pipeline de agregación.

En mi caso he optado por utilizar un dataset completamente personal que fue utilizado en mi trabajo de fin de grado (ETSIINF UPM), el cual establece las principales aplicaciones de ámbito estadístico alojadas en la tienda Google Play Store.

Se trata de un dataset de 250 valores en el que se recogen las principales características de estas aplicaciones, ya que se utilizó para realizar un análisis de datos (cálculos estadísticos, algoritmos de regresión y algoritmos de agrupamiento) en profundidad mediante el lenguaje de programación de R.

De igual manera, la limpieza del dataset fue realizada en su momento mediante la aplicación OpenRefine por lo que no deberían encontrarse valores no tipados o muchos campos vacíos (pueden encontrarse campos creados a partir del análisis de datos posterior). Sin embargo, debido a la antigüedad de este, puede que algunos valores de las características de las aplicaciones estén obsoletos y sea necesaria una actualización del mismo.

La idea principal de los ejercicios realizados es obtener algunos de los resultados que se determinaron en su momento mediante la utilización del lenguaje R, así como realizar nuevas consultas que permitan obtener nuevos resultados que en su momento no fueron tenidos en cuenta. Todo ello mediante un sistema totalmente distinto, ya que se emplea JavaScript como lenguaje de programación y JSON para acceder a los valores clave-valor del dataset.

Un resumen de las variables que pueden ser encontradas y una pequeña descripción se ofrece a continuación:

1. *Nombre*: Nombre de la aplicación que posee en la plataforma Google Play.
2. *Desarrollador/es*: Estudio/s o autor/es encargado del desarrollo de la aplicación.
3. *Contenido*: Variable que indica el contenido o tipo de aplicación.
4. *Categoría*: Variable que indica la categoría en la plataforma Google Play.
5. *Idioma*: Variable que indica el idioma de la aplicación.
6. *Elementos desplegables*: Variable asociada a la estructura que indica si la aplicación posee elementos desplegables.
7. *Barra de acción*: Variable asociada a la estructura que indica si la aplicación posee barra de acción.
8. *Sistema de navegación*: Variable asociada a la estructura que indica el modo de navegación de la aplicación.
9. *Inserción de datos*: Variable a la estructura que indica la posibilidad de inserción de datos.
10. *Transiciones entre pantallas*: Variable asociada a la estructura que indica si la aplicación ofrece transiciones entre sus pantallas.
11. *Panel lateral*: Variable asociada a la estructura que indica si la aplicación ofrece panel lateral.
12. *Usabilidad*: Variable que indica el nivel de usabilidad a la hora de utilizar la aplicación.
13. *Calidad* de la interfaz visual: Variable que indica la calidad de interfaz empleada en el desarrollo.
14. *Precio*: Variable que indica el *precio* de la aplicación.
15. *Descargas*: Variable que indica el número de *descargas* realizadas a través de la plataforma Google Play.
16. *Tamaño*: Variable que indica el *tamaño* que ocupa la aplicación.
17. *Valoraciones*: Variable que indica el total de *valoraciones* realizadas por los usuarios en la plataforma Google Play.
18. *Media de las valoraciones*: Variable que indica la puntuación *media de las valoraciones* en la plataforma Google Play.

2. Carga del dataset

Se procedió a operar con el dataset mediante el entorno de NoSQLBooster, aquí podemos importar dicha colección tanto en formato JSON como en CSV mediante el botón de *Import*. En mi caso, realicé previamente una conversión del dataset a través del programa OpenRefine para evitar posibles conflictos.

Figura 1: Barra de comandos en NoSQlBooster.

