A blue and black logo

Description automatically generated

**درس پایگاه داده پيشرفته**

**گزارش پروژه**

**تحلیل داده‌های Google Play Store**

**نام استاد: دكتر نعمت بخش**

**نام دانشجو: على ابراهيمى**

**شماره دانشجویی:‌ 4033644001**

**بهمن ١٤٠٣**

**بخش اول: جمع‌آوری و آماده‌سازی داده‌ها**

در این پروژه، داده‌های مربوط به Google Play Store از Kaggle دانلود و سپس برای استفاده در پایگاه داده PostgreSQL، پیش پردازش و آماده شدند.

**۱. دانلود مجموعه داده‌ها از Kaggle**

ابتدا، مجموعه داده‌ای از سایت Kaggle دانلود شد که شامل اطلاعات اپلیکیشن‌های موجود در Google Play Store بود.

**۲. تجزیه و تحلیل داده‌های گم‌شده**

در این مرحله، ابتدا بررسی شد که کدام ستون‌ها دارای مقادیر گم‌شده هستند و میزان گم‌شدن داده‌ها برای هر ستون محاسبه شد. این اطلاعات به ما کمک کرد تا تصمیم بگیریم که کدام داده‌ها باید پر شوند و کدام‌ها باید حذف شوند. پس از شناسایی داده‌های گم‌شده، داده‌هایی که برای تحلیل‌های اصلی ضروری بودند، اصلاح شدند.

**۳. حذف رکوردهای تکراری یا نامعتبر**

رکوردهای تکراری که ممکن بود چندین بار برای یک اپلیکیشن ثبت شده باشند، شناسایی و حذف شدند. همچنین رکوردهایی که شامل داده‌های نادرست یا ناقص در فیلدهای حیاتی مانند نام اپلیکیشن، شناسه توسعه‌دهنده، تعداد نصب‌ها و حجم بودند، حذف شدند.

**۴. استانداردسازی فرمت داده‌ها**

در این مرحله، فرمت داده‌ها به‌طور یکنواخت استاندارد شد تا وارد پایگاه داده PostgreSQL شوند. مهم‌ترین تغییرات شامل:

**تبدیل تاریخ‌ها:** مقادیر تاریخ‌های مربوط به تاریخ انتشار و آخرین به‌روزرسانی به فرمت استاندارد YYYY-MM-DD تبدیل شدند.

**تبدیل داده‌های عددی:** مقادیر فیلدهایی مانند تعداد نصب‌ها و قیمت که به صورت متنی ذخیره شده بودند، به مقادیر عددی تبدیل شدند.

**حجم اپلیکیشن‌ها:** مقادیر موجود در فیلد حجم که واحدهای مختلفی داشتند، به مگابایت تبدیل شدند.

**تبدیل مقادیر گم‌شده:** مقادیر گم‌شده در فیلدهای خاص مانند امتیاز اپلیکیشن‌ها، تعداد نصب‌ها، حجم و تاریخ‌ها پر شدند.

**۵. اصلاح داده‌های خاص**

در این مرحله، برخی از داده‌های خاص که شامل مشکلات فرمت بودند، اصلاح شدند. به‌عنوان مثال، نام اپلیکیشن‌ها و شناسه‌های توسعه‌دهندگان که حاوی کوتیشن‌های اضافی بودند، اصلاح شدند.

**۶. ذخیره‌سازی داده‌های تمیز شده**

پس از انجام عملیات پاکسازی و استانداردسازی داده‌ها، اطلاعات تمیز شده در قالب فایل‌های CSV ذخیره شدند. داده‌ها به سه فایل جداگانه تقسیم شدند:

**فایل دسته‌بندی‌ها:** شامل اطلاعات مربوط به تمامی دسته‌بندی‌ها.

**فایل توسعه‌دهندگان:** شامل اطلاعات مربوط به توسعه‌دهندگان اپلیکیشن‌ها.

**فایل اپلیکیشن‌ها:** شامل تمامی ویژگی‌های اپلیکیشن‌ها همراه با شناسه‌های دسته‌بندی‌ها و توسعه‌دهندگان.

**بخش دوم: ایجاد پایگاه داده**

در این بخش، مراحل طراحی پایگاه داده برای ذخیره‌سازی داده‌ها توضیح داده می‌شود.

**۱. طراحی اسکیما**

جدول apps که شامل اطلاعات اپلیکیشن‌هاست و به جداول دسته‌بندی‌ها و توسعه‌دهندگان مرتبط می‌شود.

جدول categories که برای ذخیره اطلاعات دسته‌بندی‌ها است.

جدول developers که اطلاعات توسعه‌دهندگان را نگهداری می‌کند.

**۲. وارد کردن داده**

بعد از طراحی اسکیما و ایجاد جداول، داده‌های تمیز شده که در قالب فایل‌های CSV ذخیره شده بودند، وارد پایگاه داده شدند. داده‌ها شامل اطلاعات اپلیکیشن‌ها، دسته‌بندی‌ها و توسعه‌دهندگان بودند.

**بخش سوم: توسعه APIهای سمت بک‌اند**

**۱. فیلترها و جستجو**

|  |  |
| --- | --- |
| استخراج مقادیر پیشفرض فیلترها از پایگاه داده | GET /filters |
| جستجو و صفحه‌بندی اپلیکیشن‌ها بر اساس فیلترها | GET /apps |

**۲. آمار و تحلیل‌های داده**

|  |  |
| --- | --- |
| GET /statistics/rating\_distribution | محاسبه توزیع امتیازات اپلیکیشن‌ها |
| GET /statistics/release\_trend | محاسبه روند انتشار اپلیکیشن‌ها |
| GET /statistics/update\_trend | محاسبه روند به‌روزرسانی اپلیکیشن‌ها |
| GET /statistics/average\_rating | محاسبه میانگین امتیاز اپلیکیشن‌ها |

**۳. مدیریت دسته‌بندی‌ها**

|  |  |
| --- | --- |
| دریافت لیست تمامی دسته‌بندی‌ها | GET /categories |
| دریافت اطلاعات یک دسته‌بندی خاص | GET /categories/{category\_id} |
| ایجاد دسته‌بندی جدید | POST /categories |
| ویرایش دسته‌بندی موجود | PUT /categories/{category\_id} |
| حذف دسته‌بندی خاص | DELETE /categories/{category\_id} |

