TI-6P4 Praktikum Probabilitas dan Statistika

P01 - Visualisasi Data

Ventje Jeremias Lewi Engel, M.T.

Seringkali data yang Anda miliki tidak mudah dimengerti sampai Anda melihat bentuk visualnya, seperti plot dan bagan.

Memvisualisasikan data dengan cepat adalah ***skill yang sangat penting***, baik untuk statistik terapan maupun nanti ketika belajar machine learning.

Pada praktikum kali ini, Anda akan belajar lima tipe plot yang perlu diketahui ketika memvisualisasi data dan cara membuatnya menggunakan Python.

Mari kita mulai!

Library Python untuk Visualisasi Data

Python memiliki beberapa library untuk membuat visualisasi data dengan berbagai fitur yang lengkap. Anda bisa membuat plot yang custom, menarik, dan bahkan interaktif menggunakan Python

Sebagai gambaran, berikut ini adalah beberapa library visualisasi data yang populer yang ada di Python:

- **Matplotlib:** dasar, plotting dengan cepat dan bisa kustomisasi
- **Pandas Visualization:** fiturnya mudah digunakan, dibangun di atas Matplotlib
- **Seaborn:** cukup high-level dengan desain default yang menarik
- **ggplot:** berdasarkan ggplot yang ada di platform R, menggunakan Grammar of Graphics
- **Plotly:** bisa membuat visualisasi dan dashboard yang interaktif

Kali ini Anda akan mencoba membuat visualisasi data dengan Matplotlib dan Seaborn. Anda akan fokus kepada cara membuat visualisasi datanya. Tugas di rumah adalah menjelaskan hasil visualisasi datanya.

Import Dataset

Di praktikum ini, Anda akan menggunakan dua dataset yang bisa dipakai dengan bebas, yaitu dataset Iris dan Wine Review, yang akan dimuat dengan fungsi *pandas read_csv*.

```
import pandas as pd
iris = pd.read_csv('iris.csv', names=['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class'])
print(iris.head())
```

:

	sepal_length	sepal_width	petal_length	petal_width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

```
wine_reviews = pd.read_csv('winemag-data-130k-v2.csv', index_col=0)
wine_reviews.head()
```

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian
4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks

Matplotlib

Matplotlib adalah library plotting Python yang paling terkenal. Biasa disebut *low-level* dengan kustomisasi yang sangat lengkap, artinya Anda bisa mengedit plot atau bagan yang dibuat sesuai keinginan. Asal Anda mau membaca dokumentasinya.

Anda bisa menginstall Matplotlib dengan pip atau conda.

```
pip install matplotlib
or
conda install matplotlib
```

Jika Anda menggunakan Anaconda, maka Matplotlib sudah terinstall secara *default.*

Matplotlib sering digunakan untuk membuat line chart, bar chart atau grafik batang , dan histogram. Import library ini dengan cara:

```
import matplotlib.pyplot as plt
```

Scatter Plot

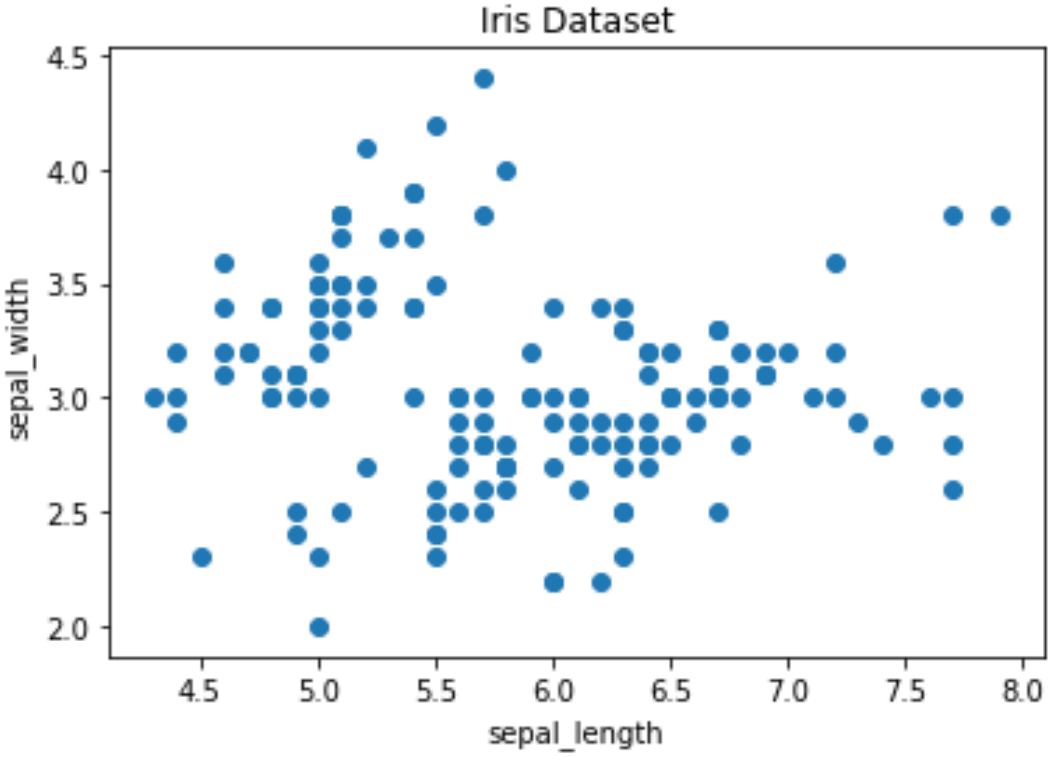
Scatter plot atau diagram pencar atau juga disebut diagram sebar adalah gambaran yang menunjukkan **kemungkinan hubungan** (korelasi) antara dua variabel dan menunjukkan keeratan hubungan antara dua variabel tersebut.

