به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی پیشرفته

تمرین دوم

طراح	امیرحسین روشن دل Roshandel2004@gmail.com
تاریخ بارگذاری	۱۴۰۳/۱۲/۲۲
مهلت ارسال	۱۴۰۴/۰۱/۱۷

فهرست

بخش نظری ۲

۱. سوال اول ۲

۲. سوال دوم ۳

۳. سوال سوم ۴

بخش عملی ۶

مقدمه ۶

توضیحات مجموعه داده ۷

سوالات ۸

۱. طراحی اسکیماهای Snowflake و Star در Pandas ۸

۲. مقایسه سرعت اجرای عملیات گروه‌بندی در Star و Snowflake Schema ۸

۳. محاسبه Roll-up: تحلیل فروش در سطوح مختلف ۸

۴. محاسبه Drill-down: تحلیل فروش روزانه ۹

۵. تحلیل فروش در شهرها و شعب با Data Cube ۹

۶. تحلیل رفتار مشتریان با Slice & Dice ۹

ملاحظات ۱۱

استفاده مسئولانه از هوش مصنوعی ۱۲

۱. هدف و اصول کلی ۱۲

۲. استفاده مجاز از LLMها ۱۲

۳. استفاده غیرمجاز از LLMها ۱۲

۴. مستندسازی ۱۳

۵. آمادگی ارائه شفاهی ۱۳

۶. پیامدهای تخلفات ۱۳

۷. موارد تکمیلی ۱۳

۸. اظهارنامه ۱۳

بخش نظری

۱. سوال اول

فرض کنید که cuboid پایه‌ای یک سامانه مدیریت بیمارستانی دارای دو سلول زیر است و می‌دانیم که:

$$p_i \neq q_i$$

$$\left(p_1, p_2, \underline{q_3}, p_4, \underline{q_5}, p_6, q_7, p_8, p_9, p_{10}\right) : 22$$

$$\left(q_1, q_2, \underline{q_3}, q_4, \underline{q_5}, q_6, p_7, q_8, q_9, q_{10}\right) : 14$$

الف) چند cuboid در این data cube وجود دارد؟

ب) در این data cube چند سلول aggregate غیرتهی دارد؟

ج) چند سلول بسته‌ی غیرتهی در این data cube موجود است؟ چه تعدادی از آنها aggregated هستند؟

د) اگر minimum support برابر ۲۵ باشد، تعداد سلول‌های aggregate غیرتهی در iceberg cube متناظر چقدر خواهد بود؟

۲. سوال دوم

یک بیمارستان اطلاعات درمان بیماران را در قالب سه بعد (زمان، بیمارستان، بخش درمانی) ذخیره می‌کند:

جدول ۱. اطلاعات درمان بیماران در بیمارستان

Patients Treated	Department	Hospital Branch	Time (Month)
100	Cardiology	New York	Jan
80	Neurology	New York	Jan
50	Orthopedic	Chicago	Jan
90	Neurology	Los Angeles	Feb
110	Cardiology	New York	Feb
130	Oncology	Chicago	Mar
40	Neurology	New York	Mar
70	Cardiology	Los Angeles	Apr

الف) طراحی و تحلیل Data Cube

۱. تمام Cuboidهای ممکن در این Data Cube را فهرست کنید.

ب) انجام عملیات OLAP (برای هر قسمت زیر نیز مشخص کنید که از کدام عملیات OLAP استفاده کرده اید.)

۱. برای شناسایی پرترافیک‌ترین بخش درمانی در بیمارستان‌ها و تعیین اینکه در کدام ماه این بخش بیشترین بیماران را درمان کرده است، چه عملیات‌های OLAP باید انجام شود؟ (Cuboid اولیه در سطح جزئیات جدول باشد.)

۲. برای بررسی روند تغییرات تعداد بیماران در بیمارستان‌های نیویورک و لس‌آنجلس در بخش‌های Cardiology و Neurology در بازه زمانی ژانویه تا مارس، چگونه می‌توان از عملیات‌های OLAP استفاده کرد؟ (Cuboid اولیه در سطح جزئیات جدول باشد.)

۳. برای مقایسه میانگین تعداد بیماران درمان‌شده در بیمارستان نیویورک در ماه مارس ۲۰۲۳ با میانگین تعداد بیماران درمان‌شده در بیمارستان لس‌آنجلس در ژانویه ۲۰۲۲، مشخص کنید که چه عملیات‌های OLAP باید انجام شود. (زمان در سطح روز و بقیه مطابق جزئیات جدول باشد.)

