به نام خدا

پیش بینی مشتریانی که نیاز به استعلام نهاب دارند

مقدمه :

تجزیه و تحلیل داده‌ها زمینه‌ای است که ما با آنالیز و بررسی داده‌های خام می‌توانیم اطلاعات مفیدی بدست آوریم. حوزه‌ی تجزیه و تحلیل داده‌ها مدت‌های زیادی است که وجود داشته است و شرکت‌ها از داده‌هایی که از طریق نظرسنجی یا داده‌هایی که از تراکنش‌های مشتری‌ها بدست می‌آورند برای استراتژی‌های بازاریابی استفاده می‌کنند. در گذشته شرکت‌ها کارشناس آمار را استخدام می‌کردند تا از داده‌ها اطلاعات مفیدی بدست آورند و پتانسیل بازار را شناسایی کنند تا مدل‌هایی برای فروش بیشتر ارائه کنند. تجزیه و تحلیل داده‌ها نسبت به سال‌های گذشته رشد گسترده‌ای داشته است. با افزایش قدرت محاسباتی، شرکت‌ها قادر به جمع آوری و ذخیره‌سازی و پردازش داده‌ها در مقیاس‌هایی هستند که در سال‌های گذشته غیرقابل تصور بود. تحلیل‌گران با استفاده از مهارتی که دارند می‌توانند مقدار خیلی وسیعی از داده‌ها را جمع آوری و پردازش کنند و با استفاده از مهارت آماری خودشان قادر هستند اطلاعات مفیدی از داده‌ها استخراج کنند.

**توضیح کلی در مورد دیتاست**

به دلیل هزینه بر بودن استعلام نهاب، لازم است برای برخی از فرآیندهای بیزینسی مانند تعریف مشتری، افتتاح سپرده، اعطای تسهیلات و صدور کارت، این استعلام انجام شود. این فرآیندها زمانبر هستند و باعث کند شدن سرویس میشوند. این دیتاست شامل اطلاعات مشتریانی است که به بانک مراجعه کرده اند و برای برخی از آنها نیاز به استعلام نهاب بوده است. هدف از جمع آوری این داده ها، تحلیل عوامل مؤثر بر استعلام نهاب میباشد. هر سطر در دیتاست نمایانگر یک مشتری است و ستونها شامل ویژگی های مختلفی هستند که میتوانند به پیشبینی نیاز یا عدم نیاز به استعلام کمک کنند.

**معرفی ستون های دیتاست**

اولین مرحله در هر پروژه تحلیل داده، جمع آوری داده های مورد نیاز است. داده های مشتری از ادغام سه فایل CSV تشکیل شدند. این فایل شامل اطلاعات زیر است :

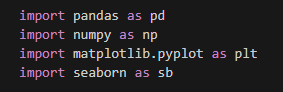
1. شماره مشتری (CIF)
2. کد پستی (PostalCode)
3. نوع مشتری (CustomerType)
4. تاریخ تولد (Birthdate)
5. کد تراکنش(TransactionCode)
6. آیا استعلام نیاز دارد یا خیر (requiredInquiry)

این دیتاست میتواند به تحلیل و شناسایی الگوهای مرتبط با نیاز به استعلام نهاب کمک کند و به محققان و کارشناسان در بهبود فرآیندهای بیزینسی یاری رساند. با بررسی دقیق هر یک از این ستونها، میتوان عوامل مؤثر بر نیاز به استعلام نهاب را شناسایی و راهکارهای مؤثری برای کاهش زمان و هزینه های مرتبط ارائه داد.

در این پروژه، با بررسی عوامل مؤثر و تجزیه و تحلیل اطلاعات، مدلهایی را اجرا میکنیم تا بهترین دقت در پیشبینی مشتریانی که نیاز به استعلام نهاب دارند را به دست بیاوریم. هدف این است که با استفاده از داده های موجود، بتوانیم مشتریان جدیدی با اطلاعات مشابه را شناسایی کنیم و پیشبینی کنیم که آیا نیاز به استعلام نهاب خواهند داشت یا خیر.

در ادامه توضیح هر بخش به طور کامل ارائه خواهد شد.

**مرحله 1 : بارگذاری کتابخانه ها و داده ها**



این کتابخانه ها ابزارهای قدرتمندی برای تحلیل داده ها و مصورسازی آنها فراهم میکنند.

