پردازش و بصری سازی پایگاه داده ی *Iris*

شارون سارونیان

پایگاه داده ی انتخابی از سایت ***UCI*** روی این پروژه پایگاه داده ی ***Iris*** می باشد. این پایگاه داده حاوی اطلاعاتی درباره ی ۳ نوع متفاوت گل های ***Iris*** می باشد.

1. *Iris Setosa*
2. *Iris Versicolor*
3. *Iris Virginica*

پایگاه داده شامل ۵ ویژگی ***(feature)*** می باشد:

1. *Sepal Length*
2. *Sepal Width*
3. *Petal Length*
4. *Petal Width*
5. *Species*

که ویژگی های ۱ تا ۴ طول و عرض گلبرگ و کاسبرگ ها به سانتی متر می باشد. ***range*** یا دامنه و واحد تمامی داده های ۱ تا ۴ یکسان بوده و بین ۰ تا ۸ سانتی متر می باشد. ویژگی پنجم نوع داده یا همان گل ها را مشخص می کند که در بالا ذکر شد و نوعی ***label*** برای داده ها می باشد.

ویژگی های ۱ تا ۴ از نوع عددی یا ***numeric*** هستند و ویژگی پنجم اسمی یا ***nominal*** است.

در این پایگاه داده ۱۵۰ نمونه (***sample***) یا سطر وجود دارد.

به نظر من این پایگاه داده می تواند به عنوان ورودی هم به الگوریتم های ***clustering*** و هم ***classification*** داده شود. یعنی ویژگی های ۱ تا ۴ به عنوان ورودی در نظر گرفته شوند. خروجی در الگوریتم های ***clustering*** می تواند مشخص کردن تعداد خوشه ها باشد. در الگوریتم های ***classification*** هم خروجی پیش بینی کردن کلاس یک داده ی جدید است.

کد ها به ضمیمه پیوست شده اما در ادامه کلیاتی درباره ی چگونگی کار و نتایج مطرح می کنیم.

ابتدا از پکیج ***pandas*** برای خواندن پایگاه داده استفاده می کنیم و پایگاه داده را چاپ می کنیم:

	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa
5	5.4	3.9	1.7	0.4	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa
8	4.4	2.9	1.4	0.2	Setosa
9	4.9	3.1	1.5	0.1	Setosa
10	5.4	3.7	1.5	0.2	Setosa
11	4.8	3.4	1.6	0.2	Setosa
12	4.8	3.0	1.4	0.1	Setosa
13	4.3	3.0	1.1	0.1	Setosa
14	5.8	4.0	1.2	0.2	Setosa
15	5.7	4.4	1.5	0.4	Setosa
16	5.4	3.9	1.3	0.4	Setosa
17	5.1	3.5	1.4	0.3	Setosa
18	5.7	3.8	1.7	0.3	Setosa
19	5.1	3.8	1.5	0.3	Setosa
20	5.4	3.4	1.7	0.2	Setosa
21	5.1	3.7	1.5	0.4	Setosa
22	4.6	3.6	1.0	0.2	Setosa
23	5.1	3.3	1.7	0.5	Setosa
24	4.8	3.4	1.9	0.2	Setosa
25	5.0	3.0	1.6	0.2	Setosa
26	5.0	3.4	1.6	0.4	Setosa
27	5.2	3.5	1.5	0.2	Setosa
28	5.2	3.4	1.4	0.2	Setosa
29	4.7	3.2	1.6	0.2	Setosa
..	...	...	...	...	...

[150 rows x 5 columns]

همان طور که می بینیم، پایگاه داده دارای ۱۵۰ سطر و ۵ ستون می باشد. این مطلب را با دستور *shape* در *pandas* نیز می توان چک کرد:

```
(150, 5)
```

با دستور *head()* می توان ۵ داده ی اول پایگاه داده را دید:

	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

قبل از این که برخی مقادیر آماری را برای پایگاه داده محاسبه کینم، بهتر است آن را از دیدگاه ***missing value*** بررسی کنیم:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
Sepal Length    150 non-null float64
Sepal Width     150 non-null float64
Petal Length    150 non-null float64
Petal Width     150 non-null float64
Species         150 non-null object
dtypes: float64(4), object(1)
memory usage: 5.9+ KB
None
```

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	50	50	50	50
Versicolor	50	50	50	50
Virginica	50	50	50	50

با دستور ***info()*** و ***data.groupby('Species').count()***/طلاعات شکل های بالا به دست می آید.

طبق این نمودارها، ***missing value*** نداریم و همه ی داده ها از نوع ***float*** هستند. پس مطمئنیم همه خانه های پایگاه داده پر هستند و خانه ی خالی نداریم. اما ممکن است بعضی خانه ها به صورت ***NA*** پر شده باشند که برای ما ***noise*** ایجاد می کنند. پس تعداد خانه های پر شده با ***NA*** در هر ستون را با دستور ***data.isnull().sum()*** بررسی می کنیم:

```
Sepal Length    0
Sepal Width     0
Petal Length    0
Petal Width     0
Species         0
dtype: int64
```

تعداد داده های *unique* برای ستون *Species* را با دستورات *data["Species"].unique()* و *data["Species"].value_counts()* می بینیم:

```
['Setosa' 'Versicolor' 'Virginica']
Virginica     50
Versicolor    50
Setosa        50
Name: Species, dtype: int64
```

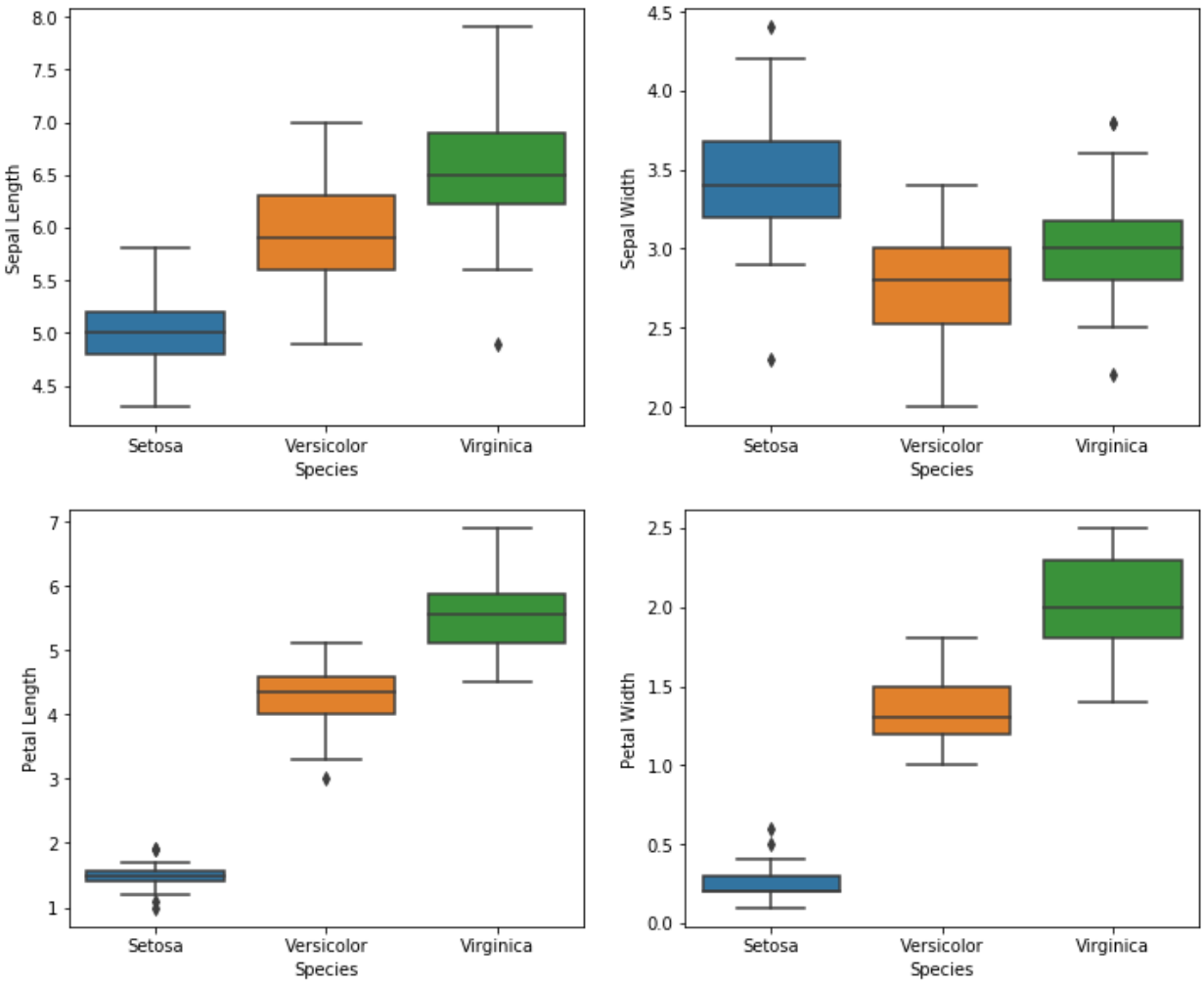
پس مطمئنیم که داده های *missing value* و *noise* نداریم.

حال به محاسبه ی تعداد داده ها ، میانگین، انحراف معیار، مینیمم، ماکزیمم و چارک های اول (۲۵٪)، دوم (۵۰٪)و سوم (۷۵٪) با دستور *describe()* می پردازیم تا پراکندگی داده ها را ببینیم:

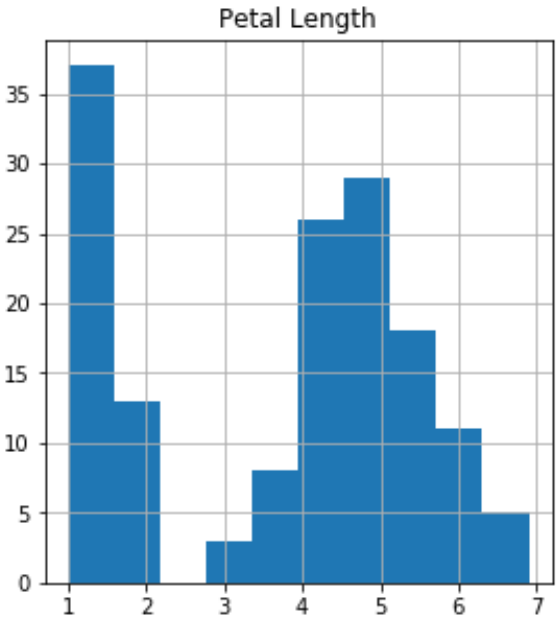
	Sepal Length	Sepal Width	Petal Length	Petal Width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

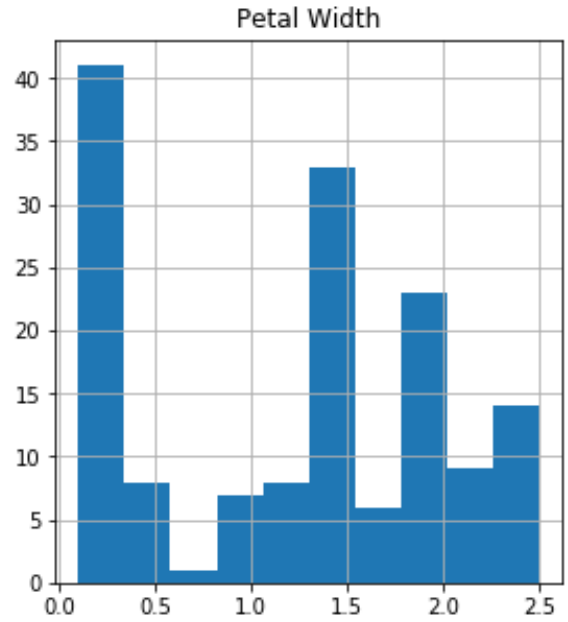
حال زمان آن رسیده تا به بصری سازی داده بپردازیم.

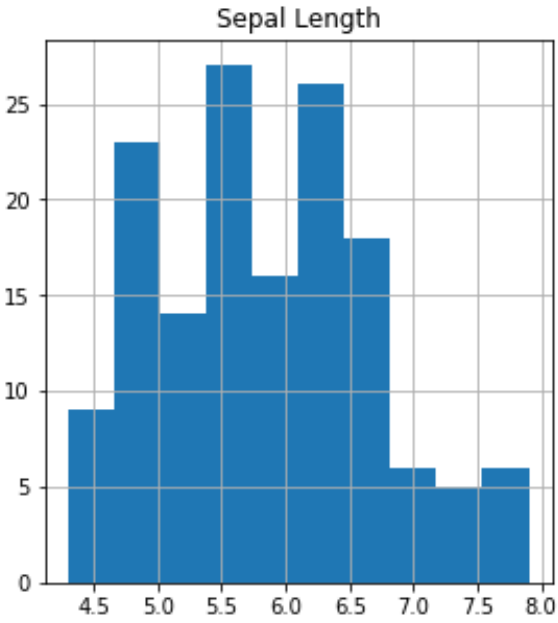
از آن جا که ۵ معیار آماری بالا را محاسبه کردیم، بهتر است ابتدا طبق آن نمودار *Boxplot* هر ویژگی را رسم کنیم تا درباره ی پراکندگی داده ها و وجود *outlier* های احتمالی دید داشته باشیم:

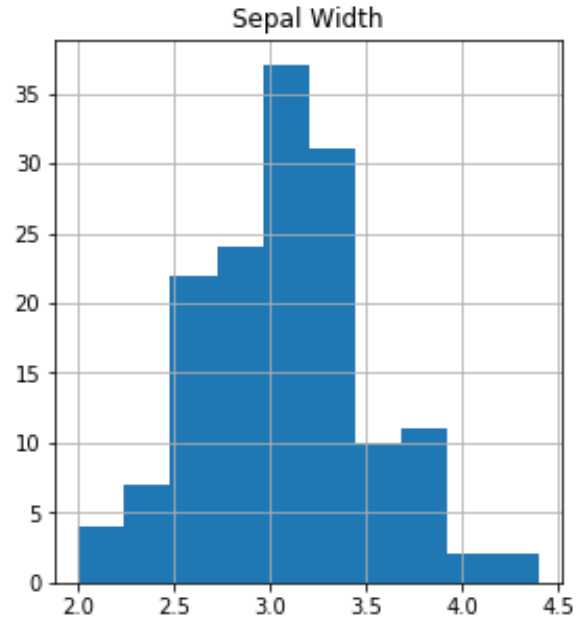


با توجه به شکل، به نظر می رسد که دارای *outlier* هایی در ویژگی ها هستیم. مطمئنیم این داده ها *noise* نیستند چون قبلاً چک کردیم که داده ی *NA* با *missing value* نداریم و در *metadata* هم ذکر شده که داده ها همگی به درستی جمع آوری و وارد پایگاه داده شده اند. حال برای بررسی بهتر پراکندگی داده ها *histogram* آن ها را رسم می کنیم:





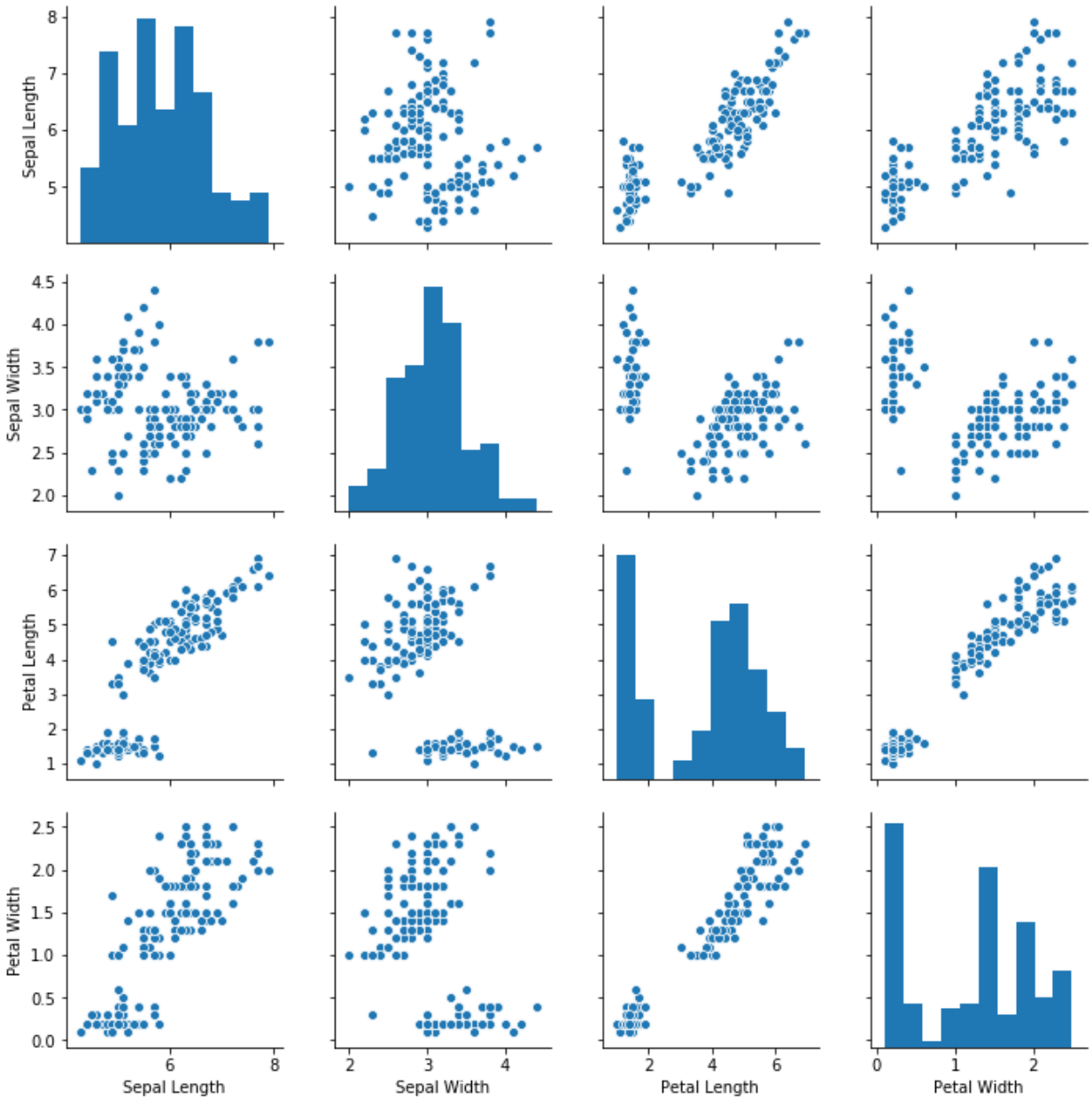




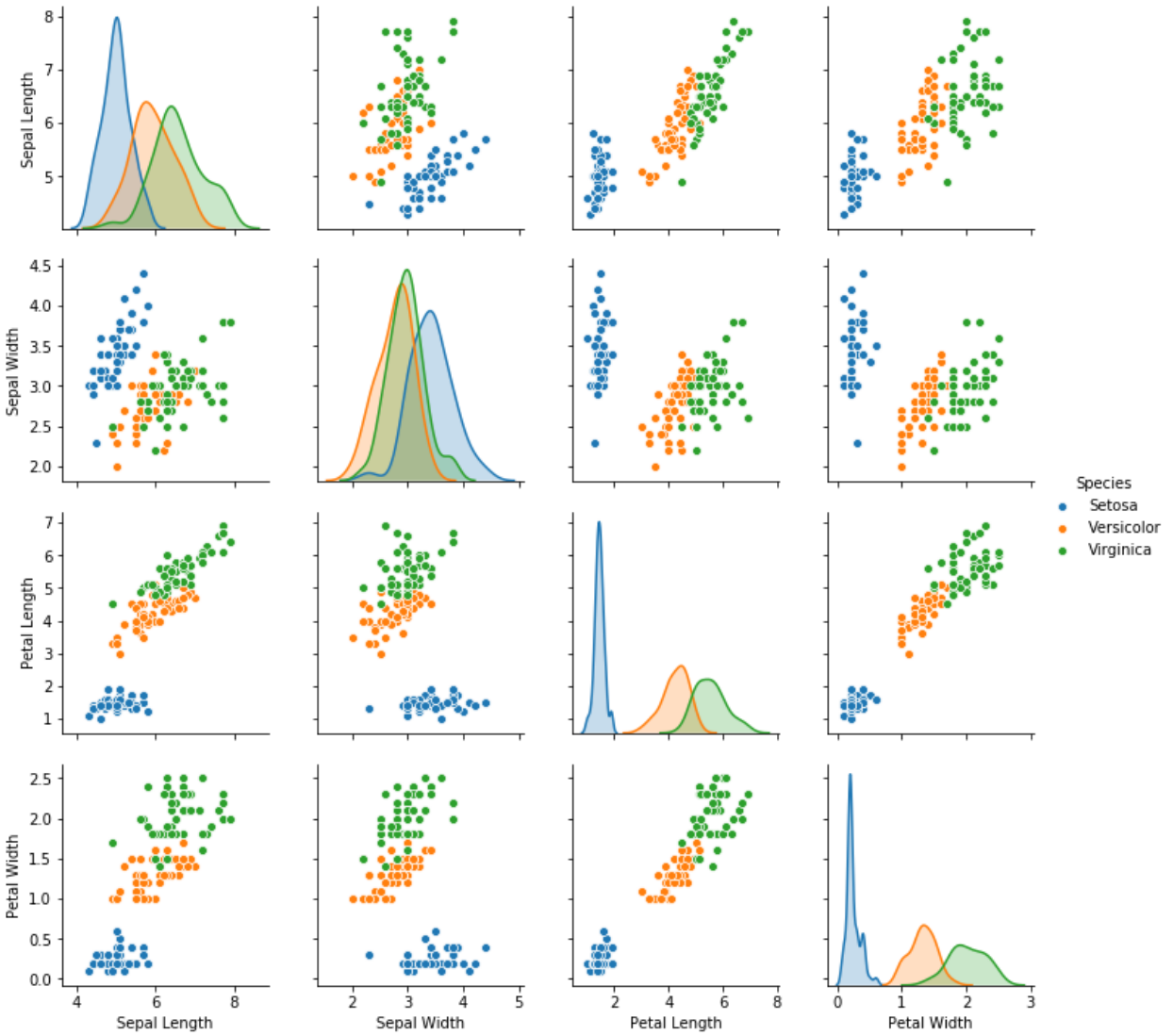
به نظر می رسد ویژگی های ***Sepal Length*** و ***Sepal Width*** توزیع گوسی دارند. این ویژگی می تواند برای الگوریتم هایی که از این فرض استفاده می کنند مفید باشد. با این ویژگی محاسبه ی میانگین و انحراف معیار هم منطقی بوده است.

پس این ویژگی ها نشان می دهند که داده های ما در این ۲ *attribute* از نوع *symmetric* هستند و هم چنین می فهمیم که *Petal Width* و *Petal Length* نه *positively skewed* هستند و نه *negatively skewed*. شاید نمودارهای بعدی دید بهتری از آن ها به ما بدهند.

در این مرحله *scatter plot* را برای هر دو جفت ویژگی برای داده ها رسم می کنیم تا ببینیم آیا روابطی بین متغیرها وجود دارد یا خیر:



بـــــــــه نظر می رسد برخی از جفت ویژگی ها دارای ویژگی ***diagonal grouping*** هستند. این می تواند نشان دهنده ی همبستگی بالا بین برخی از آن ها باشد. برای دیـــــد بهتر از ***seaborn pairplot*** استفاده می کنیم:



از این نمودار اطلاعات ارزشمندی به دست می آید. بـــــــه نظر می رسد بین دو ویژگی ***Petal Length*** و ***Petal Width*** همبستگی وجود داشته باشد. چون نقاط رسم شده از چپ به راست و از پایین به بالا به صورت قطری در آمده اند و این نشانه ی همبستگی است. (این حدس را بیشتر بررسی می کنیم.)