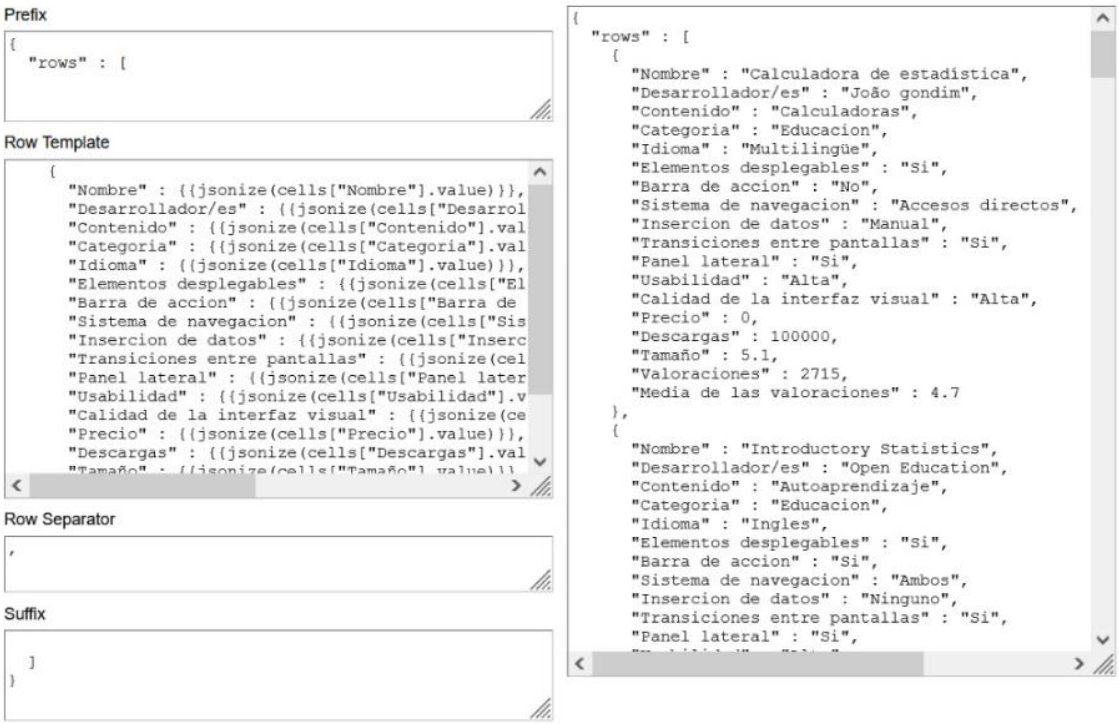

Figura 2: Template de OpenRefine para convertir a JSON.

3. Ejercicios realizados

Se ha realizado una totalidad de 11 ejercicios que recorren cada uno de los requerimientos pedidos en esta entrega. A continuación, se ofrece el código de cada una de las consultas realizadas, acompañados de una breve explicación, así como los resultados obtenidos.

Algunos resultados son ofrecidos tanto en formato JSON como en formato tablas para ayudar al lector a interpretar los mismos.

1. Conocer la aplicación más descargada y menos descargada.

```
db.dataset.find().sort({Descargas:-1}).limit(1)
db.dataset.find().sort({Descargas:+1}).limit(1)
```

Como en cualquier estudio de mercado y desde la parte de un desarrollador, nos interesa saber cuál es nuestro máximo competidor para poder aprender de sus virtudes, así como la menos descargada para poder analizar como enfocar un buen desarrollo.