1. Pandas : یکی از محبوبترین ابزارهای تحلیل داده در زبان برنامه نویسی پایتون است. این کتابخانه امکانات گسترده ای برای کار با داده های ساختاریافته و جدولی فراهم میکند. با استفاده از pandas میتوان داده ها را به راحتی بارگذاری، پاکسازی، و تحلیل کرد. این کتابخانه به ویژه برای کار با داده های سری زمانی و داده های جدولی بسیار مفید است.
2. numpy : یکی از پایه های اصلی محاسبات علمی در پایتون است. این کتابخانه امکانات گسترده ای برای کار با آرایه های چندبعدی و انجام محاسبات عددی فراهم میکند. numpy به دلیل سرعت بالا و کارایی مناسب، در بسیاری از پروژه های علمی و مهندسی مورد استفاده قرار میگیرد.
3. matplotlib : یکی از قدرتمندترین ابزارهای مصورسازی داده در پایتون است. این کتابخانه امکان ایجاد نمودارهای متنوع و سفارشی سازی آنها را فراهم میکند. با استفاده از matplotlib میتوان داده ها را به صورت گرافیکی نمایش داد و الگوها و روندهای موجود در داده ها را به راحتی شناسایی کرد.
4. seaborn : یک ابزار پیشرفته برای مصورسازی داده ها است که بر پایه matplotlib ساخته شده است. این کتابخانه امکانات بیشتری برای ایجاد نمودارهای آماری و گراف های پیچیده فراهم میکند. seaborn به ویژه برای مصورسازی داده های آماری و نمایش روابط بین متغیرها بسیار مفید است.

در این مرحله، با بارگذاری کتابخانه های pandas، numpy, matplotlib و seaborn، ابزارهای لازم برای تحلیل و مصورسازی داده ها را فراهم کرده ایم. این کتابخانه ها به ما کمک میکنند تا داده ها را به صورت کارآمد و دقیق تحلیل کرده و نتایج را به صورت گرافیکی نمایش دهیم.

بارگذاری داده ها:

داده های ما در سه دیتاست جداگانه قرار دارد که نیاز است با هم ادغام شوند.

دیتاست اول با نام nahab1DataSet شامل ستون های Cif ، TransactionCode ، Birthdate و CustomerType میباشد.در این دیتاست به ازای بعضی از cif ها چند کد تراکنش ذخیره شده و cif ستون unique نمیباشد.

دیتاست دوم با نام nahab2DataSet شامل ستون های Cif، transationCode ، Birtdate و postalcode می باشد که در این دیتاست cif یونیک می باشد.

دیتاست اول و دوم اطلاعات مشتریانی را شامل میشود که قبلا برای تراکنش مورد نظر، استعلام گرفتند.

دیتاست سوم با نام nahab3Dataset شامل اطلاعات مشتریانی است که استعلام نگرفتند.



فراخوا نی دیتاست

در این مرحله، ما از تابع read\_csv کتابخانه Pandas استفاده می کنیم تا داده ها را از فایل CSV بخوانیم. این تابع فایل را

بارگذاری کرده و داده ها را به یک DataFrame تبدیل می کند . متغیر های df\_nahab1 و df\_nahab2 حالا شامل داده های ما است و می توانیم با آن کار کنیم.

در ابتدا نیاز است دیتاست های مشتریانی که استعلام گرفتند با یکدیگر ادغام شوند. برای این کار از کد زیر استفاده میکنیم.



متغیر merged\_df ادغام دیتاست های مشتریانی که استعلام گرفتند می باشد.

پیش پردازش داده قبل ادغام فایل های اکسل :

این مرحله یکی از مراحل حیاتی در تحلیل داده است. این مرحله شامل تمیز کردن داده‌ها، حذف مقادیر مفقوده و تغییر نوع داده‌ها (در صورت نیاز) می‌شود. برای انجام این کار از کتابخانهpandas  استفاده می‌کنیم.

1. پر کردن مقدار صفر بجای null برای postalCode



1. حذف بخشdecimal مربوط به فیلد postalCode



1. حذف time بصورت 00:00:00 از ستون birthdate



1. چون کد تراکنش بصورت string است نیاز داریم آن را با مقادیر عددی map کنیم. برای این کار ، ستون TRANSACTIONCODE از دیتافریم merged\_df را گرفته و تمام مقادیر یکتا را در یک مجموعه (set) قرار میدهیم. سپس یک دیکشنری ایجاد میکنیم که در آن هر کد تراکنش به عنوان کلید و اندیس آن در مجموعه transactioinCode به عنوان مقدار (value) قرار میگیرد.



پس از نرمال کردن دیتاست، خروجی را دریک فایل جدید به اسم output نگه میداریم. این کار را با صدازدن متد to\_scv انجام میدهیم.

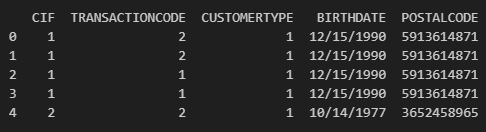


در مرحله بعد از این دیتاست برای ادغام با دیتاست مشتریانی که استعلام ندارند استفاده میکنیم.

