TI-6P4 Praktikum Probabilitas dan Statistika

P02 - Statistik Deskriptif

Ventje Jeremias Lewi Engel, M.T.

Analisis Deskriptif adalah analisis yang dilakukan untuk menilai karakteristik dari sebuah data. Karakterisitik itu banyak sekali, antara lain: nilai Mean, Median, Sum, Variance, Standar error, standar error of mean, mode, range atau rentang, minimal, maksimal, skewness dan kurtosis.

Memperoleh Data

Data dapat diunduh di Google Classroom atau diambil dari sini. Data tersebut merupakan data multivariabel. Data yang digunakan ini merupakan data nilai siswa sekolah menengah di Amerika Serikat.

Tahap Analisis : Mengimport Packages

```
In [4]: #Import packages untuk analisis deskriptif
        import pandas as pd

In [5]: import numpy as np

In [6]: import matplotlib.pyplot as plt
```

Import Packages

Sebelum melakukan analisis, praktikan harus meng-*import packages* terlebih dahulu, sebagai berikut:

Pandas : **Pandas** digunakan untuk mengolah data dengan mudah. Fitur dalam P**andas** salah satunya adalah untuk membuat *data frame.*

Numpy : **Numpy** berperan penting dalam operasi matriks atau array, modul ini juga menyediakan berbagai fungsi-fungsi yang memudahkan dalam perhitungan matematika.

Matplotlib : **Matplotlib** merupakan modul yang digunakan untuk visualisasi grafik. Grafik tersebut diperoleh melalui hasil komputasi data yang dimiliki penulis.

Tahap Analisis : Menginput Data

```
In [7]: #Memasukkan Data
        Students=pd.read_csv("StudentsPerformance.csv")
        Students.head()
```

Out[7]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Import Data

Untuk mengetahui informasi data, seperti jumlah dan tipe data, gunakanlah fungsi info().

```
In [8]: Students.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
gender                         1000 non-null object
race/ethnicity                 1000 non-null object
parental level of education    1000 non-null object
lunch                          1000 non-null object
test preparation course        1000 non-null object
math score                     1000 non-null int64
reading score                  1000 non-null int64
writing score                  1000 non-null int64
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Informasi Data

Data yang dimiliki terdiri dari 1000 data, 8 variabel yaitu 5 variabel (*gender, race/ethnicity, parental level of education, lunch,* dan *test preparation cours*) yang bertipe data objek, dan 3 variabel (*math score, reading score* , dan *writing score*) yang bertipe data *integer*.

Soal 01: Sebutkan 2 tipe data yang dipelajari di kuliah! Tipe data objek di Python berarti termasuk yang mana?

Tahap Analisis : Analisis Deskriptif dan Visualisasi Data

Analisis pertama, praktikan ingin mengetahui rata-rata nilai dari ketiga score yaitu *math, reading, dan writing.*

```
In [56]: df=pd.DataFrame(Students, columns=['math score', 'reading score','writing score']).mean()
         df

Out[56]: math score       66.089
         reading score    69.169
         writing score    68.054
         dtype: float64
```

Rata-rata Score

Nilai rata-rata tertinggi adalah nilai membaca, yaitu 69,169. Sedangkan yang terendah adalah nilai matematika yaitu 66,089. Dikarenakan nilai matematika merupakan nilai terendah, selanjutnya praktikan ingin mengetahui statistik deskriptif dari *math score* (nilai matematika). Gunakan fungi describe().

