**به نام خدا**

**پروژه های درس داده کاوی ترم دوم سال تحصیلی 1400- 1399**

**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.

1. 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 |
|  |  |  |  |  |  |  |  |  |  |

1. برای ویژگی ها (در صورت امکان)، قوانین **معتبر بودن** را تعریف نموده و میزان معتبر بودن رکوردها را براساس قوانین تعربف شده برای ویژگی ها ارزیابی کنید. چنانچه راهکاری برای برخورد با داده‌های غیر معتبر دارید، توضیح دهید.
2. 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.
3. 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؟
4. براساس دانشی که نسبت به دیتاست به دست آورده اید، 5 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 |
| 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