Untuk membuat scatter plot dengan Matplotlib kita bisa menggunakan fungsi scatter. Kita akan menggunakan plt.subplots untuk memberi judul dan label pada sumbu.

```
fig, ax = plt.subplots()


ax.scatter(iris['sepal_length'], iris['sepal_width'])

ax.set_title('Iris Dataset')
ax.set_xlabel('sepal_length')
ax.set_ylabel('sepal_width')
```



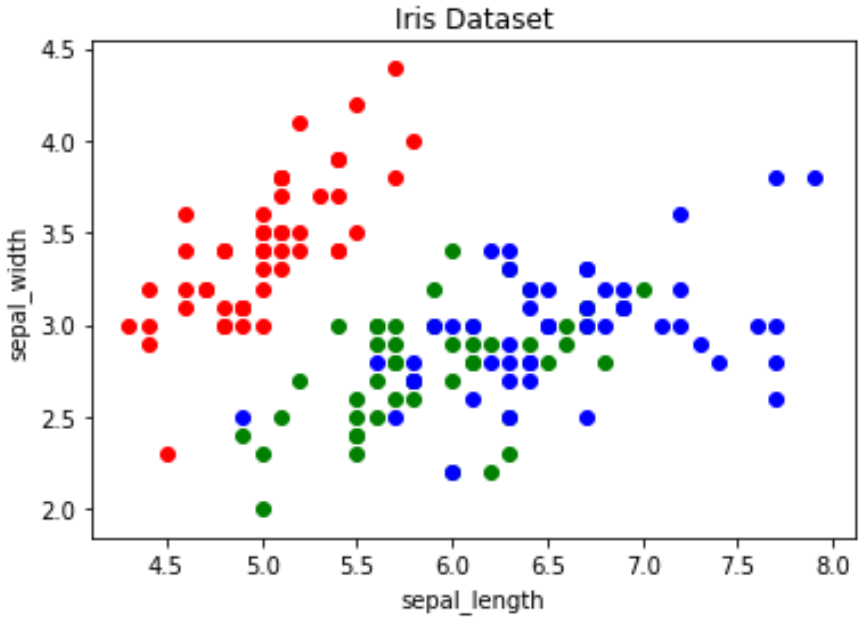
Kita bisa memberi warna untuk setiap datapoin yang ada berdasarkan sebuah kriteria, misalnya class. Pertama, kita harus buat kamus untuk memetakan class dengan warna. Kedua, kita memberi warna untuk setiap datapoin di dalam plot menggunakan for-loop.

```
colors = {'Iris-setosa':'r', 'Iris-versicolor':'g', 'Iris-virginica':'b'}

fig, ax = plt.subplots()

for i in range(len(iris['sepal_length'])):
   ax.scatter(iris['sepal_length'][i], iris['sepal_width'][i],color=colors[iris['class'][i]])

ax.set_title('Iris Dataset')
ax.set_xlabel('sepal_length')
ax.set_ylabel('sepal_width')
```



Line Chart

Grafik garis atau line chart adalah grafik yang menggunakan titik sebagai poin data dan kemudian titik tersebut disambung dengan titik berikutnya. Grafik ini merupakan salah satu grafik yang paling sering digunakan setelah grafik batang.