```
In [10]: #Analisis Deskriptif
         df['math score'].describe()

Out[10]: count    1000.00000
         mean       66.08900
         std        15.16308
         min         0.00000
         25%        57.00000
         50%        66.00000
         75%        77.00000
         max       100.00000
         Name: math score, dtype: float64
```

Analisis Deskriptif

Dari hasil tersebut, dapat dilihat bahwa data *math score* terdiri dari 1000 data, dengan rata-rata 66,0890, standar deviasi 15,16 , nilai minimum 0, nilai maksimum 100, nilai kuartil 1 57, nilai kuartil 2 (median) 66, nilai kuartil 3 77.

Selanjutnya praktikan ingin mengetahui perolehan nilai matematika siswa berdasarkan jenis kelamin (*gender*). Kemudian membuat plot horizontal bar dengan fungsi barh().

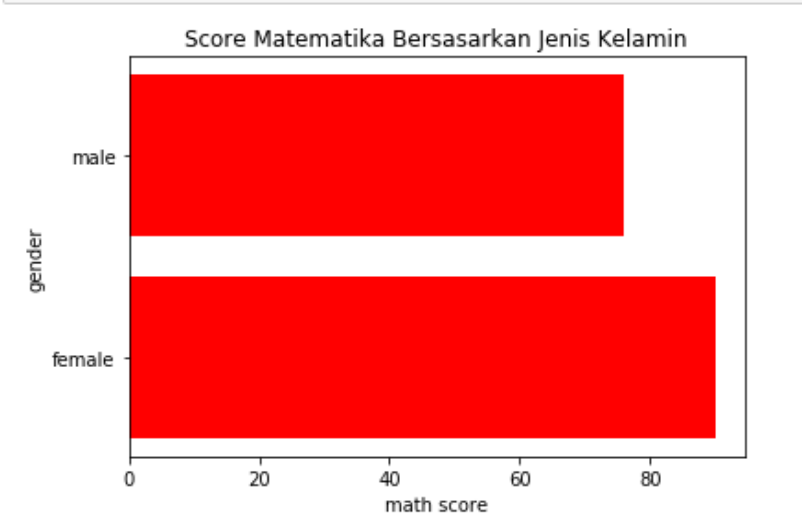
```
In [8]: #Membuat Data Frame
        df=pd.DataFrame(Students, columns=['gender','math score'])

In [32]: indeks=np.array(df['gender'])
         plt.barh(indeks[1:10],df["math score"].iloc[1:10],color="r")
         plt.title('Score Matematika Bersasarkan Jenis Kelamin')
         plt.xlabel('math score')
         plt.ylabel('gender')
Out[32]: Text(0,0.5,'gender')