Resultado:


```
{
    "_id" : ObjectId("618d6bcd21630f6b8dd8e4a9"),
    "Nombre" : "GeoGebra Calculadora Gráfica",
    "Desarrollador/es" : "International GeoGebra Institute",
    "Contenido" : "Calculadoras",
    "Categoria" : "Educacion",
    "Idioma" : "Multilingüe",
    "Elementos desplegables" : "Si",
    "Barra de accion" : "Si",
    "Sistema de navegacion" : "Pestañas",
    "Insercion de datos" : "Ambos",
    "Transiciones entre pantallas" : "Si",
    "Panel lateral" : "Si",
    "Usabilidad" : "Alta",
    "Calidad de la interfaz visual" : "Alta",
    "Precio" : 0,
    "Descargas" : 10000000,
    "Tamaño" : 15,
    "Valoraciones" : 39927,
    "Media de las valoraciones" : 4.2
}
```

```
{
    "_id" : ObjectId("618d6bcd21630f6b8dd8e4c3"),
    "Nombre" : "Statistics Pro",
    "Desarrollador/es" : "Rohit Mehta",
    "Contenido" : "Informativas",
    "Categoria" : "Personalizacion",
    "Idioma" : "Ingles",
    "Elementos desplegables" : "No",
    "Barra de accion" : "Si",
    "Sistema de navegacion" : "Accesos directos",
    "Insercion de datos" : "Ninguno",
    "Transiciones entre pantallas" : "Si",
    "Panel lateral" : "Si",
    "Usabilidad" : "Alta",
    "Calidad de la interfaz visual" : "Alta",
    "Precio" : 1.49,
    "Descargas" : 1,
    "Tamaño" : 11,
    "Valoraciones" : 0,
    "Media de las valoraciones" : null
}
```


2. Conocer la media total de las valoraciones.

```
db.dataset.aggregate([{$group:{"_id":"_id",
    "Media total de las valoraciones del conjunto":
    {$avg:"$Media de las valoraciones"}}},
    {$project: {"_id":0}}
])
```

Conocer la media de las valoraciones de los usuarios nos permite saber si es un conjunto de aplicaciones aceptado por el público, siendo 1 la peor valoración y 5 la mejor.

Resultado:


```
{
    "Media total de las valoraciones del conjunto" :  4.101025641025641
}
```

3. Conocer la cantidad de descargas totales del conjunto de aplicaciones.

```
db.dataset.aggregate([{$group:{"_id":"_id",
    "Descargas totales del conjunto":{$sum:"$Descargas"}}},
    {$project: {"_id":0}}
])
```

De igual manera, saber la cantidad de descargas realizadas puede venirnos bien para futuras operaciones estadísticas, así como para comparar este conjunto de aplicaciones con otro conjunto de otra índole y así ver su situación en el Google Play Store.

Resultado:

```
{
    "Descargas totales del conjunto" : 17937735
}
```

4. Conocer el contenido de las aplicaciones más descargadas junto a su media de valoraciones.

```
db.dataset.aggregate([
    {$group:{"_id":"$Contenido",
             "Descargas totales":{"$sum":"$Descargas"},
             "valor":{"$avg":"$Media de las valoraciones"}
            }
    },
    {$project:{"Descargas totales" : 1,
               "Media de valoraciones" : {$trunc: ["$valor",2]}
              }
    },
    {$sort:{"Descargas totales":-1}
    }
])
```

Tanto como desarrollador como cliente, saber cual es la puntuación de los distintos contenidos mas descargados puede ahorrar tiempo a la hora de determinar qué tipo de aplicación deseamos realizar, así como fiabilidad a la hora de buscar entre el contenido deseado como usuario.

Resultado:

	_id	Descargas totales	Media de valoraciones
1	Calculadoras	11.205.864 (11.2M)	4,12
2	Cheat seet	2.161.750 (2.2M)	4,24
3	Informativas	1.816.656 (1.8M)	4,12
4	Autoaprendizaje	1.722.025 (1.7M)	4,06
5	Miniljuegos	725.080 (0.73M)	4,34
6	Divulgacion	306.360 (0.31M)	3,88

5. Categoría de las aplicaciones más descargadas junto a su media de valoraciones y espacio que ocupan en la tienda Google Play Store.

```
db.dataset.aggregate([
    {$group:{"_id":"$Categoria",
             "Descargas totales":{"$sum":"$Descargas"},
             "valor":{"$avg":"$Media de las valoraciones"}
             "tam":{"$sum" : "$Tamaño"}
            }
    },
    {$project:{"Descargas totales" : 1,
               "Media de valoraciones" : {$trunc: ["$valor",2]},
               "Media de valoraciones" : {$ifNull: [ "$valor", "No evaluada"]}
               "Espacio en Google Play Store (MB)" : {$trunc: ["$tam",1]}

    }
    },
    {$sort:{"Descargas totales":-1}
    }
])
```

