درس پایگاه داده پیشرفته

گزارش پروژه

تحلیل داده‌های Google Play Store

نام استاد: دکتر نعمت بخش

نام دانشجو: علی ابراهیمی

شماره دانشجویی: ٤٠٣٣٦٤٤٠٠١

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

محاسبه توزیع امتیازات اپلیکیشن‌ها	GET /statistics/rating_distribution
محاسبه روند انتشار اپلیکیشن‌ها	GET /statistics/release_trend
محاسبه روند به‌روزرسانی اپلیکیشن‌ها	GET /statistics/update_trend
محاسبه میانگین امتیاز اپلیکیشن‌ها	GET /statistics/average_rating

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

دریافت لیست توسعه‌دهندگان با صفحه‌بندی	GET /developers
دریافت اطلاعات یک توسعه‌دهنده خاص	GET /developers/{developer_id}
ایجاد توسعه‌دهنده جدید	POST /developers
ویرایش توسعه‌دهنده موجود	PUT /developers/{developer_id}
حذف توسعه‌دهنده خاص	DELETE /developers/{developer_id}

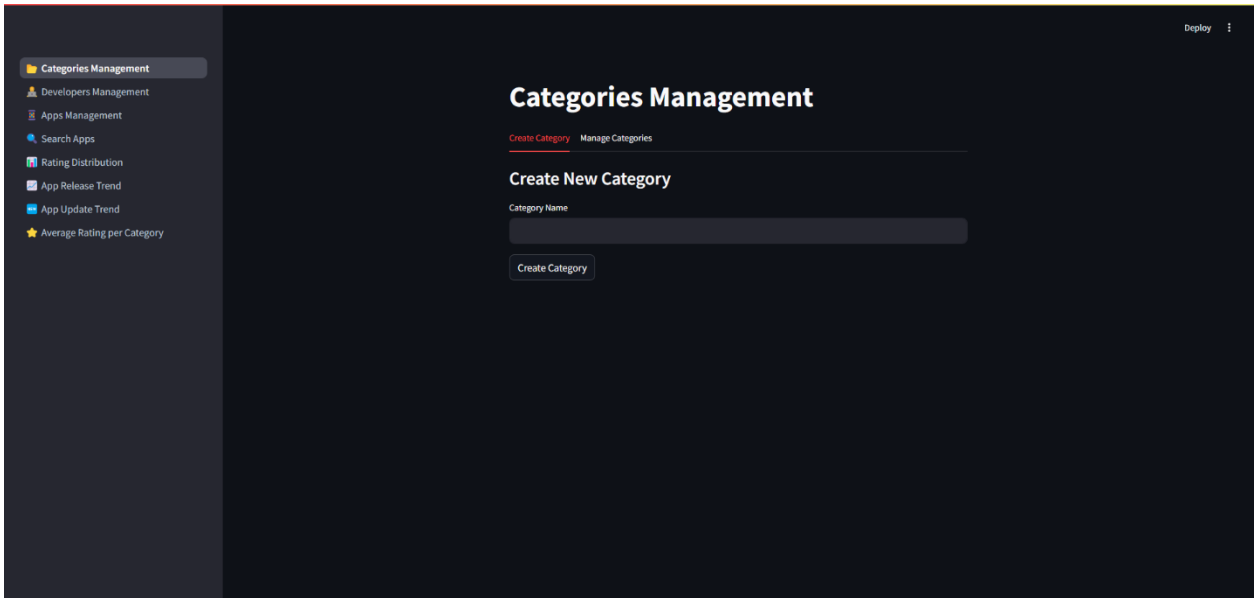
۵. مدیریت اپلیکیشن‌ها

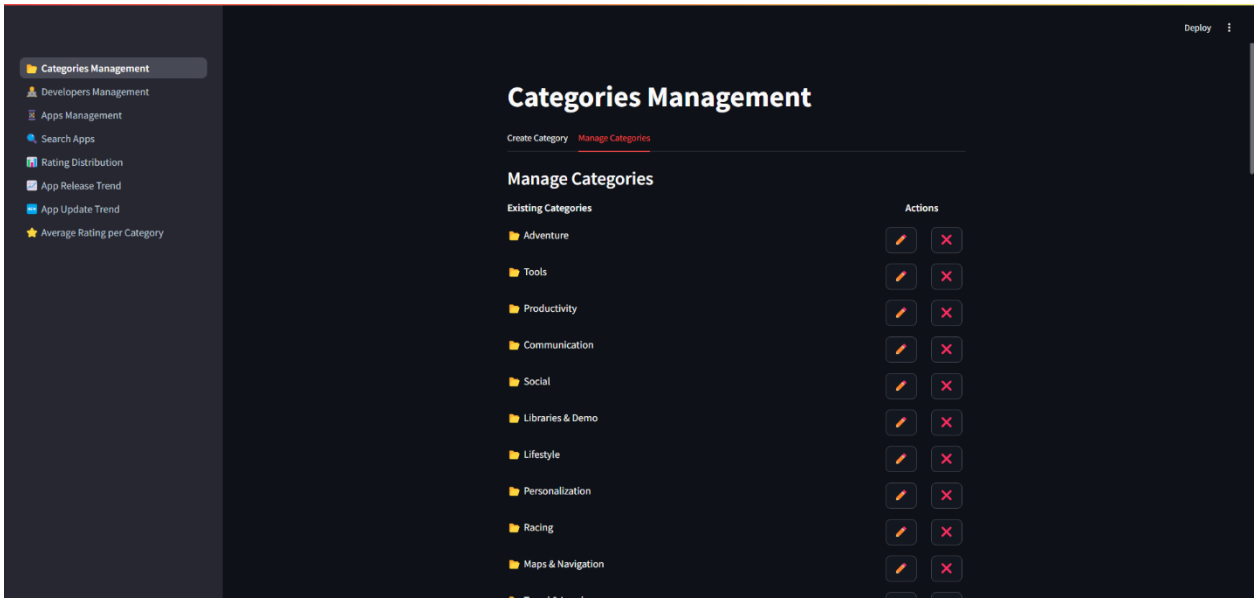
دریافت اطلاعات یک اپلیکیشن خاص	GET /apps/{app_id}
ایجاد اپلیکیشن جدید	POST /apps
ویرایش اپلیکیشن موجود	PUT /apps/{app_id}
حذف اپلیکیشن خاص	DELETE /apps/{app_id}