Secara umum grafik line digunakan untuk membandingkan suatu data berdasarkan deret waktu tertentu (time series).

Di Matplotlib, kita bisa membuat garik garis dengan fungsi plot. Kita juga bisa memplotkan lebih dari satu variabel dalam satu bidang.

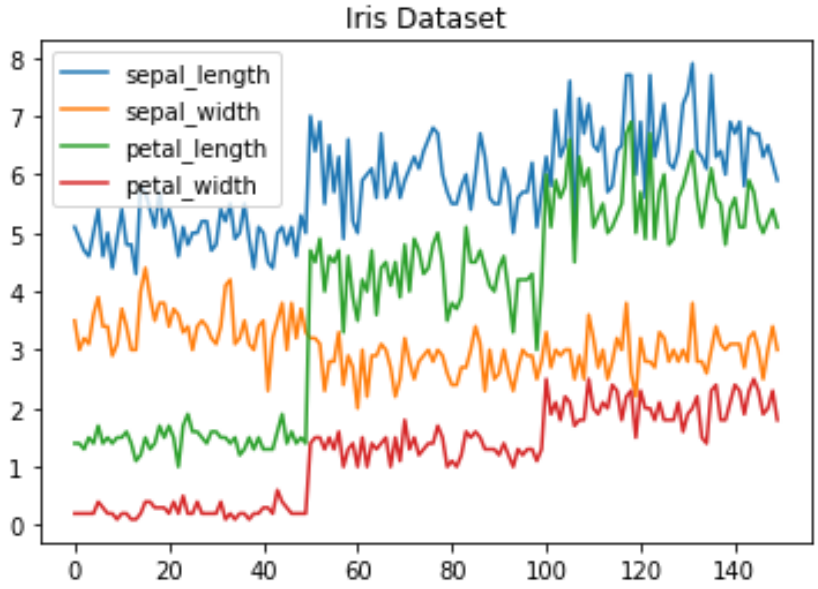
```
columns = iris.columns.drop(['class'])

x_data = range(0, iris.shape[0])

fig, ax = plt.subplots()

for column in columns:
   ax.plot(x_data, iris[column])

ax.set_title('Iris Dataset')
ax.legend()
```



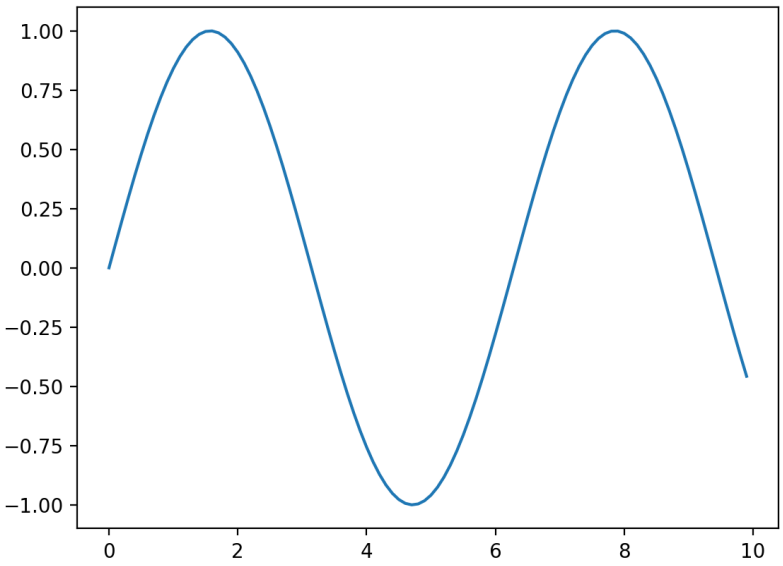
Contoh berikut ini Anda akan membuat 100 nilai bertipe float sebagai sumbu x dan fungsi sinus(x) sebagai hasil sumbu y.

```
from numpy import sin

# consistent interval for x-axis
x = [x*0.1 for x in range(100)]

# function of x for y-axis
y = sin(x)

# create line plot
pyplot.plot(x, y)
pyplot.show()
```



