Database Design Week 15 Final Presentation

- GROUP 2
  - DIVYANK AGARWAL
  - SOUMYAJIT CHAKRABORTY
  - MRUNAL PATIL
  - ADITI SONAWANE

Dataset

Apps on Google Play Store

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market!

User group 1:

People who are interested in developing or launching an app and want to analyse the market trends.

Questions:

- How successful apps of that category have been in the market?
- In which category one should develop app to be in the editor’s choice?
- Whom to hire for developing the app.

User group 2:

People who want to promote their products on apps.

Questions:

- Which app can be used to advertise their product?
- Promote products on apps which align with product category

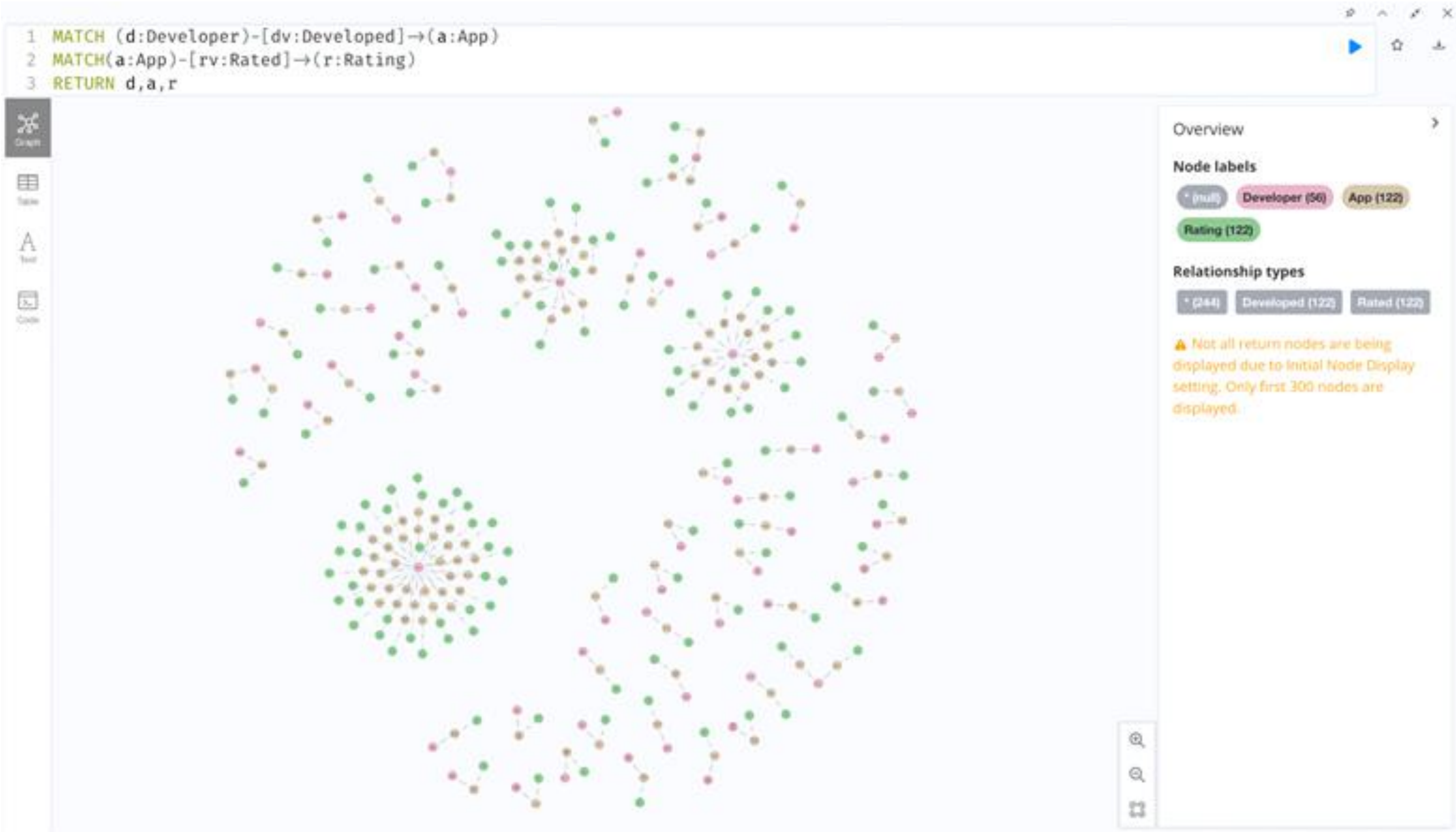
User group 3:

Google

Questions:

- Which apps performed the best and give them Best App Award.
- Which top 20 apps didn’t perform well remove them from play-store.

Dataset Graph on Neo4j



```
MATCH (d:Developer)-[dv:Developed]→(a:App)
MATCH(a:App)-[rv:Rated]→(r:Rating)
RETURN d,a,r
```

Overview

Node labels

* (null) Developer (56) App (122) Rating (122)

Relationship types

* (244) Developed (122) Rated (122)

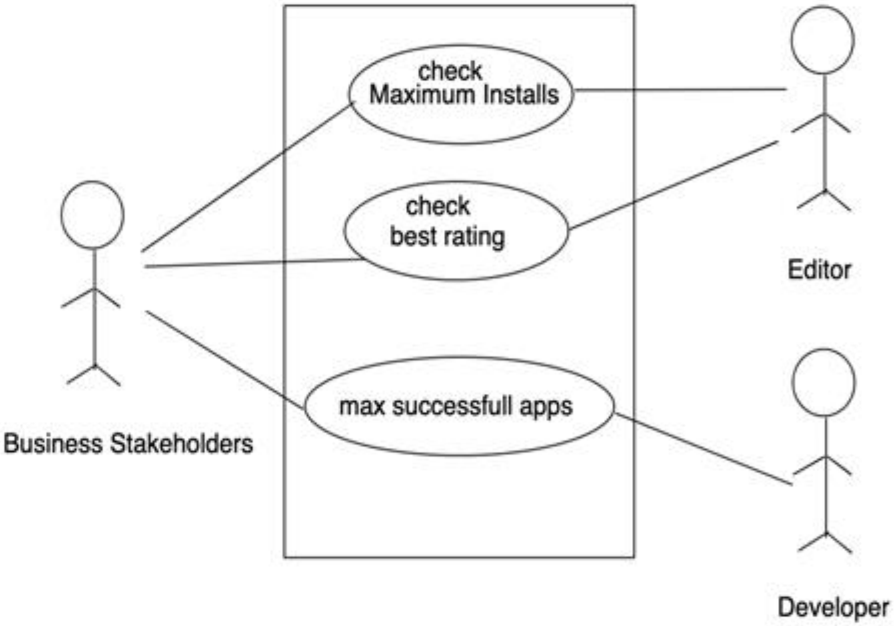
⚠ Not all return nodes are being displayed due to Initial Node Display setting. Only first 300 nodes are displayed.

How our dataset answers users’ problem statements?

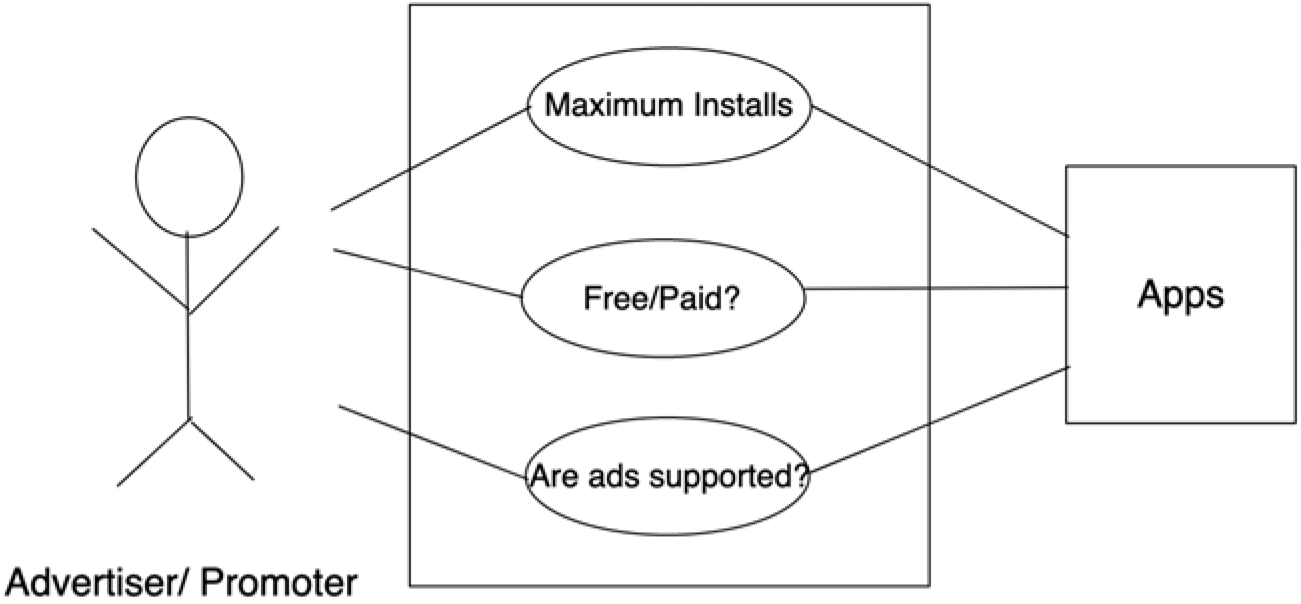
- Based on **number of installs** of the apps or **rating**, user can predict the success of apps or the categories. This information can help **business owners** to decide **which category is profitable.**
- It will also help **Google** understand which apps have performed well and nominate them for award. Apps that did not perform well(less downloads, ratings) can be removed.
- **Investors** can decide in which app they should invest in.
- The **developers** who have most success rate can be searched in the database and those can be hired for developing the app.
- People who want to **promote their product** can look for apps which have maximum in-app purchase and use such apps for **advertisement**.