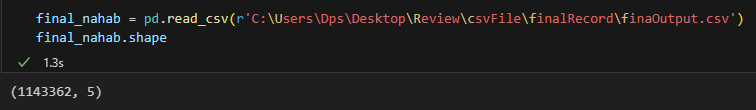


نمایش چند سطر ابتدایی داده ها

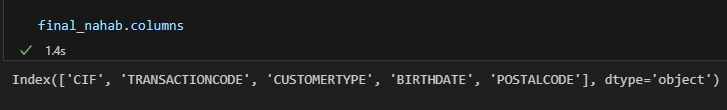
فایل نهایی ادغام شده را finalOutput می نامیم. با استفاده از متد head() ، ما میتوانیم پنج سطر اول DataFrame را مشاهده کنیم. این کار به ما کمک می کند تا نگاهی سریع به ساختار و محتوای داده ها بیندازیم و بررسی کنیم که آیا داده ها به درستی بارگذاری شدند یا خیر.

****

برای اینکه تعداد سطر و ستون ها را مشاهده کنیم از دستور df.shape استفاده کنیم . عدد اول تعداد سطرها و عدد دوم تعداد ستون ها را نمایش میدهد.

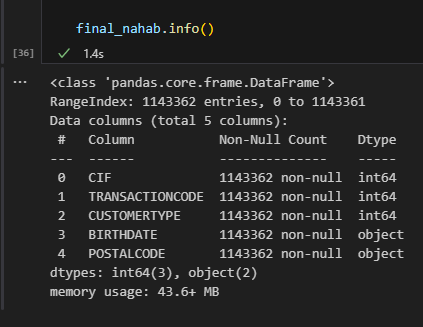


دستور df.columns در کتابخانه ی Pandas در پایتون برای نمایش نام های ستون های یک DataFrame استفاده میشود. این دستور یک Index از نام های ستون های DataFrame را برمی گرداند و به ما کمک می کند تا به نام ستون ها دسترسی پیدا کنیم و آن ها را بررسی کنیم . ممکن است در نام ستون ها خطای املایی یا کاراکتر اضافه ای باشد .



بررسی داده ها :

در این مرحله، تعداد کل سطرها و ستون ها و همچنین نوع داده های هر ستون را بررسی می کنیم . این دستور نشان میدهد که چند ستون و سطر داریم و نوع داده های هر ستون را مشخص می کند. همچنین نشان می دهد که آیا داده های مفقود داریم یا خیر.

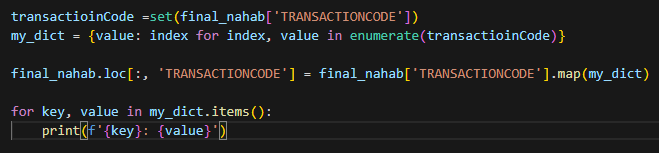
****

در حال حاضر فایل نهایی که ادغام شده و روی آن کار میکنیم finalOutput نام دارد. این فایل اکسل شامل ستون های زیر است :

1. Cif
2. transactionCode
3. customerType
4. Birthdate
5. postalCode
6. requiredInquiry

حال باید دیتاهای این فایل برای اینکه به عنوان فایل آموزشی ارسال شود ، تحلیل و بررسی شود.

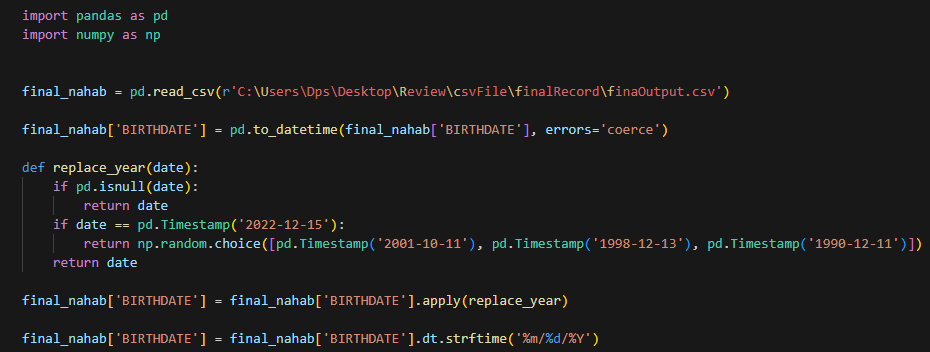
مورد اول کد تراکنش می باشد. چون کد تراکنش بصورت string ی تعریف شده بود آن را به داده ی عددی map کردیم. قطعه کدی که معادل سازی را نمایش میدهد بصورت زیر است :



و دیتای map شده معادل جدول زیر می باشد :

