پروژه نهایی

دیجی‌کالا به تازگی بخشی از دیتاستش را در اختیار عموم قرار داده است. برای پروژه‌ی نهایی درس ما از این دیتاست استفاده می‌کنیم.

دیتاست دیجی‌کالا

قسمت اول (۵۰ نمره)

- فایل‌های دیتاست دانلود شده ساختار مناسب رابطه‌ای ندارند (بعضی ستون‌ها به صورت یک رشته که مجموعه ای از key-value می باشد، ذخیره شده‌اند). در قسمت اول پروژه، می‌خواهیم این دیتابیس را دوباره طراحی کنید، به طوری که 5NF باشد (می‌دانیم اگر یک جدول 3NF باشد و تمامی کلیدها تک‌ستونه باشند، جدول 5NF است) و هیچ ستونی به صورت رشته key-value نباشد. برای گوشی‌های هوشمند، لپ‌تاپ‌ها، تبلت‌ها، کنسول‌های بازی، هدفون‌ها و اسپیکرها تمام خاصیت‌ها را استخراج کرده و جداول مربوطه را بسازید و سپس اطلاعات آنها را که در ساختار جدید به صورت رابطه‌ای قرار گرفته‌اند، از داخل فایل‌ها به دیتابیسی که طراحی کردید، لود کنید. دقت کنید که فرایند ساخت جدول‌ها و لود دیتاها توسط یک اسکریپت به صورت خودکار باید انجام شود. ساخت دستی یا لود دستی دیتا قابل قبول نیست.
- بر روی برندهای گوشی موبایل `index` مناسب قرار دهید. سپس با استفاده از آنالیز کردن کوئری بررسی کنید آیا کوئری‌های مربوط به فیلتر کردن بر اساس برند، از آن ایندکس استفاده می‌کنند یا خیر. برای این منظور ۵ کوئری مختلف که بر اساس فیلتر برند نوشته شده‌اند، ارایه دهید و نتایج تحلیل را در یک گزارش بنویسید.

نکات

- دیتاهایی که صورت key-value هستند، باید به صورت رابطه‌ای ذخیره کنید.
- برای انجام پروژه از `MySQL` یا `PostgreSQL` استفاده کنید.
- گزارش شما باید فرم یک گزارش درست را داشته باشد؛ شامل اطلاعات دانشجویی شما و فرمت مناسب باشد.

قسمت دوم (۳۰ نمره - امتیازی)

در این قسمت از شما انتظار می‌رود که بر روی دیتاست تحلیل انجام دهید. کیفیت تحلیل‌های شما مهم است و بر روی نمره دریافتی از این قسمت تاثیر زیادی دارد.

نکته مهم: همه‌ی تحلیل‌های شما باید با استفاده از نمودارها و هر ابزار دیگری که باعث درک بهتر موضوع می‌شود، باشد. به عبارت دیگر باید داده‌ها را Visualize کنید. همچنین اگر دیتاست مرتبط دیگری در دسترس دارید که به تحلیل شما می‌تواند کمک کند، با ذکر منبع استفاده کنید.

سوال‌ها

- تاثیر افزایش قیمت‌ها در سال اخیر، بر فروش دیجی‌کالا چه بوده است؟ بررسی و تحلیل کنید که چه رابطه‌ای بین میزان تورم و افزایش قیمت‌ها با فروش وجود دارد.
- در مورد مشتریان شهرهای مختلف چه می‌توان گفت؟ مثلا تعداد مشتریان هر شهر به نسبت جمعیت آن شهر در کدام شهرها بیشتر/کمتر است؟ یا کدام شهرها خریدهای گران‌تر و کدام شهرها خریدهای ارزان‌تر انجام می‌دهند؟ هر شهر چه دسته‌ای محصولات رو بیش‌تر از بقیه دسته‌ها خریداری کرده است؟
- کدام برندها، برند محبوب ایرانی‌هاست؟ محبوبیت یک برند با قیمت آن چه رابطه‌ای دارد؟
- میزان فروش در جشنواره‌های سال نو، یلدا و ... چه تفاوتی با بقیه روزها دارد؟
- پیش‌بینی شما برای دیجی‌کالا طی ده سال آینده چیست؟
- تحلیل کنید که در چه دسته‌هایی از محصولات، محصولاتی با برند متفرقه بیش‌تر فروش رفته‌اند. آیا به قیمت آن‌ها ربطی دارد؟
- محبوب‌ترین محصولات کدام‌ند؟ محبوبیت می‌تواند تابعی از میزان فروش و تعداد نظرات باشد.
- می‌خواهیم کاربرها را بر اساس نظراتی که زیر محصولات ثبت کرده‌اند، دسته‌بندی کرده و از ۱۰ نفر اول آن‌ها به نوعی قدردانی کنیم. آن‌ها را مشخص کنید.
- چه کلماتی در نظرات بیشتر از همه ظاهر شده‌اند؟ word_cloud کلمات داخل نظرات را بسازید. بدین منظور میتوانید از ابزارهایی مشابه این استفاده کنید.
- کدام دسته از محصولات بیشترین نظرات را دارند؟ در بین نظرات این دسته چقدر نظرات تکراری و یا خیلی شبیه به هم با اختلاف دو یا سه کلمه وجود دارد؟

نکات

- هر ابزار یا تکنولوژی مورد قبول است، مثلا برای Visualize کردن، می‌توانید از کتاب‌خانه هر زبانی (مثلا برای پایتون `mathplotlib`)، اکسل یا هر ابزار دیگری استفاده کنید.
- برای گرفتن نمره کامل از این بخش، حداقل باید به ۵ تا از سوالات جواب قابل قبولی داده باشید.
- بدیهی است که گزارش شما نباید دارای غلط علمی، املایی یا نگارشی باشد.
- برای آن که نسبت به نحوه‌ی گزارش دادن شهود پیدا کنید، بهتر است نمونه‌های زیر را ببینید:

Spotify Usage and Revenue Statistics (2019)

Twitter Revenue and Usage Statistics (2018)

Instagram Revenue and Usage Statistics (2018)

ارسال پروژه

```
project.zip
├── part1
│   ├── README.md
│   ├── report1.pdf
│   └── codes
│       └── data ## don't upload it
└── part2
    ├── report2.pdf
    └── codes
```

- در فایل `README.md` باید توضیحات در مورد اجرای برنامه و پیش‌نیازهای برنامه شما نوشته شده باشد.
- در پوشه‌های `codes` تمامی کدهایی را که برای رسیدن به اهداف پروژه پیاده‌سازی کرده‌اید، قرار دهید.
- در فایل‌های `report1.pdf` `report2.pdf` گزارش تحلیل‌های قسمت اول و دوم سوال را قرار دهید.
- فرض کنید دیتاست دریافتی از دیجی‌کالا در پوشه `data` قرار دارد. (نباید آن را آپلود کنید.) کدهای شما باید بر اساس این فرض اجرا شوند.

نکته: در صورتی که نیاز به اطلاعات چند جدول داشتید و اطلاعات ناقص بود، در گزارش خود بنویسید که امکان استخراج این بخش نبود و اطلاعات مربوطه به آن (مثلا ID) را ریپورت کنید.