In [33]: plt.show()
```

Membuat DataFrame dan *Plot*

Dari kode di atas, diperoleh hasil seperti berikut :

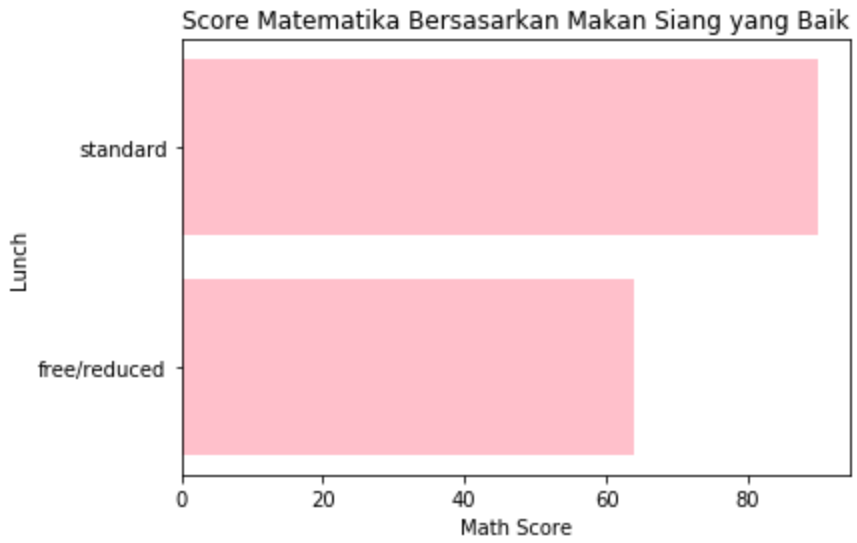


Gender terhadap Math Score

Dari plot dapat dilihat bahwa siswa berjenis kelamin perempuan memperoleh nilai matematika lebih tinggi dari pada siswa laki-laki yaitu lebih dari 80, sedangkan siswa laki-laki hanya lebih dari 60 kurang dari 80. Wow, apakah memang perempuan lebih pintar matematika daripada pria?

Soal 02: Buatlah kode agar warna grafik batang menjadi hijau!

Selanjutnya kita ingin mengetahui apakah makan siang yang baik (bergizi standar) mempengaruhi perolehan score matematika. Perhatikan hasil tersebut :

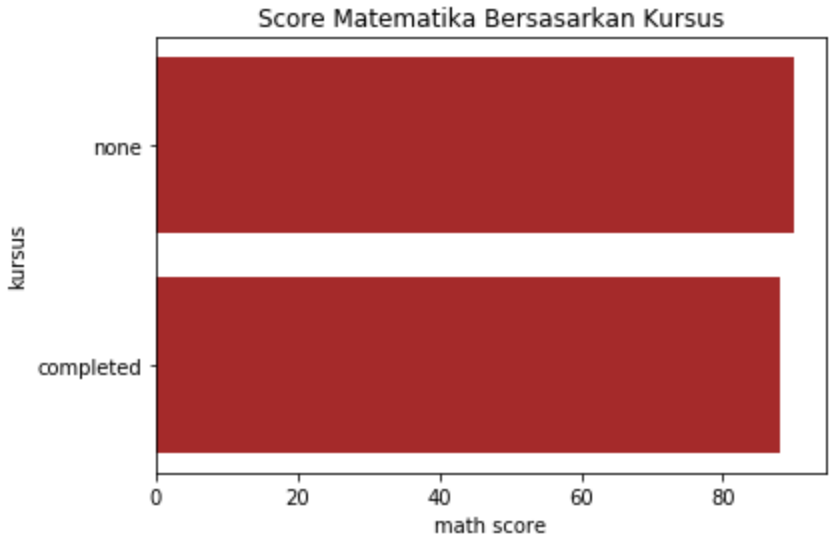


Makan Siang terhadap Nilai Matematika

Dilihat dari *plot* tersebut, bahwa makan siang yang baik (dengan gizi standar) mungkin ada korelasi dengan *score* matematika. Para siswa yang makan siangnya tidak baik *(free/reduced)* nilai matematikanya cenderung rendah.

Soal 03: Buatlah kode untuk menghasilkan plot makan siang vs. score matematika!

Setelah itu, praktikan ingin mengetahui apakah melakukan kursus sebelum ujian (*preparation course*) mempengaruhi hasil perolehan nilai matematika saat ujian. Perhatikan hasil berikut :



Kursus terhadap Math Score

Dari hasil *output* tersebut, dapat diketahui bahwa siswa yang tidak mengikuti kursus justru memperoleh nilai rata-rata tertinggi dari pada siswa yang mengikuti kursus secara lengkap.

Soal 04: Buatlah kode untuk membuat plot kursus vs score matematika!

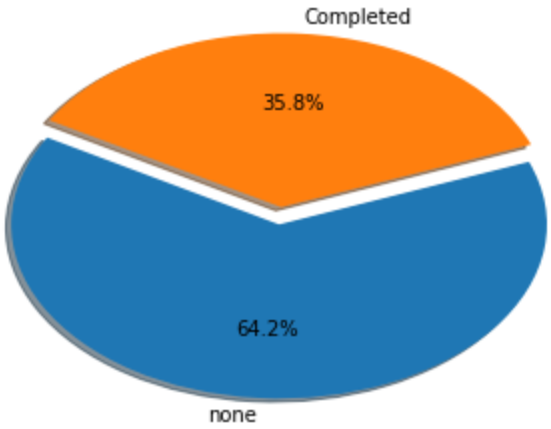
Setelah itu, kita ingin melihat persentase siswa yang mengikuti kursus dan tidak mengikuti kursus.