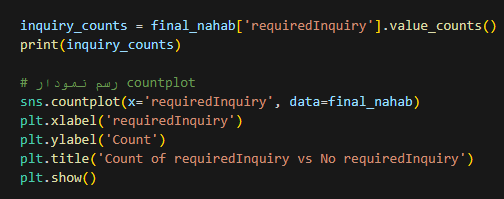
|  |  |  |
| --- | --- | --- |
| توضیح | معادلسازی | کد تراکنش |
| مشتریانی که استعلام نگرفتند | -1 |  |
| مشتریانی که کد تراکنش برای آنها ثبت نشد | 0 |  |
|  | 1 | D106601 |
|  | 2 | D109301 |
|  | 3 | C107012 |
|  | 4 | D104701 |
|  | 5 | D190058 |
|  | 6 | D130088 |
|  | 7 | E702101 |

در مرحله بعد، ستون تاریخ را بررسی میکنیم. این ستون یک سری دیتاهای تستی اشتباه با مقدار 12/15/2022 دارد. چون این تاریخ تولد برای مشتری معتبر نمی باشد بصورت رندم با مقادیر 11/10/2001 و 13/12/1998 و 11/12/1990 جایگزین کردیم. قطعه کد مربوط به آن در زیر آمده :

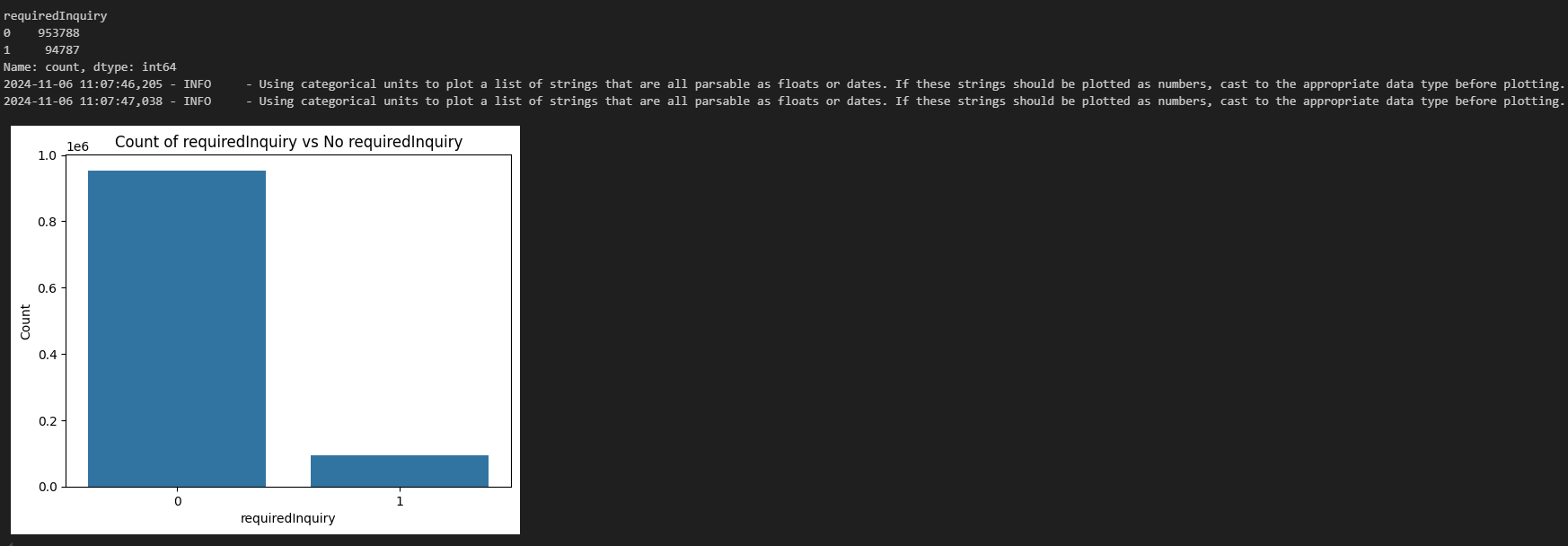


بررسی وضعیت ستون تارگت (requiredInquiry)

با استفاده از دستور counts\_value میتوانیم تعداد افرادی که استعلام گرفتند یا بلعکس را مشاهده کنیم و برای درک بهتر روی نموار countplot پیاده سازی کنیم.

