به نام خدا

پروژه های درس داده کاوی ترم دوم سال تحصیلی ۱۴۰۰- ۱۳۹۹

فاز اول پروژه: پیش پردازش

برای یک مجموعه داده که مربوط به دیتاست دیوار می باشد. موارد ذیل را برای این دیتاست ها انجام دهید.

۱. ویژگی های مجموعه داده را طبق جدول زیر توصیف نمایید. سپس با رسم نمودار Box Plot مقادیر پرت هر ویژگی را شناسایی کنید.

ردیف	نام ویژگی	نوع	بازه مقادیر	Min	Max	Mean	Mode	Median	مقادیر پرت
۲.									

برای ویژگی ها (در صورت امکان)، قوانین **معتبر بودن** را تعریف نموده و میزان معتبر بودن رکوردها را براساس قوانین تعریف شده برای ویژگی ها ارزیابی کنید. چنانچه راهکاری برای برخورد با داده‌های غیر معتبر دارید، توضیح دهید.

۳. برای هر ویژگی مشخص نمایید با چه روشی می توان **صحت داده ها** را بصورت خودکار ارزیابی کرد. (به عنوان مثال در بخش نام استان، مقدار مشهد غیر صحیح است) راه حل خود را برای خودکاری سازی تشخیص داده‌های غیر صحیح برای هر ویژگی دارید بیان کنید.

۴. با استفاده از سه روش ارائه شده در کلاس، **کامل بودن** داده ها را ارزیابی کنید. بنظر شما کدام روش ارزیابی برای این مجموعه داده مناسب تر است؟ چرا؟

۵. براساس دانشی که نسبت به دیتاست به دست آورده اید، ۵ موضوع چالشی که میتوان بر روی این دادگان بررسی نمود را بیان نمایید.

فاز دوم پروژه: شناسایی الگوهای مکرر

در این فاز سه گام کلی زیر باید صورت بپذیرند:

- استخراج الگوهای قرار دادن پست فروش محصولات در شهرهای مختلف در دیتاست دیوار
- استخراج الگوهای فروش محصولات مختلف در دیتاست دیجیکالا
- مقایسه ارتباط میان خرید محصولات در شهرهای مختلف و قرار دادن پست فروش در شهرهای مختلف
- استخراج الگوهای مکرر میان روز انتشار پست و پلتفرم استفاده شده
- استحراج الگوهای مکرر برای محصول و پلتفرم استفاده شده

فاز سوم پروژه:دسته بندی محصولات براساس میزان محبوبیت و دسته بندی مناطق براساس میزان خرید افراد

در این فاز نیز چهار بخش اساسی باید صورت بپذیرد که به شرح ذیل می باشند:

- خوشه بندی شهرها براساس کالاهایی که در آن ها پست فروش گذاشته میشود(این خوشه بندی باید براساس همه ی فیلدهای مورد نیاز صورت بپذیرد)
- خوشه بندی شهرها براساس محصولاتی که در آن ها به فروش میرسد.
- مقایسه خوشه بندی ها صورت گرفته در دو فاز قبلی
- خوشه بندی محصولات براساس قیمت آنها

فاز چهارم پروژه(اختیاری):

در این فاز باید یکی از پروژه های زیر انتخاب و در صورت تمایل پیاده سازی شوند.

- تعیین پرت بودن یا نبودن یک قیمت پیشنهاد داده شده
- تعیین قیمت برای یک محصول
- تعیین قیمت براساس توضیحات درج شده برای یک محصول

دیتاست دیوار

توصیف ویژگی‌ها در دیتاست پستهای موجود بشرح زیر است:

نام	ویژگی
archive_by_user	تعیین آرشیو شدن پست
Brand	برند محصول
Cat1	دسته بندی نوع اول برای محصول
Cat2	دسته بندی نوع دوم برای محصول
Cat2	دسته بندی نوع سوم برای محصول
City	شهر
Created_at	زمان ایجاد پست
Desc	توضیحات محصول
Id	شناسه ی پست
Image_count	تعداد تصاویر
Mileage	مسافت پیموده شده
Platform	پلتفورم ایجاد پست
Price	قیمت
Title	عنوان
Year	سال

دیتاست دیجی کالا

نکته:برای فاز اول نیازی به پیش پردازش داده های این بخش نمی باشد.

دیتاست محصولات:

نام	ویژگی
ID	شناسه جدول
product_title_fa	عنوان فارسی محصول
product_title_en	عنوان انگلیسی محصول
url_code	کد لینک
category_title_fa	دسته بندی فارسی محصول
category_title_en	دسته بندی انگلیسی محصول
category_keywords	کلمات کلیدی دسته بندی
brand_name_fa	عنوان فارسی نام تجاری محصول
brand_name_en	عنوان انگلیسی نام تجاری محصول
product_attributes	ویژگی های محصول

دیتاست تاریخچه خرید مشتریان:

نام	ویژگی
ID_Order	شناسه سفارش
ID_Customer	شناسه خریدار
ID_Item	شاسه محصول
DateTime_CartFinalize	زمان تحویل محضول
Amount_Gross_Order	میزان سفارشات ناخالص

نام	ویژگی
city_name_fa	نام فارسی شهر
Quantity_item	تعداد خریداری شده از هر محصول

دیتاست نظرات:

نام	ویژگی
product_id	شناسه محصول
confirmed_at	زمان ثبت نظر
Comment	نظرات

دیتاست تاریخچه محصولات:

نام	ویژگی
ID	شناسه
product_variant_id	شناسه متغیر محصولات
selling_price	قیمت فروش محصولات
rrp_price	قیمت فروش محصولات
base_price	قیمت پایه محصولات
buy_price	قیمت خرید محصولات
order_limit	میزان محدودیت سفارش هر محصول
Start_at	تاریخ افزوده شدن محصول
end_at	تاریخ خاتمه محصول
Tags	تگ های محصولات
show_in_price_history	نمایش محصول در تاریخچه قیمت
Active	فعال بودن یا نبودن محصول
created_at	تاریخ ایجاد محصول در سایت
product_id	شناسه محصول
marketplace_seller_id	شناسه فروشنده کلی محصول

دیتاست نظرات مشتریان و کیفیت محصولات:

نام	ویژگی
product_id	شناسه محصول
product_title	عنوان محصول
title_en	عنوان انگلیسی محصول
user_id	شناسه کاربر
Likes	میزان دوست داشتن یک محصول
Dislikes	میزان دوست نداشتن یک محصول
verification_status	تایید شدن یا رد شدن نظر داده شده
Recommend	میزان توصیه درباره محصول
Title	عنوان محصول
comment	نظرات درباره محصول
advantages	مزایای محصول
disadvantages	معایب محصول

موفق باشید

مسیح ابوالفضلی اصفهانی