```
In [64]: Students["test preparation course"].value_counts()
```

```
Out[64]: none         642
         completed    358
         Name: test preparation course, dtype: int64
```

```
In [72]: labels = ["none", "Completed"]
         explode = (0.1,0)
         jumlah = [642, 358]
```

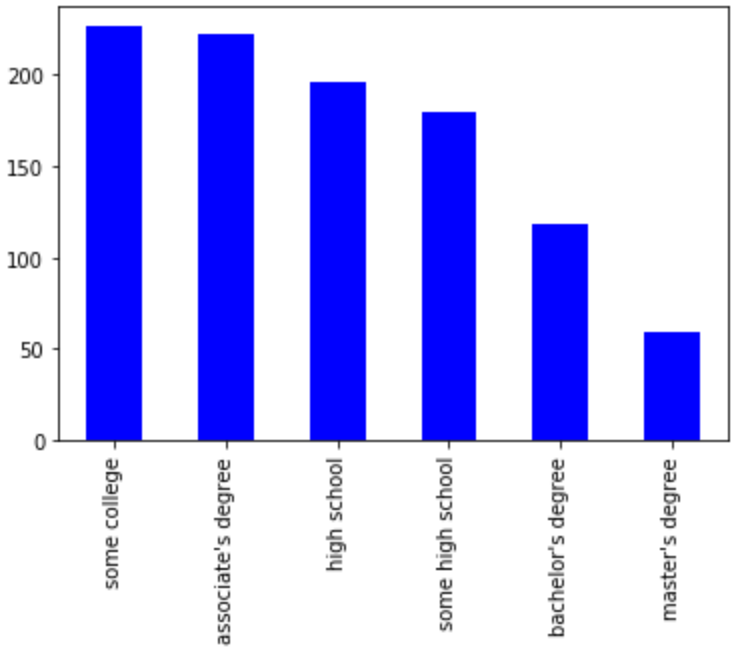
```
In [73]: plt.pie(jumlah, labels = labels, explode = explode, autopct = "%1.1f%%", shadow = True, startangle = 150)
         ...
         plt.show()
```



Pie Chart

Dari hasil pie chart, terlihat bahwa presentase siswa yang tidak mengikuti kursus sebanyak 64,2 % sedangkan untuk yang mengikuti kursus hanya 35,8 %.

Selanjutnya praktikan akan melihat frekuensi pendidikan orang tua siswa, sebagai berikut:



Pendidikan Orangtua Siswa

Dilihat dari *plot* diatas, pendidikan terakhir orangtua terbanyak adalah *some college* lebih dari 200, sedangkan yang terendah adalah *master's degree* lebih dari 50 kurang dari 100.

Soal 05 Buatlah kode untuk membuat frekuensi pendidikan orangtua siswa menjadi Pie Chart!

Sekian pembahasan analisis deskriptif dan visualisasi data.

Berikutnya di rumah, praktikan bisa menyelam lebih dalam untuk plot data menjadi Boxplot. Materi boxplot ini bisa menjadi nilai bonus untuk praktikum.

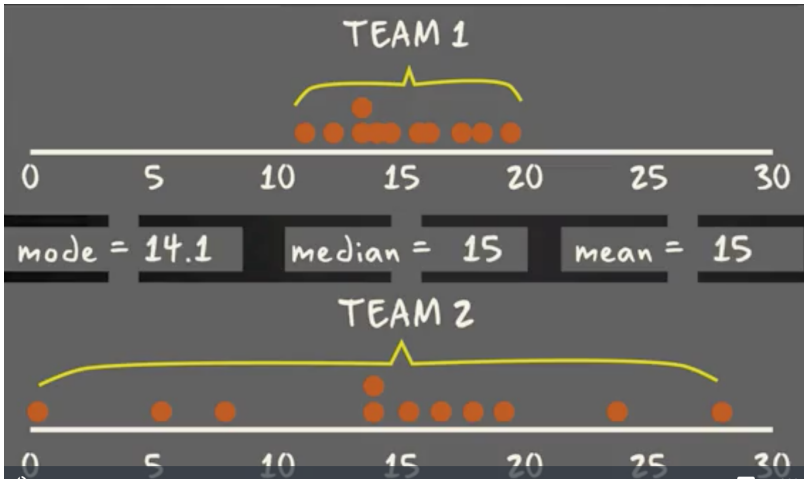
Mengenal Boxplot