Histogram

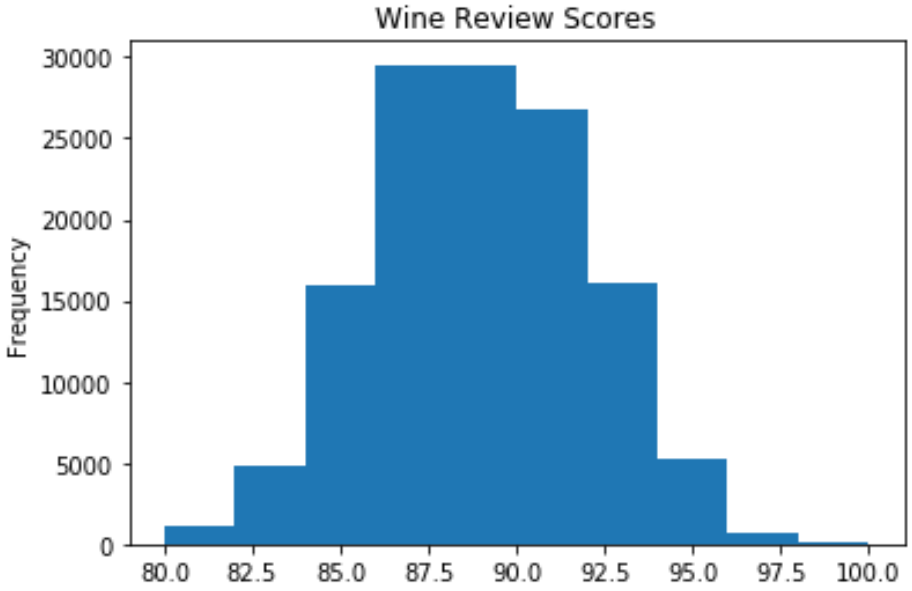
Histogram merupakan tampilan bentuk grafis untuk menunjukkan distribusi data secara visual atau seberapa sering suatu nilai yang berbeda itu terjadi dalam suatu kumpulan data.

Di Matplotlib, kita bisa membuat histogram dengan fungsi hist. Matplotlib we can create a Histogram using the hist method. Kalau kita menggunakan data kategorikal seperti variabel points di dataset wine-review maka python akan otomatis menghitung seberapa sering setiap kelas muncul dalam dataset.

```
fig, ax = plt.subplots()

ax.hist(wine_reviews['points'])

ax.set_title('Wine Review Scores')
ax.set_xlabel('Points')
ax.set_ylabel('Frequency')
```



Bar Chart

Grafik atau diagram batang (bar chart) berguna untuk menyajikan perbandingan data pada satu atau beberapa seri data.

Misalnya, angka penjualan untuk setiap bulan dalam satu tahun, atau penjualan tahun 1 dan tahun 2 di beberapa kota.

Data pada grafik batang disajikan dalam bentuk persegi panjang horizontal, yang panjangnya sesuai dengan nilai masing-masing. Dengan begitu kita bisa melihat dengan cepat dan mudah data mana yang memiliki kinerja atau nilai yang lebih tinggi dibandingkan dengan data yang lain.

Grafik batang di Matplotlib bisa dibuat dengan fungsi bar. Fungsi bar() tidak otomatis menghitung frekuensi, sehingga kita perlu menggunakan fungsi value_counts() dari pandas. Bar chart cocok untuk data dengan kategori kurang dari 20 karena bila lebih dari 20 maka tampilannya akan kacau.

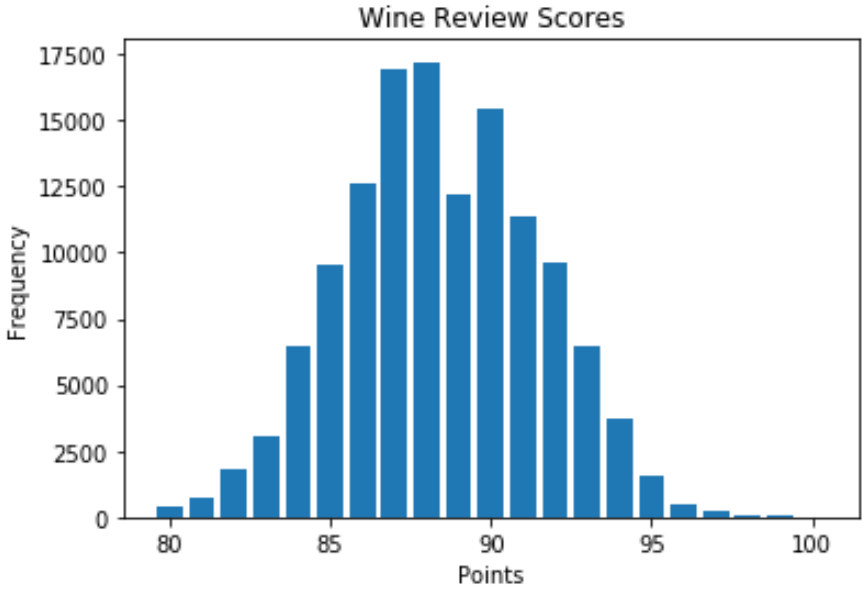
```
fig, ax = plt.subplots()

data = wine_reviews['points'].value_counts()

points = data.index
frequency = data.values

ax.bar(points, frequency)

ax.set_title('Wine Review Scores')
ax.set_xlabel('Points')
ax.set_ylabel('Frequency')
```