۳. سوال سوم

جدول زیر شامل اطلاعات پزشکان بیمارستان در سه بعد است:

جدول ۲. اطلاعات پزشکان در بیمارستان

Department	Education Level	Specialization
Cardiology	PhD	Cardiologist
Cardiology	Master	Cardiologist
Cardiology	Master	Cardiologist
Neurology	PhD	Neurologist
Neurology	Master	Neurologist
Neurology	PhD	Neurologist
Orthopedic	Bachelor	Orthopedic
Orthopedic	Master	Orthopedic
Orthopedic	Master	Orthopedic
Oncology	PhD	Oncologist
Oncology	Master	Oncologist

الف) انتخاب بهترین ترتیب پردازش ابعاد در الگوریتم BUC

۱. ترتیب پردازش ابعاد را طوری انتخاب کنید که اجرای الگوریتم BUC کمترین هزینه محاسباتی و بیشترین کارایی را داشته باشد.

۲. معیارهای خود را برای این انتخاب توضیح دهید.

۳. توضیح دهید که چگونه ترتیب انتخابی شما می‌تواند باعث کاهش زمان اجرای BUC شود.

ب) اجرای BUC و محاسبه‌ی Iceberg Cube

۱. الگوریتم BUC را روی مجموعه‌داده‌ی فوق اجرا کنید و Iceberg Cube را با شرط Minimum Support = 2 به دست آورید. سپس جدول نهایی Iceberg Cube را نمایش داده و به‌وضوح مشخص کنید که کدام سلول‌ها به دلیل نداشتن حداقل حمایت (support کمتر از ۲) حذف می‌شوند و چرا؟

۲. با توجه به جدول نهایی به‌دست‌آمده، توضیح دهید که در مقایسه با حالت Full Cube (شامل همه سلول‌هایی که در داده‌های واقعی حداقل یک نمونه دارند)، تعداد سلول‌ها در Iceberg Cube چقدر کاهش یافته است؟

بخش عملی

مقدمه

در دنیای داده‌محور امروزی، توانایی مدیریت، پردازش و تحلیل حجم عظیمی از داده‌ها، یکی از مهارت‌های کلیدی برای متخصصان داده محسوب می‌شود. در این تمرین، شما با اصول OLAP و Data Warehouse آشنا خواهید شد. این مفاهیم از اجزای اساسی سیستم‌های تصمیم‌گیری سازمانی هستند و نقش مهمی در استخراج اطلاعات ارزشمند از داده‌های حجیم ایفا می‌کنند.

در این تمرین، شما با یک مجموعه داده‌ی واقعی از فروش یک سوپرمارکت کار خواهید کرد که شامل اطلاعاتی مانند قیمت، تعداد فروش، دسته‌بندی محصولات و سایر ویژگی‌های مرتبط است. هدف این بخش، بررسی این داده‌ها از طریق شبیه‌سازی عملیات‌های OLAP و آشنایی با تحلیل چندبعدی داده‌ها است.

شما یاد خواهید گرفت که چگونه داده‌ها را در سطوح مختلف تجمیع و تفکیک کنید و با به‌کارگیری عملیات Slice، Dice، Roll-up و Drill-down، ابعاد مختلف اطلاعات را بررسی کنید. این فرایند به شما کمک می‌کند تا الگوهای پنهان در داده‌ها را شناسایی کرده و روندهای مهم را تحلیل کنید.

در نهایت، این تمرین به شما امکان می‌دهد تا با ساختارهای انبار داده و روش‌های تحلیلی OLAP آشنا شده و مهارت‌های خود را در تحلیل داده‌های تجاری و تصمیم‌گیری مبتنی بر داده ارتقا دهید.

در پایان، شما باید یافته‌های خود را مستندسازی کرده و استراتژی‌های بهینه‌سازی پیشنهادی را ارائه دهید. هدف نهایی این است که مهارت‌های خود را در طراحی، پیاده‌سازی و تحلیل انبار داده و سیستم‌های OLAP تقویت کرده و تسلط بیشتری بر روی مفاهیم پیشرفته مدیریت داده کسب کنید.توجه داشته باشید که این فرآیند بخش مهمی از ارزیابی عملکرد شما در این تمرین را تشکیل می‌دهد و در نتیجه نهایی شما تأثیر قابل‌توجهی خواهد داشت. بنابراین، سعی کنید تمامی مراحل را با دقت و جزئیات کافی در پاسخ‌های خود منعکس کنید تا توانایی تحلیل و حل مسئله شما به‌خوبی نمایان شود.

