به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده‌کاوی

تمرین عملی ۱

اسفند ماه ۱۴۰۱

*فهرست

پیش‌نیازها

برای پاسخ به این تمرین عملی باید از زبان برنامه‌نویسی **Python** استفاده‌کنید و نیاز است که پیش از شروع، یک سرور **Jupyter** بر روی سیستم نصب و راه‌اندازی شود تا بتوانید بر روی یک فایل **.ipynb** کدهای خود را اجرا کنید، همچنین راه حل جایگزین آن استفاده از **Google Colab** است.

استفاده از کتابخانه‌های **Pandas** و **Numpy** می‌تواند گزینه‌ی مناسبی برای حل مسائل پیشرو باشد، همچنین دو کتابخانه‌ی **Matplotlib** و **Seaborn** در بخش مصورسازی مجموعه‌داده‌گان مثمر ثمر واقع می‌شود.

بخش ۱ (House)

شرح دادگان

این مجموعه داده با نام house.csv در فایل فشرده dataset.zip قرار داده شده و شامل اطلاعات مربوط به خانه‌ها است، توضیحات در مورد این مجموعه‌داده در فایل data_description.txt وجود دارد.

پیش‌پردازش

پیش‌پردازش، یکی از مهم‌ترین گام‌ها در پروژه‌های داده‌کاوی است. رویکردهای مختلفی در زمینه‌ی مدیریت داده‌های گم شده و تبدیل داده‌ها به فرمت‌های دیگر مورد استفاده قرار می‌گیرد و انتخاب دقیق این رویکردها تأثیر مستقیمی در کیفیت نتایج نهایی دارد؛ لذا همواره می‌بایست بهترین رویکرد را شناسایی و اعمال نمود.

۱. ابتدا ۵ سطر ابتدایی دیتاست را در Jupyter با استفاده از کتابخانه Pandas نمایش دهید.
۲. برای هر کدام از ستون‌های این مجموعه‌داده تعداد مقادیر گمشده[1] را گزارش کنید.
۳. چه رویکردی برای هر کدام از ستون‌ها مناسب است دلیل آن را ذکر کنید. (راه‌حل‌ها می‌تواند شامل حذف ردیف‌های شامل داده گمشده یا درج کردن داده در آن مکان باشد)
۴. با استفاده از روش‌های انتخاب شده در سؤال ۳ مشکل داده‌های گمشده را برطرف کنید.
۵. در این قسمت داده‌های پرت[2] را شناسایی کنید و روش برخورد را بیان کنید. (در این قسمت حذف ردیف‌ها یا نرمال‌سازی داده‌ها بخشی از روش‌های در اختیار شماست)
۶. بررسی کنید آیا مقادیری وجود دارند که دارای تناقض باشند، دو مورد را ذکر کنید. (تناقض به این منظور که خانه‌ای با متراژ پایین گران‌تر از خانه‌ای با متراژ بالا باشد، برای این مثال خاص چند خانه با این شرایط را ذکر و همچنین برای دلیل این تناقض فرضیه سازی کنید)

[1] Missing value
[2] Outliers

نمایش دادگان

مصورسازی داده‌ها با استفاده از نمودارهای مناسب دید بهتری نسبت به اطلاعات موجود در مجموعه‌داده را ایجاد می‌کند و همچنین می‌تواند باعث شود تحلیل‌گران و مدیران تصمیمات بهتری اتخاذ کنند.

۱. میانگین قیمت خانه در هر محله را با نمودار میله‌ای[1] نمایش دهید و همچنین مقدار میانگین در نمودار بر روی هر میله با فونت و رنگ مناسب نوشته شده باشد. (برای نمایش داده‌ها با استفاده از نمودار میله‌ای زمانی که تعداد میله‌ها زیاد باشد، بهتر است که میله‌ها به‌صورت افقی باشند، همچنین مرتب کردن مقدارها به‌صورت نزولی، دید بهتری را به بیننده می‌دهد)

۲. پنج محله با بیشترین تعداد خانه و پنج محله با گران‌ترین خانه‌ها را مشخص کنید.

۳. تعداد اتاق خواب‌های یک خانه چه تأثیری بر روی قیمت آن دارد با نمودار مناسب این تغییر را نمایش دهید.

۴. بررسی کنید کدام یک از متغیرها با یکدیگر ارتباط بیشتری دارند و این ارتباط را با نمودار مناسب نمایش دهید (یک مورد کافیست و صرفاً بیشترین ارتباط را نمایش دهید، از معیار Correlation استفاده‌کنید نمودارheatmap نیز در بخش می‌تواند برای نمایش زوج ارتباطات مناسب باشد)

۵. توزیع قیمت خانه‌ها را با نمودارهای جعبه‌ای[2]، فراوانی[3] و توزیعی[4] نمایش دهید. کدام نمودار اطلاعات بیشتری را به بیننده می‌دهد (این نمودارها دارای پارامترهایی هستند که ممکن است در این بخش کمک کننده باشند. یک از این پارامترها، Scale نمودار است؛ روش‌های متفاوت را برای آن امتحان کنید و بررسی کنید کدام یک مناسب‌تر است.)

