پروژه درس داده‌کاوی – ترم پاییز 1401 – فاز 1

سید محمد طاها طباطبایی – ملیکا ذبیحی نیشابوری

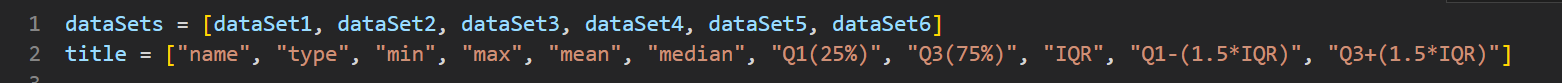
1) توضیح کد:

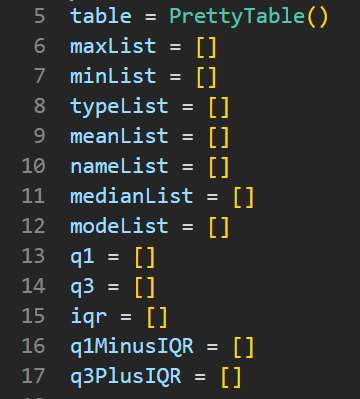
در سلول اول، چهار کتابخانه مورد نیاز برای پردازش های این فاز، ایمپورت شده است.

* کتابخانه pandas برای پردازش روی داده‌های دیتاست.
* کتابخانه prettyTable برای رسم جداول.
* کتابخانه seaBorn برای رسم نمودار های باکس پلات.
* کتابخانه matplotlib برای تصویرسازی های مورد نیاز.

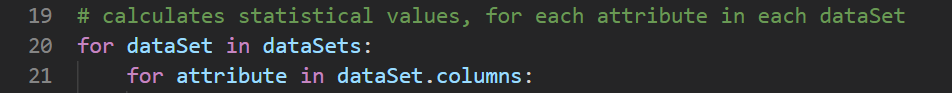
در سلول دوم، 6 فایل csv دیتاست به ترتیب لود می‌شوند.

\* فایل آخر با انکودینگ دیفالت که utf-8 بود، لود نمیشد، به همین دلیل از انکودینگ cp1252 استفاده کردیم.

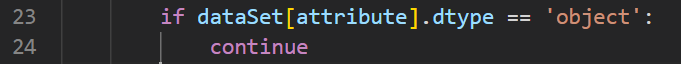


در سلول سوم، دو لیست می‌سازیم. لیست dataSets شامل 6 دیتاست موجود و لیست title، عناوین ستون های جدولی است که برای نمایش مقادیر خواسته شده در صورت تمرین، استفاده می‌کنیم.

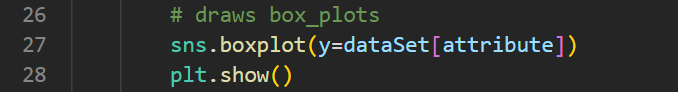
در شکل بالا، لیست های ساخته شده ای مشاهده می‌شود، که برای نگهداری مقادیر ستون ها، از آن‌ها استفاده می‌کنیم.



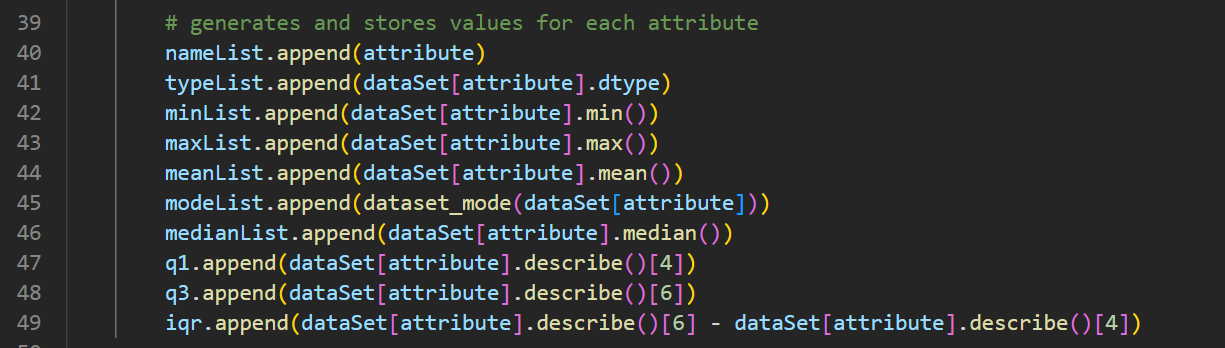
در دو حلقه تو در تو، هر ویژگی از هر جدول را بررسی می‌کنیم.