Desde la visión del cliente conocer el éxito de las categorías que ofrece Google Play Store puede asegurarnos que la aplicación descargada perteneciente al mejor contenido ofrecerá mejores resultados. De igual manera saber el tamaño total puede servir como variable de estudio para posibles análisis de agrupamientos.

No se muestra la totalidad de los resultados debido al espacio de la imagen.

Resultado:

	_id	Descargas totales	Media de valoraciones	Espacio en Google Play Store (MB)
1	Educacion	14.139.888 (14.1M)	4,101	1036,8 (1.0K)
2	Herramientas	1.781.506 (1.8M)	4,0233	157,4
3	Libros y obras de consulta	1.080.650 (1.1M)	3,9929	248,1
4	Cartas	500.000 (0.50M)	4,5	0,2
5	Deportes	201.430 (0.20M)	4,0833	269,8
6	Puzles	100.050 (0.10M)	4,8	65,7
7	Medicina	51.010 (51.0K)	4,2	22,7
8	Productividad	32.800 (32.8K)	4,1857	60,8
9	Entretenimiento	21.010 (21.0K)	4,3667	25
10	Musica y audio	10.000 (10.0K)	3,6	3,7
11	Casino	7500 (7.5K)	3,725	58,6
12	Empresa	5100 (5.1K)	4,75	113
13	Comunicacion	5000 (5.0K)	4	29
14	Educativos	1000 (1.0K)	4	2,7
15	Mapas y navegacion	500	No evaluada	35

6. Cantidad de descargas de las aplicaciones gratuitas y de pago.

```
db.dataset.aggregate([
  {
   $facet:{
    "Gratis":[
        {$match: { "Precio" : {$eq: 0}}},
        {$group:{"_id":"_id","Aplicaciones gratuitas":{$sum:"$Descargas"}}},
        {$project: {"_id":0}},
        ],
    "Pago":[
        {$match: { "Precio" : {$gt: 0}}},
        {$group:{"_id":"_id","Aplicaciones de pago":{$sum:"$Descargas"}}},
        {$project: {"_id":0}}
        ]
     }
  },
    {
    $project: {
            "Descargas totales":{
                $setUnion: ["$Gratis","$Pago"]
            }
        }
     }
])
```

A la hora de realizar los análisis de regresión se determinó que la variable descargas está influenciada de manera negativa por el precio de la aplicación. De esta manera podemos observar que una alta cantidad de descargas están asociadas a aplicaciones gratuitas, ya que el público prefiere optar por este tipo de aplicaciones.

Resultado:


```
{
    "Descargas totales" : [
        {
            "Aplicaciones de pago" : 117015
        },
        {
            "Aplicaciones gratuitas" : 17820720
        }
    ]
}
```


7. Media de la cantidad de valoraciones de aplicaciones con más de 5.000 descargas y menos de 5.000.

```
db.dataset.aggregate([
  {
   $facet:{
    "a":[
        {$match: { "Descargas" : {$lte: 5000}}},
        {$group:{"_id":"_id","x":{$avg:"$Valoraciones"}}},
        {$project:{"_id":0,
   "Aplicaciones con menos de 5000 descargas" : {$trunc: ["$x",0]}}}
        ],
    "b":[
        {$match: { "Descargas" : {$gt: 5000}}},
        {$group:{"_id":"_id","y":{$avg:"$Valoraciones"}}},
        {$project:{"_id":0,
    "Aplicaciones con mas de 5000 descargas" : {$trunc: ["$y",0]}}}
        ]
     }
  },
    {
    $project: {"Media de comentarios":{$setUnion: ["$a","$b"]}}
    },
    {$unwind: "$Media de comentarios"}
])
```

El análisis de regresión logística realizado con anterioridad nos permitió conocer que hay una gran influencia de los comentarios a la hora de descargar una aplicación, ya que el cliente siente mayor fiabilidad.

