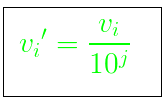
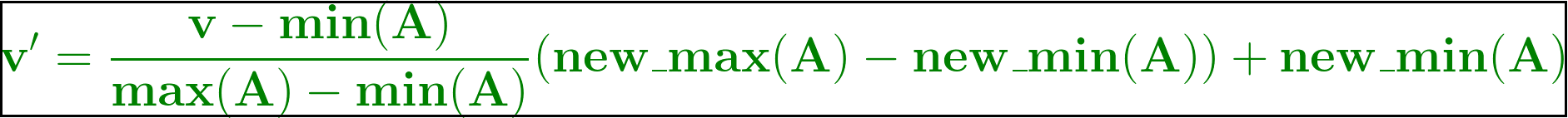
**سوال اول:**

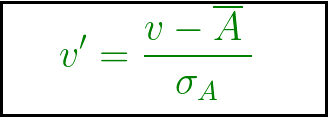
* Dimension: ساختاری که برای دسته­بندی داده­ها مورد استفاده قرار می­گیرد و کار را برای پیدا کردن و استخراج الگوی داده­ها آسانتر می­کند.
* Outlier: داده­هایی هستند که از داده­های معمول ما به دور هستند و به آنها شباهت ندارند. این داده می­توانند بسیار مفید واقع شوند.
* Independent variable: متغیرهایی که مستقل هستند و به عنوان متغیرهایی ورودی مدل مورد استفاده قرار می­گیرند، گفته می­شود. از این متغیرها می­توان، متغیرهای وابسته را بدست آورد.
* Dependent variable: ترجمه آن متغیر وابسته است و از روی متغیر مستقل قابل محاسبه و پیش­بینی است. معمولا هدف ما پیدا کردن مقدار این متغیر­ها است بنابراین به آنها متغیر هدف نیز می­گویند.
* Stratified sampling: در این نمونه­برداری ابتدا داده­ها را به دسته های کوچکتر می­شکنیم و سپس از هر دسته تعدادی را انتخاب می­کنیم (در هر دسته داده­های مشابه قرار می­گیرند).

**سوال دوم:**

* Decimal Scaling: با جابجایی نقطه اعشار، مقادیر داده­ها نرمال می­شود. برای اجرای این تکنیک، هر مقدار داده را بر حداکثر مقدار مطلق داده­ها تقسیم می­کنیم.



* Min-Max Normalization: در این تکنیک نرمال­سازی داده­ها، تبدیل خطی روی داده­های اصلی انجام می­شود. حداقل و حداکثر مقدار از داده ها پیدا می­شود و هر مقدار مطابق فرمول زیر جایگزین می­شود.
* z-Score Normalization: در این تکنیک، مقادیر بر اساس میانگین و انحراف معیار داده A نرمال می­شوند. فرمول استفاده شده به صورت زیر است:



**سوال سوم:**

در الگوریتم ChiMerge ابتدا یک مرحله مقداردهی اولیه وجود دارد. در این مرحله با مرتب سازی نمونه­ها و مجموعه­ها براساس مقدارشان برای ویژگی­های در حال گسسته سازی و سپس تشکیل گسسته سازی اولیه مقدار دهی اولیه می­شوند. سپس در انتها یک فرآیند مبتنی بر ادغام از پایین به بالا است، که در آن فواصل از طریق آماره­های محاسبه می­شود و بازه­های مجاور با کمترین مقدار فاصله ادغام می­شوند تا وقتی که یک شرط خاتمه برآورده شود.

**سوال چهارم:**

*Cosine similarity:*

*Correlation:*

*Euclidean distance:*

**سوال پنجم:**

روش کاهش داده­ها ممکن است به توصیف فشرده­ای از داده­های اصلی دست یابد که از نظر کمیت بسیار کوچکتر است اما کیفیت داده­های اصلی را حفظ می­کند.