**۴. مدیریت توسعه‌دهندگان**

|  |  |
| --- | --- |
| GET /developers | دریافت لیست توسعه‌دهندگان با صفحه‌بندی |
| GET /developers/{developer\_id} | دریافت اطلاعات یک توسعه‌دهنده خاص |
| POST /developers | ایجاد توسعه‌دهنده جدید |
| PUT /developers/{developer\_id} | ویرایش توسعه‌دهنده موجود |
| DELETE /developers/{developer\_id} | حذف توسعه‌دهنده خاص |

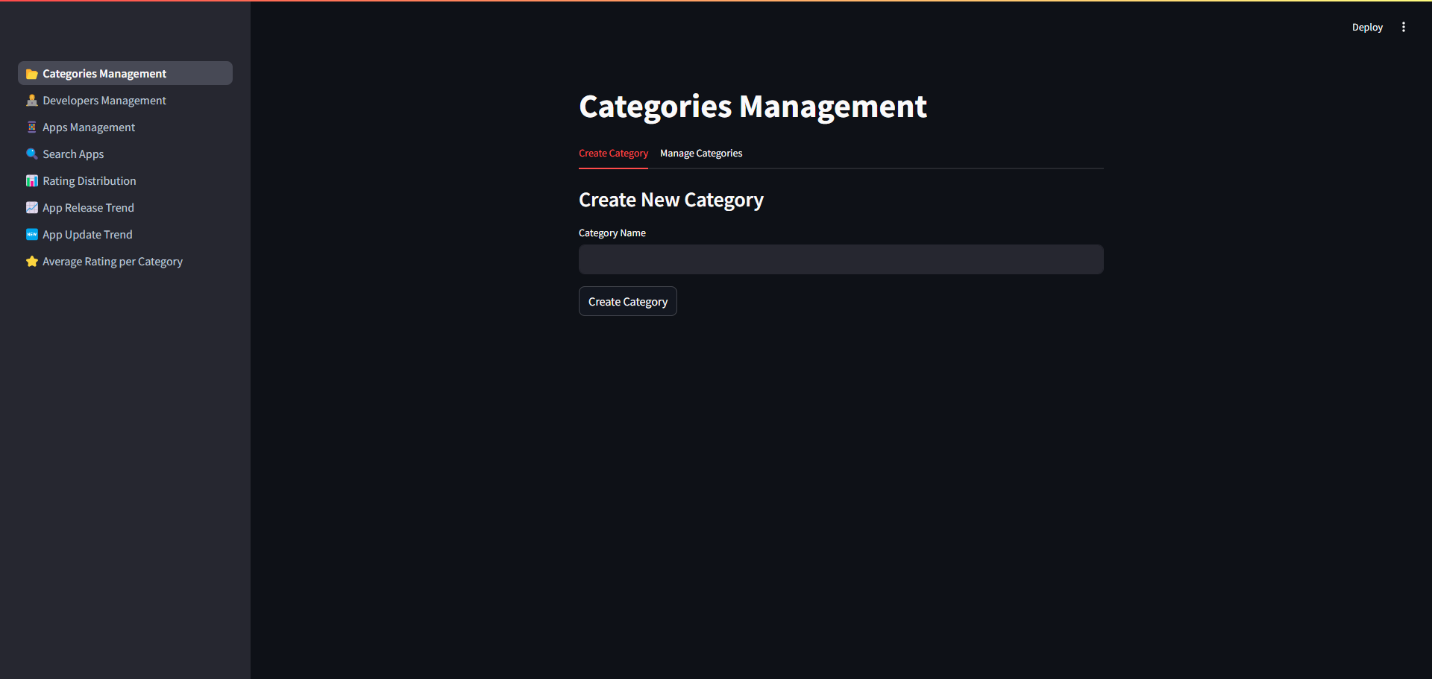
**۵. مدیریت اپلیکیشن‌ها**

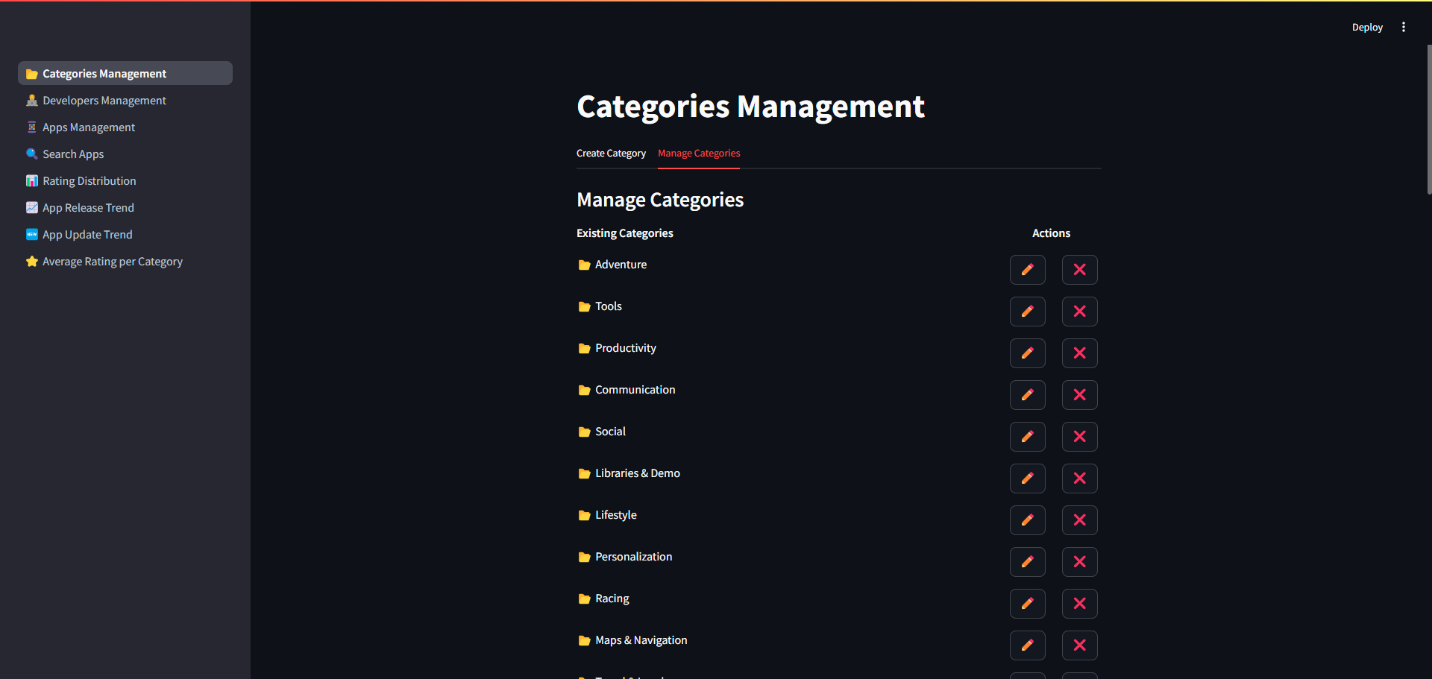
|  |  |
| --- | --- |
| دریافت اطلاعات یک اپلیکیشن خاص | GET /apps/{app\_id} |
| ایجاد اپلیکیشن جدید | POST /apps |
| ویرایش اپلیکیشن موجود | PUT /apps/{app\_id} |
| حذف اپلیکیشن خاص | DELETE /apps/{app\_id} |