Design Models

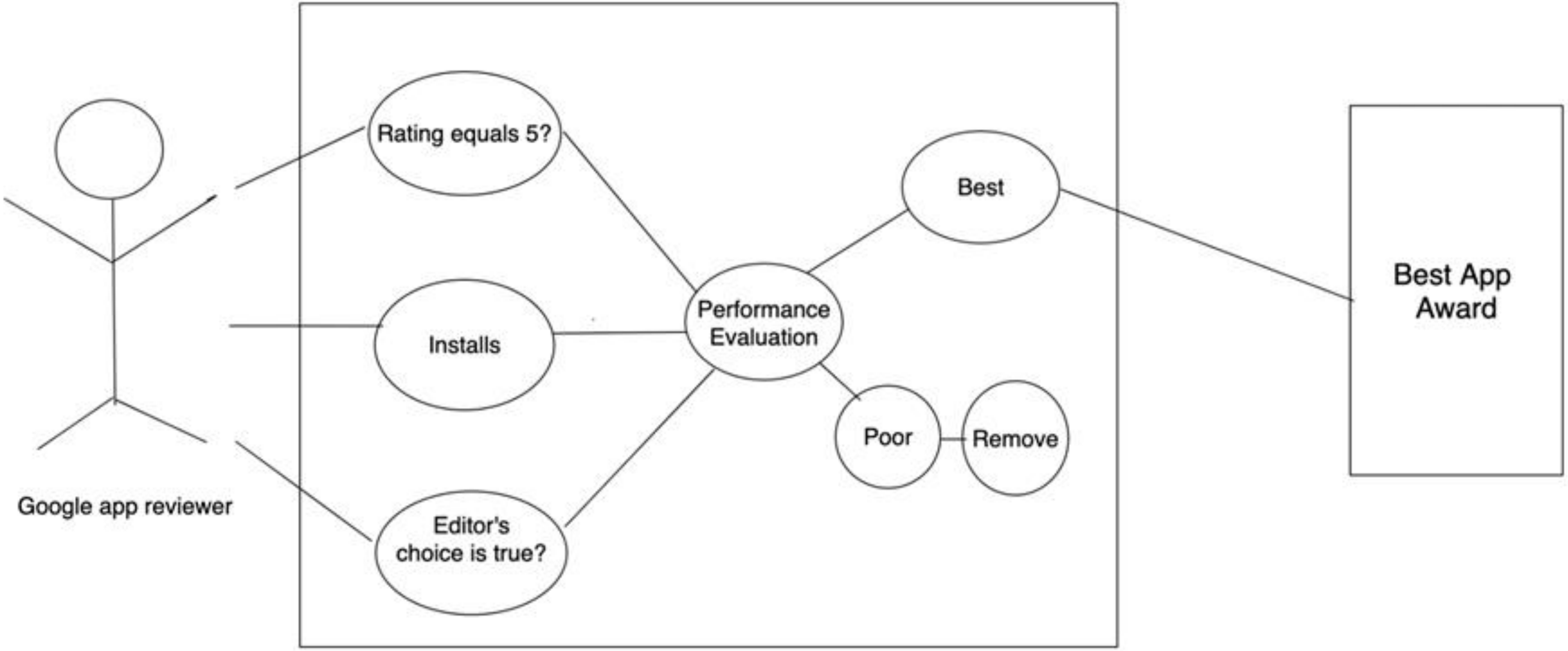
1.Use case diagram for User 1:



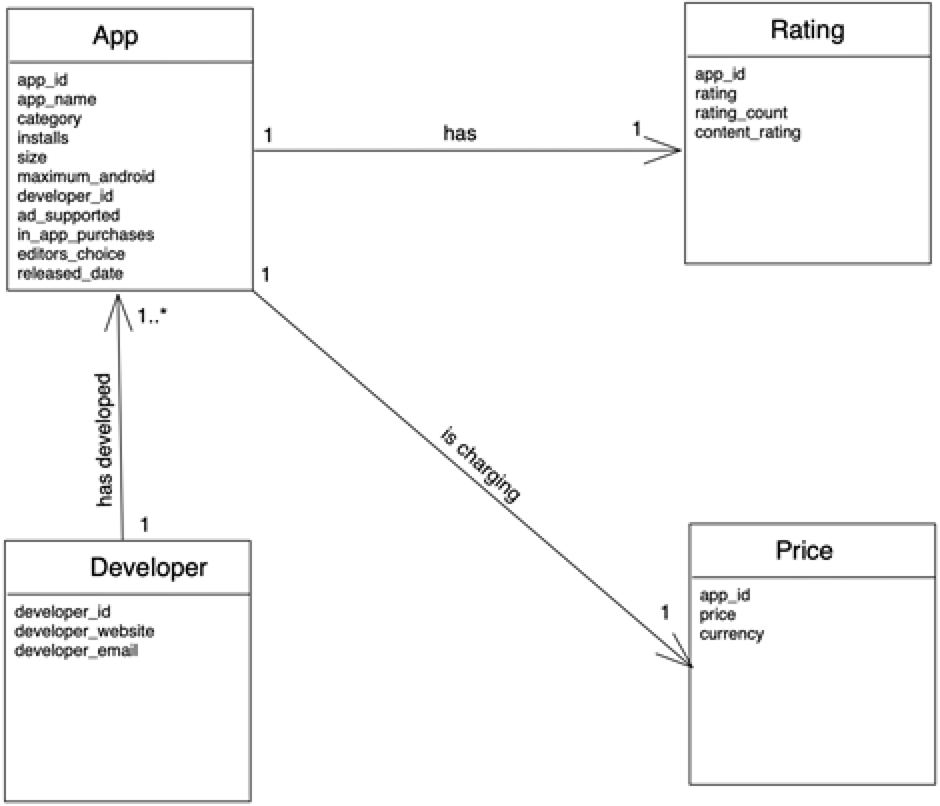
2. Use case diagram for User 2:



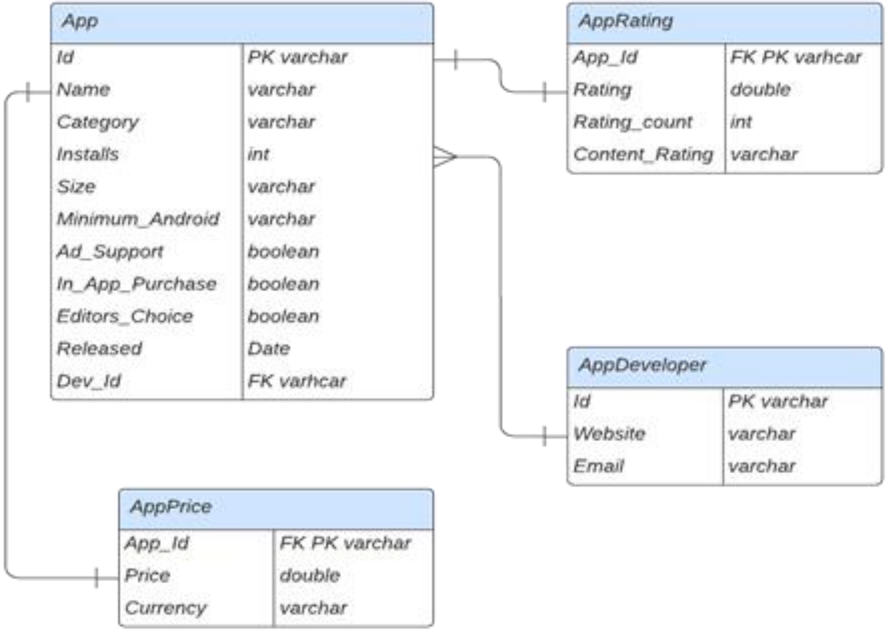
3. Use case diagram for User 3:



Class Diagram:



Entity Relationship Diagram:



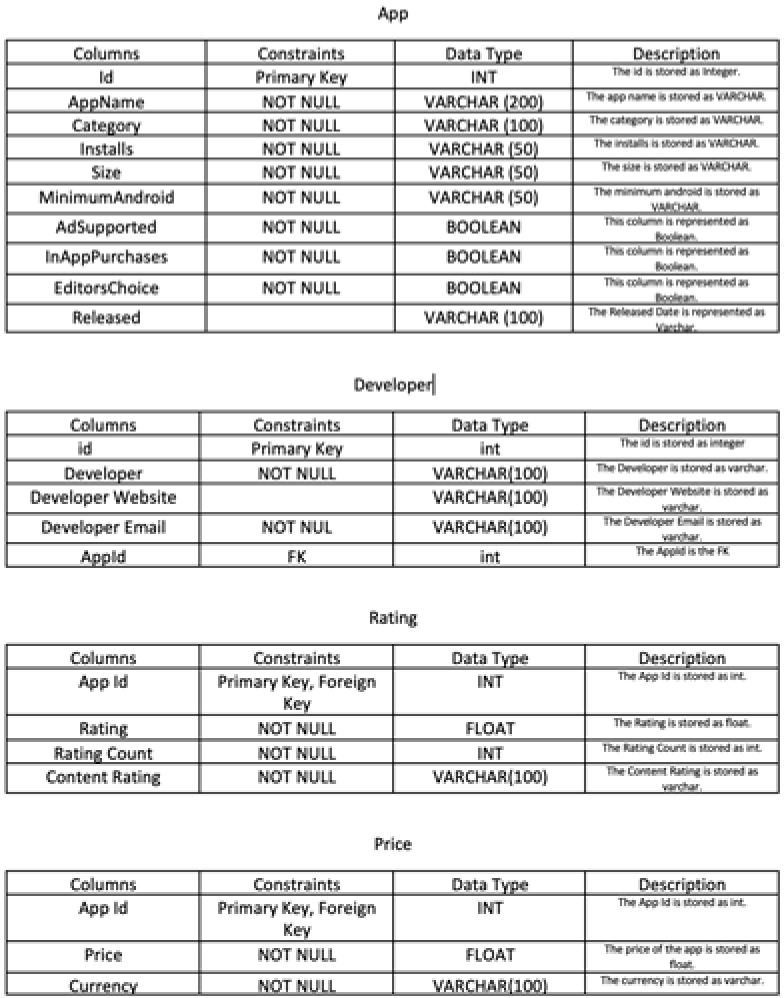
App	
Id	PK varchar
Name	varchar
Category	varchar
Installs	int
Size	varchar
Minimum_Android	varchar
Ad_Support	boolean
In_App_Purchase	boolean
Editors_Choice	boolean
Released	Date
Dev_Id	FK varhcar