Box plot atau boxplot (juga dikenal sebagai diagram box-and-whisker) merupakan suatu diagram ringkasan distribusi sampel yang bisa menggambarkan bentuk distribusi data (*skewness*), ukuran pemusatan data, dan ukuran penyebaran (keragaman) data. Boxplot adalah salah satu cara dalam statistik deskriptif untuk menggambarkan secara grafik dari data numeris melalui lima ukuran (five number summary) sebagai berikut:

1. Nilai minimum: nilai observasi terkecil
2. Q1: kuartil terendah atau kuartil pertama
3. Q2: median atau nilai pertengahan
4. Q3: kuartil tertinggi atau kuartil ketiga
5. Nilai maksimum: nilai observasi terbesar.

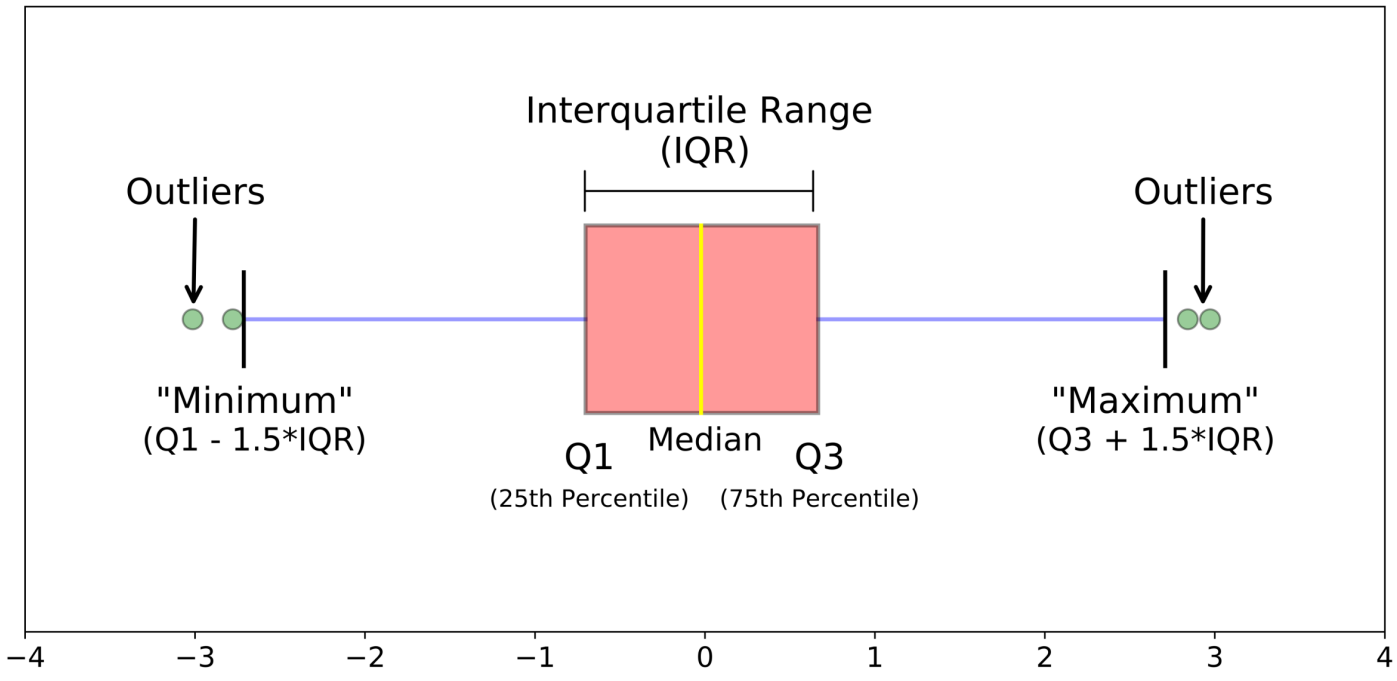
Mengapa Perlu Boxplot?

Untuk beberapa distribusi dan dataset, Anda akan menemukan bahwa kita memerlukan lebih banyak informasi daripada ukuran pemusatan data (median, rata-rata, dan modus). Perhatikan ilustrasi berikut.



Kedua Tim Memiliki Ukuran Pemusatan Data yang Sama Tetapi Berbeda Distribusi

Boxplot berguna untuk memberikan gambaran penyebaran data, selain itu juga dapat menunjukkan ada tidaknya nilai **outlier** dan nilai **ekstrim** dari data.



Boxplot dan Keterangannya

Nilai yang berada di luar minimum sampai maksimum bisa diidentifikasi sebagai outliers.

Suatu nilai dikatakan outlier jika:

Q3 + (1.5 x IQR) < outlier $\leq$ Q3 + (3 x IQR)

atau Q1 − (1.5 x IQR) > outlier $\geq$ Q1 − (3 x IQR)

Selanjutnya, suatu nilai dikatakan ekstrim jika lebih besar dari Q3 + (3 x IQR) atau lebih kecil dari Q1 − (3 x IQR)

Boxplot dengan Python

Kita akan menggunakan dataset Tips.csv

Menggunakan Pandas

Kita bisa menggunakan pandas untuk menggambar boxplot, tepatnya fungsi boxplot().

