

أمامكم بيانات مبيعات لشركة افتراضية، تحتوي على 5000 صف و24 عمود. البيانات تشمل تفاصيل الطلبات، العملاء، المنتجات، الأسعار، والشحن. الأعمدة تحتوي على بيانات كمية (مثل الأسعار والكميات)، بيانات نوعية (مثل المدينة ونوع العميل)، وتواريخ الطلب والشحن. بعض الأعمدة المالية مخزنة ك object وتحتاج تحويل قبل التحليل.

البيانات المقدمة:

https://www.kaggle.com/datasets/rajneesh231/retail-insights-a-comprehensive-sales-dataset?utm_source=chatgpt.com

الأهداف:

1. تنظيف البيانات وتجهيزها للتحليل.
2. استكشاف البيانات (Exploratory Data Analysis – EDA) وفهم الأنماط الأساسية.
3. تطبيق أساليب Data Mining على البيانات، بما في ذلك التصنيف، الانحدار، التجميع، قواعد الترابط، وتحليل البيانات الزمنية.

المطلوب:

الجزء 0: Know Your Data

قبل البدء بأي تحليل أو نموذج، تعرف على البيانات:

1. الأعمدة وأنواعها.
2. التحقق من القيم المفقودة.
3. تحويل أنواع البيانات.
4. الكشف عن القيم الشاذة: البحث عن أسعار أو كميات غير منطقية.
5. الفهم العام للبيانات: عدد الصفوف والأعمدة، توزيع العملاء، المنتجات الأكثر مبيعًا وربحية، الأنماط الموسمية في المبيعات.

الجزء 1: Data Cleaning

- إزالة أو معالجة القيم المفقودة.
- تحويل الأعمدة المالية إلى أرقام.
- التأكد من تنسيق التواريخ.
- معالجة القيم الشاذة أو الخارجة عن النطاق الطبيعي

الجزء 2: Exploratory Data Analysis

- عرض ملخص إحصائي للأعمدة الرقمية.
- رسم توزيعات المبيعات، الربحية، الخصومات.
- تحليل فروقات المبيعات بين المدن، أنواع العملاء، وفئات المنتجات.

الجزء 3: Data Mining Tasks

1- Classification

- التنبؤ بـ Customer Type أو Order Priority بناءً على خصائص الطلب.

2- Regression

- التنبؤ بـ Profit Margin أو Order Total من الكمية، السعر، الخصم، وطريقة الشحن.

3- Clustering

- تجميع العملاء حسب سلوك الشراء أو الموقع الجغرافي.
- تجميع المنتجات حسب الربحية أو الشعبية.

4- Association Rules

- اكتشاف المنتجات التي غالبًا ما تُشترى معًا.
- تحليل تأثير الخصومات على اختيار المنتجات المشتركة.

Time Series Analysis -5

- استخدام Order Date و Ship Date لتحليل المبيعات الشهرية أو الموسمية.
- محاولة التنبؤ بمستوى المبيعات للأشهر القادمة.

الجزء4: Report

- وثق جميع خطواتك: تنظيف البيانات، التحليل الاستكشافي، الرسوم البيانية.
- عرض نتائج التصنيف، الانحدار، التجميع، قواعد الترابط، والتحليل الزمني.
- استنتاجات: المنتجات الأكثر ربحًا، العملاء الأكثر نشاطًا، تأثير الخصومات، الأنماط الموسمية.

يمكنك استخدام مكتبات Python مثل:

pandas, numpy, matplotlib, seaborn, scikit-learn, mlxtend.

وأيضا تسليم الشغل كـ Jupyter Notebook
ورفعه على github.

Nora Hussain