AppRating	
App_Id	FK PK varhcar
Rating	double
Rating_count	int
Content_Rating	varchar

AppDeveloper	
Id	PK varchar
Website	varchar
Email	varchar

AppPrice	
App_Id	FK PK varchar
Price	double
Currency	varchar

Implementation model: Data Dictionary



App

Columns	Constraints	Data Type	Description
Id	Primary Key	INT	The id is stored as Integer.
AppName	NOT NULL	VARCHAR (200)	The app name is stored as VARCHAR.
Category	NOT NULL	VARCHAR (100)	The category is stored as VARCHAR.
Installs	NOT NULL	VARCHAR (50)	The installs is stored as VARCHAR.
Size	NOT NULL	VARCHAR (50)	The size is stored as VARCHAR.
MinimumAndroid	NOT NULL	VARCHAR (50)	The minimum android is stored as VARCHAR.
AdSupported	NOT NULL	BOOLEAN	This column is represented as Boolean.
InAppPurchases	NOT NULL	BOOLEAN	This column is represented as Boolean.
EditorsChoice	NOT NULL	BOOLEAN	This column is represented as Boolean.
Released		VARCHAR (100)	The Released Date is represented as Varchar.

Developer

Columns	Constraints	Data Type	Description
id	Primary Key	int	The id is stored as integer
Developer	NOT NULL	VARCHAR(100)	The Developer is stored as varchar.
Developer Website		VARCHAR(100)	The Developer Website is stored as varchar.
Developer Email	NOT NUL	VARCHAR(100)	The Developer Email is stored as varchar.
AppId	FK	int	The AppId is the FK

Rating

Columns	Constraints	Data Type	Description
App Id	Primary Key, Foreign Key	INT	The App Id is stored as int.
Rating	NOT NULL	FLOAT	The Rating is stored as float.
Rating Count	NOT NULL	INT	The Rating Count is stored as int.
Content Rating	NOT NULL	VARCHAR(100)	The Content Rating is stored as varchar.

Price

Columns	Constraints	Data Type	Description
App Id	Primary Key, Foreign Key	INT	The App Id is stored as int.
Price	NOT NULL	FLOAT	The price of the app is stored as float.
Currency	NOT NULL	VARCHAR(100)	The currency is stored as varchar.

Data Integrity in our Dataset

- ON DELETE CASCADE
- ON UPDATE CASCADE
- Primary Keys
- Foreign Keys
- Not null

Challenges with selected dataset

- Handling Null values
- Duplicate Entries
- Data cleaning (Commas in Number of Installs)

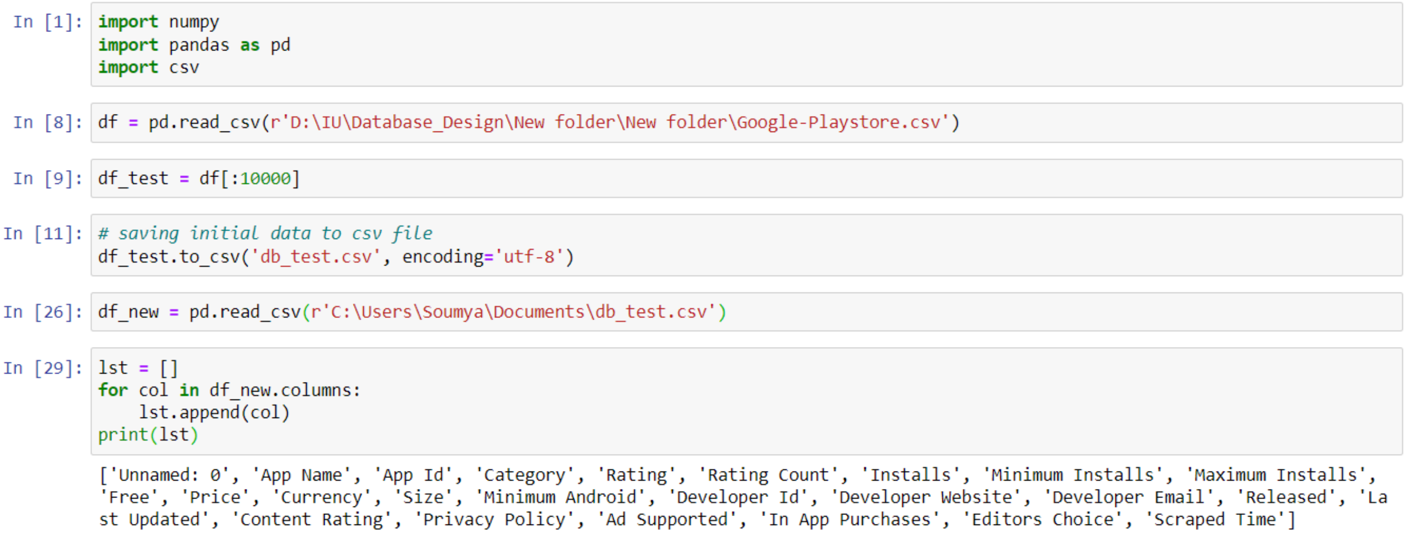
Ethics or Privacy issues:

- The data has been derived from an open source website known as Kaggle.
- Various licensed datasets are present in Kaggle.
- Hence, the data which has been used for this project is licensed.

Data Preprocessing and Feature Engineering

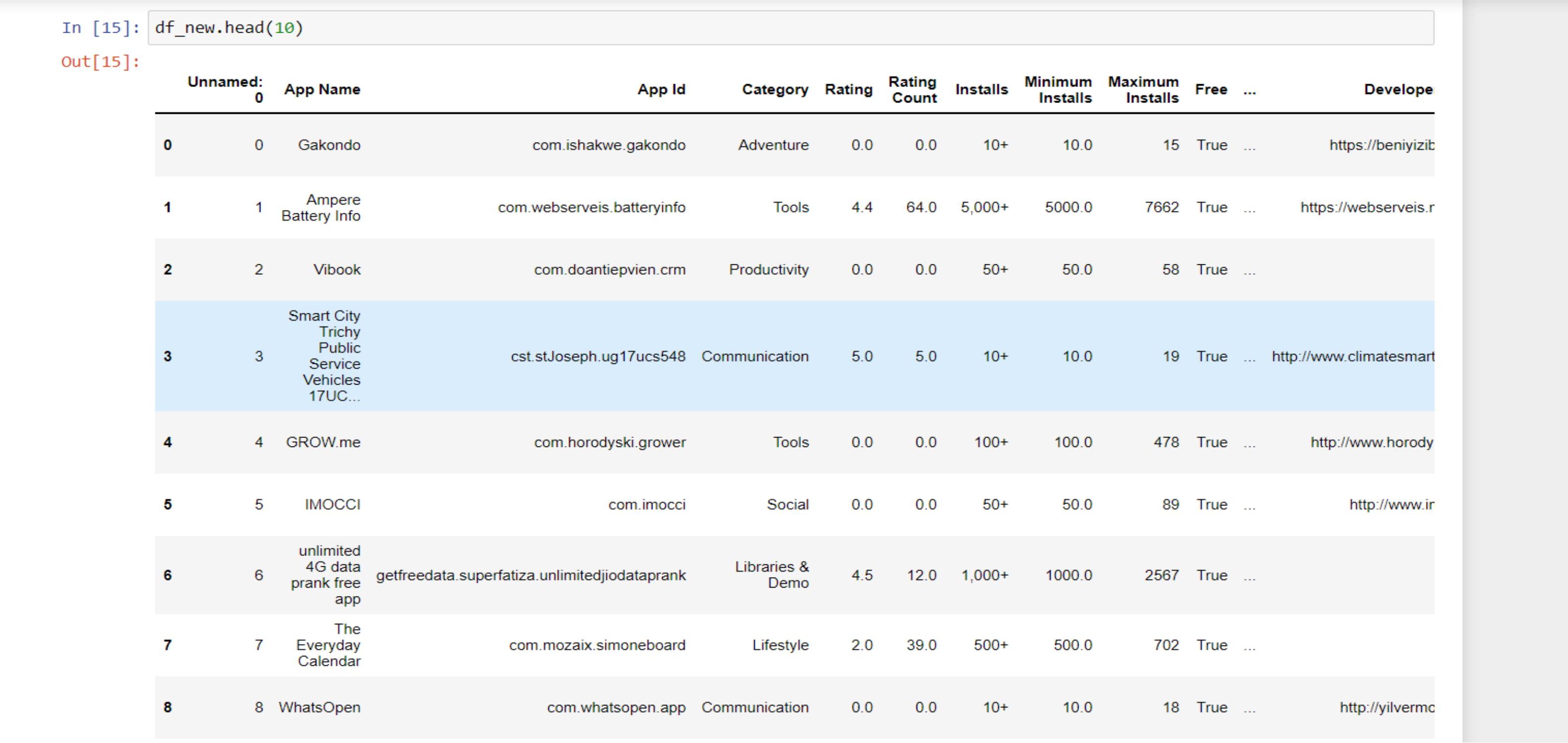
About the dataset -

- The dataset was collected from Kaggle.
- The size of the data was more than 23 million rows and hence it was impossible to process for query without reducing the size.
- The size of the reduced data is 10000 * 25.



```
In [1]: import numpy
        import pandas as pd
        import csv

In [8]: df = pd.read_csv(r'D:\IU\Database_Design\New folder\New folder\Google-Playstore.csv')

In [9]: df_test = df[:10000]

In [11]: # saving initial data to csv file
         df_test.to_csv('db_test.csv', encoding='utf-8')

In [26]: df_new = pd.read_csv(r'C:\Users\Soumya\Documents\db_test.csv')

In [29]: lst = []
         for col in df_new.columns:
             lst.append(col)
         print(lst)

         ['Unnamed: 0', 'App Name', 'App Id', 'Category', 'Rating', 'Rating Count', 'Installs', 'Minimum Installs', 'Maximum Installs',
         'Free', 'Price', 'Currency', 'Size', 'Minimum Android', 'Developer Id', 'Developer Website', 'Developer Email', 'Released', 'La
         st Updated', 'Content Rating', 'Privacy Policy', 'Ad Supported', 'In App Purchases', 'Editors Choice', 'Scraped Time']
```