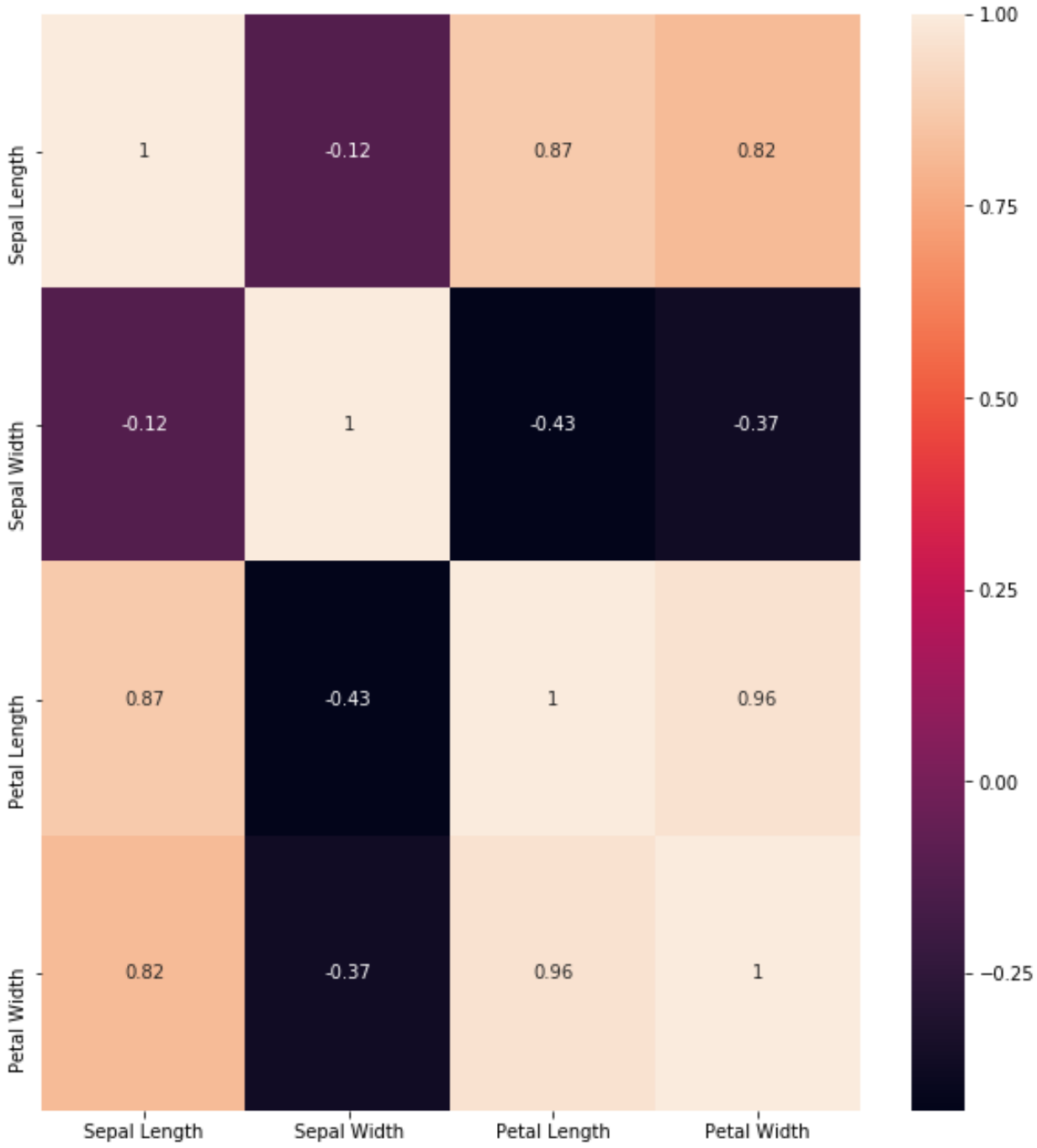
هم چنین به نظر می رسد داده های کلاس ***setosa*** در همه ی نمودارها از بقیه کلاس ها جدا رسم شده اند. این می تواند نشان دهنده ی این نکته باشد کـــه کلاس ***setosa*** جداپذیر خطی از سایر کلاس هاست ولی بقیه کلاس ها این ویژگی را ندارند چون به طور شهودی هم دیده می شود که بسیار در هم ریخته هستند و قابل جدا شدن با یک خط نیستند. (درستی این حدس در ***metadata*** هم دیده می شود و این نکته عملاً ذکر شده ! ولی با یک ***plot*** ساده به زیبایی این را دیدیم.)

به نظر می رسد مقداری ***overlap*** بین کلاس های ***virginica*** و ***versicolor*** وجود داشته باشد پس احتمالاً نمی شود به یک ***classification rate*** بی نقص رسید.

حال حدس همبستگی را دقیق تر و با محاسبه ی ماتریس ***correlation*** با متد ***Pearson*** و هم چنین محاسبه ی ***covariance*** می کنیم (این محاسبات هم چنین برای ***data integration*** در پیش پردازش هم استفـــاده می شوند تا ***redundancy*** ها را کشف کنیم و در صورت امکان در ***feature reduction*** آن ها را حذف کنیم:

	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	1.000000	-0.117570	0.871754	0.817941
Sepal Width	-0.117570	1.000000	-0.428440	-0.366126