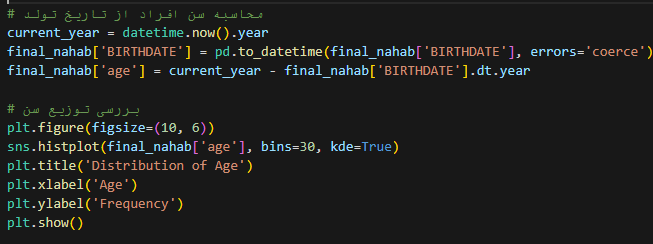


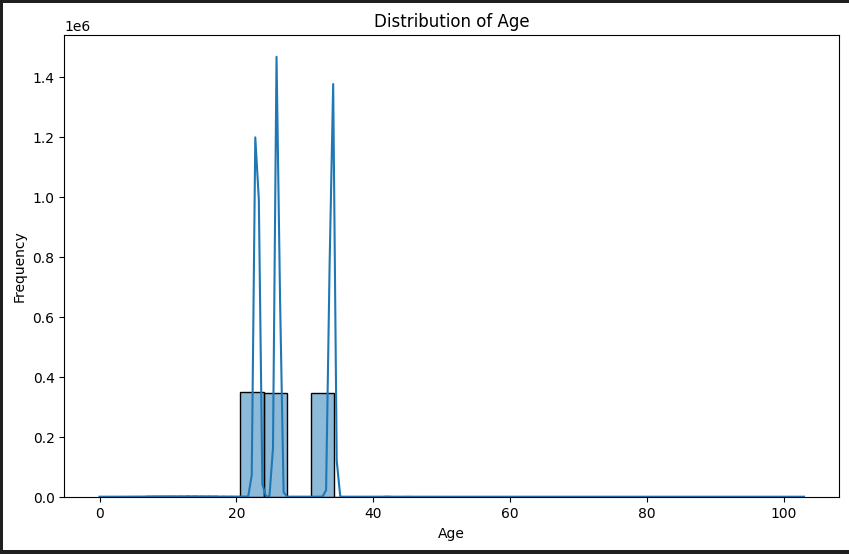
خروجی :



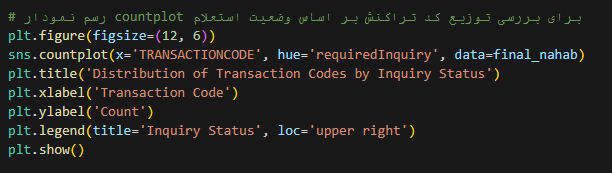
بررسی توزیع ویژگی های مختلف و شناسایی الگوهای موجود در داده ها :

توزیع سن: توزیع سن افراد را بررسی می کنیم تا ببینیم آیا سن در استعلام تاثیر دارد یا خیر.