```
% matplotlib inline
```

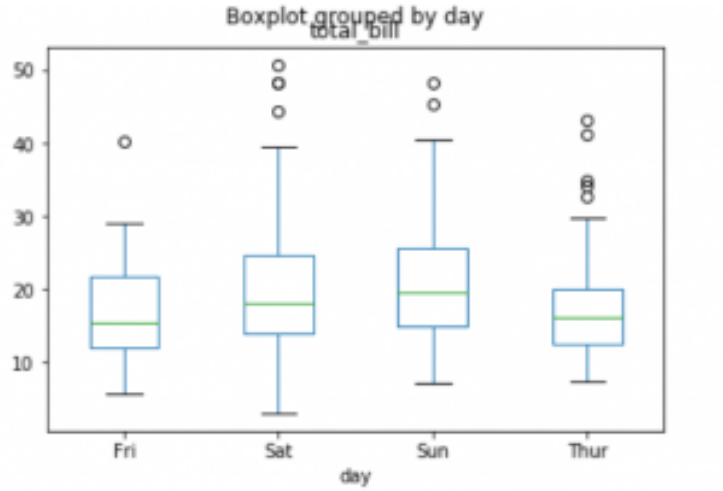
```
tips = pd.read_csv( "Tips.csv" )
```

```
tips.head()
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

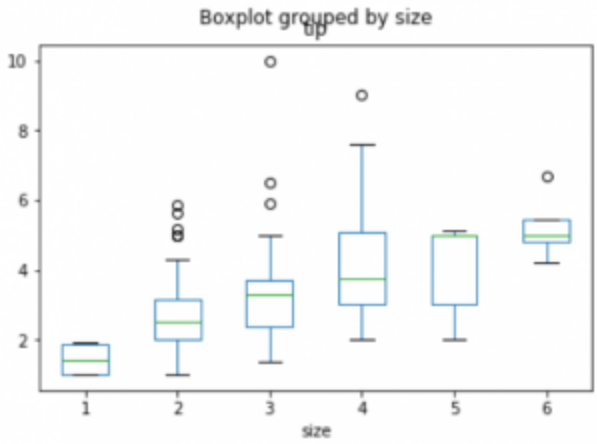
Boxplot dengan memperhatikan total_bill sebagai sumbu Y dan dikelompokkan oleh day.

```
df.boxplot(by = 'day', column = ['total_bill'], grid = False)
```



Boxplot dengan memperhatikan tip sebagai sumbu Y dan dikelompokkan oleh size.

```
df.boxplot(by = 'size', column = ['tip'], grid = False)
```