Seaborn

Seaborn adalah library visualisasi data yang dibangun di atas Matplotlib. Seaborn memberikan interface fungsi yang lebih sederhana dengan tampilan grafik/plot yang menarik.

Saya lebih suka menggunakan seaborn karena kemudahan integrasinya dengan pandas.

Cara import seaborn:

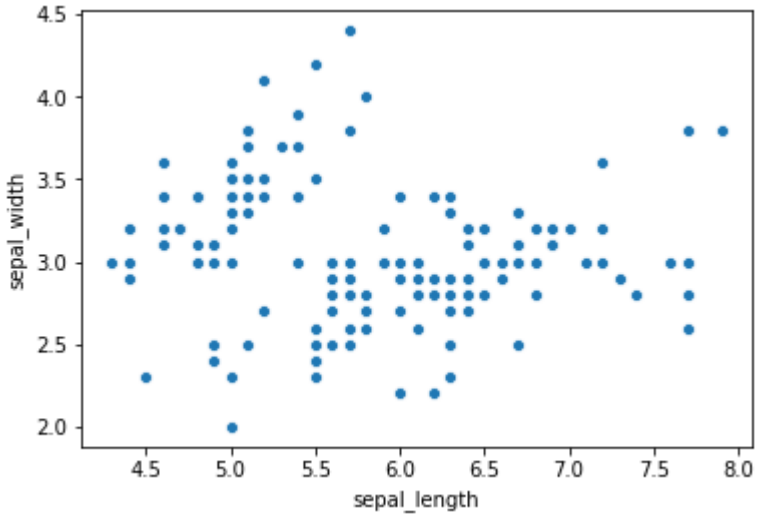
```
import seaborn as sns
```

Seaborn secara default sudah terinstall ketika menggunakan Anaconda.

Scatter plot

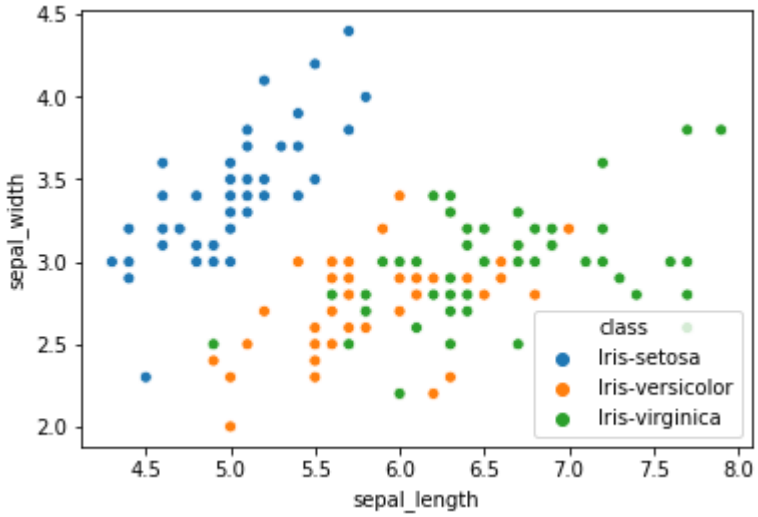
Kita bisa memanggil fungsi scatterplot() untuk membuat diagram sebar dengan seaborn. Kita bisa langsung menentukan data yang akan dipanggil sebagai sumbu x dan sumbu y di dalam parameter.

```
sns.scatterplot(x='sepal_length', y='sepal_width', data=iris)
```



Kita juga bisa langsung memberikan warna pada setiap kategori data berdasarkan class. Tinggal memanggilnya di dalam parameter. Hal ini lebih mudah daripada menggunakan Matplotlib.