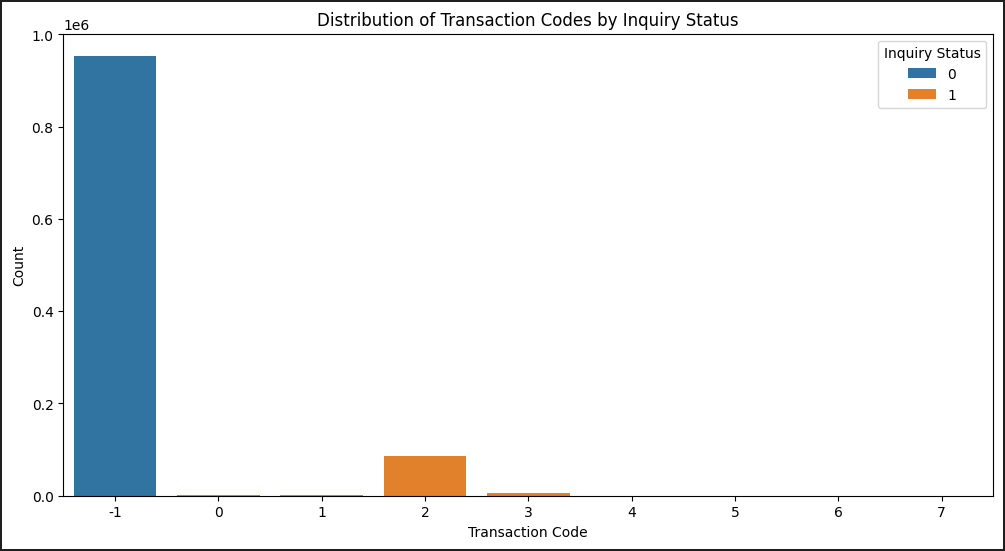
****

****

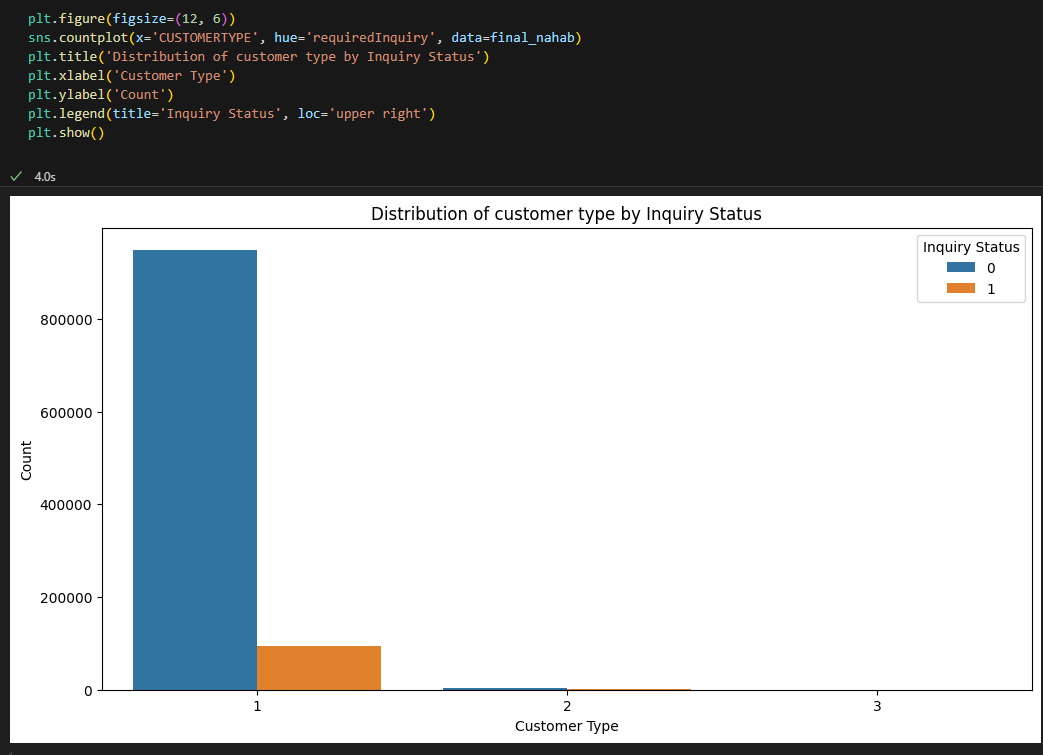
توزیع کد تراکنش : رسم نمودار countplot برای بررسی توزیع کد تراکنش بر اساس وضعیت استعلام

****

طبق نمودار مشخص شده که مشتریانی که کد تراکنش 2 معادل D109301 داشتند، بیشترین آمار استعلام نهاب را شامل می شوند.

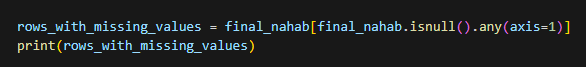
****

توزیع نوع مشتری : رسم نمودار countplot برای بررسی توزیع نوع مشتری بر اساس وضعیت استعلام

****

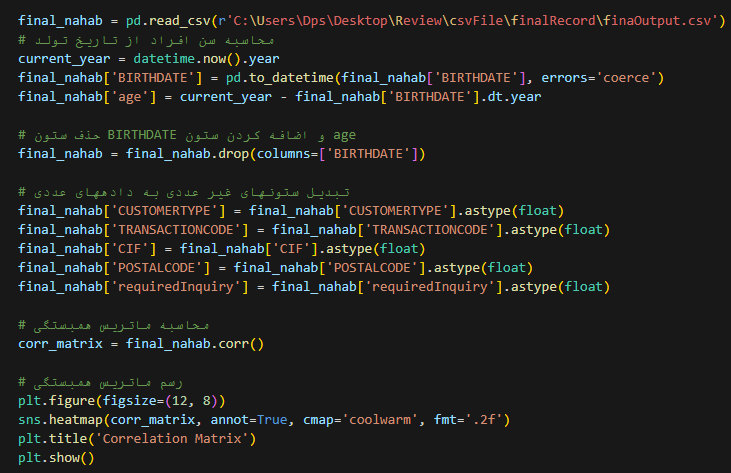
ماتریس همبستگی :

ابتدا بررسی میکنیم اگر مقدار NULL وجود دارد حذف یا با دیتای صحیح چایگزین شود. پیدا کردن مقادیر NULL