Petal Length	0.871754	-0.428440	1.000000	0.962865
Petal Width	0.817941	-0.366126	0.962865	1.000000
	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	0.685694	-0.042434	1.274315	0.516271
Sepal Width	-0.042434	0.189979	-0.329656	-0.121639
Petal Length	1.274315	-0.329656	3.116278	1.295609
Petal Width	0.516271	-0.121639	1.295609	0.581006

هر دو ماتریس حدس ما برای همبستگی بین ***Petal Length*** و ***Petal Width*** را تائید می کنند. با یک ***heatmap*** (که ورودی آن ماتریس ***correlation*** است) این موضوع بهتر بصری سازی می شود:

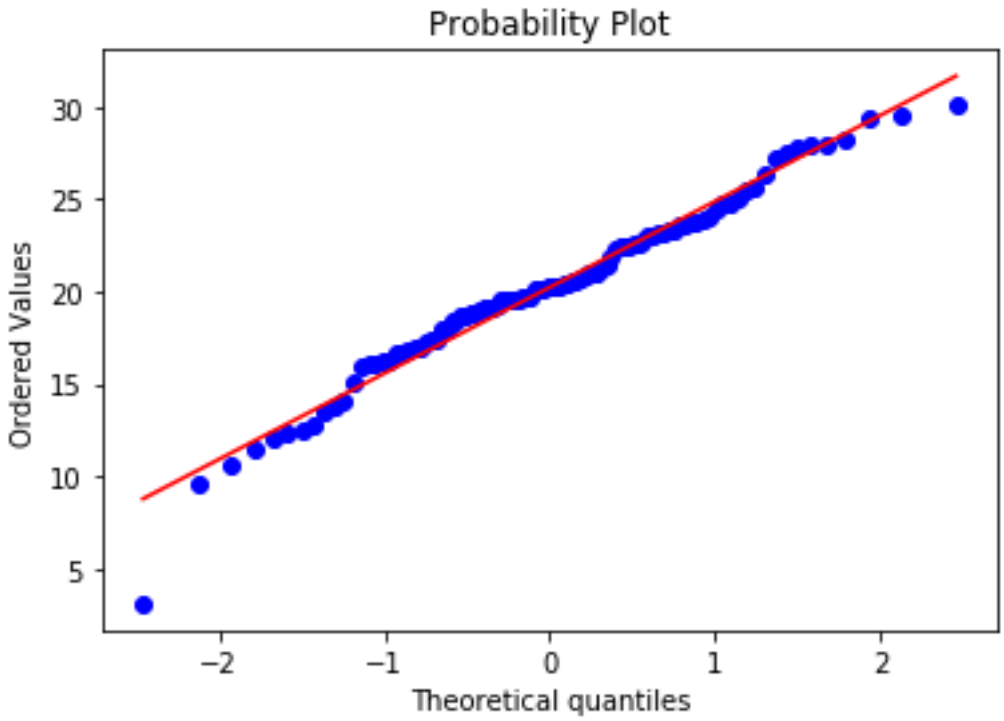


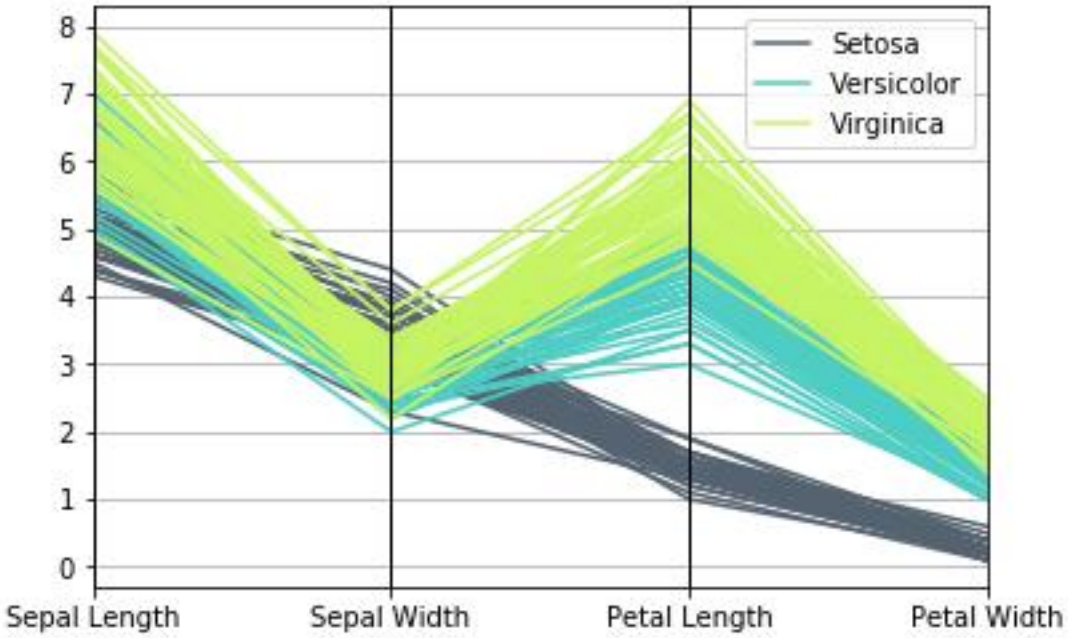
همان طور که می بینیم، خانه ی مربوط به ***Petal Length*** و ***Petal Width*** با مقدار 0.96 نزدیک ترین مقدار به 1 است که همبستگی آن ها را تائید می کند.

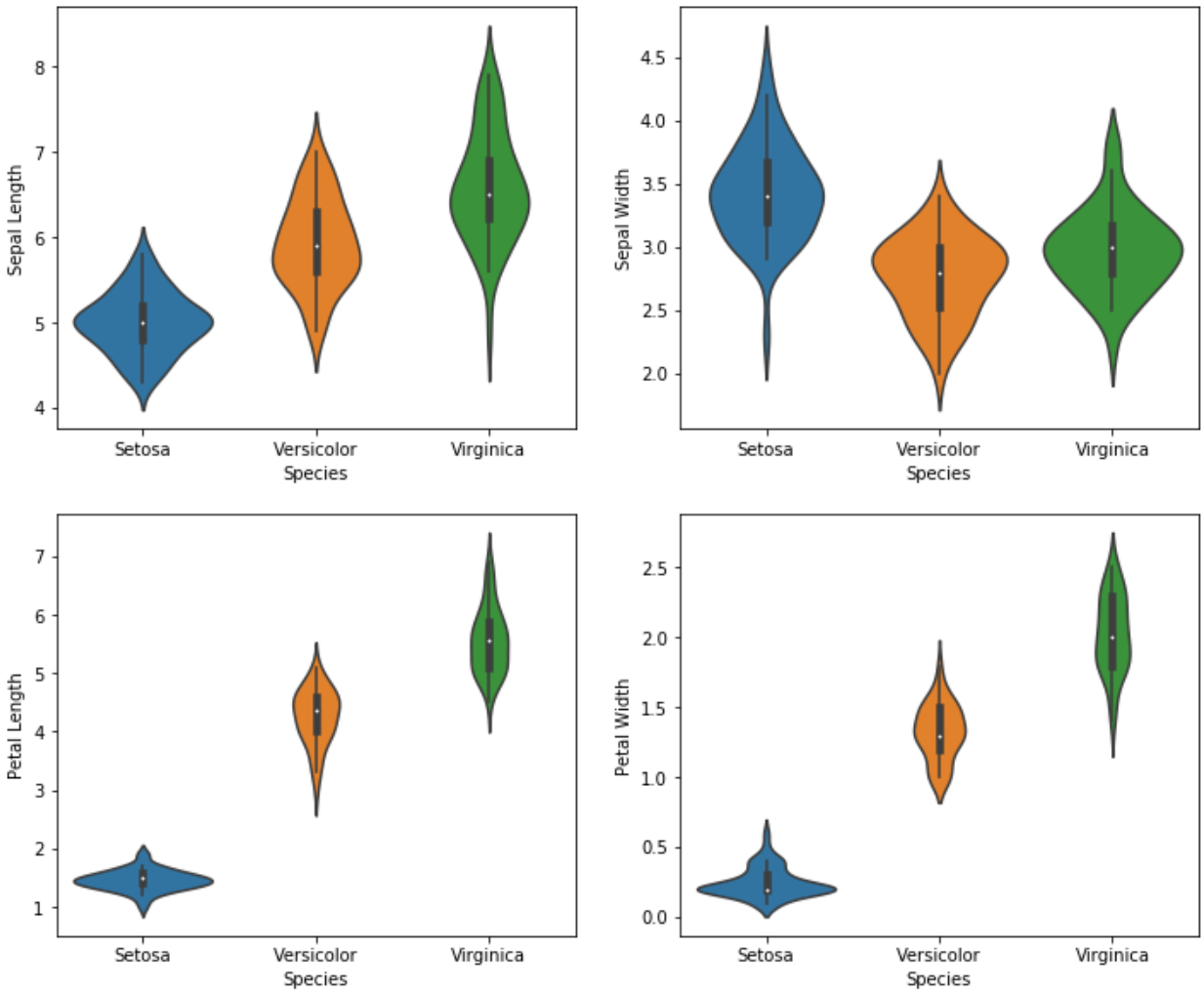
توجه به مقدار *Sepal Length* و *Petal Length* که 0.87 است و هم چنین نمودار *scatter* آن ها که تقریباً حالت قطری دارد، نشان می دهد که بین این دو ویژگی هم احتمالاً همبستگی وجود دارد.

از همه این بحث ها حدس می زنیم که شاید بشود ۲ ویژگی از این ویژگی ها را *reduce* کرد کـه این حدس ایده ی استفاده از روش *pca* در *data reduction* را به ما می دهد.