روش های مختلفی برای این کار وجود دارد که به شرح زیر هستند:

* Data Cube Aggregation
* Dimension reduction: زمانی که به داده‌ای برخورد می‌کنیم که اهمیت کمتری دارد، فقط از ویژگی مورد نیاز برای تحلیل خود استفاده می‌کنیم. با این کار اندازه داده­ها را کاهش می­دهیم زیرا ویژگی های منسوخ یا اضافی را حذف می­کنیم.
  + Step-wise Forward Selection
  + Step-wise Backward Selection
  + Combination of forwarding and Backward Selection
* Data Compression: تکنیک فشرده­سازی داده­ها با استفاده از مکانیزم­های مختلف رمزگذاری، حجم فایل­ها را کاهش می­دهد.
  + Lossless Compression
  + Lossy Compression
* Numerosity Reduction: در این روش، داده­های واقعی با مدل­های ریاضی یا نمایش کوچکتر آنها جایگزین می­شوند.
* Discretization & Concept Hierarchy Operation
  + Top-down discretization
  + Bottom-up discretization

**سوال ششم:**

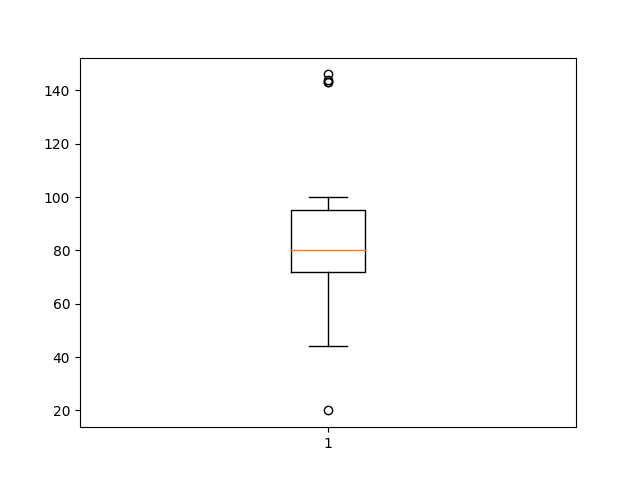
به هر چیزی که بتوانیم برای یک داده محاسبه کنیم، ویژگی گفته می­شود. در استخراج ویژگی یاfeature extraction تمام عوامل قابل محاسبه برای داده­ها را بدست می­آوریم. در انتخاب ویژگی یا feature selection از بین مجموعه ویژگی­های موجود، یک زیرمجموعه ویژگی انتخاب می­شود که معمولاً این زیرمجموعه از ویژگی­ها، از همه مفیدتر هستند. به عبارت دیگر ابتدا استخراج ویژگی انجام شده و سپس از نتیجه­ی آن برای انتخاب ویژگی استفاده می­شود.

تبدیل موجک یک تکنیک پردازش سیگنال است که سیگنال­های خطی را تبدیل می­کند. هنگامی که این روش اعمال می­شود، بردار داده به یک بردار عددی متفاوت متشکل از ضرایب موجک تبدیل می­شود.

موجک تبدیل در کاهش داده­ها نیز مفید است. اگر بخش کوچکی از قوی ترین ضرایب موجک را ذخیره کنیم، آنگاه می­توان تقریب فشرده­ی داده­های اصلی را به دست آورد.

**سوال هفتم:**

نمودار جعبه­ای به شکل زیر می­باشد:

****

* Min = 44
* Max = 144
* Outliers = 20, 146

**سوال هشتم:**

* noise به طور پیش فرض نامطلوب است، زیرا مقدار اصلی ویژگی را تحریف می­کند. outlierها به طور بالقوه می­توانند درست باشند، و حتی شناسایی آنها می­تواند هدف اصلی برخی از وظایف داده کاوی باشد. بنابراین، outlierها به طور بالقوه می­توانند جالب و یا مطلوب باشند، اما نویز (طبق تعریف) اینطور نیست.
* وجود نویز در ویژگی­ها می تواند داده­ها را تصادفی­تر یا غیرعادی­تر به نظر برساند. بنابراین، ممکن است برخی از نمونه­ها در داده های پر نویز به صورت outlier ظاهر شوند.
* outlierها می­توانند اشیاء درست باشند که به نظر می­رسد به مجموعه داده تعلق ندارند. آنها معمولاً به عنوان نویز طبقه بندی نمی­شوند.
* نویز در داده­ها می­تواند به طور تصادفی برخی از مقادیر درست را غیرعادی و یا برخی نقاط پرت را به عنوان اشیاء درست جلوه دهد.

