worldskills

Session 1

Software Applications Development

Time: 2:00

چشم‌انداز

یک آژانس **گردشگری**، طی سال‌های اخیر داده‌های گسترده‌ای درباره‌ی **جاذبه‌ها**، **مشتریان** و **بازدیدها** جمع‌آوری کرده است. این داده‌ها اطلاعات ارزشمندی درباره‌ی **رفتار مشتریان**، **محبوبیت جاذبه‌ها** و **روندهای گردشگری** در مناطق مختلف کشور ارائه می‌دهند.

در این پروژه، شما باید از مهارت‌های خود در **تحلیل** و **بصری‌سازی داده** استفاده کنید تا این مجموعه‌داده را بررسی، پاک‌سازی و مدل‌سازی کنید و از آن اطلاعات ارزشمندی استخراج نمایید که بتواند به تصمیم‌گیری‌های راهبردی و بهبود عملکرد آژانس و تجربه‌ی مشتریان کمک کند.

نتایج شما به تیم مدیریتی این آژانس ارائه خواهد شد تا با دیدگاهی مبتنی بر داده، به بهبود **برنامه‌ریزی گردشگری**، **تعامل مشتریان** و **تلاش‌های بازاریابی** بپردازند.

داده‌ها

این پروژه شامل فایل‌های زیر است:

داده‌های آژانس گردشگری (فایل‌های csv)

- **attractions.csv**
- **customers.csv**
- **visits.csv**

فایل Data Dictionary (شامل توضیحات و اطلاعات ستونهای جداول موجود در فایل‌های داده)

- **DataDictionary.xlsx (Excel):**

نکات مهم

تمام داده‌ها، نمودارها و خروجی‌ها باید به صورت واضح، قابل فهم و حرفه‌ای ارائه شوند، به گونه‌ای که Style Guide در رنگ‌ها و عناوین رعایت شده باشد. تحلیل‌ها، محاسبات و خروجی‌های کد باید دقیق، قابل بازتولید و بدون خطا باشند.

فایل‌های خروجی نهایی باید نام و فرمتی مطابق با موارد ذکر شده در صورت سوال داشته باشند. در غیراینصورت، نمره‌ی آن بخش به فرد تعلق نخواهد گرفت.

جزئیات پیاده‌سازی

۱.۱ ایجاد فایل نوت‌بوک و نوشتن کدهای تحلیل داده

یکی از گزینه‌های زیر را برای پیاده‌سازی انتخاب کنید:

- ایجاد یک فایل نوت‌بوک با فرمت ipynb.
- ایجاد یک فایل اسکریپت با فرمت py.

تمام کدها، خروجی‌ها و مراحل تحلیل بخش‌های بعدی را مرحله به مرحله در همین فایل بنویسید.

نام فایل:

Session1_DataAnalysis.ipynb (در صورت استفاده از نوت‌بوک)

یا

Session1_DataAnalysis.py(در صورت استفاده از اسکریپت)

۱.۲ وارد کردن داده‌ها و بررسی اولیه

واردکردن داده‌ها

– فایل‌های CSV ارائه شده را در محیط تحلیل داده خود وارد کنید.

بررسی اولیه

– پنج ردیف اول هر DataFrame را نمایش دهید تا ساختار و محتوای آن‌ها را درک کنید.
– وجود مقادیر گمشده، ناسازگاری‌ها و خطاهای احتمالی ورود داده را بررسی کنید.

خروجی نهایی

نام فایل:

Session1_DataExploration.pdf

محتوای فایل:

برای **هر یک** از سه فایل CSV موارد زیر را در فایل قرار دهید:

- جدولی شامل پنج ردیف اول داده‌ها
- گزارشی شامل **میانگین، میانه، بیشترین** و **کمترین** مقدار برای همه‌ی ستون‌های حاوی داده‌های عددی
- ذکر دقیق نام ستون‌های دارای ناسازگاری‌ها و ناهنجاری‌ها (در صورت وجود) و تعداد این ناهنجاری‌ها:
  - ستون‌های شامل مقادیر **گمشده** و **خالی** به همراه تعداد ردیف‌های دارای مشکل
  - ستون‌های عددی حاوی مقادیری خارج از بازه‌ی مشخص شده به همراه تعداد ردیف‌های دارای مشکل
  - مثال:
    - ستون income: دارای ۳۰ سطر با مقادیر گشمده است.
    - ستون age: دارای ۳۰ سطر با مقادیر منفی است.

۱.۳ پاک‌سازی داده‌ها

تصحیح مقادیر Rating **در جدول** attractions

– تمام مقادیر ستون Rating که بالاتر از ۵ هستند به ۵ تغییر یابند.

– تمام مقادیر ستون Rating که پایین‌تر از ۰ هستند به ۰ تغییر یابند.

تبدیل نوع داده‌ها (Data Type Conversion)

– ستون‌های مربوط به تاریخ و زمان در فایل‌های داده به نوع‌داده‌ی تاریخ (datetime) تبدیل شوند.