Por lo que una de las ecuaciones obtenidas en esta regresión determinó que hay una mayor posibilidad de que una aplicación alcance las 5.000 descargas si esta posee una gran cantidad de comentarios.

Vamos a determinar la cantidad media de comentarios de las aplicaciones que superan esta cantidad y las que no lo logran, para así ver si es correcta esta afirmación.

Resultado:

```
{
    "Media de comentarios" : {
        "Aplicaciones con mas de 5000 descargas" : 1481
    }
},

{
    "Media de comentarios" : {
        "Aplicaciones con menos de 5000 descargas" : 18
    }
}
```

8. Modificación del campo Idioma String a Array (valor "Multilingüe" a múltiples idiomas).

```
// Añadimos un campo al valor idioma y sobreescribimos en nuestro dataset
db.dataset.aggregate([{$addFields:{"Idioma":{"$split":["$Idioma",","]}}},
{$out:"dataset"}])

// Sustituimos Multilingüe por el conjunto de idiomas y realizamos la
actualizacion oportuna
var query = {"Idioma":"Multilingüe"}
var operacion =
{$push:{"Idioma":{$each:["Ingles","Español","Frances","Hindi",
"Ruso","Italiano","Portugues","Aleman"]}}}
db.dataset.updateMany(query,operacion)

// Finalmente extraemos el valor Multilingüe del campo Idioma
var query = {}
var operacion = {$pull:{"Idioma": "Multilingüe"}}
db.dataset.updateMany(query,operacion)
```

El dataset importado en R no admitía atributos multivalorados para la propiedad de idioma, por lo que aquellas aplicaciones que poseían varios idiomas fueron definidas como *Multilingüe*. El formato JSON permite trabajar con múltiples valores para una sola clave, por lo que es conveniente actualizar este campo para así conocer los idiomas de todas las aplicaciones.

Resultado:


```
{
    "acknowledged" : true,
    "matchedCount" : 18,
    "modifiedCount" : 18
}

{
    "_id" : ObjectId("618d6bcd21630f6b8dd8e3ef"),
    "Nombre" : "Calculadora de estadística",
    "Desarrollador/es" : "João gondim",
    "Contenido" : "Calculadoras",
    "Categoria" : "Educacion",
    "Idioma" : [
        "Ingles",
        "Español",
        "Frances",
        "Hindi",
        "Ruso",
        "Italiano",
        "Portugues",
        "Aleman"
    ],
    "Elementos desplegables" : "Si",
    "Barra de accion" : "No",
    "Sistema de navegacion" : "Accesos directos",
    "Insercion de datos" : "Manual",
    "Transiciones entre pantallas" : "Si",
    "Panel lateral" : "Si",
    "Usabilidad" : "Alta",
    "Calidad de la interfaz visual" : "Alta",
    "Precio" : 0,
    "Descargas" : 100000,
    "Tamaño" : 5.1,
    "Valoraciones" : 2715,
    "Media de las valoraciones" : 4.7
}
```


Así pasamos de una variable en formato String a una variable en formato Array, donde albergaremos todos los posibles idiomas que tiene la aplicación. Se ha determinado (para ahorrar tiempo de creación de código) que todas las aplicaciones cuyo lenguaje era multilingüe posean un array de todos los idiomas encontrados en el dataset.