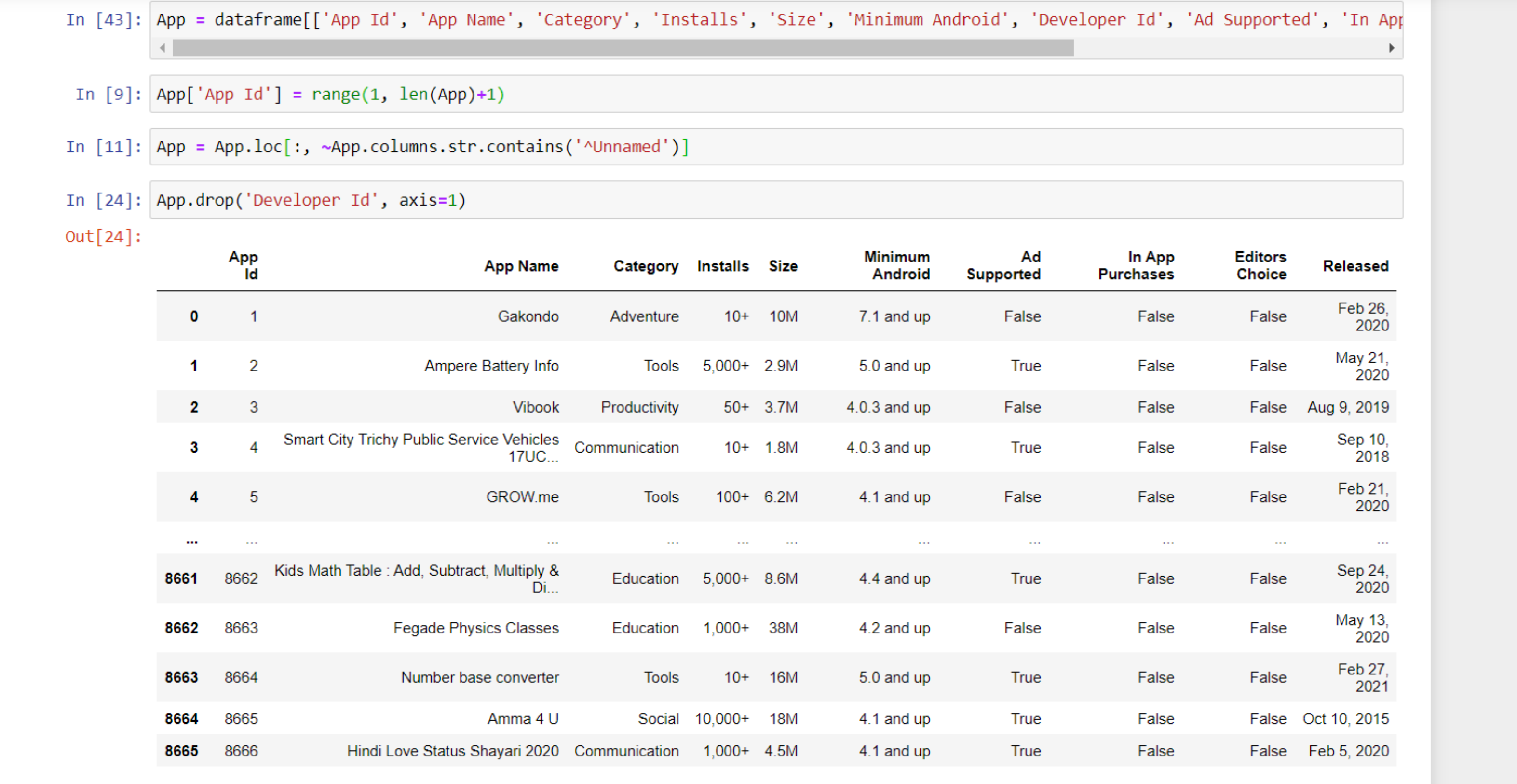
Original Dataset



```
In [15]: df_new.head(10)
Out[15]:
```

	Unnamed: 0	App Name	App Id	Category	Rating	Rating Count	Installs	Minimum Installs	Maximum Installs	Free	...	Develope
0	0	Gakondo	com.ishakwe.gakondo	Adventure	0.0	0.0	10+	10.0	15	True	...	https://beniyizib
1	1	Ampere Battery Info	com.webserveis.batteryinfo	Tools	4.4	64.0	5,000+	5000.0	7662	True	...	https://webserveis.r
2	2	Vibook	com.doantiepvien.crm	Productivity	0.0	0.0	50+	50.0	58	True	...	
3	3	Smart City Trichy Public Service Vehicles 17UC...	cst.stJoseph.ug17ucs548	Communication	5.0	5.0	10+	10.0	19	True	...	http://www.climatesmart
4	4	GROW.me	com.horodyski.grower	Tools	0.0	0.0	100+	100.0	478	True	...	http://www.horody
5	5	IMOCCI	com.imocci	Social	0.0	0.0	50+	50.0	89	True	...	http://www.ir
6	6	unlimited 4G data prank free app	getfreedata.superfatiza.unlimitedjiodataprank	Libraries & Demo	4.5	12.0	1,000+	1000.0	2567	True	...	
7	7	The Everyday Calendar	com.mozaix.simoneboard	Lifestyle	2.0	39.0	500+	500.0	702	True	...	
8	8	WhatsOpen	com.whatsopen.app	Communication	0.0	0.0	10+	10.0	18	True	...	http://yilvermc

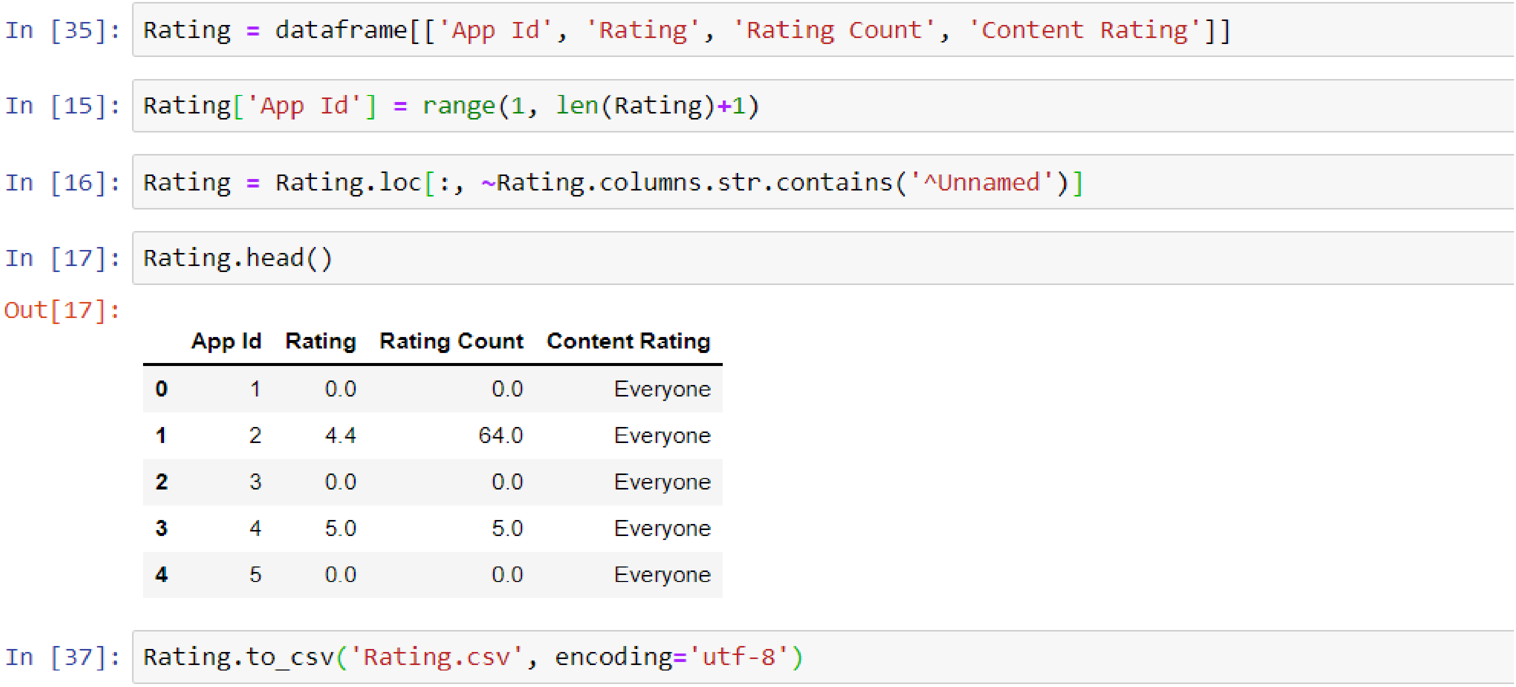
App



```
In [43]: App = dataframe[['App Id', 'App Name', 'Category', 'Installs', 'Size', 'Minimum Android', 'Developer Id', 'Ad Supported', 'In App

In [9]: App['App Id'] = range(1, len(App)+1)

In [11]: App = App.loc[:, ~App.columns.str.contains('^Unnamed')]

In [24]: App.drop('Developer Id', axis=1)
Out[24]:
```