**سوال نهم:**

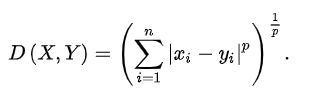
* مقادیر ممکن بین 1- تا 1 یا بین 0 تا 1 است.
* خیر، ممکن است دو داده ضریبی از یکدیگر باشند و برابر نباشند.
* اگر میانگین دو داده برابر صفر باشد آنگاه می­تواند گفت Cosine Similarity و Correlation آنها باهم برابرند.

**سوال دهم:**

در نمودار quantile، ما بین داده­ها و یک ویژگی آنها نتیجه­گیری کنیم ولی در نمودار quantile-quantile دو مجموعه داده مجزا با یکدیگر مقایسه می­شوند.

**سوال یازدهم:**

برای داده­های عددی می­توانیم از فاصله Minkowski استفاده کنیم که فرمول آن همانند زیر است:

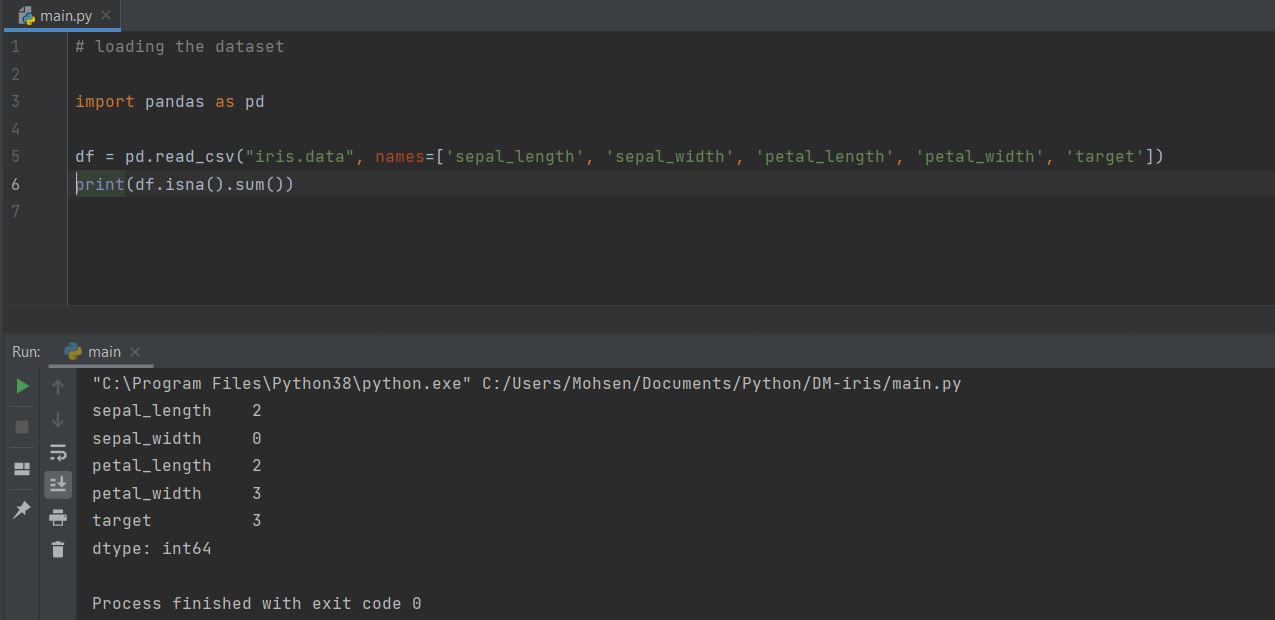


برای داده­های اسمی باید ابتدا آنها را به داده­های باینری تقسیم کنیم و سپس از روش های دیگر محاسبه فاصله همانند جاکارد استفاده کنیم.