اما قبل از رفتن به سراغ این روش، خالی از لطف نیست کـــــــــه نمودارهای *qq-plot*، *violin-plot* و *parallel coordinates* را هم در بحث بصری سازی داشته باشیم:




8
7
6
5
4
3
2
1
0
Setosa
Versicolor
Virginica
Sepal Length
Sepal Width
Petal Length
Petal Width



حال به بحث پیش پردازش می رسیم. از بین مراحل پیش پردازش، ***handle*** کردن ***noise*** و ***missing value*** را قبلاً بررسی کرده ایم. در بحث ***integration*** نیز بحث های مفصلی درباره ی همبستگی و ***redundancy*** داشتیم. حال به دو بحث ***data reduction*** و ***data transformation*** می پردازیم.

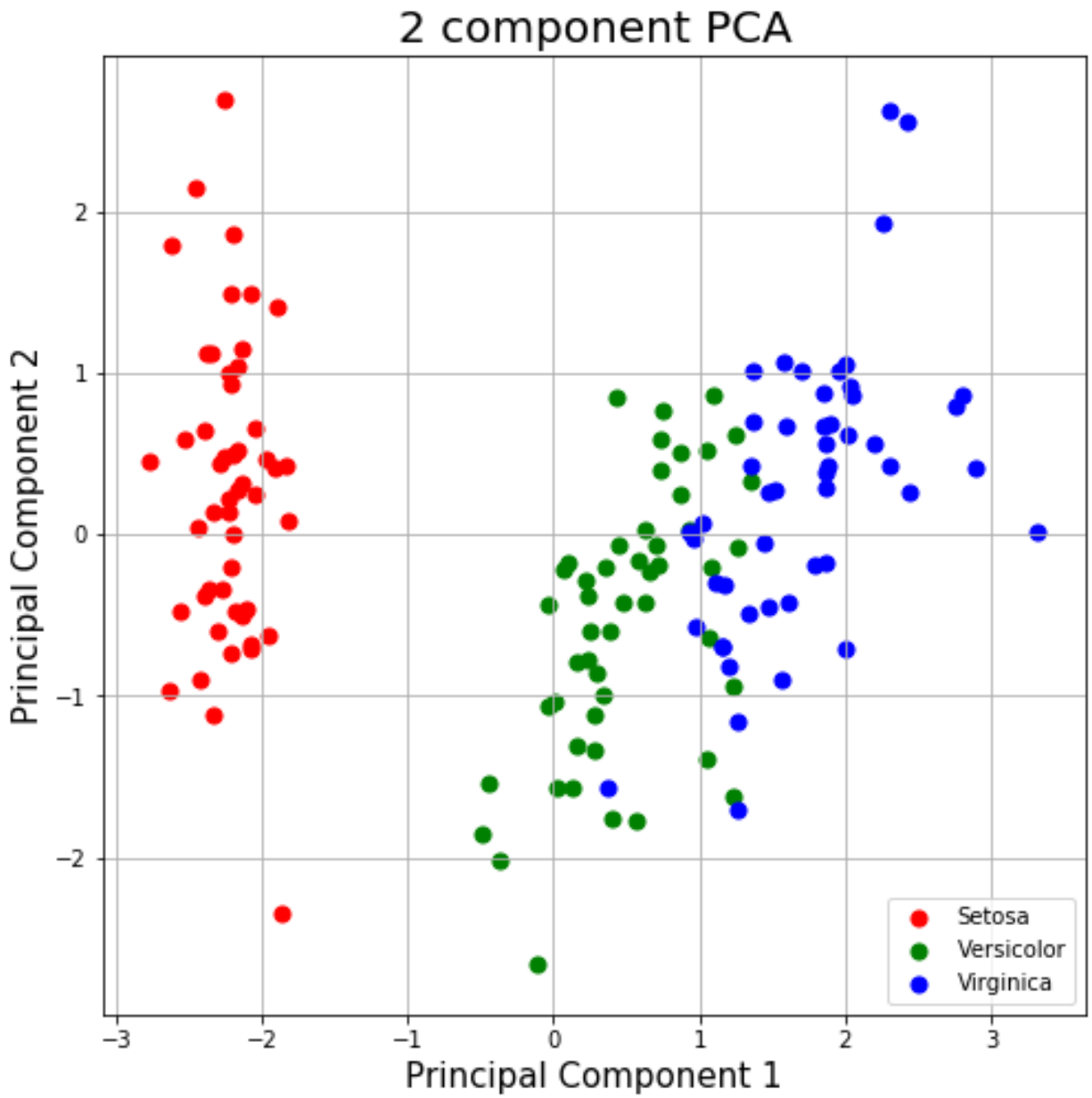
در ابتدای الگوریتم ***pca*** می دانیم کـــــه داده ها را به نحوی باید ***scale*** کنیم. این کار یعنی ***scale*** کردن و نرمال سازی را برای بحث ***data transformation*** هم نیاز داریم. در این جا چون قبلاً هم دیدیم که ۲ تا از ویژگی ها توزیع گوسی دارند، ترجیح می دهیم از بین انـــــــواع روش های استاندارد سازی مثل ***min-max*** و ***z-score*** (استاندارد) و ...، از ***standard scaler*** استفاده کنیـــم چون این ***scaler*** فرض می کند که توزیع داده های ما گوسی است و ما هم تقریباً این فرض را در ۲ ویژگی داریم. این ***scaler***، داده ها را طوری ***rescale*** می کند که میانگین ویژگی ها ۰ بوده و انحراف معیار ۱ داشته باشند:

```
[[-9.00681170e-01  1.01900435e+00 -1.34022653e+00 -1.31544430e+00]
 [-1.14301691e+00 -1.31979479e-01 -1.34022653e+00 -1.31544430e+00]
 [-1.38535265e+00  3.28414053e-01 -1.39706395e+00 -1.31544430e+00]
 [-1.50652052e+00  9.82172869e-02 -1.28338910e+00 -1.31544430e+00]
 [-1.02184904e+00  1.24920112e+00 -1.34022653e+00 -1.31544430e+00]
 [-5.37177559e-01  1.93979142e+00 -1.16971425e+00 -1.05217993e+00]
 [-1.50652052e+00  7.88807586e-01 -1.34022653e+00 -1.18381211e+00]
 [-1.02184904e+00  7.88807586e-01 -1.28338910e+00 -1.31544430e+00]
 [-1.74885626e+00 -3.62176246e-01 -1.34022653e+00 -1.31544430e+00]
 [-1.14301691e+00  9.82172869e-02 -1.28338910e+00 -1.44707648e+00]
 [-5.37177559e-01  1.47939788e+00 -1.28338910e+00 -1.31544430e+00]
 [-1.26418478e+00  7.88807586e-01 -1.22655167e+00 -1.31544430e+00]
 [-1.26418478e+00 -1.31979479e-01 -1.34022653e+00 -1.44707648e+00]
 [-1.87002413e+00 -1.31979479e-01 -1.51073881e+00 -1.44707648e+00]
 [-5.25060772e-02  2.16998818e+00 -1.45390138e+00 -1.31544430e+00]
 [-1.73673948e-01  3.09077525e+00 -1.28338910e+00 -1.05217993e+00]
 [-5.37177559e-01  1.93979142e+00 -1.39706395e+00 -1.05217993e+00]
 [-9.00681170e-01  1.01900435e+00 -1.34022653e+00 -1.18381211e+00]
 [-1.73673948e-01  1.70959465e+00 -1.16971425e+00 -1.18381211e+00]
 [-9.00681170e-01  1.70959465e+00 -1.28338910e+00 -1.18381211e+00]
 [-5.37177559e-01  7.88807586e-01 -1.16971425e+00 -1.31544430e+00]
 [-9.00681170e-01  1.47939788e+00 -1.28338910e+00 -1.05217993e+00]
 [-1.50652052e+00  1.24920112e+00 -1.56757623e+00 -1.31544430e+00]
 [-9.00681170e-01  5.58610819e-01 -1.16971425e+00 -9.20547742e-01]
```

بعد از استاندارد کردن داده ها روش ***pca*** را روی آن ها ***fit*** می کنیم:

	principal component 1	principal component 2
0	-2.264703	0.480027
1	-2.080961	-0.674134
2	-2.364229	-0.341908
3	-2.299384	-0.597395
4	-2.389842	0.646835
5	-2.075631	1.489178
6	-2.444029	0.047644
7	-2.232847	0.223148
8	-2.334640	-1.115328
9	-2.184328	-0.469014
10	-2.166310	1.043691
11	-2.326131	0.133078
12	-2.218451	-0.728676
13	-2.633101	-0.961507
14	-2.198741	1.860057
15	-2.262215	2.686284
16	-2.207588	1.483609
17	-2.190350	0.488838
18	-1.898572	1.405019
19	-2.343369	1.127849
20	-1.914323	0.408856
21	-2.207013	0.924121
22	-2.774345	0.458344
23	-1.818670	0.085559
24	-2.227163	0.137254
25	-1.951846	-0.625619
26	-2.051151	0.242164
27	-2.168577	0.527150
28	-2.139563	0.313218
29	-2.265261	-0.337732
..	...	...

حاصل به دست آمدن ۲ ویژگی از روی ۴ ویژگی قبلی است پس عمل ***reduction*** را با موفقیت انجام دادیم. حال به رسم ***plot*** پایگاه داده می پردازیم:



برای این که مطمئن شویم ***pca*** به درستی کار کرده از درصد ***explained variance*** استفاده می کنیم. این معیار به ما می گوید که چقدر اطلاعات (واریانس) بـــه هر ویژگی استخراج شده نسبت داده می شود. چون ما می دانیم که در کاهش بعد از ۴ به ۲ به هر حال مقداری از اطلاعات را از دست می دهیم. با استفاده از درصد ذکر شده در بالا می بینیم که ویژگی اول ۷۲٪ واریانس و ویژگی دوم ۲۳٪ واریانس را شامل شده. با هم، این ۲ ویژگی ۹۵٪ اطلاعات را حفظ کرده اند:

```
0.9581320720000164
```

پس با این کار خیالمان راحت می شود که هم ***data reduction*** انجام داده ایم و هم نسبت خوبی از اطلاعات را حفظ کردیم.

یکی از روش های دیگر کاهش بعد، ***attribute subset selection*** است کـــــــه یکی از روش های آن ***feature selection*** با استفاده از ***decision tree classifier*** است.

با این که این روش برای ***classify*** کردن استفاده می شود اما می توان از آن برای ***data reduction*** هم بهره برد. کد مربوط به ***decision tree*** پیوست شده است.

```
(150, 4)
[0.09868611 0.04949553 0.48412402 0.36769434]
(150, 2)
```

می بینیم قبل از اعمال این روش بعد (۴ ,۱۵۰) بود و بعد از انجام این روش به (۲, ۱۵۰) رسیدیم یعنی این روش ویژگی ها را بر اساس اهمیت کلاسه بندی کرده و ۲ ویژگی را که در درخت ظاهر نشده حذف کرده. پس با این روش هم می توان کاهش بعد انجام داد.

از بررسی های بالا ابتدا به این ایده رسیدیم که روش ***pca*** را برای داده ها به کار گیریم و در صورت امکان ۲ ویژگی را حذف کنیـــــــــــم. دلیل این که ***pca*** را بـــــه دیگـــــــــر روش های ***reduction*** مانند ***attribute subset reduction*** ترجیح می دهیم این است که ***pca*** ۲ ویژگی را مستقیماً حذف نمی کند بلکه با ترکیب آن ها داده ها را به فضایی کوچک تر نگاشت می کند.

با این حال ممکن است هنگام مدل سازی به این نتیجه برسیم که ***decision tree*** و یا یکی دیگر از روش های ***reduction*** مفیدتر است و بهتر کار می کند. پس در این مرحله از این فراتر نمی توان رفت.

در بحث ***data transformation***، یکی از موارد استاندارد کردن داده ها بود کـــــــــه این کار را انجام دادیم. هم چنین، آنالیز داده به روش رسم ***histogram*** در هر دو مبحث ***reduction*** و ***transformation*** ذکر شده که قبلاً این کار را انجام دادیم. پس با تمامی توضیحات داده می توان استناد کرد که بصری سازی و پیش پردازش کاملی روی پایگاه داده ی ***iris*** انجام دادیم و داده های آماده ی تقسیم شدن به مجموعه های ***test*** و ***train*** و مدل سازی شدن هستند.