فاز اول پروژه

درس مبانی داده کاوی

نیمسال اول تحصیلی ۱۴۰۳

بخش اول – شناخت مجموعه داده

با توجه به مجموعه داده ای که در اختیار دارید، موارد زیر را برای آن انجام دهید.

۱. ویژگی های مجموعه داده را طبق جدول زیر برای داده های عددی بدست بیاورید.

❋ برای مثال ستون "Sales" را مورد بررسی قرار دهید.

نام ویژگی	نوع	بازه مقادیر	Min	Max	Mean	Mode	Median	Outlier

۲. با رسم نمودار Box Plot مقادیر پرت (Outlier) هر یک از ویژگی های بالا را شناسایی کنید.

۳. با رسم نمودار Scatter بین هر دو ستون مشخص شده، بررسی کنید آیا رابطه ای بین این ستون ها وجود دارد و در صورت وجود آن را توضیح دهید.

ویژگی اول	ویژگی دوم
Shipping cost	Sales
Shipping cost	Profit
Sales	Profit

بخش دوم - ارزیابی کیفیت داده

با بررسی مقادیر گم‌شده (missing) ، داده‌های پرت (outlier)، ناهمسانی‌ها و خطاهای موجود، کیفیت هر دو مجموعه داده را ارزیابی کنید و برای این مرحله سعی کنید موارد ذیل را برای آن انجام دهید.

۱. با توجه به مدل کیفیت ISO ۲۵۰۱۲ و بعد ذاتی آن، برای داده هایی که در اختیار دارید، کیفیت آن را با توجه به فاکتورهای کیفیت مربوطه ارزیابی نمایید و برای هر کدام چه درصدی از کیفیت حاصل می‌شود.

نام ویژگی	تعداد رکورد	تعداد مقدار Null	Accuracy	Completeness	Validity	Currentness	Consistency

* Consistency : باید دو ستون که به یکدیگر مرتبط هستند را سازگاری آنها را بررسی کنید. برای مثال بررسی کنید مقادیر ستون City با Country سازگار است یا خیر.
* Currentness : بررسی ستون "Order.Date" و خرید هایی که از سال ۲۰۱۱ تا الان هستند.
* Validity : بررسی مقادیر داده های هر ستون بر اساس فرمت داده شده در فایلdictionary
* Accuracy : به نسبت داده های معتبر و دارای مقدار به کل تعداد داده ها گفته می‌شود.

۲. با توجه به موارد زیر در جدول، اشکالات در دیتاست ها وجود دارد، را مشخص کنید و به صورت مختصر درباره هر کدام توضیح دهید.

Single-Instance	Single-Schema

۳. برای بهبود کیفیت داده مورد نظر، راهکارهای خود را ارائه نمایید.

بخش سوم - پیش پردازش(Preprocessing)

در این بخش شما باید داده هایی که در اختیار دارید را به فرمی ساختار یافته و تمیز تبدیل نمایید و در یک قالب مناسب برای تجزیه و تحلیل تبدیل کنید.

❋ در نظر داشته باید پیش پردازش شما به گونه ای باشد که در فاز بعدی برای موارد مطرح شده زیر مناسب باشد :

- باید بتوانید محصولاتی که با یکدیگر بیشتر خریداری شده اند را پیدا کنید.
- خوشه بندی کردن مشتریان بر اساس خرید هایی که انجام داده اند.
- محصولات فروخته شده بر اساس ویژگی فصل را پیدا کنید.

موارد زیر برخی از اقداماتی است که در این بخش باید انجام دهید.

۱. Missing value ها هندل شوند.

با توجه به مقادیر ستون های دیتاست، با استفاده از روش هایی مانند میانگین، مد، میانه و یا رگرسیون مقادیر ناموجود را مقداردهی کنید. در صورتی که ستونی بیش از میزان مجاز مقدار تاموجود داشت می توانید آن ستون را حذف کنید.

۲. تبدیل داده (data conversion)

برای برخی از داده های موجود در دیتاست عملیات نرمالسازی را انجام دهید.

۳. ساخت ویژگی های جدید

برای دستیابی به دانش بیشتر برخی از ستون ها را ترکیب کرده و به عنوان ستونی جدید در دیتاست ذخیره و نگه داری کنید.

٤. برای دادها های عددی outlier را شناسایی کنید و از دیتاست حذف کنید.

۵. در صورت نیاز از تکنیک های data reduction استفاده کنید.
۶. در صورت نیاز داده های عددی به داده های categorical تبدیل شوند .
۷. مصور سازی دیتاست بر اساس مقادیر موجود الزامی است.

- **بخش های زیر را تحلیل کنید**

- سه تا از بهترین شهرها بر اساس مجموع فروش را بیابید.
- کدام مشتری بیشترین تعداد خرید را داشته است.
- بیشترین میانگین هزینه ارسال مربوط به کدام شهر می‌باشد.
- سفارش هایی با اولیت "High" بیشتر با چه نحوه ارسالی، ارسال شده اند.
- با مقایسه segment و profit بگویید کدام نوع خریداران، سود بیشتری را رقم زده اند.

نکات تحویل :

- پروژه در گروه های **حداکثر** دو نفری پیاده سازی شود.
- فایل ها باید در قالب studentOneName-studentTwoName-Phase۱.zip ارسال شود.
- ارسال فایل تنها از طریق سامانه vu مورد قبول بوده و فایل های ارسال شده در تلگرام و... تصحیح نخواهد شد.
- تمامی مراحل انجام پروژه، نتایج تحلیل ها و پیش پردازش انجام شده را به صورت کامل و در یک **گزارش** ارائه دهید.

موفق باشید