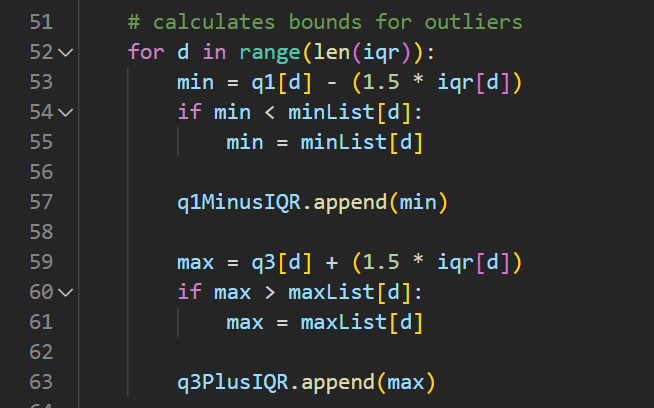
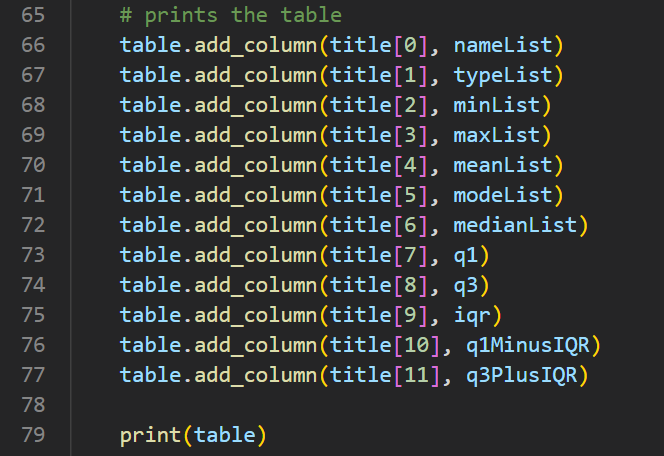


در همان ابتدای تحلیل، با توجه به اینکه ستون هایی با مقادیر غیر عددی، از لحاظ آماری و برای بررسی ویژگی های عددی، مناسب نیستند، کنار گذاشته می‌شوند. پس در این حالت، ادامه حلقه داخلی اجرا نمی‌شود.

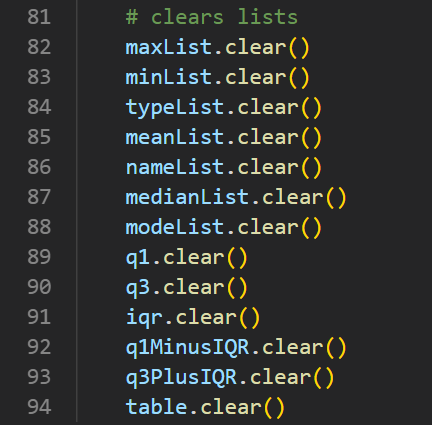


طبق خواسته مرحله دوم سوال، نمودار باکس پلات را برای هر ویژگی رسم می‌کنیم.

در این بخش کد، مقادیر لازم برای محاسبه داده‌های آماری خواسته شده را، استخراج می‌کنیم. به طور مثال در خط 32، میانگین مقادیر هر ویژگی، به لیست minList افزوده می‌شود.

برای محاسبه مرز داده های پرت، طبق فرمول معرفی شده در اسلاید ها عمل می‌کنیم. مرز پایین در شرط اول و مرز بالا در شرط دوم، به ازای هم ویژگی تعیین می‌شود.

در این بخش، جدول اطلاعات آماری را می‌سازیم و آن را چاپ می کنیم. 6 جدول در خروجی خواهیم داشت که در فایل های جداگانه، همراه کد ارسال شده اند.

در انتهای حلقه داخلی، لیست های استفاده شده را پاک می‌کنیم تا در پیمایش بعدی، مجددا مورد استفاده قرار بگیرند.

2) تحلیل و بررسی:

متاسفانه دیتاست های موجود از کیفیت پایینی برخوردار هستند. مقادیر رکورد ها برای بسیاری از ویژگی ها، null است، که عملا تحلیل و بررسی بر اساس آن ستون را بی‌معنی می‌کند.

همچنین رسم باکس پلات و محاسبه شاخص های آماری برای ویژگی هایی که فرمت هایی مثل آیدی یا تاریخ دارند، عملا بی اهمیت است، هرچند در کد ما، نادیده گرفته نشدند.

3) ایده تحلیلی:

پیش‌بینی مقادیر ستون returnAmvalToAnbar که "وضعیت اموال برگشتی به انبار" را نشان می‌دهد.

شرح:

ستون returnAmvalToAnbar ، وضعیت اموال برگشتی به انبار را مشخص می‌کند. برای این ویژگی، 4 حالت وجود دارد:

1- مستعمل قابل بهره برداری

2- مستعمل ملزم به تعمیر

3- اسقاط قابل تفکیک

4-اسقاط غیر قابل تفکیک

مقادیر این ستون برای رکورد های دیتاست در دو جدول productInstance و transfer\_item\_D کاملا null است. به عنوان یک پیشنهاد برای تحلیل داده، می‌توان با حل یک مساله خوشه‌بندی، وضعیت اموال را در 4 حالت فوق، پیش‌بینی کرد.