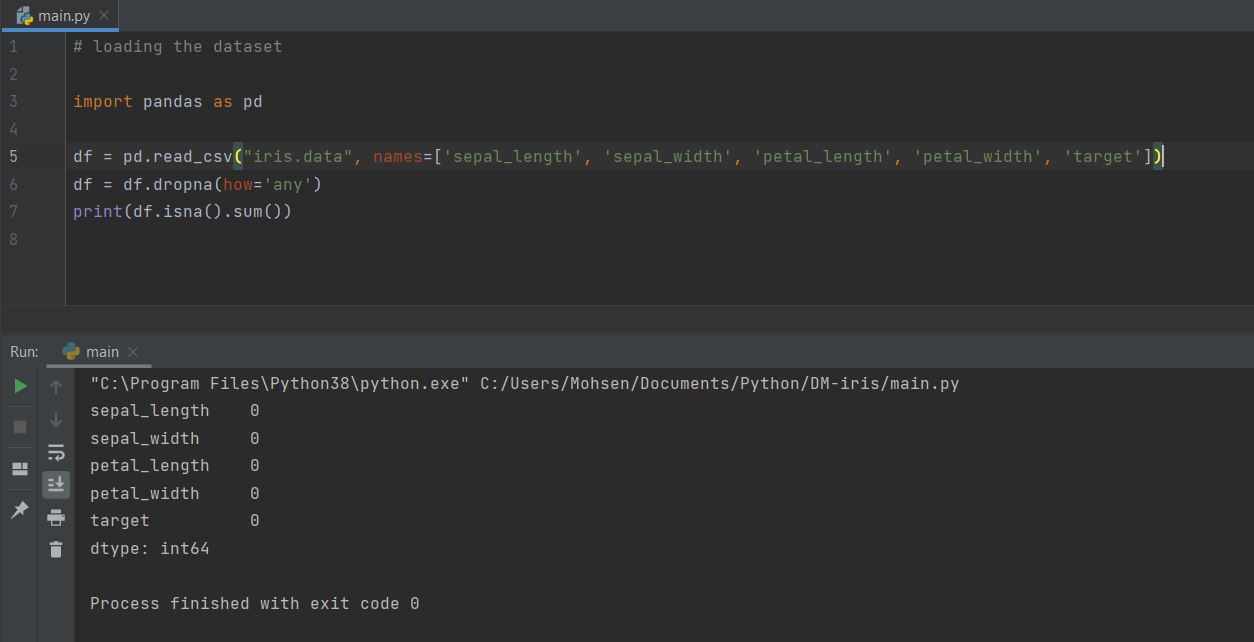
**بخش پیاده­سازی:**

تمامی اسکرین­شات ها در پوشه Screenshots قرارداده شده­اند.

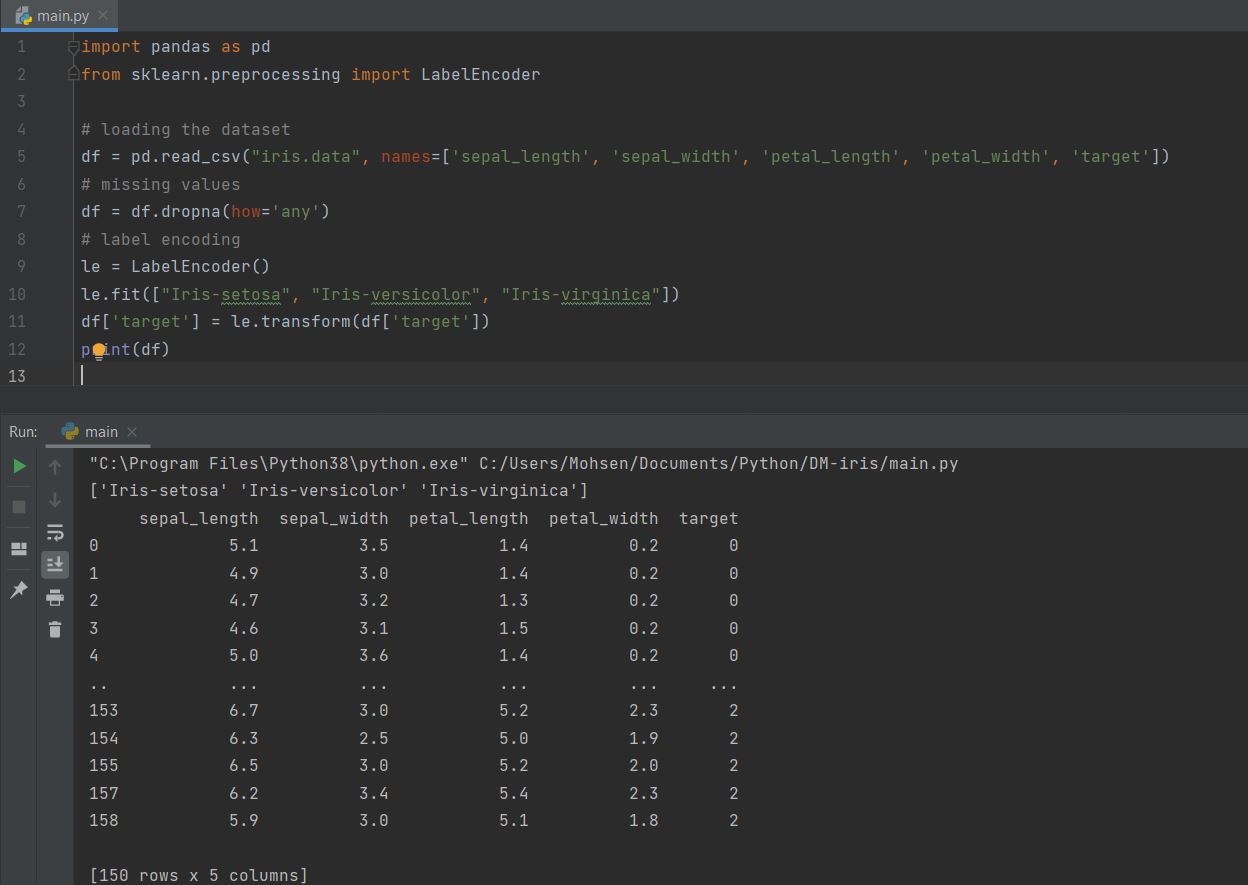
ابتدا تعداد ردیف­های دارای مقدار NaN را برای هر ویژگی پیدا می­کنیم:



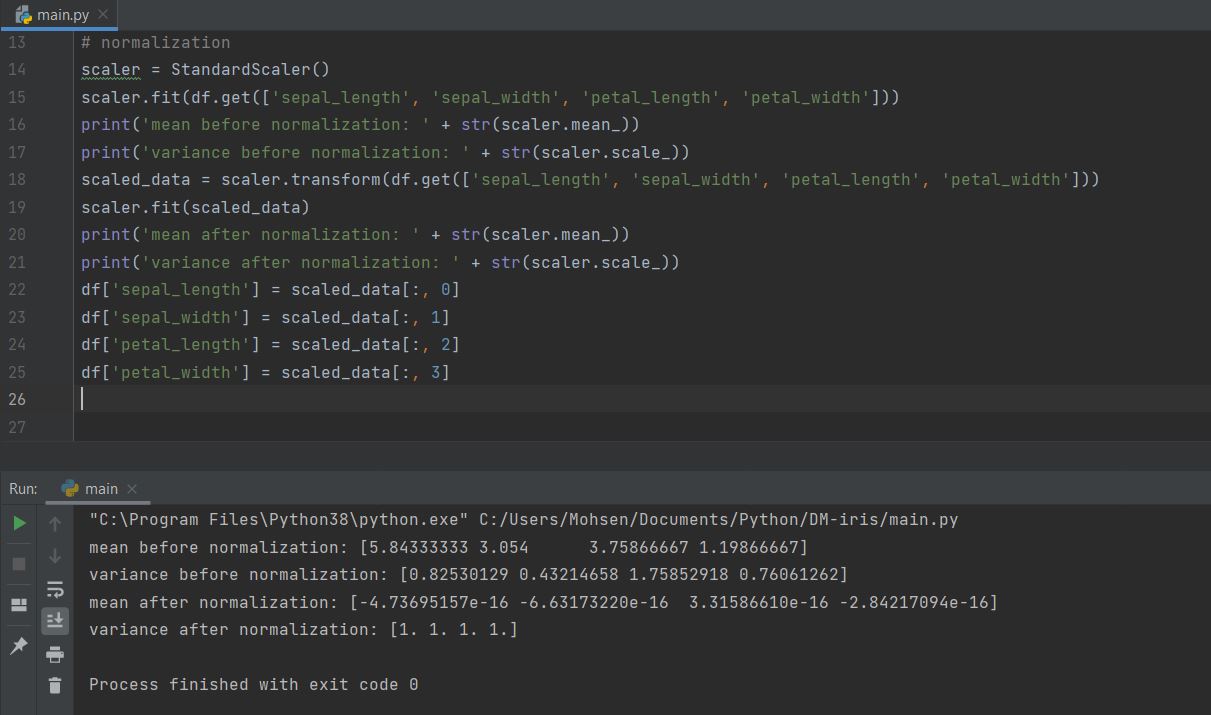
سپس آنها را از مجموعه حذف می­کنیم:



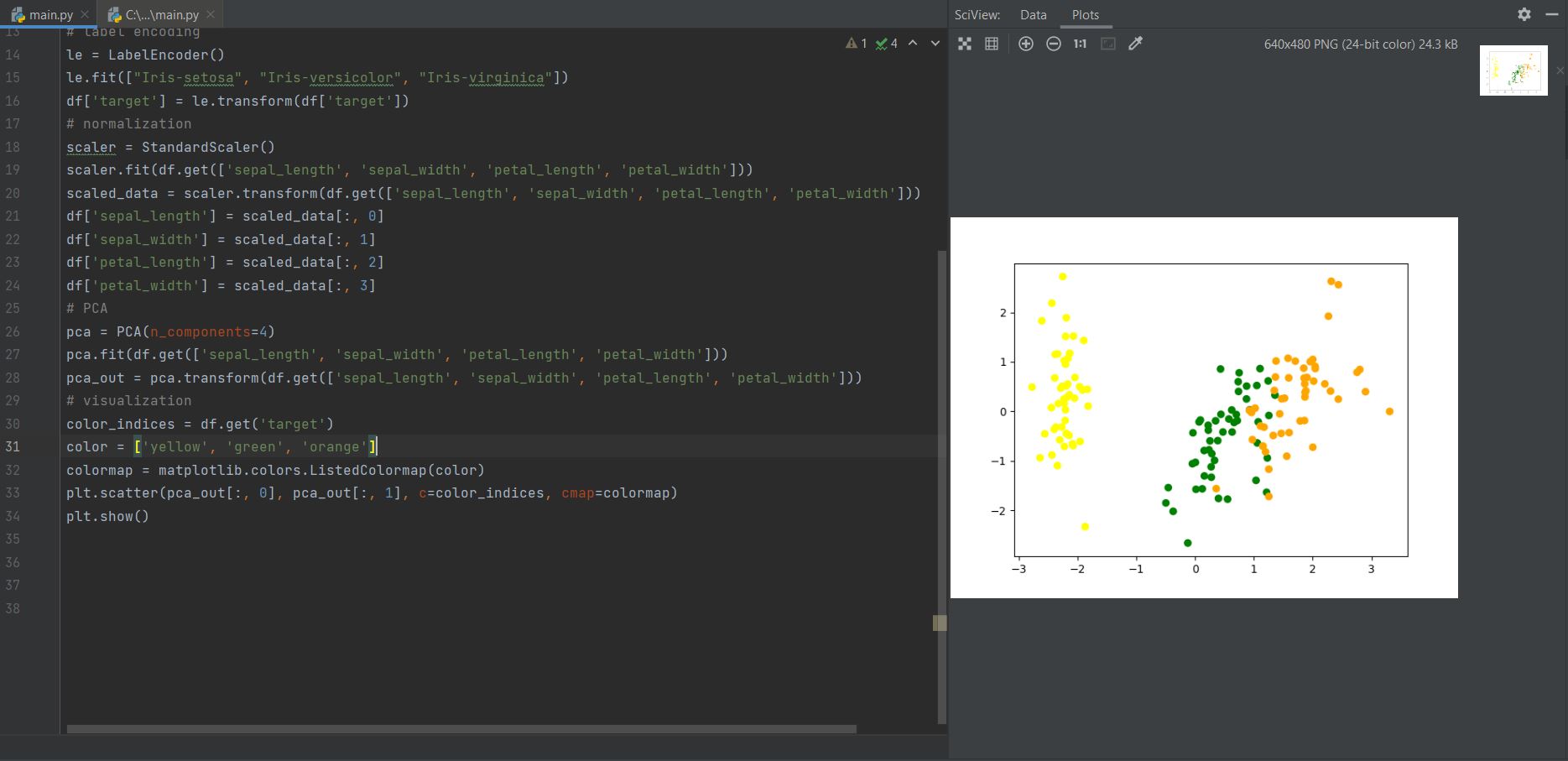
در ادامه داده­های غیرعددی را کد می­کنیم:



در بخش بعدی داده­ها را نرمال می­کنیم:



در انتها الگوریتم PCA را اجرا می­کنیم و داده را پلات می­کنیم:



نمودار­های جعبه­ای به ترتیب ویژگی­ها در پوشه Screenshots قرارداده شده­اند.