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

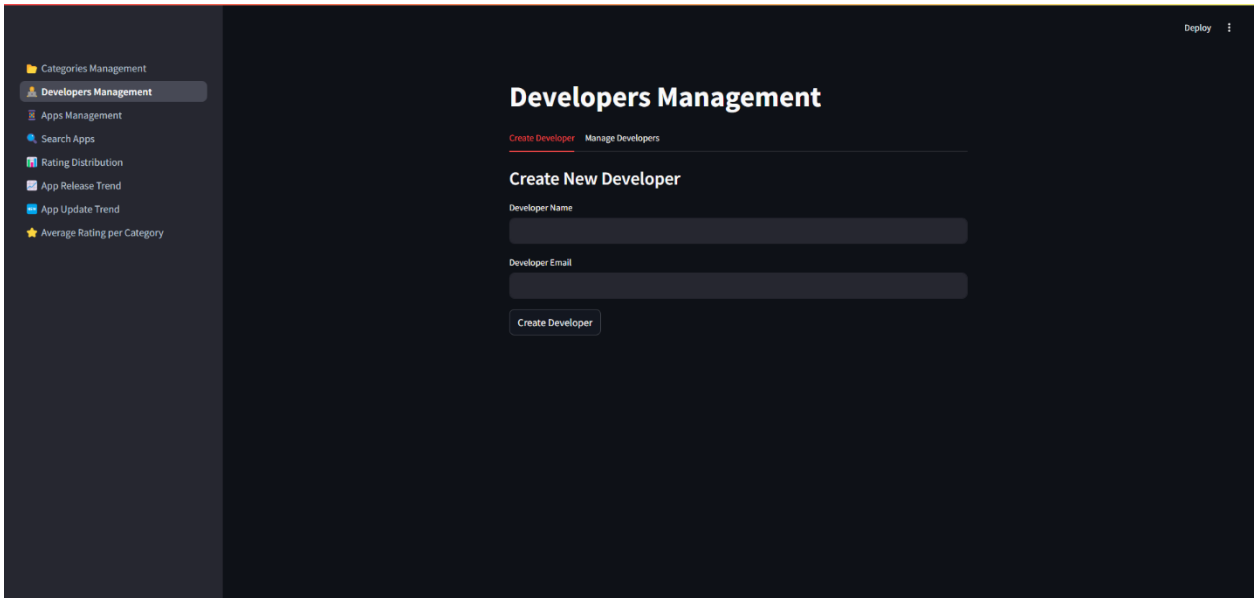


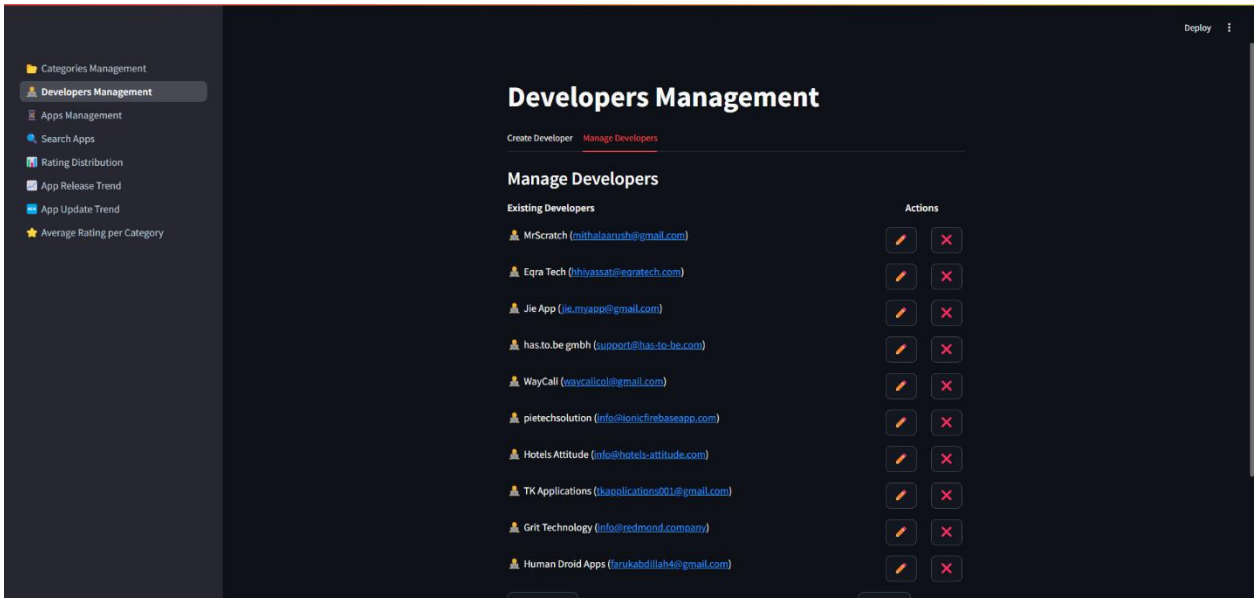


قابلیت‌ها:

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

۲. مدیریت توسعه‌دهندگان





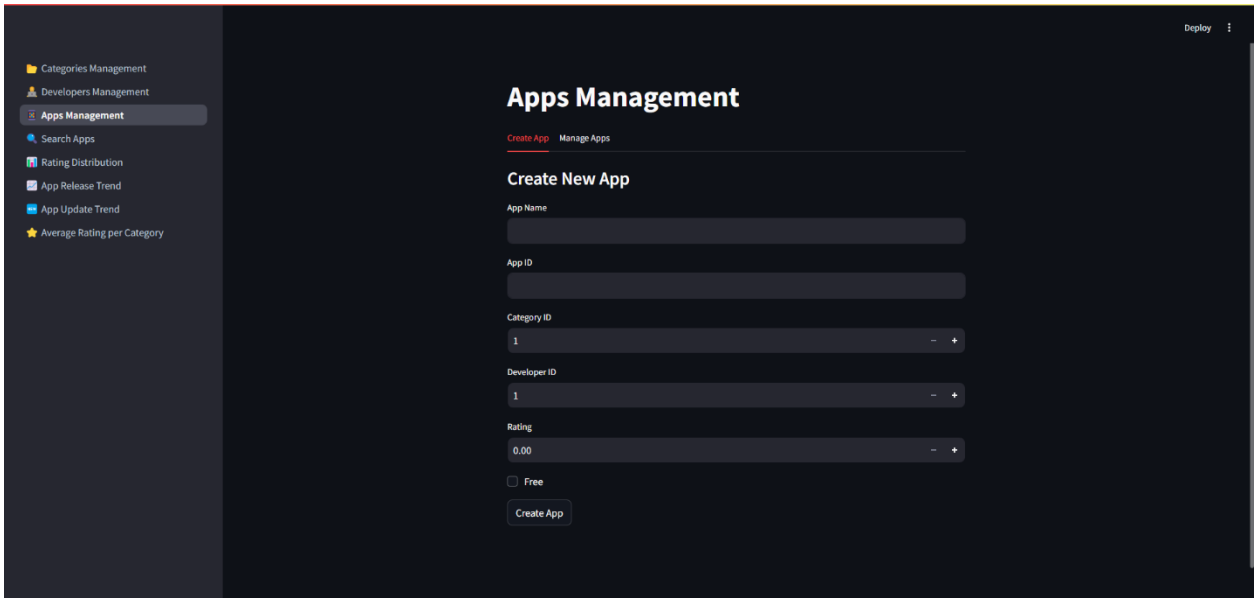
فیلترها:

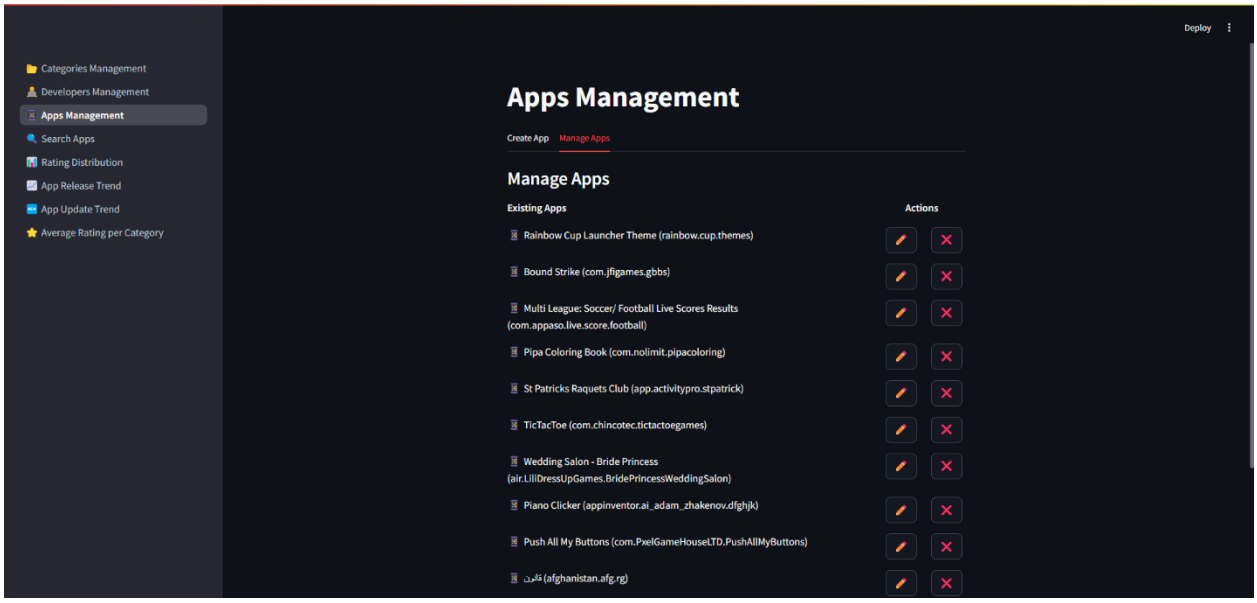
- صفحه‌بندی (شماره صفحه، تعداد نتایج در هر صفحه)

قابلیت‌ها:

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

۳. مدیریت اپلیکیشن‌ها





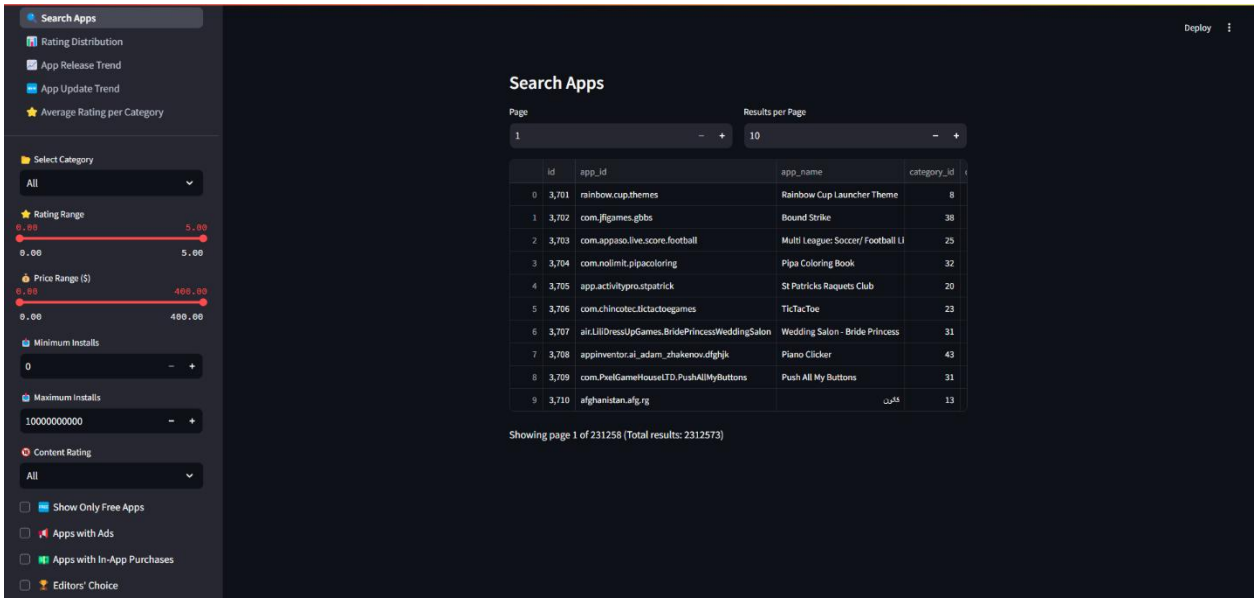
فیلترها:

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

قابلیت‌ها:

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

۴. جستجوی اپلیکیشن‌ها



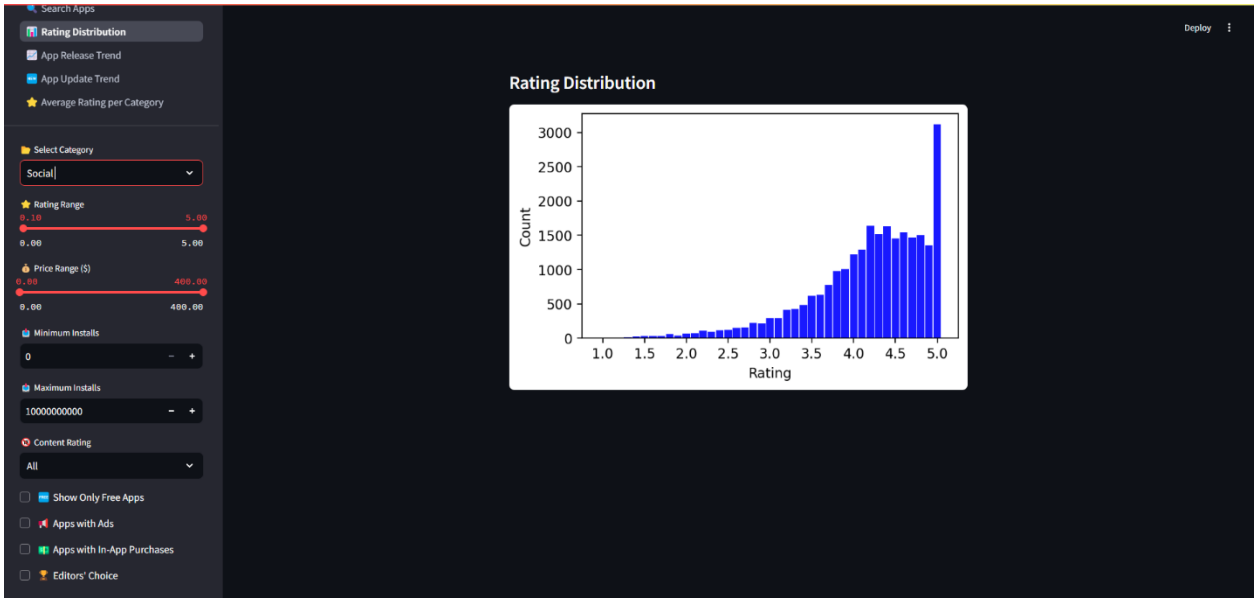
فیلترها:

- 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



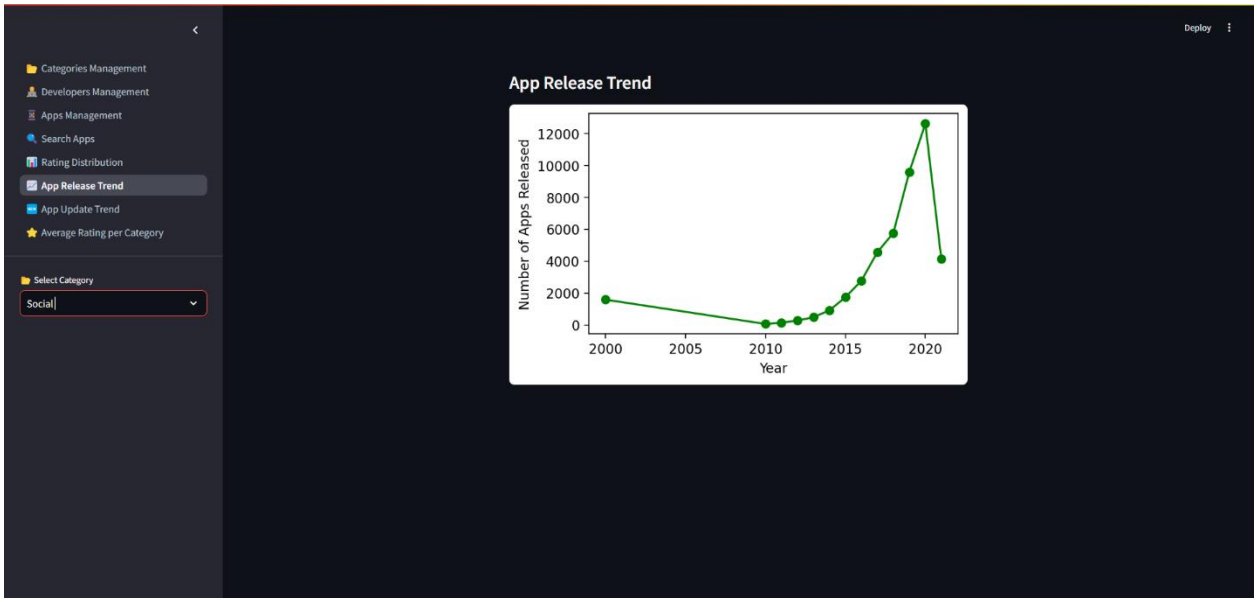
فیلترها:

- 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



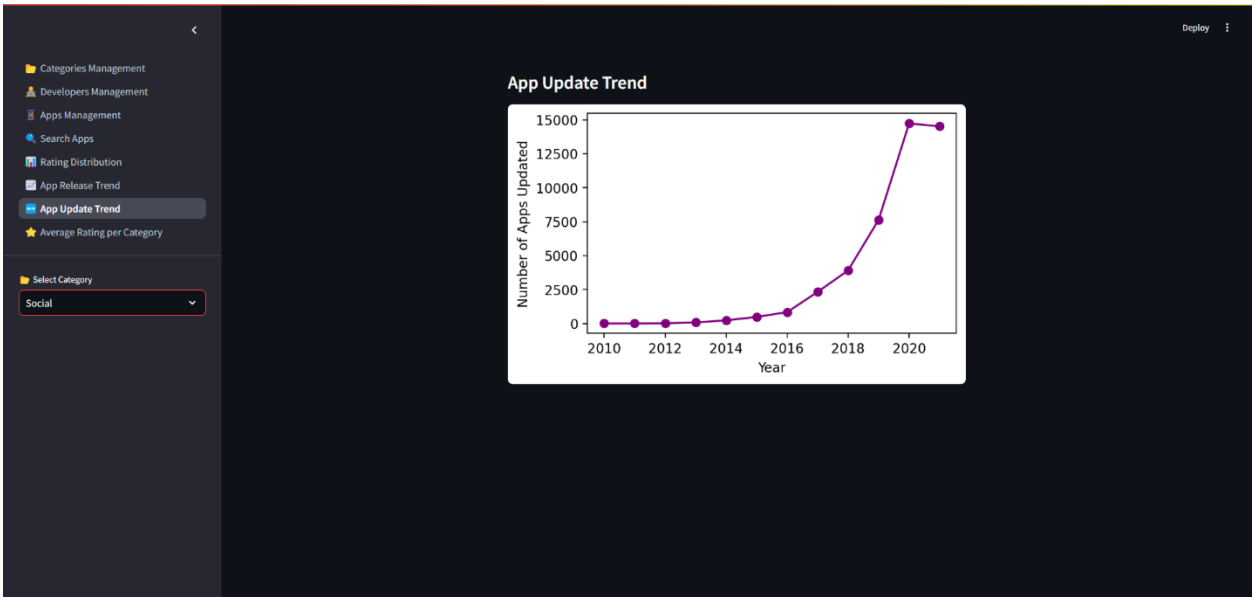
فیلترها:

- دسته‌بندی

نمایش:

- نمودار خطی تعداد اپلیکیشن‌های منتشر شده در هر سال

۷. روند به‌روزرسانی اپلیکیشن‌ها



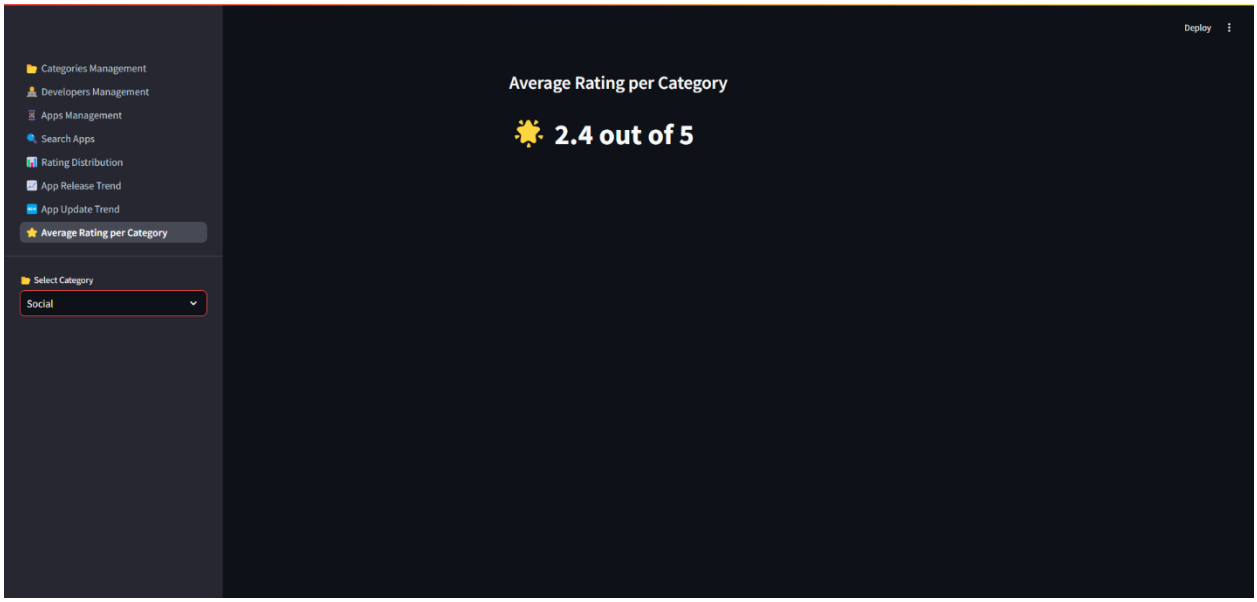
فیلترها:

- 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 ۱۸۰۰ 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 ۱۶۰۰ درصد بهبود، از ۶.۵۳۲۶ 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