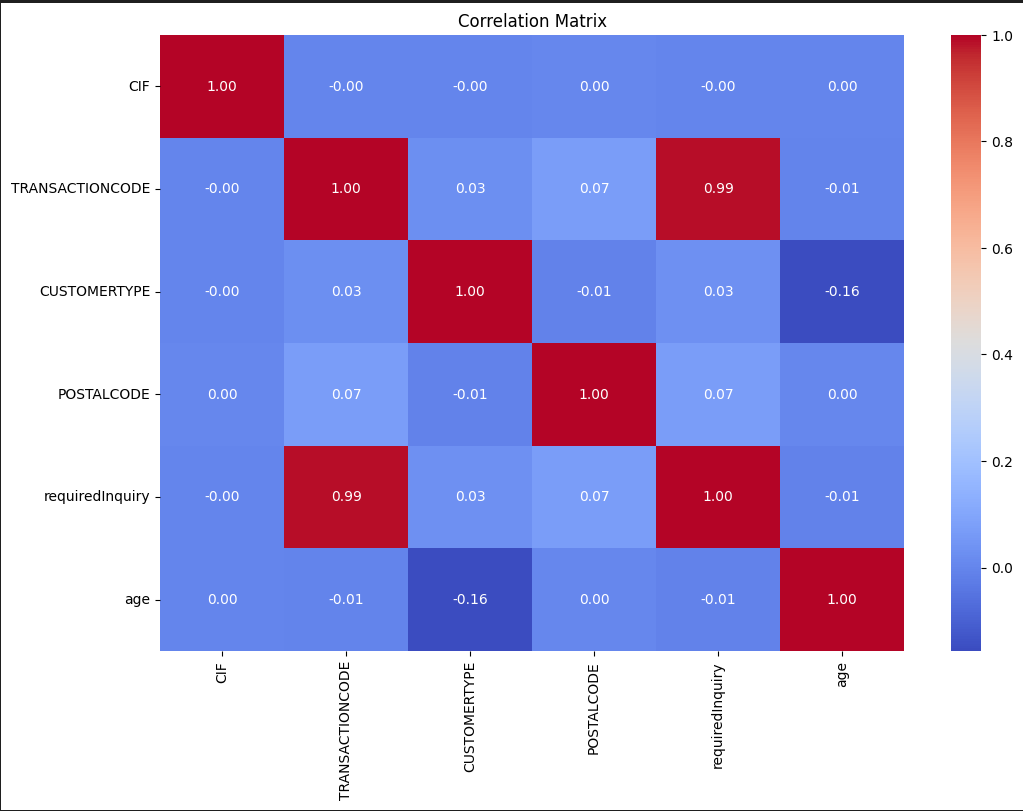
****

کد مربوط به ماتریس همبستگی :

corr\_matrix = final\_nahab.corr : این خط از کد ماتریس همبستگی را محاسبه میکند. ماتریس همبستگی یک جدول است که نشان میدهد چگونه متغیرهای مختلف با یکدیگر همبستگی دارند. مقدار همبستگی بین -1 و 1 متغیر است که 1 نشاندهنده همبستگی کامل مثبت، -1 نشاندهنده همبستگی کامل منفی و 0 نشاندهنده عدم همبستگی است.

****

خروجی :

****

تحلیل ماتریس همبستگی :

1: همبستگی کامل مثبت. به این معنا که با افزایش یکی از متغیرها، دیگری نیز به همان نسبت افزایش مییابد.

0: عدم همبستگی. به این معنا که هیچ رابطه خطی بین دو متغیر وجود ندارد.

-1: همبستگی کامل منفی. به این معنا که با افزایش یکی از متغیرها، دیگری به همان نسبت کاهش مییابد.

تفسیر اعداد در ماتریس همبستگی:

مقادیر مثبت (مثلاً 0.5): نشاندهنده همبستگی مثبت است. به این معنا که با افزایش یکی از متغیرها، دیگری نیز افزایش مییابد.

مقادیر منفی (مثلاً -0.5): نشاندهنده همبستگی منفی است. به این معنا که با افزایش یکی از متغیرها، دیگری کاهش مییابد.

مقادیر نزدیک به 0 (مثلاً 0.1 یا -0.1): نشاندهنده همبستگی ضعیف یا عدم همبستگی است.

برای مثال در ماتریس همبستگی بالا، مقدار همبستگی بین CustomerType و Age برابر با -0.16 است. این مقدار نشاندهنده یک همبستگی منفی ضعیف بین این دو متغیر است. به عبارت دیگر، با افزایش سن، نوع مشتری به طور ضعیفی کاهش مییابد، اما این رابطه خیلی قوی نیست.

