

**گروه کامپیوتر** – **دانشکده مهندسی** – **دانشگاه فردوسی مشهد**

**گزارش فاز اول پروژه**

**درس: مبانی داده کاوی**

**ارائه دهندگان:**

**سبحان مرادیان دقیق - فاطمه چاجی**

**استاد:**

**دکتر بهشید بهکمال**

**پاییز 1399**

# **تعریف قوانین معتبر بودن داده ها**

# **ارزیابی صحت داده ها**

* 1. **دیتاست محصولات**
* **ويژگي کد لینک**

مقدار این ویژگی را با توجه به ویژگی عنوان فارسي محصول بدست می‌آوریم. بدین صورت که در بین کلمات موجود در ویژگی،

"-" قرار می‌گیرد.

* 1. **دیتاست تاریخچه خرید مشتریان**
* **ویژگی های شناسه خرید، شناسه سفارش، شناسه محصول**

مقادیر این شناسه ها باید یکتا باشند و تکراری بودن یک مقدار در یکی از ویژگی ها، نشان دهنده غیر صحیح بودن آن است.

* **ویژگی زمان تحویل محصول**

از آنجا که تاریخ تاسیس سایت دیجی کالا سال 2012 می‌باشد، تاریخ سفارش ثبت شده پیش از این زمان غیر صحیح می‌باشد که این امر در کد زده شده نیز پوشش داده شده است.

* **ویژگی نام شهر**

برای بررسی صحت مقدار وارد شده برای این ویژگی، يك مقدار آستانه تعريف مي‌كنيم. از آن جا كه اكثر اسامي بكار رفته براي اين ويژگي به ميزان زياد تكرار شده اند، اسامي تكرار شده كمتر از آستانه مساوي 2، صحيح نمي‌باشند و به درستي وارد نشده اند.

* 1. **دیتاست تاریخچه محصولات**
* **ویژگی های شناسه، شناسه متغیر محصولات، شناسه محصول و شناسه فروشنده کلی محصول**

مقادیر این شناسه ها باید یکتا باشند و تکراری بودن یک مقدار در یکی از ویژگی ها، نشان دهنده غیر صحیح بودن آن است.

* **ویژگی میزان محدودیت سفارش هر محصول**

با توجه به قیمت فروش هر محصول، در صورتی که قیمت فروش بالا باشد محدودیت سفارش محصول نمی‌تواند بالا باشد. برای مثال محصول با قیمت 10 میلیون، محدودیت سفارش بالای 5 نخواهد داشت چرا که منطقا به این میزان در انبار فروشگاه نیز موجود نخواهد بود. بنابراین با توجه به قیمت محصول، می‌توانیم برای میزان محدودیت سفارش، آستانه تعیین کنیم.

* **ویژگی هاي selling\_price و rrp\_price**

با مقایسه اين دو ويژگي و با توجه به تعریف rrp\_price[[1]](#footnote-1)، می‌توان نتیجه گرفت همواره رابطه زیر بین مقادیر وارد شده برای این ویژگی برقرار است:

selling\_price >= rrp\_price

* **ویژگی هاي قیمت پایه و قیمت خرید محصولات**

با توجه به تعريف دو ویژگی buy\_priceو base\_price می‌توان نتیجه گرفت همواره رابطه زیر بین مقادیر وارد شده برای این ویژگی برقرار است:

buy\_price <= base\_price

* **ویژگی های تاریخ افزوده شدن محصول و تاریخ ایجاد محصول در سایت**

با توجه به تعريف این دو ویژگی می‌توان نتیجه گرفت همواره رابطه زیر بین مقادیر وارد شده برای این ویژگی برقرار است:

start\_at >= created\_at

* 1. **دیتاست کیفیت محصولات**
* **ویژگی شناسه محصول**

مقادیر این شناسه باید یکتا باشند و تکراری بودن یک مقدار در اين ویژگی، نشان دهنده غیر صحیح بودن آن است.

* **ویژگی عنوان انگليسي محصول**

مقادير مربوط به اين ويژگي همگي دو كاراكتري هستند اما الزاما هر دو كاراكتري براي اين ويژگي مقداري صحيح نمي‌باشند. براي بررسي صحيح بودن مقادير اين ويژگي، بايد از ديتاست ديگري از ديجي كالا كه شامل عناوين انگليسي محصولات موجود مي‌باشد، كمك گرفت.

* **میزان دوست داشتن و نداشتن یک محصول**

از آنجا كه ممكن است ديتاست به طريقي مثلا توسط يك بدافزار دستكاري شده باشد، براي ميزان دوست داشتن و نداشتن محصول يك آستانه تعريف مي‌كنيم. بدين صورت كه هر عددي بالاتر از 3 برابر ميزان فروش محصول، مقداري صحيح براي ميزان دوست داشتن و نداشتن نمي‌باشد.

* 1. **دیتاست نظرات**
* **ویژگی شناسه محصول**

مقادیر این شناسه باید یکتا باشند و تکراری بودن یک مقدار در اين ویژگی، نشان دهنده غیر صحیح بودن آن است.

* **ویژگی زمان ثبت نظر**

از آنجا که تاریخ تاسیس سایت دیجی کالا سال 2012 می‌باشد، تاریخ نظر ثبت شده پیش از این زمان غیر صحیح می‌باشد.

# **ارزیابی کامل بودن داده ها**

1. recommended retail price: the price that the company that makes a product says it should be sold for [↑](#footnote-ref-1)