۶. با استفاده از نمودار پراکندگی[5] می‌توانیم ارتباط دو متغیر با یکدیگر در فضای دو بعدی را ترسیم کنیم، آیا نمودار دیگری یا حالت خاصی از این نمودار وجود دارد که بتواند در فضای دو بعدی ارتباط بیشتر از دو متغیر را نمایش دهد؟ در صورت وجود این نمودار را برای بیش از دو پارامتر به‌شکل معناداری ترسیم کنید.

[1] Bar plot
[2] Box plot
[3] Histogram
[4] Distribution plot
[5] Scatter plot

بخش ۲ (Purchase)

شرح دادگان

مجموعه‌داده‌ی این بخش با نام purcahses.csv به‌همراه users.csv قرار داده‌شده است این مجموعه‌داده‌ها شامل:

1. Purchases

Columns	جزئیات
User_id	شناسه‌ی خریدار
SKU	شناسه محصول خریداری‌شده
AddedTime	تاریخ خرید
Price	مبلغ خرید
CurrencyISO	واحد پول خرید

2. Users

Columns	جزئیات
User_id	شناسه‌ی کاربر
CountryCode	کد کشور محل سکونت کاربر
RegisterTime	تاریخ ثبت‌نام کاربر

پیش‌پردازش

۱. تمامی ارزها را به یک واحد یکسان ترجیحاً دلار تبدیل کنید. (برای انجام این مرحله می‌توانید از این API کمک بگیرید)

۲. در ادامه نیاز داریم که برای هر خرید بدانیم که کاربر مربوط به آن در چه تاریخی ثبت‌نام کرده و در چه کشوری سکونت دارد این اطلاعات را بر اساس شناسه کاربر به جدول خریدها اضافه کنید.

نمایش دادگان

۱. با استفاده از نمودار مناسب مجموع خریدها را بر اساس تاریخ نمایش دهید. (به این صورت که مجموع خریدها را در هر روز جمع بزنید)
۲. در کدام هفته بیشترین خرید انجام‌شده؟ (هفته از دوشنبه شروع شود و اسم هر هفته با تاریخ اولین روز آن مشخص شود) پنج هفته پر خرید را با نمودار مناسب نمایش دهید
۳. با استفاده از نمودار مناسب در کدام یک از روزهای هفته کاربران بیشترین خرید را دارند؟
۴. کاربر با بیشترین خرید مربوط به کدام کشور می‌شود؟
۵. تعداد کاربران خریدار یکتای هر کشور را نمایش دهید.
۶. پنج کشور با بیشترین مبلغ خرید را نمایش دهید.
۷. کدام کشور خریداران بهتری دارد؟
۸. تبلیغات هزینه دارد! اگر بخواهیم با استفاده از تبلیغات تعدادی کاربر جدید از یکی از کشورهای این مجموعه داده جذب کنیم بهتر است از کدام کشور انتخاب شوند؟ استدلال شما برای انتخاب این کشور چیست؟
۹. تعداد کاربران ثبت نام شده هر کشور را بر روی نقشه نمایش دهید.

ملاحظات (حتماً مطالعه شود)

- تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA1_StudentID تحویل داده شود، که این فایل فشرده شامل یک فایل کد با فرمت ipynb. و یک فایل گزارش pdf. است.
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. انجام پیاده‌سازی بدون گزارش و تحلیل فاقد اعتبار است. برای خوانایی بهتر کد فایل Jupyter خود را بخش‌بندی کنید.
- آدرس هر دیتاست را در یک متغیر ذخیره کنید و برای لود کردن آن از متغیر استفاده‌کنید این عمل در اولین بلوک فایل Jupyter به‌صورت زیر انجام شود.

In [ ]:
```
houseDatasetPath = 'YourPath'
usersDatasetPath = 'YourPath'
purchaseDatasetPath = 'YourPath'
```

- مهلت تحویل تمرین به هیچ‌عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته‌می‌شود که جریمه تأخیر تحویل تمرین تا یک هفته ۳۰ درصد است.
- توجه کنید این تمرین باید به‌صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت‌کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد. همچنین نوشته نشدن کدها توسط هوش‌مصنوعی نیز بررسی می‌شود!
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto: mj.kamyab@ut.ac.ir

مهلت تحویل بدون جریمه: ۱۴۰۱/۱۲/۲۲

مهلت تحویل با تأخیر، با جریمه ۳۰ درصد: ۱۴۰۱/۱۲/۲۹