**بخش چهارم: توسعه داشبورد با Streamlit**

داشبورد Streamlit شامل چندین صفحه است که به کاربران امکان مدیریت داده‌ها، جستجوی اپلیکیشن‌ها و مشاهده تحلیل‌های مختلف را می‌دهد. در ادامه، صفحات مختلف و قابلیت‌های هر یک توضیح داده شده است.

**۱. مدیریت دسته‌بندی‌ها**

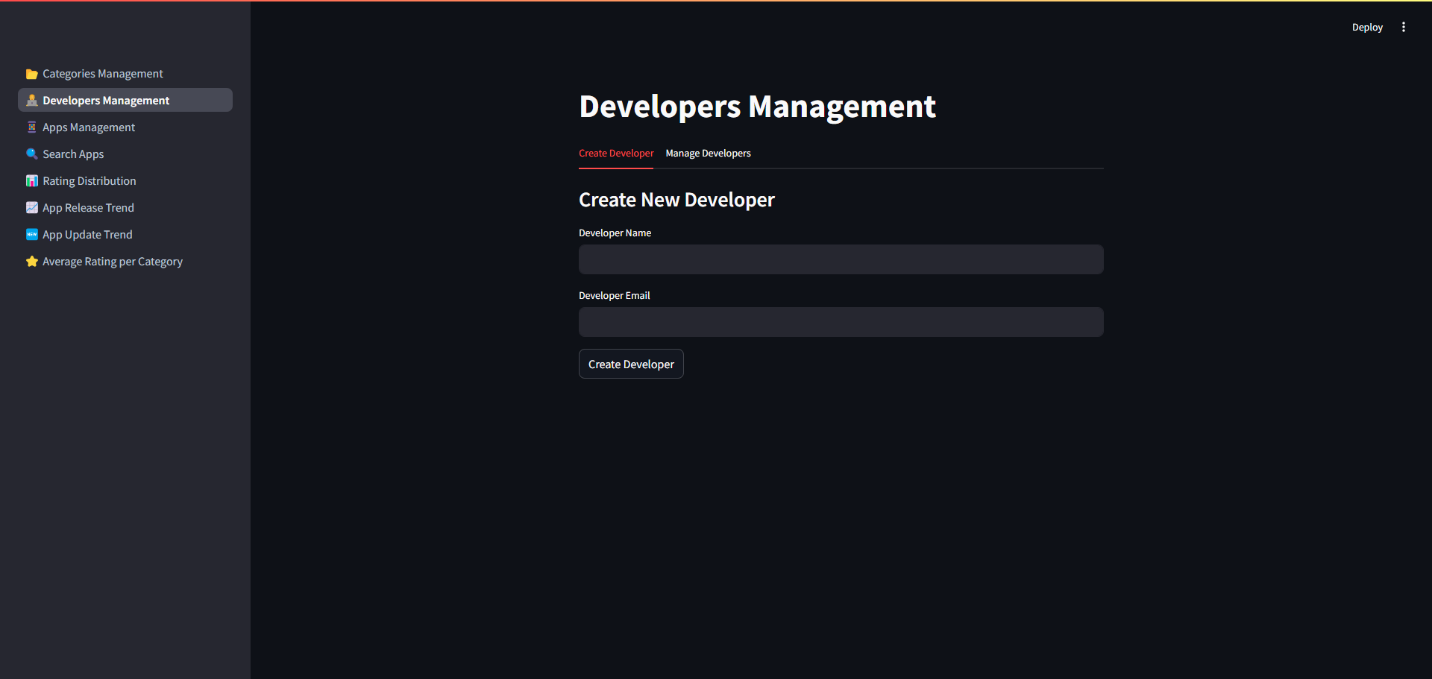