	App Id	App Name	Category	Installs	Size	Minimum Android	Ad Supported	In App Purchases	Editors Choice	Released
0	1	Gakondo	Adventure	10+	10M	7.1 and up	False	False	False	Feb 26, 2020
1	2	Ampere Battery Info	Tools	5,000+	2.9M	5.0 and up	True	False	False	May 21, 2020
2	3	Vibook	Productivity	50+	3.7M	4.0.3 and up	False	False	False	Aug 9, 2019
3	4	Smart City Trichy Public Service Vehicles 17UC...	Communication	10+	1.8M	4.0.3 and up	True	False	False	Sep 10, 2018
4	5	GROW.me	Tools	100+	6.2M	4.1 and up	False	False	False	Feb 21, 2020
...	...	...	...	...	...	...	...	...	...	...
8661	8662	Kids Math Table : Add, Subtract, Multiply & Di...	Education	5,000+	8.6M	4.4 and up	True	False	False	Sep 24, 2020
8662	8663	Fegade Physics Classes	Education	1,000+	38M	4.2 and up	False	False	False	May 13, 2020
8663	8664	Number base converter	Tools	10+	16M	5.0 and up	True	False	False	Feb 27, 2021
8664	8665	Amma 4 U	Social	10,000+	18M	4.1 and up	True	False	False	Oct 10, 2015
8665	8666	Hindi Love Status Shayari 2020	Communication	1,000+	4.5M	4.1 and up	True	False	False	Feb 5, 2020

Rating

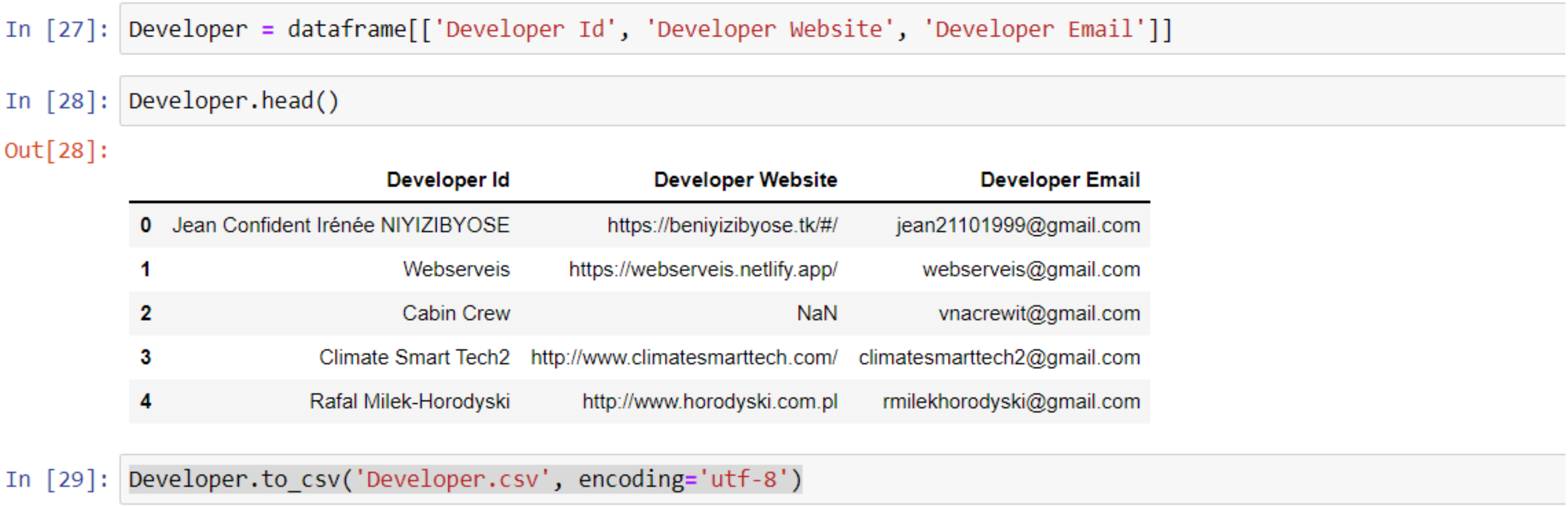


```
In [35]: Rating = dataframe[['App Id', 'Rating', 'Rating Count', 'Content Rating']]

In [15]: Rating['App Id'] = range(1, len(Rating)+1)

In [16]: Rating = Rating.loc[:, ~Rating.columns.str.contains('^Unnamed')]

In [17]: Rating.head()

Out[17]:
```

	App Id	Rating	Rating Count	Content Rating
0	1	0.0	0.0	Everyone
1	2	4.4	64.0	Everyone
2	3	0.0	0.0	Everyone
3	4	5.0	5.0	Everyone
4	5	0.0	0.0	Everyone

```
In [37]: Rating.to_csv('Rating.csv', encoding='utf-8')
```

Developer



In [27]:
```
Developer = dataframe[['Developer Id', 'Developer Website', 'Developer Email']]
```

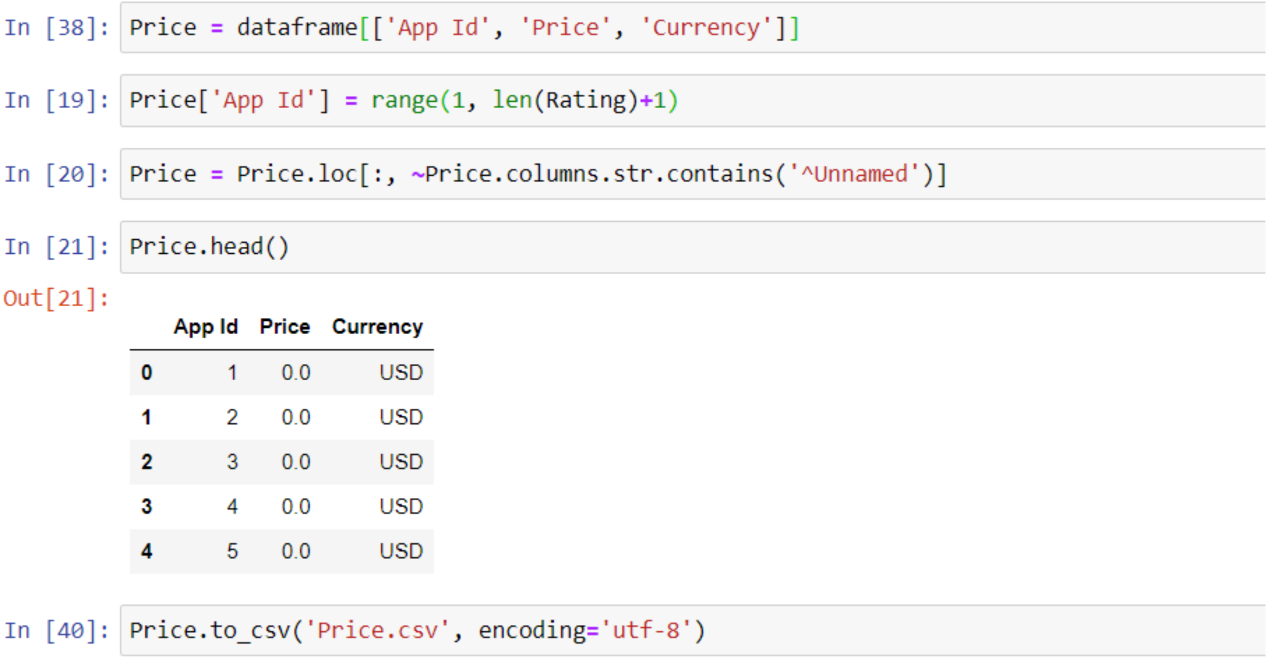
In [28]:
```
Developer.head()
```

Out[28]:

	Developer Id	Developer Website	Developer Email
0	Jean Confident Irénée NIYIZIBYOSE	https://beniyizibyose.tk/#/	jean21101999@gmail.com
1	Webserveis	https://webserveis.netlify.app/	webserveis@gmail.com
2	Cabin Crew	NaN	vnacrewit@gmail.com
3	Climate Smart Tech2	http://www.climatesmarttech.com/	climatesmarttech2@gmail.com
4	Rafal Milek-Horodyski	http://www.horodyski.com.pl	rmilekhorodyski@gmail.com

In [29]:
```
Developer.to_csv('Developer.csv', encoding='utf-8')
```

Price



```
In [38]: Price = dataframe[['App Id', 'Price', 'Currency']]

In [19]: Price['App Id'] = range(1, len(Rating)+1)

In [20]: Price = Price.loc[:, ~Price.columns.str.contains('^Unnamed')]

In [21]: Price.head()

Out[21]:
```

	App Id	Price	Currency
0	1	0.0	USD
1	2	0.0	USD
2	3	0.0	USD
3	4	0.0	USD
4	5	0.0	USD