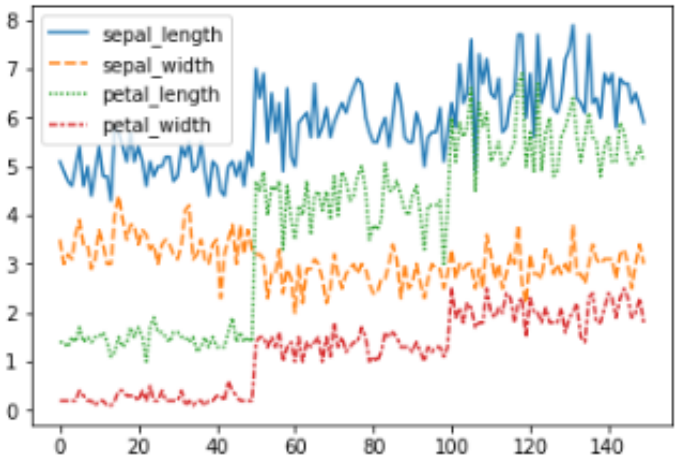
```
sns.scatterplot(x='sepal_length', y='sepal_width', hue='class', data=iris)
```



Line chart

Untuk membuat grafik garis, kita perlu memanggil fungsi lineplot() dan memasukkan data sebagai parameternya. Kita juga bisa menggunakan fungsi kdeplot() untuk membuat grafik garis yang smooth ketika ada outliers di dalam dataset.

```
sns.lineplot(data=iris.drop(['class'], axis=1))
```

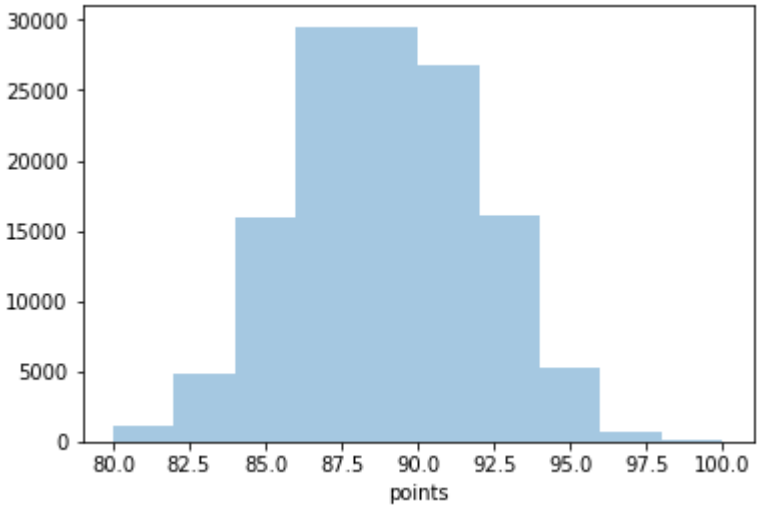


Histogram

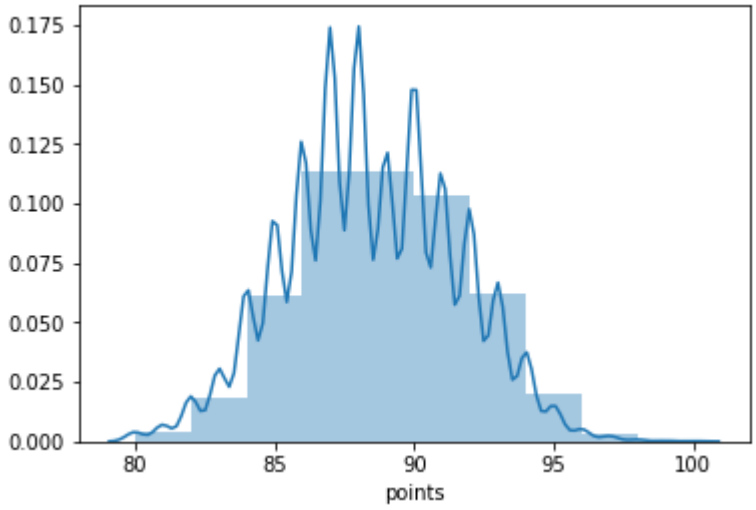
Di Seaborn, kita memanggil fungsi distplot() untuk membuat histogram. Kita perlu menentukan kolom/variabel yang akan dibuat histogram. Kita bisa juga menambahkan parameter bin. Parameter bin menentukan banyaknya kelas interval yang akan dibangun. Perhatikan penjelasan dosen di kelas.