توضیحات مجموعه داده

جدول ۳. شرح دادگان موجود در مجموعه داده فروش سوپرمارکت

نام ستون	توضیحات
Invoice ID	شناسه یکتای فاکتور خرد
Branch	شعبه‌ای که خرید در آن انجام شده‌است
City	شهری که شعبه در آن قرار دارد
Customer type	نوع مشتری (عضو یا عادی)
Gender	جنسیت مشتری
Product line	دسته‌بندی محصولی که خریده شده است
Unit price	قیمت هر واحد محصول (به دلار)
Quantity	تعداد محصول خریداری شده
Tax 5%	مقدار مالیات ۵ درصدی اعمال شده به خرید
Total	مبلغ کل پرداخت شده پس از خرید
Date	تاریخ خرید
Time	زمان خرید
Payment	روش پرداخت (نقدی، کارت اعتباری، کیف پول الکترونیکی)
cogs	هزینه کالای فروخته شده (Cost of Goods Sold)
gross margin percentage	درصد حاشیه سود ناخالص
gross income	سود ناخالص حاصل از خرید
Rating	امتیاز مشتری به تجربه خرید

سوالات

۱. طراحی اسکیماهای **Snowflake** و **Star** در **Pandas**

أ. داده‌ها را به صورت Fact Table و Dimension Tables تقسیم‌بندی کنید و DataFrameهای لازم را در Pandas ایجاد کنید:

Fact Table شامل کلیدهای خارجی به هر یک از Dimensionها (مانند مشتری، محصول، شعبه، زمان و روش پرداخت) و همچنین مقادیر عددی قابل اندازه‌گیری مانند Total خواهد بود.

راهنمایی: Dimension Tables شامل جدول‌هایی برای مشتری (Customer)، محصول (Product)، شعبه (Branch)، زمان (DateTime)، مالی (Financial) و روش پرداخت (Payment) هستند.

مثال ساخت Dimension Table مشتری:

```
customers_dim_star = df[['Customer type', 'Gender']].drop_duplicates().reset_index(drop=True)
customers_dim_star['Customer_ID'] = customers_dim_star.index
```

به همین روش، Dimension Tableهای دیگر را نیز ایجاد کنید.

ب. میزان استفاده از حافظه را برای هر مدل با استفاده از دستور زیر مقایسه کنید:

```
df.memory_usage(deep=True).sum()
```

۲. مقایسه سرعت اجرای عملیات گروه‌بندی در **Star** و **Snowflake Schema**

أ. مجموع فروش (Total) به ازای هر دسته محصول (Product line) را در مدل Star Schema محاسبه کرده و زمان اجرای این عملیات را اندازه‌گیری کنید.

ب. برای مدل Snowflake نیز مجموع فروش (Total) را برای هر دسته محصول محاسبه کرده و زمان اجرا را اندازه‌گیری کنید.

توجه کنید که در Snowflake Schema ممکن است نیاز به عملیات JOIN بین جداول باشد.

زمان‌های به دست آمده را مقایسه و تحلیل کنید:

- کدام مدل سریع‌تر بود و چرا؟

۳. محاسبه **Roll-up**: تحلیل فروش در سطوح مختلف

برای این سؤال می‌توانید از هرکدام از مدل‌های داده‌ای طراحی شده (Star یا Snowflake) استفاده کنید:

أ. مجموع فروش را در سطح ماهانه و سالانه محاسبه کنید.

ب. بررسی کنید که بیشترین فروش مربوط به کدام سال است؟ روند کلی فروش سالانه چگونه است (افزایشی یا کاهشی)؟

ج. تغییرات فروش در سطح ماهانه را تحلیل کنید و نمودار آن را رسم کنید. ماه‌هایی که تغییرات شدید در فروش داشته‌اند را مشخص کنید.

۴. محاسبه **Drill-down**: تحلیل فروش روزانه

برای این سؤال نیز از هرکدام از مدل‌های Star یا Snowflake می‌توانید استفاده کنید:

أ. مجموع فروش روزانه را برای هر یک از شعبه‌ها محاسبه کرده و با نمودار مناسب ارائه دهید.