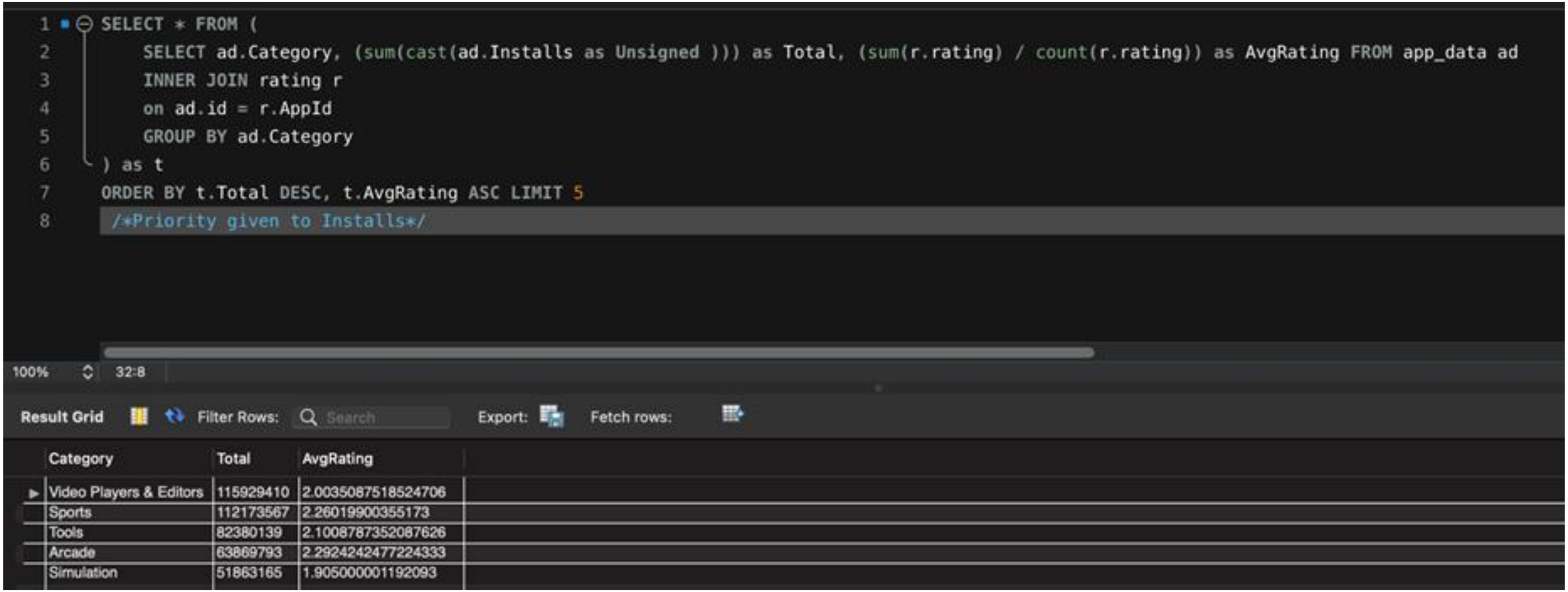
```
In [40]: Price.to_csv('Price.csv', encoding='utf-8')
```

How Dataset can be queried to answer user questions?

User 1 Query 1

Question : How successful apps of that category have been in the market?

-
  ```
  SELECT * FROM (
  SELECT ad.Category, (sum(cast(ad.Installs as Unsigned ))) as Total, (sum(r.rating) / count(r.rating)) as AvgRating FROM app_data ad
  INNER JOIN rating r on ad.id = r.AppId
  GROUP BY ad.Category) as t
  ORDER BY t.Total DESC, t.AvgRating ASC LIMIT 5
  ```

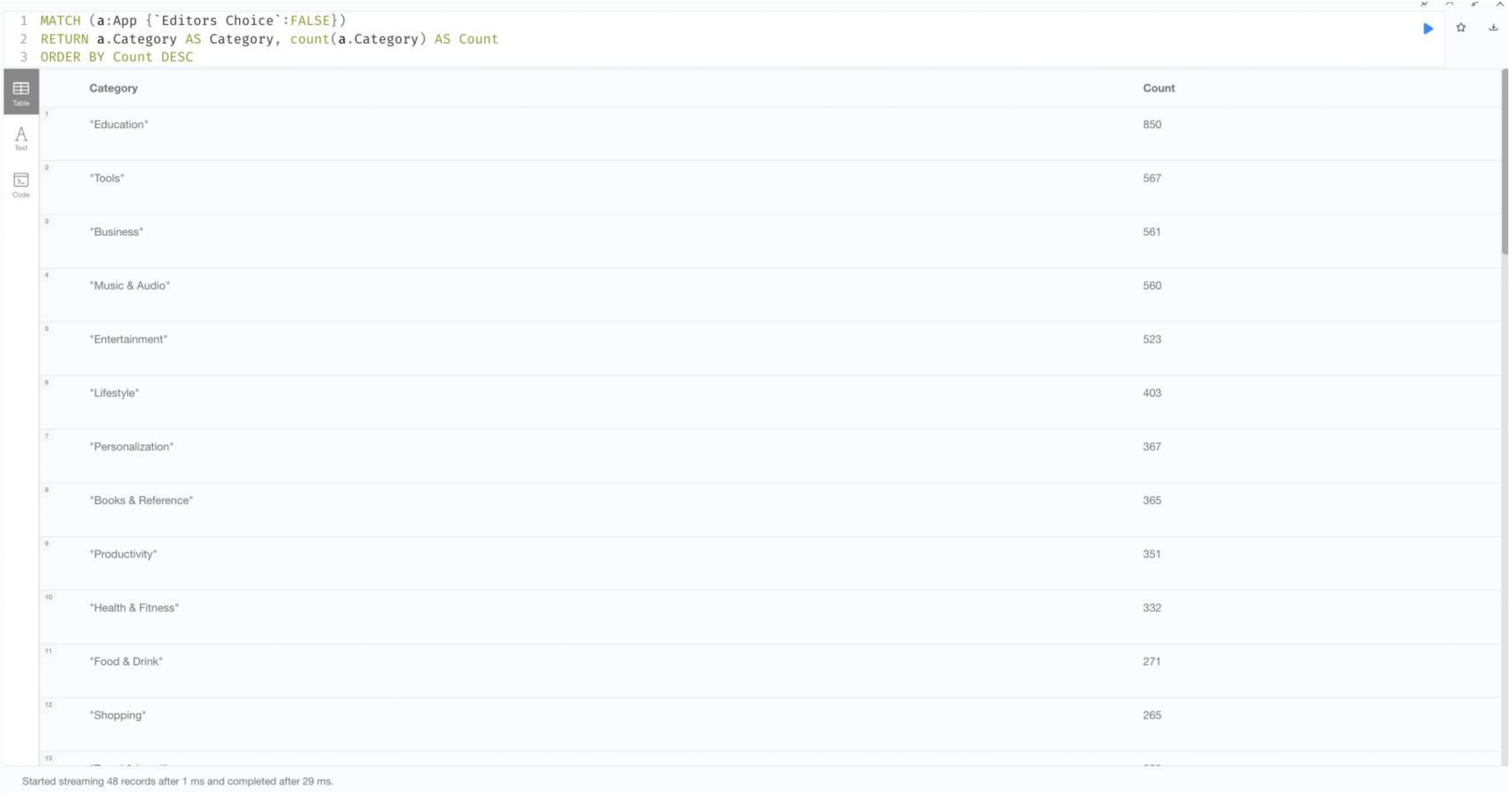


```
SELECT * FROM (
    SELECT ad.Category, (sum(cast(ad.Installs as Unsigned ))) as Total, (sum(r.rating) / count(r.rating)) as AvgRating FROM app_data ad
    INNER JOIN rating r
    on ad.id = r.AppId
    GROUP BY ad.Category
) as t
ORDER BY t.Total DESC, t.AvgRating ASC LIMIT 5
 /*Priority given to Installs*/
```

Category	Total	AvgRating
Video Players & Editors	115929410	2.0035087518524706
Sports	112173567	2.26019900355173
Tools	82380139	2.1008787352087626
Arcade	63869793	2.2924242477224333
Simulation	51863165	1.905000001192093

User 1 Query 2 - Neo4j

Question : In which category one should develop app to be in the editor’s choice?



```
MATCH (a:App {`Editors Choice`:FALSE})
RETURN a.Category AS Category, count(a.Category) AS Count
ORDER BY Count DESC
```

