• ID: در صورتیکه طول و فرمت معتبر شناسه را میداشتیم میتوانیستیم صحت هر یک از اطلاعات این ستون را بررسی کنیم از آنجاییکه این اطلاعات در دسترس ما نیست پس حالت کلی اینکه شناسه بزرگتر از صفر باشد را بررسی میکنیم.

برای بررسی صحت داده میتوان از هر یک از 4 روش زیر استفاده کرد:

1. مقادیر min و max

برای متغیر های continuous بررسی مقادیر min و max میتواند محدوده ای که داده ها در آن صحیح به شمار میروند را بدست آورد. این مقادیر را میتوان از طریق field expert و یا هم با تخمین درست به دست آورد، وابسته به پیچیدگی و اهمیت متغیر است.

به طور مثال برای Amount\_Gross\_Order کمتیرین مقدار میتواند صفر و بیشترین مقدار 100 میلیارد مد نظر میگیریم اگر مقدار داده ای خارج از این محدوده باشد غلط است.

1. بررسی مقادیر تهی

آسانترین روش برای بررسی این موضوع استفاده از تابع count است.

1. بررسی مقادیر متغیرهای categorical

در هنگام جمع آوری داده مشکلات و بی دقتی های زیاد میتواند منجر به خطا شود. بک راه سریع برای بررسی صحت مقادبر این متفیر ها تهیه لیستی از مقادیر مجاز برای crosscheck کردن با متغیر متناظر آن است. مانند city\_name\_fa برای بررسی آن میتوان یک لیست مجاز شهر های ایران را تهیه کرد و برای هر Entry چک کرد که آیا در لیست موجود است و یا خیر.

1. بررسی نرخ تکرار مقادیر متغیرهای binary

در صورتیکه در دتاست خود از این نوع متغیر داریم میتوانیم با استقاده از mean(incidence rate) بررسی کنیم که آیا این متغیر به درستی مقدار دهی شده است یا خیر. یک خطای معمول برای این نوع از متفیر ها تعیین مقدار 1 و null عوض 1 و 0 است، به آسانی قابا شناسایی است زبرا Rate خود متغیر با 1 یکسان میشود.