Kadang menggunakan bin bisa membuat rancu distribusi sebaran yang sebenarnya. Kita bisa juga memplot *gaussian kernel density estimate* sebagai pembanding. Apa itu?

```
sns.distplot(wine_reviews['points'], bins=10, kde=False)
```



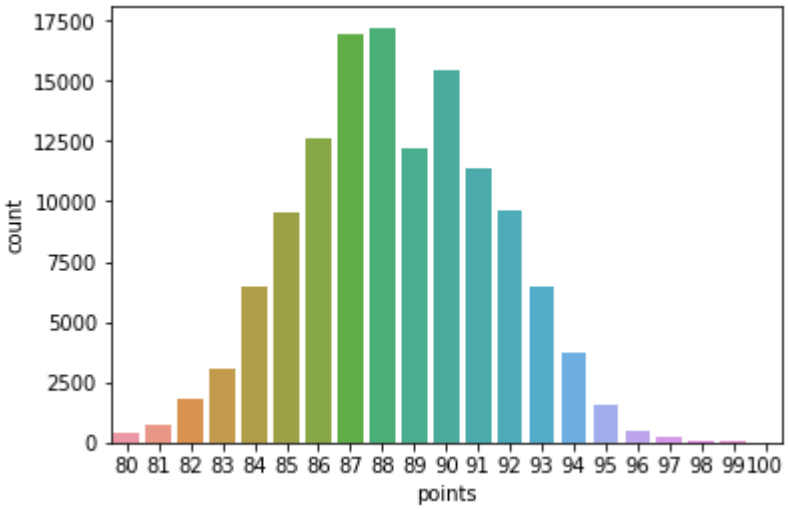

```
sns.distplot(wine_reviews['points'], bins=10, kde=True)
```



Bar chart

Untuk membuat grafik batang kita menggunakan fungsi countplot().

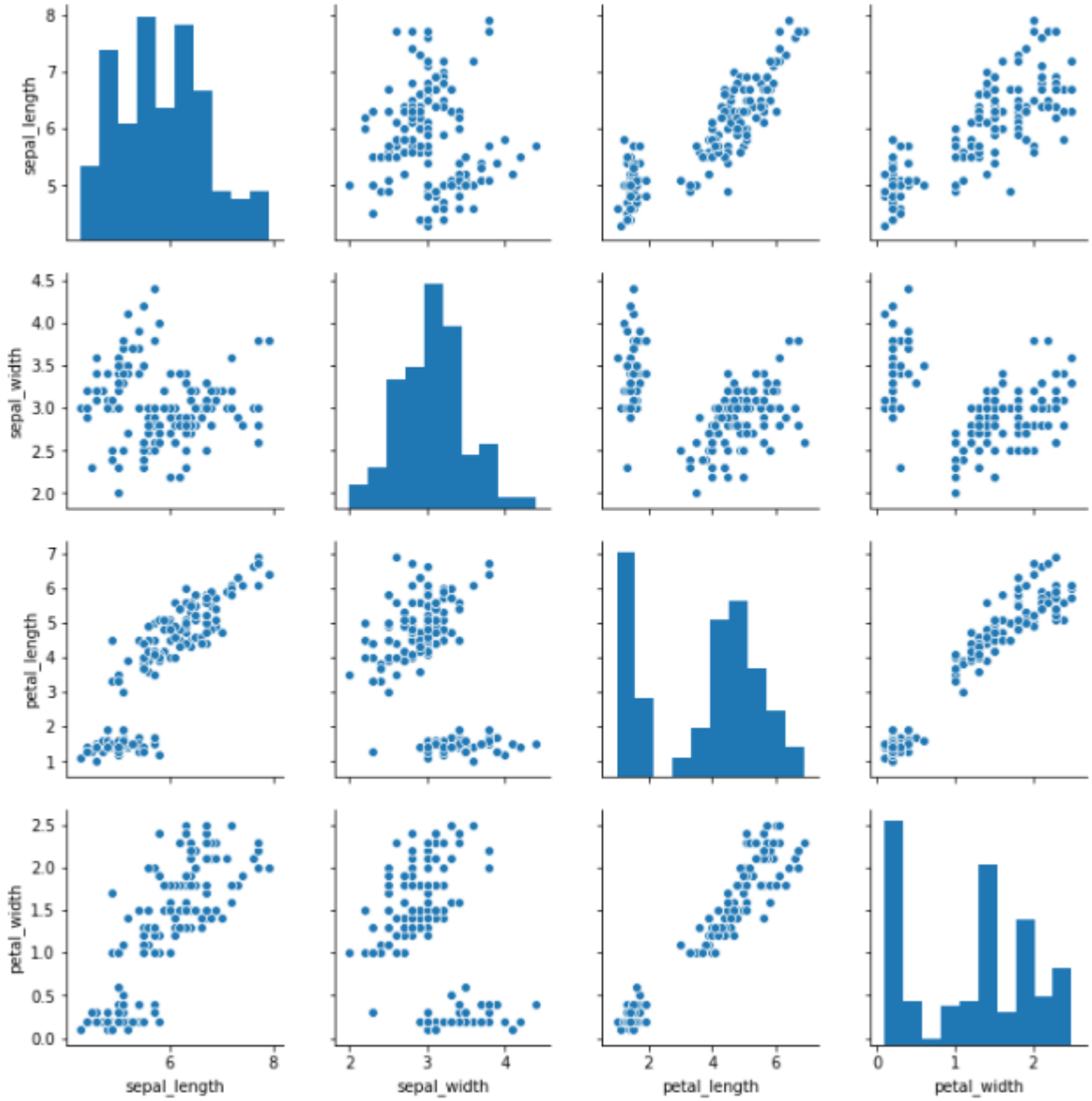
```
sns.countplot(wine_reviews['points'])
```