ب. یکی از شعبه‌ها را به انتخاب خودتان در نظر بگیرید:

- مشخص کنید در کدام روزها بیشترین و در کدام روزها کمترین فروش را داشته است.
- تأثیر روزهای هفته (شنبه تا جمعه) را بر میزان فروش تحلیل کنید و بررسی کنید آیا روز خاصی از هفته فروش بهتری داشته است یا خیر.

۵. تحلیل فروش در شهرها و شعب با **Data Cube**

أ. با استفاده از تابع pivot_table یک Data Cube را شبیه‌سازی کنید که مجموع فروش را به تفکیک شهر و شعبه نمایش دهد.

ب. بر اساس این Data Cube مشخص کنید که کدام شهر و کدام شعبه بیشترین میزان فروش را دارند.

ج. یک Data Cube دیگر شبیه‌سازی کنید که میزان فروش محصولات را در شهرهای مختلف نشان دهد. تحلیل کنید هر محصول بیشترین فروش را در کدام شهر داشته است.

۶. تحلیل رفتار مشتریان با **Slice & Dice**

أ. مشتریانی را که دسته محصول «Electronic accessories» را خریداری کرده‌اند شناسایی کنید.

- تحلیل کنید که آیا این مشتریان ویژگی خاص مشترکی (مانند نوع مشتری یا جنسیت) دارند؟

ب. تحلیل کنید مشتریانی که عضو (Member) هستند بیشتر به کدام دسته محصولات تمایل دارند و دلیل احتمالی آن را ذکر کنید.

ج. رفتار مشتریان عادی (Normal) و عضو (Member) را تحلیل کنید:

- آیا تفاوت معناداری در ترجیحات محصولی یا میزان خرید این دو گروه وجود دارد؟
- فروشگاه چگونه می‌تواند از این تحلیل برای بهبود استراتژی بازاریابی و فروش خود استفاده کند؟

نکته: تا جای ممکن تحلیل ها، مشکلات، روش هایی که برای رفع هر مشکل با آن برخورد کردید یا حتی اگر مشکلی وجود داشت و با ابزار های موجود نمیتوانستید آن را برطرف کنید را مستند کنید.

ملاحظات

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA2_StudentID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- بخش اصلی نمره به گزارش شما تعلق می‌گیرد و دستیاران الزامی برای اجرای تمام کدهای شما در صورتی که در گزارش به آن‌ها اشاره‌ای نکرده باشید ندارند. لطفا تمام موارد مورد نیاز را در گزارش ذکر کنید.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- برای تحویل تمارین از چارچوب قرارداده شده در سامانه، سایت درس به آدرس dm-ut.github.io و یا گروه تلگرام استفاده کنید.
- در صورت قصد ارسال تمرین به صورت دیگر (انگلیسی، latex و ...)، لطفا پیش از ارسال با دستیار مسئول تمرین هماهنگ کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (هم‌فکری خارج از چارچوب و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تخلف برای همه‌ی افراد مشارکت کننده، نمره تمرین، صفر در نظر گرفته خواهد شد.
- در صورت استفاده از ابزارهای هوش مصنوعی، قوانین استفاده در پایان تمرین را مطالعه کنید.
- در پایان گزارش ارسالی خود، اظهارنامه بند ۸ از قوانین استفاده مسئولانه از هوش مصنوعی را قرار دهید.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:roshandel2004@gmail.com

مهلت تحویل: ۱۷ فروردین ۱۴۰۴

مهلت تحویل با تاخیر: ۲۴ فروردین ۱۴۰۴

استفاده مسئولانه از هوش مصنوعی

۱. هدف و اصول کلی

هدف

- ترویج استفاده اخلاقی و مسئولانه از LLMها (مانند ChatGPT، Deepseek) به عنوان ابزار کمکی
- اطمینان از مشارکت فعال دانشجویان در تکالیف و درک راه‌حل‌های آن‌ها
- حفظ صداقت علمی در عین بهره‌گیری از ابزارهای مدرن هوش مصنوعی

اصول کلی

- تمرین باید نتیجه تلاش و زحمت شخصی شما باشد.
- باید به تمام بخش‌های تمرین، اعم از پیاده‌سازی و تحلیل نتایج مسلط باشید.
- تمامی کدها باید توسط خود شما اجرا شده و نتایج قابل مشاهده باشند.
- تمام مراحل انجام تمرین باید مستند و قابل پیگیری باشد.
- هرگونه نتیجه‌گیری و تحلیل باید بر اساس درک شخصی شما باشد.
- LLMها ممکن است پاسخ‌های نادرست یا قدیمی تولید کنند، اولویت با مطالب و کارگاه‌های درس است.