****

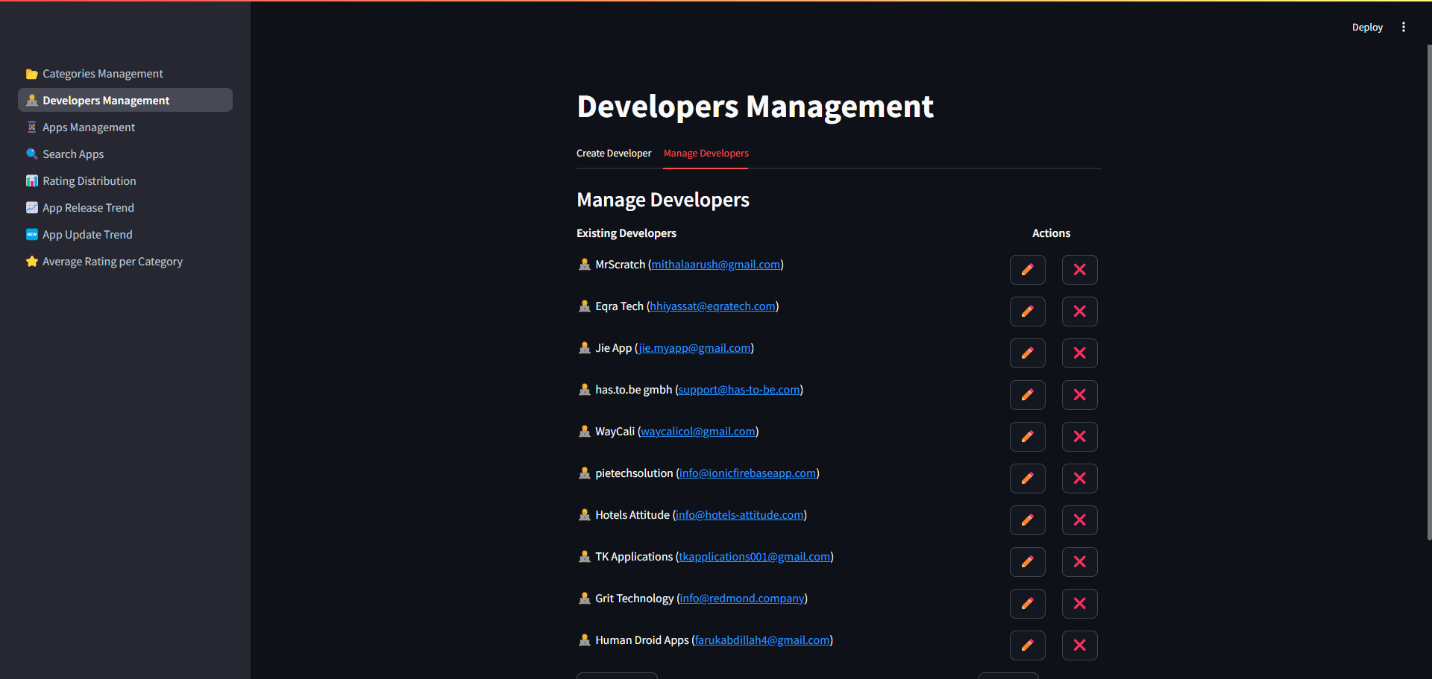
****

**قابلیت‌ها:**

* نمایش لیست دسته‌بندی‌ها
* افزودن دسته‌بندی جدید
* ویرایش نام دسته‌بندی
* حذف دسته‌بندی

**۲. مدیریت توسعه‌دهندگان**

****

****

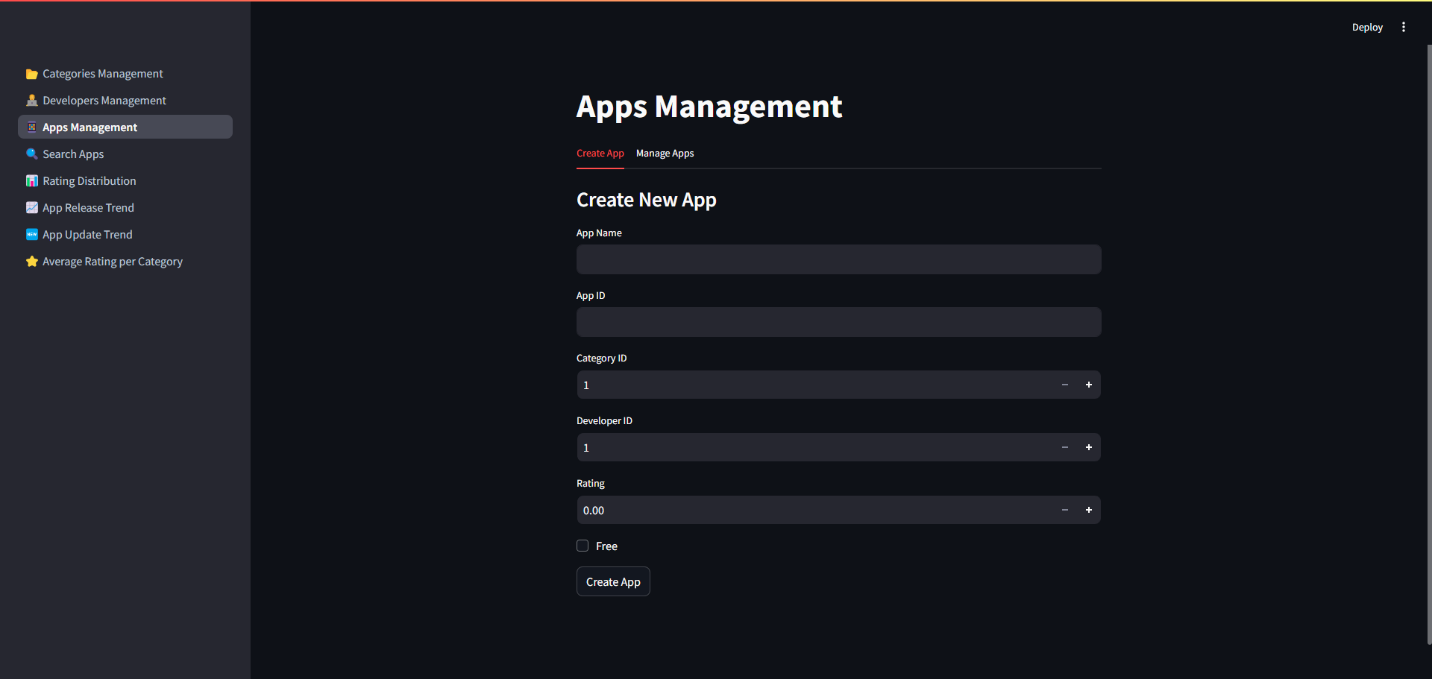
**فیلترها:**

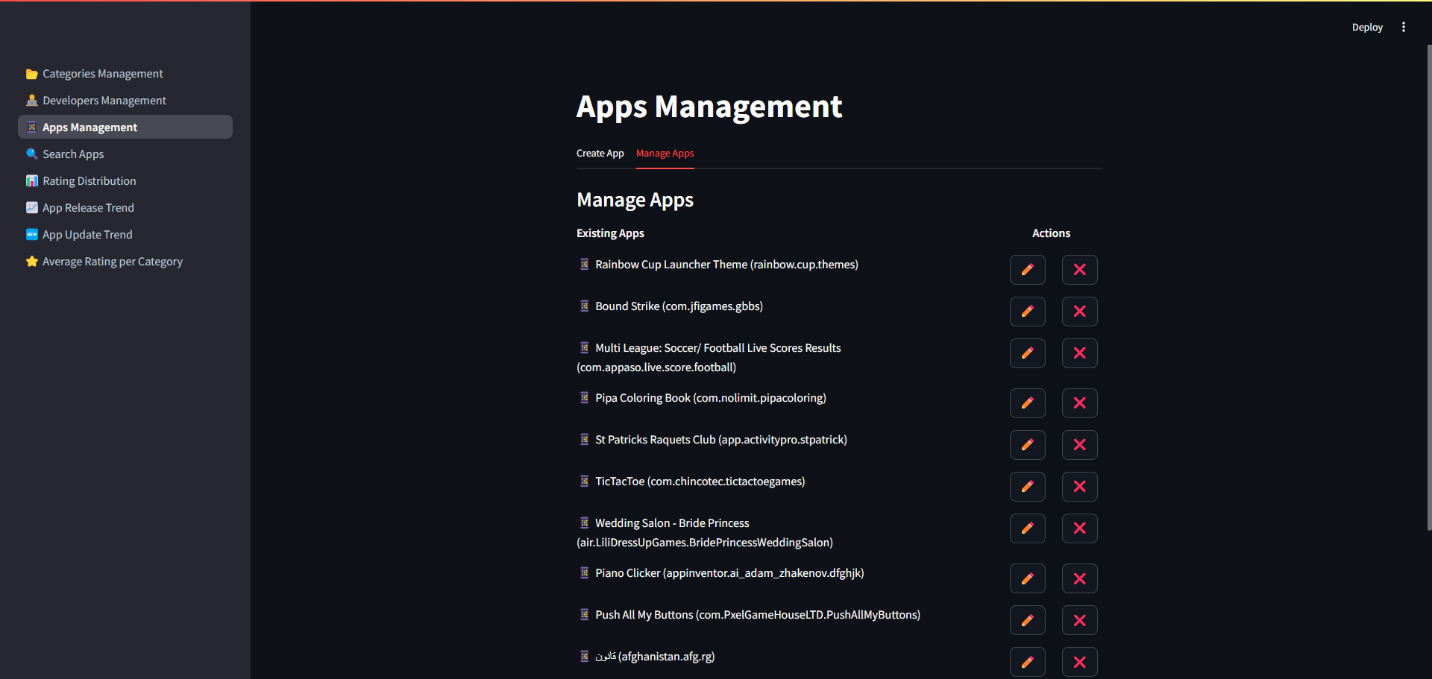
* صفحه‌بندی (شماره صفحه، تعداد نتایج در هر صفحه)