مدیریت مقادیر خالی (Missing Values)

– مقادیر خالی در ستون‌های عددی در فایل‌های داده با میانگین همان ستون جایگزین شوند.

– مقادیر خالی در ستون RegistrationDate از فایل customers.csv با یک تاریخ معتبر تصادفی در بازه‌ی تاریخ‌های موجود در مجموعه داده جایگزین شوند.

استانداردسازی داده‌ها (Data Standardization)

– تمام شماره‌های تلفن در فایل customers.csv به فرمت 0ddd-ddd-dddd تبدیل شوند. (مثال: 0912-123-4567)

خروجی نهایی

پس از انجام مراحل مرحله‌ی قبل، جداول **پاکسازی شده** را در فایل های زیر ذخیره کنید.

- customers_cleaned.csv
- attractions_cleaned.csv

- visits_cleaned.csv

۱.٤ تحلیل روندهای بازدید و درآمد در طول زمان

محاسبه شاخص‌های ماهانه سال ۲۰۲٤

- **درآمد کل ماهانه** (جمع هزینه‌های ورودی تمامی بازدیدها در هر ماه)
- **تعداد کل بازدیدها در هر ماه**
- **میانگین هزینه ورودی به ازای هر بازدید در ماه**

بصری‌سازی

- ایجاد **نمودار خطی (Line chart)** برای نمایش روند و نحوه‌ی تغییر شاخص‌های فوق در طول زمان
- استفاده از عناوین مناسب برای محورهای نمودار (مثال: Total Monthly Sales یا Total Monthly Visits)

شناسایی ماه‌های برتر

- تعیین **سه ماه برتر** از نظر درآمد کل.
- نمایش آن‌ها در یک جدول خلاصه با ستون‌های: **ماه میلادی** و **درآمد کل**

خروجی نهایی

- **نام فایل:**

Session1_TourismTrends.pdf

- **محتوای گزارش (بازه‌ی زمانی تمامی گزارش‌ها باید سال ۲۰۲٤ باشد):**
  - نمودار خطی: درآمد کل ماهانه
  - نمودار خطی: تعداد بازدیدها در هر ماه
  - نمودار خطی: میانگین هزینه ورودی هر جاذبه در هر ماه
  - جدول: سه ماه برتر بر اساس درآمد کل (ستون‌ها: نام ماه میلادی و درآمد کل)

۱.۵ تحلیل مشتریان

۱. **دسته‌بندی بر اساس سن:**

مشتریان را به گروه‌های سنی زیر تقسیم کنید:

۱. 18–24
۲. 25–34
۳. 35–44
۴. 45+

هر مشتری را به گروه سنی مربوطه اختصاص دهید.

۲. **میانگین بازدیدها به ازای هر گروه سنی:**

– میانگین تعداد بازدید مشتریان در هر گروه سنی را محاسبه کنید.

– نتایج را با استفاده از **نمودار میله‌ای** (Bar Chart) نمایش دهید تا میزان بازدید میانگین مشتریان در گروه‌های سنی مختلف قابل مقایسه باشد.

۳. **توزیع تعداد مشتریان در هر گروه سنی:**

– تعداد مشتریان در هر گروه سنی را محاسبه کنید.

– نتایج را با **هیستوگرام** (Histogram) نمایش دهید تا توزیع جمعیتی مشتریان مشخص شود.

– یک **نمودار دایره‌ای** (Pie Chart) نیز رسم کنید تا سهم هر گروه سنی از کل مشتریان به صورت درصدی مشخص شود.

خروجی نهایی

- **نام فایل:**

Session1_CustomerAgeAnalysis.pdf

- **محتوای گزارش:**
  - نمودار میله‌ای: میانگین بازدیدها به ازای هر گروه سنی
  - هیستوگرام: تعداد مشتریان در هر گروه سنی
  - نمودار دایره‌ای: سهم هر گروه سنی از کل مشتریان

۱.۶ تقسیم‌بندی بازدیدکنندگان(Segmentation)

۱. مقادیر زیر را برای هر مشتری محاسبه کنید:

- **total_visits** (تعداد کل جاذبه‌هایی که هر مشتری بازدید کرده است)
- **avg_entrance_fee** (میانگین هزینه ورودی جاذبه‌هایی که مشتری بازدید کرده است)

۲. با استفاده از الگوریتم **K-Means**، بازدیدکنندگان را بر اساس مقادیر محاسبه شده، به **۳ خوشه(cluster)** تقسیم‌بندی کنید.

۳. خوشه‌های حاصل را روی یک **نمودار پراکندگی(scatter plot)** نمایش دهید. total_visits در محور X و avg_entrance_fee در محورY

٤. برای هر خوشه رنگ متمایز انتخاب کرده و **راهنما (legend)** و **عنوان مناسب محورها** را اضافه کنید.

خروجی نهایی

- **نام فایل :**

Session1_VisitorSegmentation.pdf

- **محتوای گزارش:**
  - نمودار پراکندگی :**تقسیم‌بندی بازدیدکنندگان با K-Means**
    - محورX: Total Visits
    - محورY: Average Entrance Fee
    - رنگ: گروه خوشه