9. Conocer el porcentaje de las aplicaciones según el idioma.

```
db.dataset.aggregate([
  {$unwind: "$Idioma"},
        {$group:{"_id":"$Idioma","Cantidad":{$sum:1}}},
        {$project: {
        "Cantidad" : true,
        "Porcentaje (%)" : {
            $divide: [{$multiply: ["$Cantidad",100]},
            a =  db.dataset.find().count()]}
        }
    },
    {$sort: {"Porcentaje (%)": -1}}
])
```

Al transformar en la anterior consulta el formato del campo Idioma, podemos obtener con exactitud el porcentaje de los idiomas empleados para desarrollar las aplicaciones. Claramente el Inglés es el más empleado, por lo que como desarrolladores un gran éxito de nuestra aplicación puede residir en realizarla en esta idioma ya que llegaremos a una mayor cantidad de público.

Resultado:

	_id	Cantidad	Porcentaje (%)
1	Ingles	217	86,8
2	Español	39	15,6
3	Frances	22	8,8
4	Italiano	21	8,4
5	Aleman	19	7,6
6	Hindi	19	7,6
7	Portugues	19	7,6
8	Ruso	18	7,2
9	Arabe	1	0,4
10	Indonesio	1	0,4

10. Conocer las características de la estructura de las aplicaciones realizando una comparación entre la mejor valorada y la peor valorada por el público.

```
db.dataset.aggregate([
  {
   $facet:{
    "a":[
         {$match: {"Usabilidad" : "Alta",
         "Calidad de la interfaz visual" : "Alta"}},
         {$project: {"_id" : 0 ,"Nombre" : 1, "Elementos desplegables" :1,
         "Barra de accion" :1, "Sistema de navegacion" :1,
         "Insercion de datos" : 1,
         "Transiciones entre pantallas" :1, "Panel lateral" :1}}
         {$sort: {Descargas : -1, Valoraciones : -1,
         "Media de las valoraciones" : -1}}
         {$limit: 1},
        ],
    "b":[
         {$match: {"Usabilidad" : "Baja",
         "Calidad de la interfaz visual" : "Baja"}},
         {$project: {"_id" : 0 ,"Nombre" : 1, "Elementos desplegables" :1,
         "Barra de accion" :1, "Sistema de navegacion" :1,
         "Insercion de datos" : 1,
         "Transiciones entre pantallas" :1, "Panel lateral" :1}}
         {$sort: {Descargas : -1, Valoraciones : -1,
         "Media de las valoraciones" : -1}}
         {$limit: 1},
        ]
     }
  },
  {$project: {"Tabla Comparativa":{$setUnion: ["$a","$b"]}}},
  {$unwind: "$Tabla Comparativa"}
])
```

Esta consulta esta especialmente pensada para los posibles desarrolladores que accedan al mercado de las aplicaciones de estadística y probabilidad. Aquí podremos insertar los parámetros deseados para obtener una comparativa de las cualidades de estructura entre dos aplicaciones. De esta manera imitar el planteamiento de estructura de las mejores aplicaciones puede ser un buen punto de partida para comenzar.

De igual manera, el estudio realizado determinó que la contentación del cliente estaba directamente relacionada con una buena estructura de aplicación, una buena movilidad entre sus elementos y la capacidad de realizar transiciones de manera sencilla entre sus pantallas… etc.

Aquí la aplicación *Calculadora de estadística* es exitosa, y posee una mayoría de elementos como podemos observar en la tabla.

Resultado:

Tabla Comparativa

	Nombre	Elementos desplegables	Barra de accion	Sistema de navegacion	Insercion de datos	Transiciones entre pantallas	Panel lateral
1	Calculadora de estadística	Si	No	Accesos directos	Manual	Si	Si
2	Curso de Estadistico	Si	Si	Accesos directos	Ninguno	No	No

11. Conocer si hay alguna empresa que ha realizado más de una aplicación y extraer cuales ha realizado.

```
db.dataset.aggregate([
      {$group: {
          _id : "$Desarrollador/es",
          Aplicaciones: {$push: "$Nombre"},
          Count: {$sum:1}
                }
      },
      {$match: {
          Count: {"$gt" : 1}
      }},
      {$project: {
          "Count" : 0
      } },
      {$sort: {"Count" : -1}}
])
```

Finalmente, y de manera general esta consulta pretende establecer una visión de las empresas que participan en la creación de aplicaciones de este entorno. Conocerlas puede resultar de utilidad si deseamos invertir en este campo (en tal caso de que no seamos desarrolladores).