**قابلیت‌ها:**

* نمایش لیست توسعه‌دهندگان
* افزودن توسعه‌دهنده جدید
* ویرایش نام و ایمیل توسعه‌دهنده
* حذف توسعه‌دهنده

**۳. مدیریت اپلیکیشن‌ها**

****

****

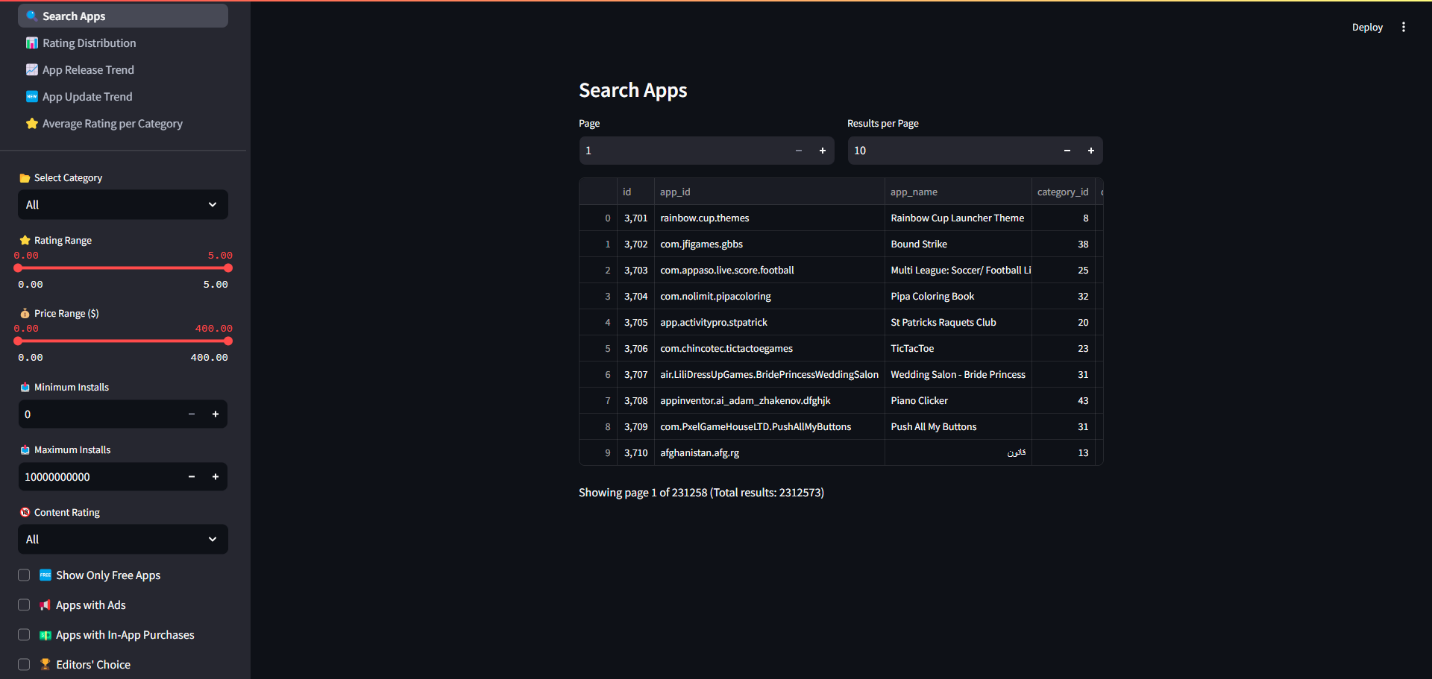
**فیلترها:**

* صفحه‌بندی (شماره صفحه، تعداد نتایج در هر صفحه)

**قابلیت‌ها:**

* نمایش لیست اپلیکیشن‌ها
* افزودن اپلیکیشن جدید
* ویرایش مشخصات اپلیکیشن
* حذف اپلیکیشن

