

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 داده های 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 داده های 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 می توان 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
- استخراج 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 می شود (این خوشه بندی باید براساس همه ی فیلدهای مورد نیاز صورت بپذیرد)
- خوشه بندی شهرها براساس محصولاتی که در آن ها به فروش می رسد.
- مقایسه خوشه بندی ها صورت گرفته در دو فاز قبلی
- خوشه بندی محصولات براساس قیمت آن ها

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:

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

دیتاست دیجی کالا

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

دیتاست محصولات:

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

دیتاست تاریخچه خرید مشتریان:

ویژگی	نام
شناسه سفارش	ID_Order
شناسه خریدار	ID_Customer
شناسه محصول	ID_Item
زمان تحویل محصول	DateTime_CartFinalize
میزان سفارشات ناخالص	Amount_Gross_Order
نام فارسی شهر	city_name_fa
تعداد خریداری شده از هر محصول	Quantity_item

دیتاست نظرات:

ویژگی	نام
شناسه محصول	product_id
زمان ثبت نظر	confirmed_at
نظرات	Comment

دیتاست تاریخچه محصولات:

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

دیتاست نظرات مشتریان و کیفیت محصولات:

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

موفق باشید