موارد ذکر شده در ادامه این سند، به عنوان راهنمایی بیشتر برای انجام تمرین آورده شده‌اند. با این حال، مسئولیت تطبیق کار با اصول کلی فوق بر عهده شماست. توجه داشته باشید که ممکن است مواردی در ادامه ذکر نشده باشند که با اصول کلی ذکر شده در تضاد باشند. در چنین مواردی به تشخیص دستیار آموزشی و دستیار مسئول، شما موظف به پاسخ‌گویی در قبال تمرین خود هستید. عدم رعایت هر یک از اصول فوق می‌تواند منجر به کسر نمره یا عدم پذیرش تمرین شود.

۲. استفاده مجاز از LLMها

شما می‌توانید از LLMها برای موارد زیر استفاده کنید:

- روشن‌سازی مفاهیم (مثال: "خوشه‌بندی DBSCAN چگونه کار می‌کند؟")
- کمک در اشکال‌زدایی (مثال: شناسایی خطاهای گرامری یا منطقی در کد)
- ایده‌پردازی رویکردها (مثال: "روش‌های مدیریت داده‌های missing را پیشنهاد دهید")

الزامات استفاده مجاز:

- ثبت تعاملات اصلی: (به بخش ۴ مراجعه کنید.)
- درک راه‌حل: باید قادر به توضیح هر خط کد یا منطق استفاده شده باشید.

۳. استفاده غیرمجاز از LLMها

اقدامات ممنوع شامل:

- کپی-پیست مستقیم خروجی‌های LLM بدون تغییر
- استفاده از LLMها برای حل اصلی مسائل (مثال: "این سؤال تکلیف را برای من حل کن")

- گرفتن کد از سایر دانشجویان به هر شکل غیر مجاز است، تغییر و پارافریز کردن کد دیگران توسط LLM نیز قابل قبول نیست.
- هرگونه استفاده که منجر به عدم احاطه شما به موضوع تمرین شود.

۴. مستندسازی

ارجاع به مشارکت‌های LLM: افزودن پانویس یا توضیح (مثال: کد با رعایت قوانین به کمک ChatGPT نوشته شده است.)

- نیازی به اشتراک گذاری پرامپت‌ها و سابقه چت نیست.
- مستندسازی تک تک تعاملات با هوش مصنوعی هدف این بخش نیست. اشاره کوتاه و کلی در بخش‌های مورد استفاده کافی است. در نظر داشته باشید که مستندسازی به معنای رفع مسئولیت نبوده و باید اصول کلی را رعایت کنید.

۵. آمادگی ارائه شفاهی

آماده دفاع از کار خود باشید: در صورت درخواست دستیار تمرین در بازه زمانی اعلام شده برای ارائه شفاهی، باید:

- رویکرد، کد یا نتایج خود را توضیح دهید.
- درک مفاهیم کلیدی را نشان دهید (مثلاً چرا یک الگوریتم خاص انتخاب شده است)

عدم توضیح کافی کار شما ممکن است منجر به جریمه شود (بخش ۶)

۶. پیامدهای تخلفات

- تخلفات جزئی (مثل مستندسازی ناقص): کاهش نمره
- تخلفات عمده (مثل کپی-پیست بدون تغییر): نمره ۵۰- در تکلیف
- تخلفات مکرر: نمره ۵۰- در تکلیف و گزارش به استاد

۷. موارد تکمیلی

- از LLMها به عنوان معلم استفاده کنید، نه پاسخ‌نامه تمرین‌ها: اولویت را به مهارت‌های حل مسئله خود بدهید.
- خروجی‌ها را متقابلاً تأیید کنید: پیشنهادات LLM را با کتاب مرجع درس، اسلایدها و کارگاه‌ها مقایسه کنید.
- از دستیاران آموزشی کمک بگیرید: اگر پاسخ LLM یا نحوه استفاده شما را گیج می‌کند، در ساعات متعارف از دستیاران آموزشی کمک بگیرید.

۸. اظهارنامه

این عبارت را در تکلیف ارسالی خود قرار دهید:

"تأیید می‌کنم که از LLMها مطابق با دستورالعمل‌های بارگذاری شده در سامانه Elearn درس به طور مسئولانه استفاده کرده‌ام. تمام اجزای کار خود را درک می‌کنم و آماده بحث شفاهی درباره آنها هستم."