۴**. جستجوی اپلیکیشن‌ها**

****

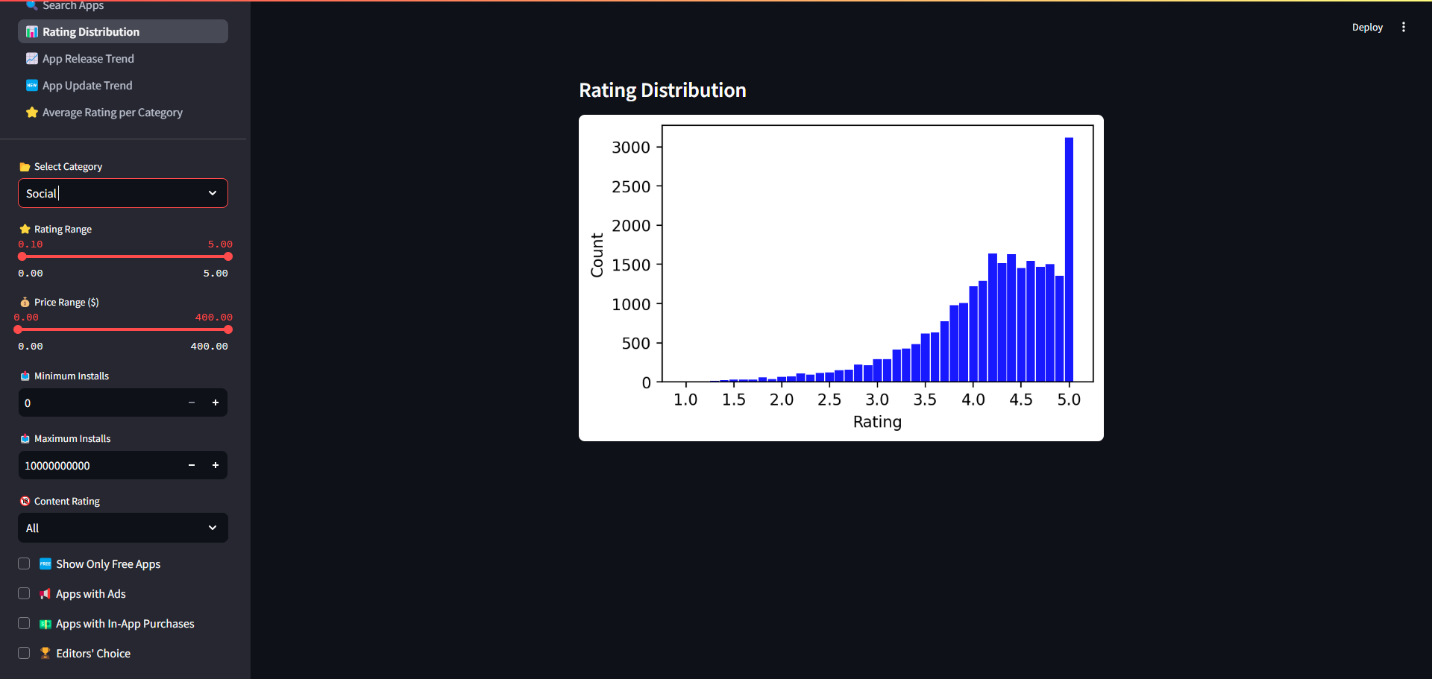
**فیلترها:**

* دسته‌بندی
* محدوده امتیاز
* محدوده قیمت
* محدوده تعداد نصب‌ها
* رده‌بندی سنی
* وضعیت رایگان بودن، پشتیبانی از تبلیغات، خریدهای درون‌برنامه‌ای، انتخاب سردبیر
* صفحه‌بندی (شماره صفحه، تعداد نتایج در هر صفحه)

**قابلیت‌ها:**

* جستجو و نمایش اپلیکیشن‌ها بر اساس فیلترهای انتخابی
* نمایش تعداد کل نتایج و تعداد صفحات

**۵. توزیع امتیازات**

****

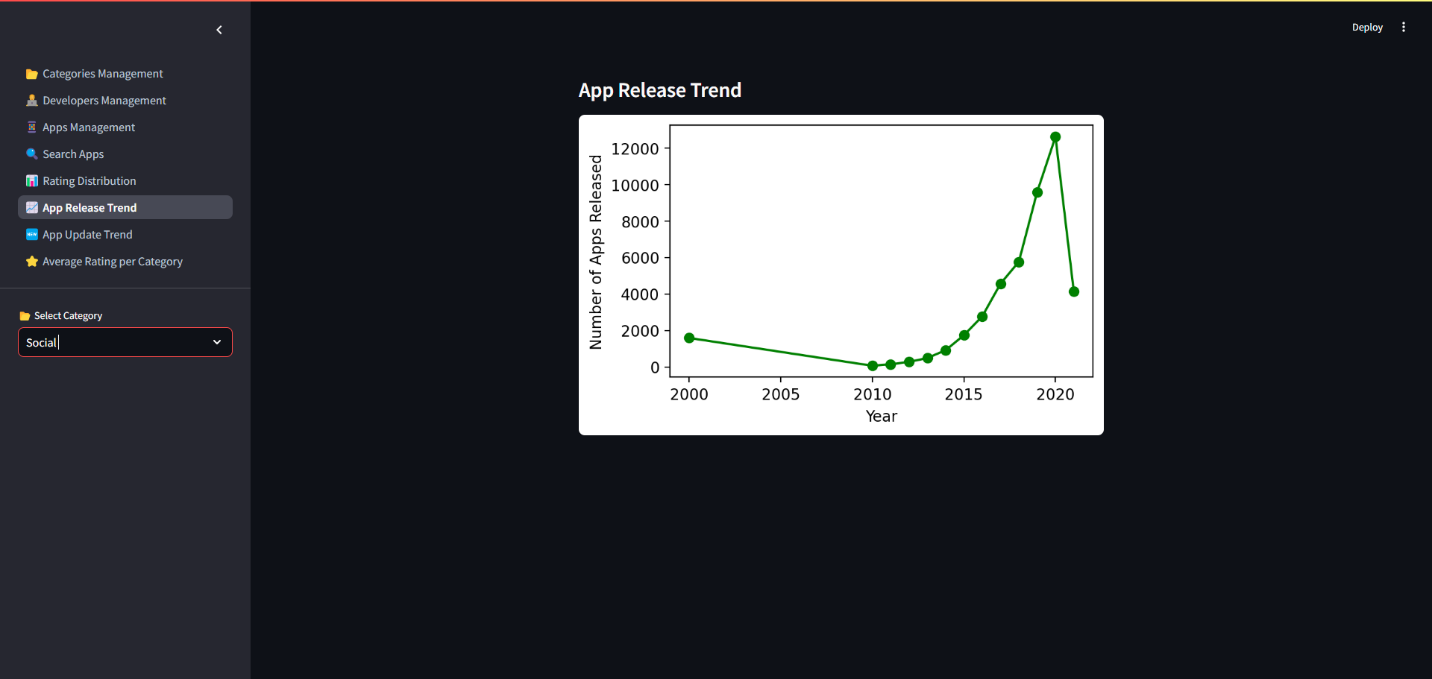
**فیلترها:**

* دسته‌بندی
* محدوده امتیاز
* محدوده قیمت
* محدوده تعداد نصب‌ها
* رده‌بندی سنی
* وضعیت رایگان بودن، پشتیبانی از تبلیغات، خریدهای درون‌برنامه‌ای، انتخاب سردبیر