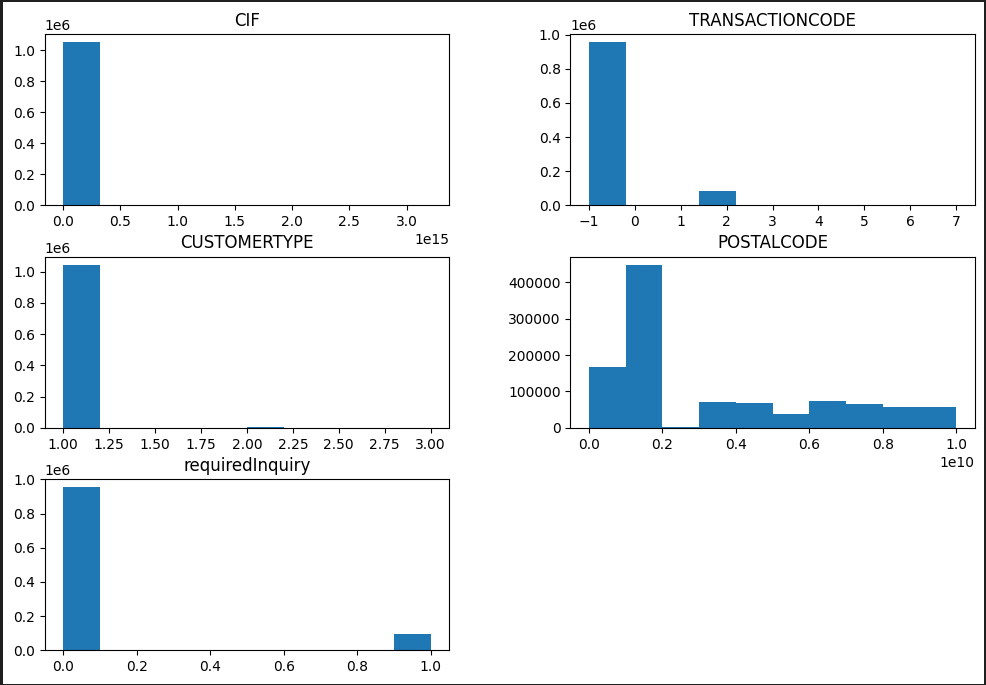
مقدار همبستگی 0.99 بین کد تراکنش (TransactionCode) و تارگت استعلام (requiredInquiry) نشاندهنده یک همبستگی بسیار قوی و مثبت بین این دو متغیر است. به عبارت دیگر، این مقدار نشان میدهد که تغییرات در کد تراکنش به شدت با تغییرات در تارگت استعلام مرتبط است.

**مشاهده نمودار هیستوگرام داده ها به صورت یکجا**

برای درک بهتر وضعیت داده های موجود در ستون ها با استفاده از نمودار هیستوگرام میتوانیم توزیع هر ستون را مشاهده کنیم.



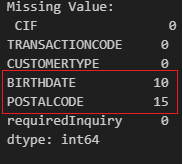
خروجی :



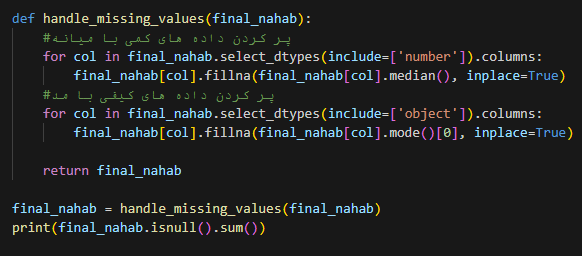
مدیریت داده های مفقوده Data Missing

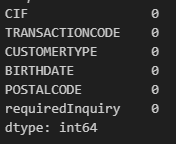
در این مرحله، ما به بررسی داده های مفقود می پردازیم. ابتدا تعداد داده های مفقود در هر ستون (NaN) را شناسایی می کنیم. سپس برای جایگزینی داده های مفقود: برای ویژگی های کیفی (داده های دسته ای)، از مد (mode) و برای ویژگی های کمی، از میانه (median) استفاده میکنیم. این کار به ما کمک میکند تا از حذف دادهها جلوگیری کنیم و اطلاعات بیشتری را حفظ کنیم . این کد تعداد داده های مفقود در هر ستون را نمایش می دهد تا بدانیم چه تعداد داده در هر ستون ناقص است.





در این مرحله داده های مفقود را مدیریت می کنیم. برای ستون های عددی از میانه و برای ستون های غیرعددی از مد استفاده می کنیم . مقادیر مفقود در ستون های عددی با میانه و مقادیر مفقود در ستون های غیرعددی با مد پر می شوند. پس از پر کردن، تعداد داده های مفقود مجدداً بررسی می شود تا مطمئن شویم تمامی داده های مفقود برطرف شده اند .





**انتخاب ویژگی ها و برچسب ها**

در این مرحله، داده ها به دو بخش تقسیم می شوند: ویژگی ها (X)و متغیر هدف (Y). این کار برای آماده سازی داده ها برای آموزش مدل ها ضروری است. متغیر هدف نشان می دهد که آیا فرد نیاز به استعلام دارد یا خیر. ستون target به عنوان متغیر وابسته در نظر گرفته می شود و سایر ستون ها به عنوان متغیرهای مستقل برای مدل سازی استفاده خواهند شد .