Menggunakan Seaborn

Syntax :

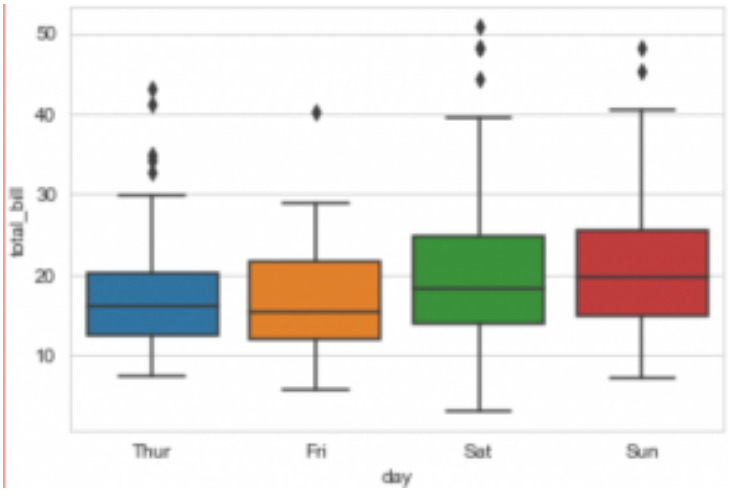
```
seaborn.boxplot(x=None, y=None, hue=None, data=None, order=None, hue_order=None,
orient=None, color=None, palette=None, saturation=0.75, width=0.8, dodge=True, fliersize=5,
linewidth=None, whis=1.5, notch=False, ax=None, **kwargs)
```

Parameters:
x = feature of dataset
y = feature of dataset
hue = feature of dataset
data = datafram or full dataset
color = color name

Boxplot dengan memperhatikan total_bill sebagai sumbu Y dan dikelompokkan oleh day.

```
sns.set_style( "whitegrid" )
```

```
sns.boxplot(x = 'day' , y = 'total_bill' , data = tips)
```

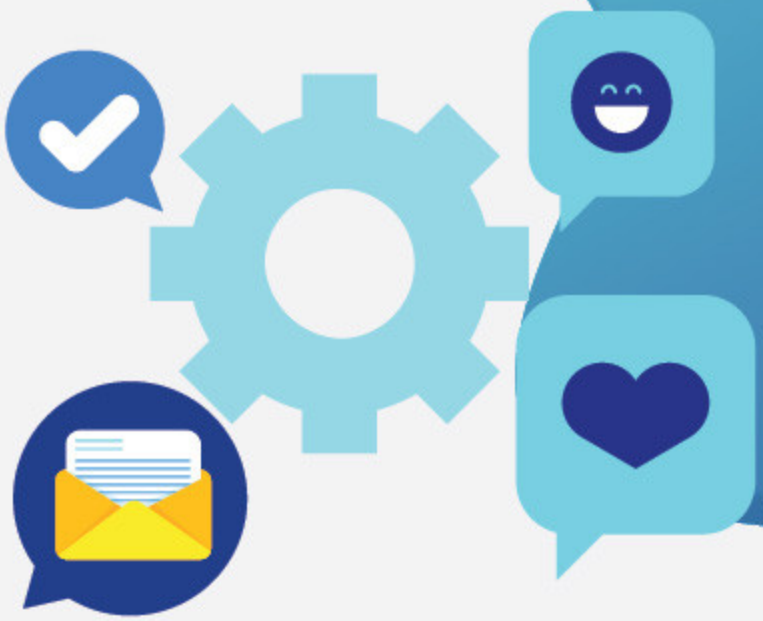


Soal Bonus: Jelaskan arti boxplot total_bill dikelompokkan oleh day.

Demikian pembahasan tentang Boxplot. Pastikan Anda mengerti sebaran dan pemusatan data ketika melihat sebuah diagram boxplot.

Stay Strong and Cheers !!

Ventje Jeremias Lewi Engel, M.T.



Do something that your future self will thank you for