**نمایش:**

* نمودار میله‌ای توزیع تعداد اپلیکیشن‌ها بر اساس امتیاز

**۶. روند انتشار اپلیکیشن‌ها**

****

**فیلترها:**

* دسته‌بندی

**نمایش:**

* نمودار خطی تعداد اپلیکیشن‌های منتشر شده در هر سال

**۷. روند به‌روزرسانی اپلیکیشن‌ها**

**A screen shot of a graph

Description automatically generated**

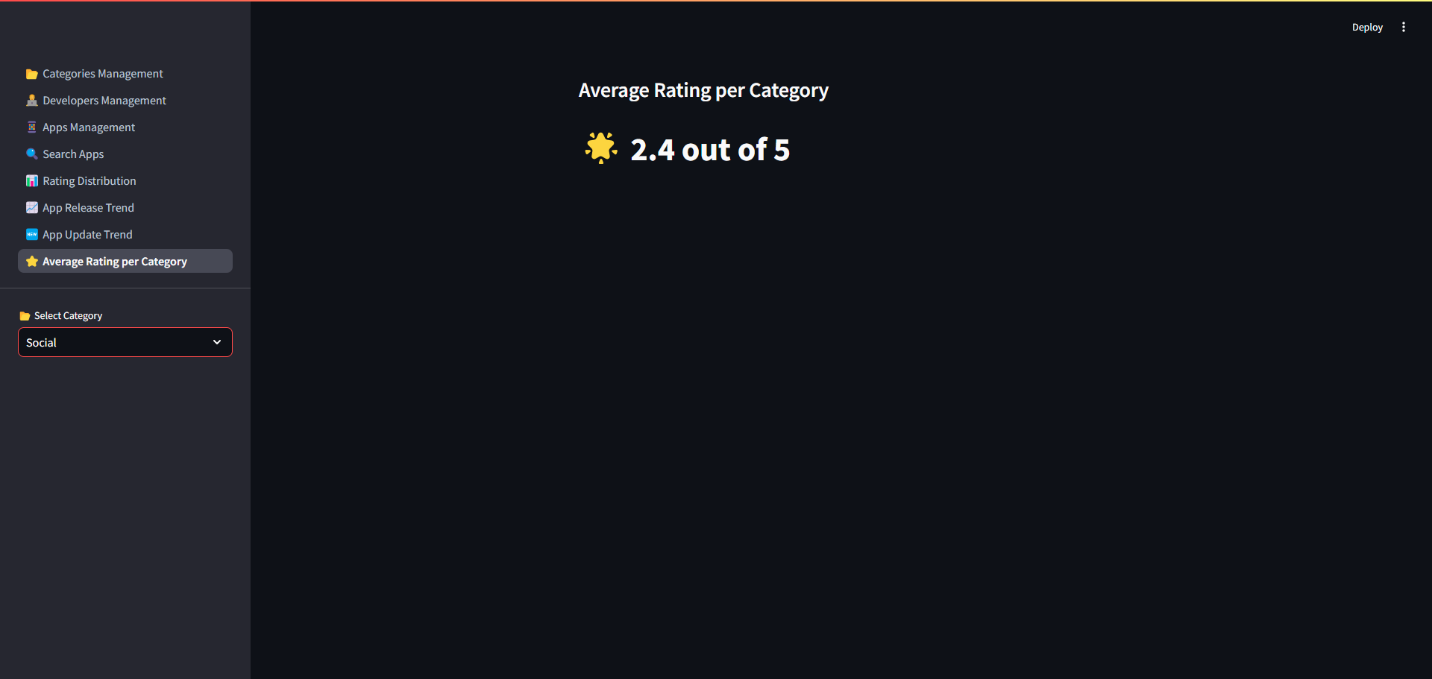
**فیلترها:**

* دسته‌بندی

**نمایش:**

* نمودار خطی تعداد اپلیکیشن‌هایی که در هر سال به‌روزرسانی شده‌اند

**۸. میانگین امتیاز اپلیکیشن‌ها**

****

**فیلترها:**

* دسته‌بندی

**نمایش:**

* مقدار میانگین امتیاز اپلیکیشن‌های یک دسته‌بندی خاص یا همه دسته‌بندی‌ها

**بخش پنجم: بهینه‌سازی پایگاه داده**

در این بخش، ایندکس‌های ایجادشده در پایگاه داده PostgreSQL و تاثیر آن‌ها بر عملکرد جستجوها بررسی می‌شود. هر ایندکس برای بهبود سرعت فیلترها، مرتب‌سازی‌ها و نمودارهای تحلیلی در داشبورد Streamlit طراحی شده است.

**۱. ایندکس‌های تک‌ستونی**

idx\_apps\_category\_id (category\_id)

* بهینه‌سازی جستجو و فیلتر اپلیکیشن‌ها بر اساس دسته‌بندی
* بهبود فیلتر دسته‌بندی در جستجوی اپلیکیشن‌ها، توزیع امتیازات، روند انتشار، روند به‌روزرسانی، میانگین امتیاز

idx\_apps\_rating (rating)

* افزایش سرعت مرتب‌سازی و فیلتر اپلیکیشن‌ها بر اساس امتیاز
* بهبود فیلتر امتیاز در جستجوی اپلیکیشن‌ها، توزیع امتیازات، میانگین امتیاز

idx\_apps\_price (price)

* بهبود جستجو و فیلتر اپلیکیشن‌ها بر اساس محدوده قیمت
* بهبود فیلتر قیمت در جستجوی اپلیکیشن‌ها

idx\_apps\_installs (installs)

* افزایش سرعت مرتب‌سازی و جستجوی اپلیکیشن‌ها بر اساس تعداد نصب‌ها.
* بهبود فیلتر نصب‌ها در جستجوی اپلیکیشن‌ها

idx\_apps\_content\_rating (content\_rating)

* بهینه‌سازی جستجو بر اساس امتیاز محتوا (رده‌بندی سنی)
* بهبود فیلتر امتیاز محتوا در جستجوی اپلیکیشن‌ها

idx\_apps\_free (free)

* افزایش سرعت فیلتر اپلیکیشن‌های رایگان یا پولی
* بهبود فیلتر اپلیکیشن‌های رایگان در جستجوی اپلیکیشن‌ها

idx\_apps\_ad\_supported (ad\_supported)

* بهینه‌سازی فیلتر اپلیکیشن‌هایی که تبلیغات دارند
* بهبود فیلتر تبلیغات در جستجوی اپلیکیشن‌ها

idx\_apps\_in\_app\_purchases (in\_app\_purchases)