De igual manera el análisis no tuvo en cuenta a los desarrolladores de este entorno, por lo que esta vez si obtenemos información de las empresas punteras en este sector.

Podría aplicarse de igual manera otro filtro para conocer la puntuación de sus aplicaciones realizadas y así saber si son empresas con una aceptación por parte del público.

No se muestra la totalidad de los resultados debido al espacio de la imagen.

Resultado:

	_id	Aplicaciones
1	Hayava Inc	["Basketball Stats", "Estadísticas de fútbol"]
2	Binary Tuts	["Probability and Statistics", "Statistics", "Basics of Statistics"]
3	FlowApp	["Statistics Calculator", "Statistics Calculator Premium"]
4	Rohit Mehta	["Statistics (2)", "Statistics Pro"]
5	SmartMedi.co	["Medical Statistics Basics", "Medical Statistics Basics (1)"]
6	Francisco Aparisi	["StatSuite (Statistics Suite)", "Power of Hypothesis Testing"]
7	Sanjeev Mehta	Array[3]
8	KAIBITS Software GmbH	["Estadisticas", "Estadisticas Pro"]
9	Renato Félix da Silva Souza	["CTI Estatística", "CTI Estatística PRO"]
10	Innovative Sun	["Statistics 12th", "Statistics TextBook 11th"]
11	IT-SUPERNOVAS	["Probability Math", "Probability Math (AR)"]
12	Windows of Vision	["Learn SPSS Manual 18 statistic", "Manual SPSS Learn 19"]
13	Galaxy Production	["Statistics Book Free", "Applied Statistics Book Free"]
14	FORMULAS.XYZ	["Interactive Statistics", "Interactive Statistics PRO"]
15	Nick Pierson	["Stats Calculator Free", "Stats Calculator (Pro)"]
16	Fernando Haro	["Estadística Descriptiva", "Estadística Descriptiva Pro", "Control Estadístico Procesos"]

4. Conclusiones

Las consultas realizadas con anterioridad intentan reflejar de la manera más objetiva lo obtenido en el análisis con el lenguaje de programación R. Las conclusiones reflejan que el cliente prefiere optar por un tipo de contenido de aplicación cercano a las calculadoras, ya que para los son una buena herramienta para la estadística descriptiva.

Por lo que sabiendo que el cliente accede a este mercado con el fin de poder resolver ciertos cálculos estadísticos, debemos de tomar como iniciativa plantear un desarrollo orientado a este tipo de aplicaciones y mejorar los posibles errores que posean las ya creadas (a través de la comparativa de estructuras, por ejemplo).

La media del conjunto de estas ofrece un buen entorno para el cliente, ya que con una puntuación de 4.1 y una cantidad de descargas de 17.937.735 lo convierte en un conjunto totalmente válido para obtener soluciones estadísticas.

No obstante, hemos de comparar este entorno con otras categorías en Google Play Store: Aplicaciones matemáticas, de Ciencias sociales.. para poder así comprar la cantidad de descargas, evaluaciones, opiniones.. y ver que realmente no es un área que puede estar cerca de quedarse obsoleta.

De todas maneras, se debería también orientar este estudio a diferentes plataformas o tiendas de aplicaciones ya que únicamente teniendo referencia de Google Play Store nos dejamos atrás muchas comparaciones posibles.

El trabajo con JSON se hace más práctico para algunos campos, especialmente aquellos con formato Array, pero la utilización del lenguaje R es mucho más útil para este tipo de análisis por lo que se recomienda facilitar las consultas y operar con el dataset sobre este formato ya que además nos permite obtener gráficos mucho mas comprensibles para el lector.