Pairplot

Terakhir, kita akan membuat plot pairwise (berdampingan) semua variabel yang ada dalam dataset menggunakan pairplot. Fungsi pairplot akan membantu kita agar cepat melihat kemungkinan relasi antarvariabel yang ada dalam dataset. Ingat, Anda perlu hati-hati dalam menginterpretasi hasilnya.

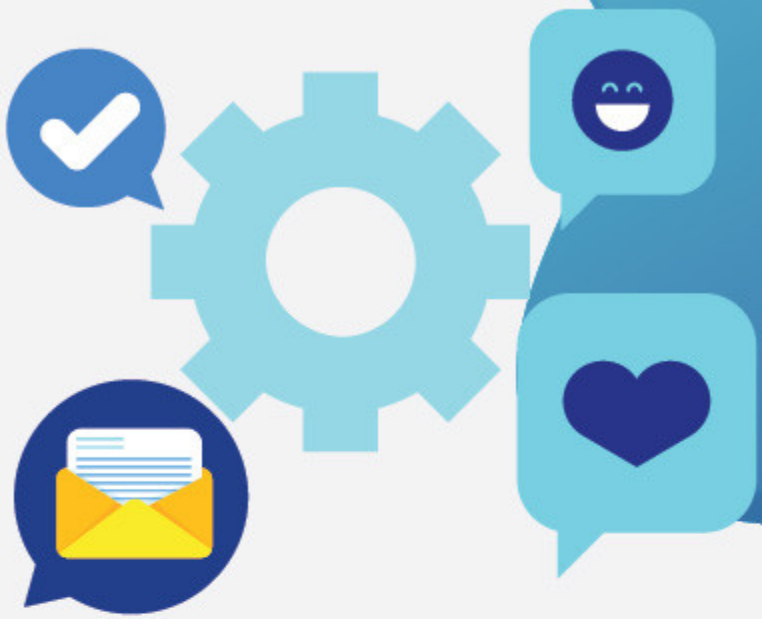
```
sns.pairplot(iris)
```



As you can see in the images above these techniques are always plotting two features with each other. The diagonal of the graph is filled with histograms and the other plots are scatter plots.

Pertanyaan

1. Jelaskan kemungkinan hubungan yang terjadi secara pairwise variabel-variabel yang ada di dataset Iris berdasarkan hasil pairplot!
2. Buatlah histogram di seaborn menggunakan bin=5, bin=10, bin=12, dan bin=15. Jelaskan mengapa histogramnya bisa berbeda!
3. Tuliskan berapa jumlah variabel dan datapoin di dataset Iris!
4. Tuliskan berapa jumlah variabel dan datapoin di dataset Wine Review!



The enemy of greatness is lazyness