* افزایش سرعت فیلتر اپلیکیشن‌های دارای خرید درون‌برنامه‌ای
* بهبود فیلتر خرید درون‌برنامه‌ای در جستجوی اپلیکیشن‌ها

idx\_apps\_editors\_choice (editors\_choice)

* بهینه‌سازی جستجو برای نمایش اپلیکیشن‌های منتخب سردبیر
* بهبود فیلتر انتخاب سردبیر در جستجوی اپلیکیشن‌ها

۲. **ایندکس‌های چندستونی**

idx\_apps\_category\_released (category\_id, released)

* افزایش سرعت جستجو بر اساس ترکیب دسته‌بندی و تاریخ انتشار
* بهبود روند انتشار اپلیکیشن‌ها

idx\_apps\_category\_last\_updated (category\_id, last\_updated)

* بهبود جستجو برای ترکیب دسته‌بندی و تاریخ آخرین به‌روزرسانی
* بهبود روند به‌روزرسانی اپلیکیشن‌ها

idx\_apps\_category\_rating (category\_id, rating)

* بهینه‌سازی مرتب‌سازی و فیلتر اپلیکیشن‌ها بر اساس دسته‌بندی و امتیاز
* بهبود جستجوی اپلیکیشن‌ها، میانگین امتیاز

**۳. نتایج ایندکس‌ کردن داده‌ها**

در ویدیوی دمو، لاگ زمان اجرای کوئری‌های مختلف روی داده‌های ایندکس شده و همچنین بعد از حذف ایندکس‌ها بررسی شده‌است. با استفاده از ایندکس روی ستون‌ها،‌ میانگین زمان اجرای کوئری‌ها از ۲ تا ۱۰ ثانیه به حدود ۰.۰۰۱ تا ۰.۱ ثانیه رسیده است.

**نمونه‌ها‌:**

**۱. کوئری جستجوی اپلیکیشن‌ها در دسته‌بندی Social با رده‌بندی سنی Teen**

* قبل از اعمال ایندکس

DEBUG:database:Executing SQL Query: SELECT count(\*) AS count\_1

FROM (SELECT apps.id AS apps\_id, apps.app\_id AS apps\_app\_id, apps.app\_name AS apps\_app\_name, apps.category\_id AS apps\_category\_id, apps.developer\_id AS apps\_developer\_id, apps.rating AS apps\_rating, apps.rating\_count AS apps\_rating\_count, apps.installs AS apps\_installs, apps.min\_installs AS apps\_min\_installs, apps.max\_installs AS apps\_max\_installs, apps.free AS apps\_free, apps.price AS apps\_price, apps.currency AS apps\_currency, apps.size AS apps\_size, apps.min\_android AS apps\_min\_android, apps.released AS apps\_released, apps.last\_updated AS apps\_last\_updated, apps.content\_rating AS apps\_content\_rating, apps.ad\_supported AS apps\_ad\_supported, apps.in\_app\_purchases AS apps\_in\_app\_purchases, apps.editors\_choice AS apps\_editors\_choice, apps.scraped\_time AS apps\_scraped\_time

FROM apps

WHERE apps.category\_id = 17 AND apps.content\_rating = Teen) AS anon\_1

DEBUG:database:Query executed in 5.5680 seconds

INFO:     127.0.0.1:59142 - "GET /apps?category=Arcade&content\_rating=Teen&free=False&ad\_supported=False&in\_app\_purchases=False&editors\_choice=False&page=1&per\_page=10 HTTP/1.1" 200 OK

* بعد از اعمال ایندکس

DEBUG:database:Executing SQL Query: SELECT count(\*) AS count\_1

FROM (SELECT apps.id AS apps\_id, apps.app\_id AS apps\_app\_id, apps.app\_name AS apps\_app\_name, apps.category\_id AS apps\_category\_id, apps.developer\_id AS apps\_developer\_id, apps.rating AS apps\_rating, apps.rating\_count AS apps\_rating\_count, apps.installs AS apps\_installs, apps.min\_installs AS apps\_min\_installs, apps.max\_installs AS apps\_max\_installs, apps.free AS apps\_free, apps.price AS apps\_price, apps.currency AS apps\_currency, apps.size AS apps\_size, apps.min\_android AS apps\_min\_android, apps.released AS apps\_released, apps.last\_updated AS apps\_last\_updated, apps.content\_rating AS apps\_content\_rating, apps.ad\_supported AS apps\_ad\_supported, apps.in\_app\_purchases AS apps\_in\_app\_purchases, apps.editors\_choice AS apps\_editors\_choice, apps.scraped\_time AS apps\_scraped\_time

FROM apps

WHERE apps.category\_id = 17 AND apps.content\_rating = Teen) AS anon\_1

DEBUG:database:Query executed in 0.0030 seconds

INFO:     127.0.0.1:59142 - "GET /apps?category=Arcade&content\_rating=Teen&free=False&ad\_supported=False&in\_app\_purchases=False&editors\_choice=False&page=1&per\_page=10 HTTP/1.1" 200 OK

زمان اجرا با بیش از ۱8۰۰ درصد بهبود، ‌از ۵.۵۶۸۰ ثانیه به ۰.۰۰۳۰ ثانیه رسیده است.

**۲. کوئری میانگین امتیاز دسته بندی**

* قبل از اعمال ایندکس

DEBUG:database:Executing SQL Query: SELECT avg(apps.rating) AS avg\_1

FROM apps

WHERE apps.category\_id = 5

DEBUG:database:Query executed in 6.5326 seconds

INFO:     127.0.0.1:64691 - "GET /statistics/average\_rating/?category\_name=Social HTTP/1.1" 200 OK

* بعد از اعمال ایندکس

DEBUG:database:Executing SQL Query: SELECT avg(apps.rating) AS avg\_1

FROM apps

WHERE apps.category\_id = 5

DEBUG:database:Query executed in 0.0040 seconds

INFO:     127.0.0.1:64691 - "GET /statistics/average\_rating/?category\_name=Social HTTP/1.1" 200 OK

زمان اجرا با بیش از ۱6۰۰ درصد بهبود، ‌از 6.5326 ثانیه به ۰.۰۰۴۰ ثانیه رسیده است.

**بخش ششم: اطلاعات تکمیلی**

**۱. لینک ویدیوی دمو:**

<https://drive.google.com/file/d/1r8GfgGVyj56qIhsO5HK8IE80V5kjlx4d/view?usp=sharing>

**۲. لینک ریپوی گیتهاب:**

<https://github.com/AliSK81/google-playstore-analysis>