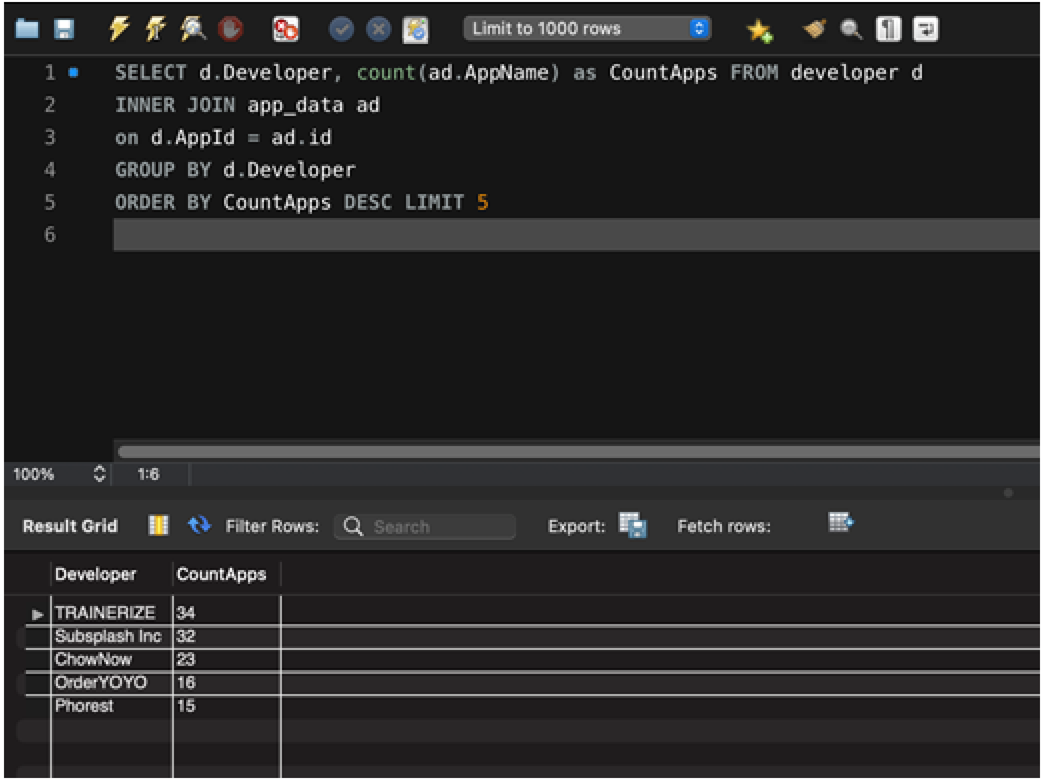
	Category	Count
1	"Education"	850
2	"Tools"	567
3	"Business"	561
4	"Music & Audio"	560
5	"Entertainment"	523
6	"Lifestyle"	403
7	"Personalization"	367
8	"Books & Reference"	365
9	"Productivity"	351
10	"Health & Fitness"	332
11	"Food & Drink"	271
12	"Shopping"	265
13		

Started streaming 48 records after 1 ms and completed after 29 ms.

User 1 Query 3

Question : Whom to hire for developing the app?

- SELECT d.Developer, count(ad.AppName) as CountApps FROM developer d
  INNER JOIN app_data ad
  on d.AppId = ad.id
  GROUP BY d.Developer
  ORDER BY CountApps DESC LIMIT 5



```
SELECT d.Developer, count(ad.AppName) as CountApps FROM developer d
INNER JOIN app_data ad
on d.AppId = ad.id
GROUP BY d.Developer
ORDER BY CountApps DESC LIMIT 5
```

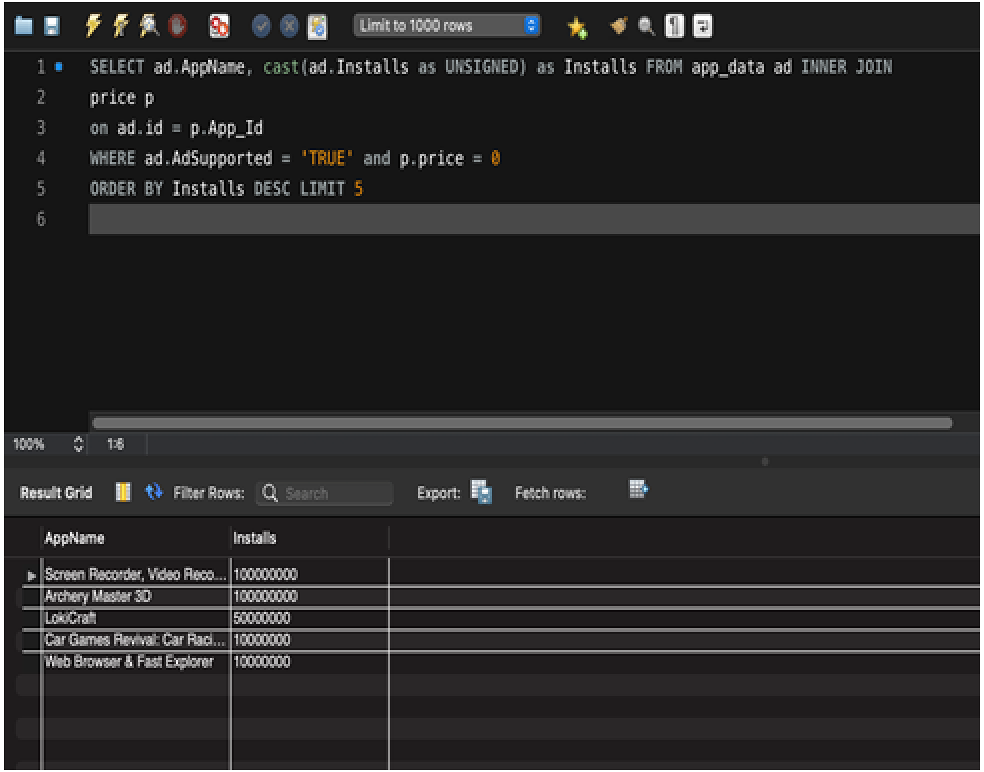
Developer	CountApps
TRAINERIZE	34
Subsplash Inc	32
ChowNow	23
OrderYOYO	16
Phorest	15

User 2 Query 1

Question : Which app can be used to advertise their product?

- SELECT ad.AppName, cast(ad.Installs as UNSIGNED) as Installs FROM app_data ad INNER JOIN
  price p
  on ad.id = p.App_Id
  WHERE ad.AdSupported = 'TRUE' and p.price = 0
  ORDER BY Installs DESC LIMIT 5

- In Neo4j:
  MATCH (a:App{Price:0.0,`Ad Supported`:TRUE})
  RETURN a.`App Name`,a.Installs
  ORDER BY a.Installs DESC
  LIMIT 5



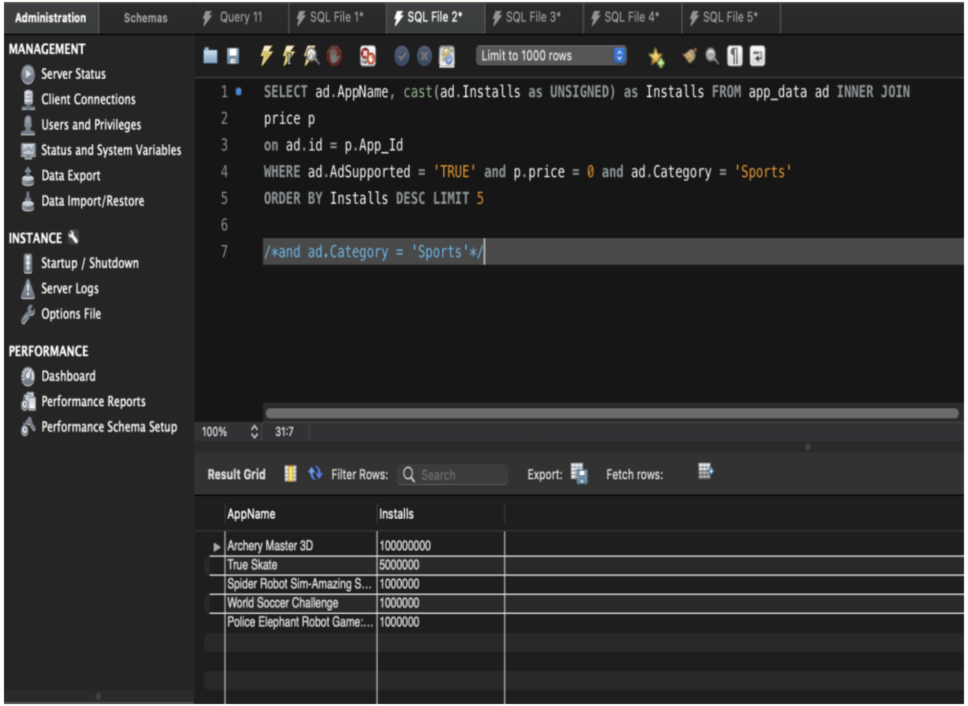
```sql
SELECT ad.AppName, cast(ad.Installs as UNSIGNED) as Installs FROM app_data ad INNER JOIN
price p
on ad.id = p.App_Id
WHERE ad.AdSupported = 'TRUE' and p.price = 0
ORDER BY Installs DESC LIMIT 5
```

AppName	Installs
Screen Recorder, Video Reco...	100000000
Archery Master 3D	100000000
LokiCraft	50000000
Car Games Revival: Car Raci...	10000000
Web Browser & Fast Explorer	10000000

User 2 Query 2

Question : Promote products on apps which align with product category

- SELECT ad.AppName, cast(ad.Installs as UNSIGNED) as Installs FROM app_data ad INNER JOIN
- price p
- on ad.id = p.App_Id
- WHERE ad.AdSupported = 'TRUE' and p.price = 0 and ad.Category = 'Sports'
- ORDER BY Installs DESC LIMIT 5



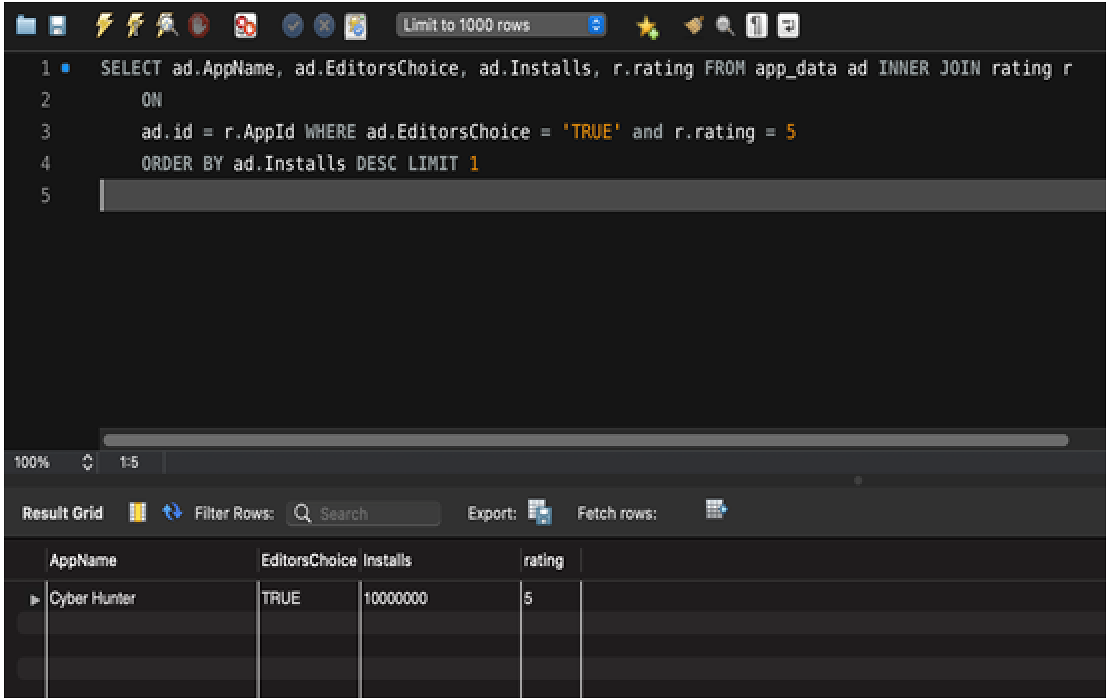
```
SELECT ad.AppName, cast(ad.Installs as UNSIGNED) as Installs FROM app_data ad INNER JOIN
price p
on ad.id = p.App_Id
WHERE ad.AdSupported = 'TRUE' and p.price = 0 and ad.Category = 'Sports'
ORDER BY Installs DESC LIMIT 5

/*and ad.Category = 'Sports'*/
```

AppName	Installs
Archery Master 3D	100000000
True Skate	5000000
Spider Robot Sim-Amazing S...	1000000
World Soccer Challenge	1000000
Police Elephant Robot Game:...	1000000

User 3 Query 1

Question : Which apps performed the best and give them Best App Award.

- SELECT ad.AppName, ad.EditorsChoice,
  ad.Installs, r.rating FROM app_data ad
  INNER JOIN rating r
  ON ad.id = r.AppId
  WHERE ad.EditorsChoice = 'TRUE'
  AND r.rating = 5
  ORDER BY ad.Installs DESC
  LIMIT 1



Limit to 1000 rows

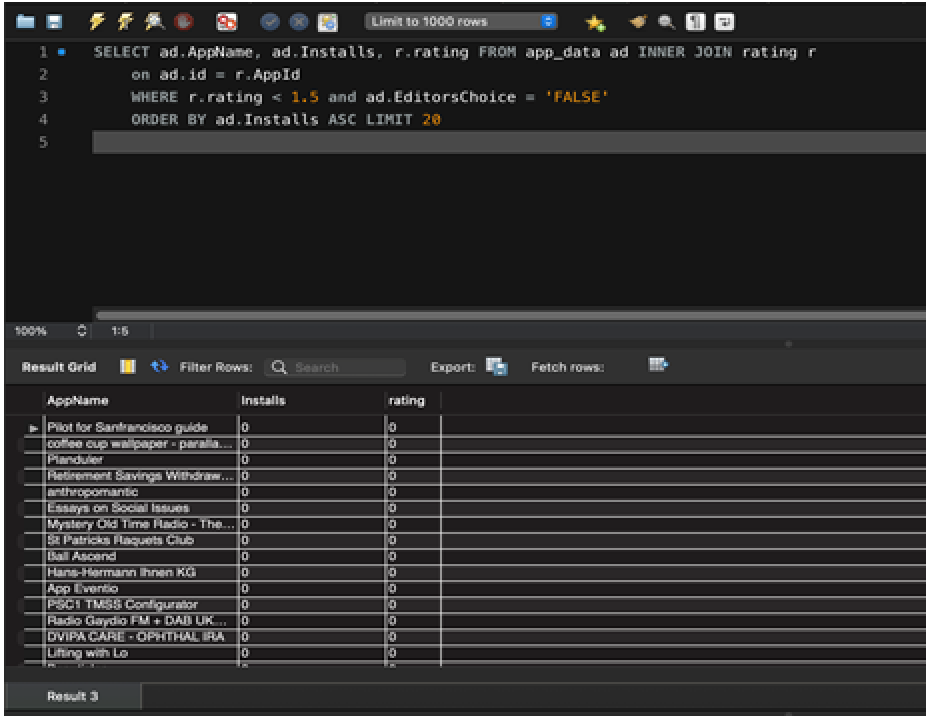
```
SELECT ad.AppName, ad.EditorsChoice, ad.Installs, r.rating FROM app_data ad INNER JOIN rating r
    ON
    ad.id = r.AppId WHERE ad.EditorsChoice = 'TRUE' and r.rating = 5
    ORDER BY ad.Installs DESC LIMIT 1
```

AppName	EditorsChoice	Installs	rating
Cyber Hunter	TRUE	10000000	5

User 3 Query 2

Question : Which top 20 apps didn’t perform well remove them from play-store.

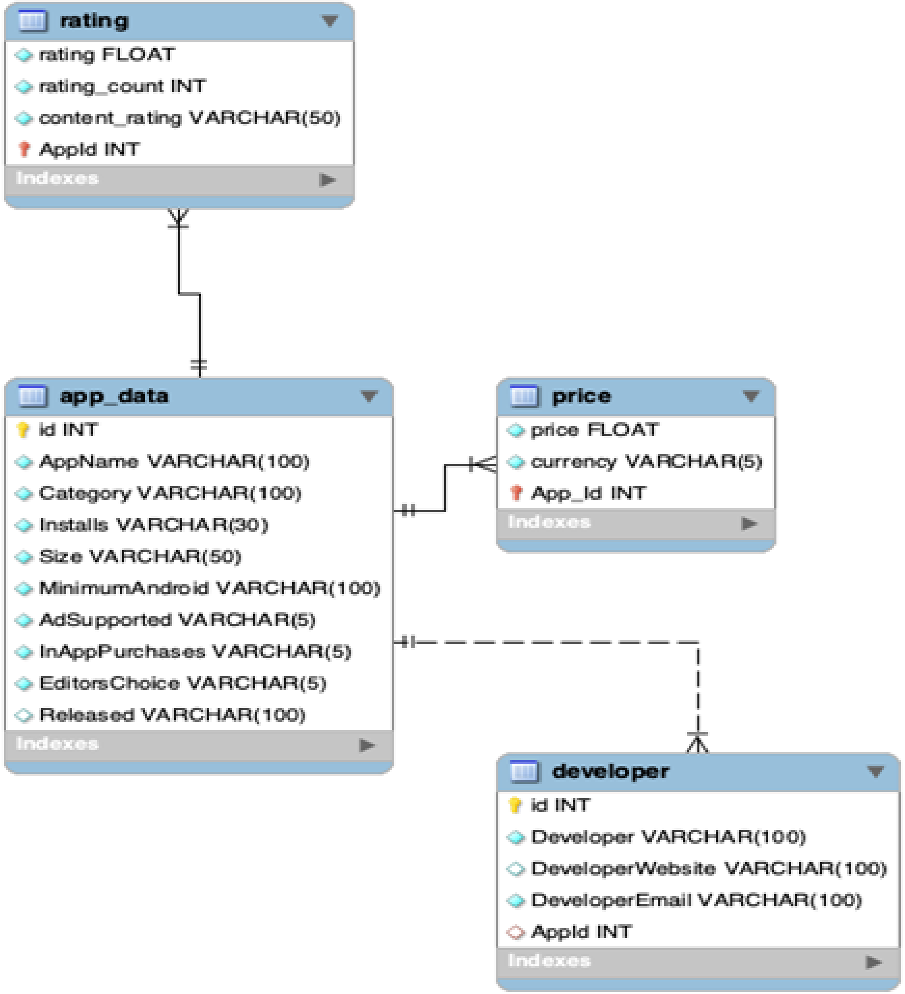
- SELECT ad.AppName, ad.Installs, r.rating
  FROM app_data ad INNER JOIN rating r
  ON ad.id = r.AppId
  WHERE r.rating < 1.5 AND
  ad.EditorsChoice = 'FALSE'
  ORDER BY ad.Installs ASC LIMIT 20



```
SELECT ad.AppName, ad.Installs, r.rating FROM app_data ad INNER JOIN rating r
    on ad.id = r.AppId
    WHERE r.rating < 1.5 and ad.EditorsChoice = 'FALSE'
    ORDER BY ad.Installs ASC LIMIT 20
```

AppName	Installs	rating
Pilot for Sanfrancisco guide	0	0
coffee cup wallpaper - paralla...	0	0
Planduler	0	0
Retirement Savings Withdraw...	0	0
anthropomantic	0	0
Essays on Social Issues	0	0
Mystery Old Time Radio - The...	0	0
St Patricks Raquets Club	0	0
Ball Ascend	0	0
Hans-Hermann Ihnen KG	0	0
App Eventio	0	0
PSC1 TMSS Configurator	0	0
Radio Gaydio FM + DAB UK...	0	0
DVIPA CARE - OPHTHAL IRA	0	0
Lifting with Lo	0	0

Reverse Engineering Diagram



rating
- rating FLOAT
- rating_count INT
- content_rating VARCHAR(50)
- AppId INT

app_data
- id INT
- AppName VARCHAR(100)
- Category VARCHAR(100)
- Installs VARCHAR(30)
- Size VARCHAR(50)
- MinimumAndroid VARCHAR(100)
- AdSupported VARCHAR(5)
- InAppPurchases VARCHAR(5)
- EditorsChoice VARCHAR(5)
- Released VARCHAR(100)

price
- price FLOAT
- currency VARCHAR(5)
- App_Id INT

developer
- id INT
- Developer VARCHAR(100)
- DeveloperWebsite VARCHAR(100)
- DeveloperEmail VARCHAR(100)
- AppId INT

Thank You !